Statistical Inference:
777 Exercises, 77 Stories, and Solutions

Nils Lid Hjort
University of Oslo

Emil Aas Stoltenberg
BI Norwegian Business School

– This version of PART ONE, EXERCISES, last touched by Nils, 12-August-2024 –

Some technical stuff

ISBN - Numbers numbers

The Kioskvelter Project

1234-5678

To my somebody
– N.L.H.

To my somebody
– E.A.S.

Preface

This book builds on Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

(xx then three-four paragraphs here, on the carrying ideas behind and structure of the book: *exercises* and *stories*. a partly flipped classroom, with direct participation from the first pages of each chapter. there will be *solutions to all exercises*, not physically placed inside the book, but rather on the book's website-to-be, perhaps url'd `www.mn.uio.no/math/english/research/projects/HjortStoltenberg`. That website will also have all datasets and code is `R` and `python` to carry out all analyses, for the construction of each of the book's figures, etc.

the Hjort-Stoltenberg book website

(xx if we're clever with the 777 exercises, 77 stories, we should mention Stigler's 7 pillars. x)

(xx briefly on on prerequisties: linear algebra, with matrix theory, etc.; calcululs, with functions of one or more variables, partial derivatives, etc.; programming, in `R` or `Python` or other appropriate language, both for running common algorithms inside relevant pacakages, and for programming one's own functions, for simulation, etc.)

(xx crisp clear prose here, regarding segments of readers and how they can manouevre through the material. overall: from beginning master's level, in statistics, probability theory, data science, machine learning, and upwards, to PhD level and more. xx) (i) The Linear Readers, who will benefit from having the stamina to work through chapter by chapter (ideally also exercise for exercise), and appropriate subsets of our stories. These readers will be at a high master or PhD level. (ii) The Statistical Stories Readers, for those who already know the basics on statistical models, parameter estimation and testing, some Bayes, etc. (iii) Our book is also for the specialists inside certain themes, who wish to learn even more.

(xx crisp clear prose here, regarding courses and teaching. below we help readers and instructors by also providing short lists of relevant stories, for the different types of

courses using our book. xx) Several types of courses can be taught from this book.

several courses which can be taught from our book

(i) Hard-core statistical inference, with parametric models, etc.: Chs. 1, half of 2, then most of 3, 4, half of 5, 6, 7; a selection of Stories.

(ii) Large-sample theory, the careful probablity theory leading to CLT and more, with applications in statistics: Chs. 1, 2, 5; half of 9, a selection of Stories.

(iii) Empirical processes, convergence, approximations, applications in statistics: Chs. 1, 2, 5, 9; a selection of Stories.

(iv) Survival and event history analysis: Chs. 1, the essence of 2, 3, 4, 5, then the full 9; a selection of Stories.

(v) Model selection and model averaging: Chs. 1, the essence of 2, 3, 4, 5, then the full 11; a selection of Stories.

(vi) Bayesian statistics and confidence distributions: Chs. 1, the essence of 3, 5, then the full 7, 8, parts of 15; a selection of Stories.

(vii) Statistics with applications: a special course can be taught with little emphasis on the theoretical details, but illustrating concepts, models, methods, inference views through a selection of perhaps fifty of our Stories.

The authors owe special thanks to Céline Cunen, Gudmund Hermansen, Tore Schweder, for having contributed significantly to several of our Statistical Stories, and also for always pleasant and inspiring long-term collaborations. Deep thanks are also due to a long list of colleagues and friends, who have taken part in discussions and rounds of clarification of relevance to various exercises and stories in our book: Marthe Aastveit, Patrick Ball, Bear Braumoeller, Gerda Claeskens, Aaron Clauset, Dennis Cristensen, Ingrid Dæhlen, Arnoldo Frigessi, Ingrid Glad, Håvard Hegre, Aliaksandr Hubin, Ingrid Hobæk Haff, Kristoffer Hellton, Bjørn Jamtveit, Martin Jullum, Vinnie Ko, Alexander Koning, Ian McKeague, Per Mykland, Per August Moen, Jonas Moss, Håvard Mokleiv Nygård, Lars Olsen, Steven Pinker, Sam Power, Oskar Høgberg Simensen, Catharina Stoltenberg, Gunnar Taraldsen, Ingunn Fride Tvete, Ingrid Van Keilegom, Lars Walløe, Jonathan Williams, Lan Zhang.

We have also benefitted, directly and indirectly, through the collective efforts of several grander wide-horizoned funded projects: the *FocuStat: Focus Driven Statistical Inference with Complex Data* 2014-2019 project (led by Hjort) at the Department of Mathematics, University of Oslo, funded by the Norwegian Research Council; the *Stability and Change* 2022-2023 project (led by Hjort and Hegre), funded by and hosted at the Centre for Advanced Study (CAS), Academy of Science and Letters, Oslo; and *Integreat: The Norwegian Centre for Knowledge-Driven Machine Learning* 2023-2033 Centre of Excellence (led by Frigessi and Glad), Oslo, funded by the Norwegian Research Council. We finally acknowledge with gratitude a partial support stipend from the Norwegian Non-Fiction Writers and Translators Association (Norsk faglitterær forfatter- og oversetterforening).

Nils Lid Hjort and Emil Aas Stoltenberg
Blindern, some day in 2025

Contents

Part I

Short & crisp

I.1

Statistical models

In this chapter we study families of distributions and densities that we are to meet time and again in this book. A partial list includes the uniform, normal and multinormal, chi-squared, the t and the F, Gamma, exponential, Weibull, Beta, Dirichlet, Poisson, compound Poisson, binomial, multinomial, geometric, Pareto, Gumbel, logistic. These families have parameters, with values to be set for certain studies or illustrations, or for purposes of confidence setting and tests; more generally these parameters are estimated from data, as we return to in several later chapters. We also learn fruitful ways of extending and mixing given families of distributions; in such fashions the classical models can be building blocks for forming new ones. Mathematical techniques for deriving crucial properties include those of moment-generating functions, convolutions, and double expectation.

Key words: distributions, models, moments, moment-generating functions, parameters, quantiles, sums, transformations

The aim of this chapter is to go through a generous list of parametric statistical models, from the well-known distributions connected with the normal model, to the Beta and the Gamma, to the binomial, Poisson, and negative binomial for discrete data, etc., along with deriving their basic properties. These models turn up repeatedly in later chapters and in our Statistical Stories, with variations, as direct models for data, or as building blocks for more complicated constructions. The normal and multinormal distributions play important roles, also because these become fruitful simple-to-use approximations to sometimes much more complicated exact distributions.

These models, for probability theory and statistics, rely on deeper mathematical constructions and considerations, with random variables being measurable functions on probability spaces, measure and integration theory, etc. For this book it has been practical to organise that body of mathematical theory in Appendix A. For the present chapter on models we take certain notions and basic definitions for granted, with background and more detail in that appendix. Thus we deal here with classes of distributions, parameters, probability densities, cumulative distribution functions, conditional and marginal distributions, means and variances, quantiles, correlations, and so on. In particular, any nonnegative function $f(y)$ integrating to 1 over some interval is a probability density over that interval; it has a cumulative distribution function (c.d.f.) $F(y) = \int_{-\infty}^{y} f(y')\,\mathrm{d}y'$; the

mean of a random variable Y drawn from this distribution is $\mathrm{E}\, Y = \int y f(y)\, \mathrm{d}y$; its median is $F^{-1}(\frac{1}{2})$; its variance $\mathrm{Var}\, Y = \mathrm{E}\,(Y - \mathrm{E}\, Y)^2$; the covariance between two random variables X and Y with means a and b is $\mathrm{cov}(X, Y) = \mathrm{E}\,(X - a)(Y - b) = \mathrm{E}\, XY - ab$; etc. We also deal with sums of random variables, drawn from the same or different distributions. In one of its classical forms, the density $h(z)$ of $Z = X + Y$, where X and Y are independent with densities $f(x)$ and $g(y)$, is

$$h(z) = \int f(z - y) g(y)\, \mathrm{d}y = \int f(x) g(z - x)\, \mathrm{d}x, \tag{1.1}$$

see Ex. A.17 for more. There are also occasions in this chapter where the rules for double expectation and variance

$$\mathrm{E}\, X = \mathrm{E}\,\mathrm{E}\,(X \,|\, Y) \quad \text{and} \quad \mathrm{Var}\, X = \mathrm{E}\,\mathrm{Var}\,(X \,|\, Y) + \mathrm{Var}\,\mathrm{E}\,(X \,|\, Y), \tag{1.2}$$

see Ex. A.23, come in handy.

In addition to defining and presenting a list of useful models, and diving into their properties and inter-connections, we develop certain tools, useful also in later chapters. These include transformations (see Ex. 1.12), moment-generating functions (see Ex. 1.30, with more in Ex. A.31), characteristic functions (see Ex. 1.33), conditional distributions, mixtures, and simulation. (xx make sure we have a little bit on simulation. xx) Also included is material on the general exponential family class, which has several of the classic models as special cases (see Ex. 1.50, with follow-up material in Ch. 4).

Importantly, several of the central models worked with in this introduction chapter find uses inside wider contexts, e.g. for regression situations, as we shall return to in later chapters. To indicate that direction of model building, below we learn about a random variable Y having a gamma distribution with parameters (a, b), which we write as $Y \sim \mathrm{Gam}(a, b)$; see Ex. 1.9. Now suppose there is a dataset consisting of measurements $(x_{i,1}, x_{i,2}, y_i)$ for individuals $i = 1, \ldots, n$, where the main outcome y_i is influenced by the covariates $x_{i,1}, x_{i,2}$. Then a gamma regression model could take the form $Y_i \,|\, (x_{i,1}, x_{i,2}) \sim \mathrm{Gam}(a_i, b)$, with $a_i = \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})$. The traditional multiple linear gression model is of a similar type, with Y_i given the covariates having normal distributions, with mean function linear in the covariates. Tools developed in later chapters may then be applied to estimate parameters, with confidence intervals, testing, comparisons, prediction, etc.

(xx also include: negative binomial, logarithmic, Poisson compound, hypergeometric, excentric hypergeometric. briefly generating functions $G(s) = \mathrm{E}\, s^X$ too. Agree on ϕ and Φ as fixed notation for the standard normal density and c.d.f. And check that we most of the time write c.d.f. check in a while *the title* we choose for the short & crisp sections, here and in all later chapters. xx)

(xx just a few pointers to later chapters. CLT. normal approximations. estimation, testing. calibrate with what's in the abstract. we may point to more complex models, making clear that these classic families of distributions are often used as stepping stones. could point to Markov chains etc., but not really touching these in this chapter. also: take care with mentions of limit distributions and CLT, which we may choose to touch here and there, but details come in Ch. 2. xx)

(xx as of 12-August-2024, we have a little fortellerproblem: we do mgf and characteristic functions in Appendix A, good, but then need just a bit of basic models there, complete with $M(t)$ formulae for the normal, the binomial, just a few more. but those models are more formally introduced here in Ch1. so how to deal with this. xx)

Normal, bi- and multinomial, exponential, gamma, mixing

Ex. 1.1 *The normal distribution.* The perhaps most famous and broadly useful distribution in probability theory and statistics is the normal distribution, also called the Gaußian distribution. It is also a building block for various inferred and related models and distributions, as we learn later in the chapter. In its standard form, before we add on two more parameters, the normal density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x^2) \quad \text{on the real line.}$$

We call this the standard normal distribution, and write $X \sim \mathrm{N}(0,1)$ to indicate this. It is standard in statistics and probability theory to use $\phi(x)$ for its density and $\Phi(x)$ for its cumulative distribution function (c.d.f.).

(a) There are myriad ways of demonstrating that $1/(2\pi)^{1/2}$ is the correct constant here, i.e. that $I = \int \exp(-\frac{1}{2}x^2)\,\mathrm{d}x = (2\pi)^{1/2}$. You are allowed to take this for granted, but attempt to show it via expressing I^2 as a double integral, featuring $\exp\{-\frac{1}{2}(x^2+y^2)\}$, and then substituting $x = r\cos\theta$ and $y = r\sin\theta$, followed by the use of double integration tools from calculus.

(b) Show that for X a standard normal, its mean is zero and its variance is one.

(c) With X a standard normal, consider $Y = \mu + \sigma X$, with μ any number and σ positive. Show that its mean and standard deviation are μ and σ, and that its density
Gauß can be written

$$f(y) = \phi\Big(\frac{y-\mu}{\sigma}\Big)\frac{1}{\sigma} = \frac{1}{(2\pi)^{1/2}\sigma} \exp\Big\{-\tfrac{1}{2}\Big(\frac{y-\mu}{\sigma}\Big)^2\Big\}.$$

We write $Y \sim \mathrm{N}(\mu, \sigma^2)$ to indicate this distribution. Show that $\Pr(\mu - 1.96\,\sigma \le Y \le \mu + 1.96\,\sigma) = 0.95$. Find the c such that $\Pr(|Y-\mu| \le c\sigma) = 0.50$.

(d) With X a standard normal, consider $Z = X^2$. Find its distribution, and show that its density becomes $g(z) = (2\pi)^{-1/2}\exp(-\frac{1}{2}z)/\sqrt{z}$. We learn about the chi-squared distribution in Ex. 1.43; this X^2 has such a chi-squared distribution, with degrees of freedom equal to 1, which we write as $X^2 \sim \chi^2_1$.

(e) Consider X_1, X_2, X_3 being independent and standard normal. Work out the means and variances of X_1^2, $X_1^2 + X_2^2$, $X_1^2 + X_2^2 + X_3^2$. Simulate say 10^4 realisations of these distributions, check their histograms, and describe their different behaviour close to zero.

(f) Consider the enigmatic density $f(x) = e^{-\pi x^2}$ on the real line, featuring and combining the eternal mathematical constants e and π, integrating to 1. What is its standard deviation, and what is the probability that an X with this distribution is inside $[-1,1]$?

(g) For X a standard normal, and for x becoming large, show that $\Pr(X \geq x) \doteq \phi(x)/x$, in the sense that the ratio $\{1 - \Phi(x)\}/\{\phi(x)/x\}$ tends to 1. This is the Mills ratio. Make a plot of this ratio, to see how it converges to 1, and to assess the implied approximation. (xx footnote, to be returned to, with hazards. xx) Show from this that $\Pr(X \in [x, x+\varepsilon] \,|\, X \geq x) \doteq x\varepsilon$ for growing x, and give this an interpretation.

(h) (xx some pointers, placed here or elsewhere. point to mgf already, for the linear combination property. then a simple question to illustrate this. xx)

Ex. 1.2 *Normal sums.* Sums of independent normals have themselves normal distributions. This is clearly easiest to demonstrate via moment-generating functions (m.g.f.s), see Ex. 1.31 below, but it is worth doing this via convolution formulae too.

(a) Let X and Y be independent standard normals. Show that $X+Y \sim \mathrm{N}(0, 2)$, via the convolution formula (1.1). With a bit more algebraic work, show that if $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(b) Generalise this: show that if $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, m$, and these are independent, then $Z = \sum_{i=1}^m a_i X_i$ is also normal, with mean $\sum_{i=1}^m a_i \mu_i$ and variance $\sum_{i=1}^m a_i^2 \sigma_i^2$.

(c) Sometimes 2 + 2 might be 5, for extremely high values of 2. If independent X and Y are 2 and 2, but observed with with some Gaussian noise on top, of standard deviation level σ, find a formula $p(\sigma)$ for the probability that $X + Y$ is outside $[3.5, 4.5]$, i.e. 3 or 5 or even farther away from 4 when rounded off to the nearest integer. Plot this function.

Ex. 1.3 *Binomial distribution.* One of the more old, classic, and deservedly famous distributions in probability and statistics is the binomial. If there is a fixed probability $p = P(A)$ of a certain event A taking place, in a certain type of experiment, then the number Y of times A is seen, in n independent experiments, is the binomial, which we write as $Y \sim \mathrm{binom}(n, p)$.

the binomial distribution

(a) Show that

$$\Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \quad \text{for } y = 0, 1, \ldots, n.$$

This involves the essential combinatorial fact that the number of ways one may place precisely y 1s in a total of n possible position is $\binom{n}{y} = n!/\{y!\,(n-y)!\}$. Explain that Y can be expressed as $X_1 + \cdots + X_n$, where X_i is a simple 0-1 variable, with $\Pr(X_i = 1) = p$, and where these are independent. Such X_i are called *Bernoulli variables*. Use this to prove the classic formulae np and $np(1-p)$ for mean and variance. Also, deduce the $\Pr(Y = y)$ formula from the $Y = \sum_{i=1}^n X_i$ description.

Bernoulli variables

(b) If the first question to ask concerning a distribution is about its centre (its mean, or perhaps its median), and the second is about its spread (its standard deviation, or perhaps a different measure, like its interquartile range), then the third question would be about its skewness, the degree of asymmetry. The classical skewness definition of a distribution, or equivalently of a random variable Y having that distribution, is skew $= \mathrm{E}\, W^3$, where $W = (Y - \mathrm{E}\, Y)/(\mathrm{Var}\, Y)^{1/2}$ is the normalised version of Y, i.e. linearly transformed to

have mean zero and standard deviation one. Show for the binomial (n, p) case that its skewness is the skewness

$$\text{skew} = \mathrm{E}\Big[\frac{Y - np}{\{np(1-p)\}^{1/2}}\Big]^3 = \frac{1-2p}{\{np(1-p)\}^{1/2}},$$

which for fixed p goes to zero with rate $1/\sqrt{n}$ (i.e. $\sqrt{n}$ skew tends to a positive constant). Briefly discuss what this entails regarding the degree of asymmetry for the binomial distribution.

the kurtosis

(c) After the skewness comes the so-called kurtosis, defined as $\text{kurt} = \mathrm{E}\, W^4 - 3$, with W as in the previous point. The minus 3 is there in order for the kurtosis to be zero for the normal distribution; show that this is the case. Then show that

$$\text{kurt} = (1/n)[1/\{p(1-p)\} - 6],$$

for the binomial, and comment.

Ex. 1.4 *Trinomial probabilities.* (xx emil looks it over and checks if this is suitable here in Ch1, perhaps before Ex. 1.5. if not in App A. xx) Consider the so-called trinomial distribution for a random pair (X, Y), with probability mass function

$$f(x,y) = \frac{n!}{x!\,y!\,(n-x-y)!} p^x q^y (1-p-q)^{n-x-y} \quad \text{for } x \geq 0, y \geq 0, x+y \leq n.$$

Here n is the total count number, p, q the probabilities of events of type One and Two in repeated experiments, with $p + q < 1$. With $Z = n - X - Y$ representing the number of events of type Three (not One, not Two), this is a model for the numbes of events One, Two, Three in n independent experiments; hence the trinomial name. See also Ex. 1.5.

(a) Verify that what is here called the probability mass function is the same as the density of the distribution with respect to counting measure on the set of (x, y) with $x \geq 0$, $y \geq 0$, $x + y \leq n$ – or, for that matter, with respect to counting measure on the set of all pairs (x, y) with $x \geq 0$, $y \geq 0$.

(b) Show by summing over the y that the distribution of X becomes a binom(n, p) from Ex. 1.3.

(c) Show that $Y \,|\, (X = x) \sim \text{binom}(n - x, q/(1-p))$. Give a formula for $\mathrm{E}\,(Y \,|\, X = x)$, and deduce the formula for $\mathrm{E}\, X$ from this. Find also the covariance between X and Y, using this scheme of conditioning with respect to $X = x$ first. Deduce that the correlation between them is $-\{p/(1-p)\}^{1/2}\{q/(1-q)\}^{1/2}$.

(d) Find a formula for $\Pr(X \leq x_0, Y \leq y_0)$, expressed as a sum over $x \in \{0, 1 \ldots, x_0\}$ (as opposed to a double sum over lots of (x, y) pairs). For a setup with $n = 50$, $(p, q) = (0.22, 0.33)$, compute the probability $\Pr(X \leq 15, Y \leq 15)$.

Ex. 1.5 *The multinomial model.* The binomial model, with basic properties treated in Ex. 1.3, is about sorting and counting events in two categories; if $Y \sim \text{binom}(n, p)$, then also $n - Y \sim \text{binom}(n, 1-p)$. The *multinomial model* is the natural extension to

more than two categories. Suppose there are n independent experiments, where each time one (and only one) of the events $A_1, \ldots, A_k$ takes place, with the same probabilities $p_1, \ldots, p_k$ for each experiment. Let then $Y = (Y_1, \ldots, Y_k)$, with Y_j counting the number of times A_j occurred, for $j = 1, \ldots, k$. Of course $Y_1 + \cdots + Y_k = n$, and $p_1 + \cdots + p_k = 1$, so there are $k - 1$ free parameters in the model.

(a) Show that $Y_j \sim \text{binom}(n, p_j)$, and deduce that we already know $\mathrm{E}\, Y_j = np_j$ and $\mathrm{Var}\, Y_j = np_j(1-p_j)$, even before we start working on the joint distribution of $(Y_1, \ldots, Y_k)$.

(b) Show that the joint probability distribution becomes

the multinomial model

$$f(y_1, \ldots, y_k) = \Pr(Y_1 = y_1, \ldots, Y_k = y_k) = \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_k^{y_k}$$

for nonnegative $(y_1, \ldots, y_k)$ with sum n. The first factor $n!/(y_1! \cdots y_k!)$ is a combinatorial one, the number of different ways one may place '1' in y_1 positions, '2' in y_2 positions, etc., up to 'k' in y_k positions. Note that this generalises the classic $n!/(y_1!\, y_2!) = \binom{n}{y_1}$ for the binomial case, the number of ways one may place '1' in y_1 ways (and hence '2' in $n - y_1$ ways) in a list $1, \ldots, n$.

(c) Show that each pair has a trinomial distribution, e.g.

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \frac{n!}{y_1!\, y_2!\, (n - y_1 - y_2)!} p_1^{y_1} p_2^{y_2} (1 - p_1 - p_2)^{n - y_1 - y_2}$$

for $y_1 \geq 0, y_2 \geq 0, y_1 + y_2 \leq n$. Note that formulae from Ex. 1.4 therefore apply to pairs (Y_i, Y_j) here. For $i \neq j$, show that $\mathrm{cov}(Y_i, Y_j) = -np_ip_j$, and find the correlation between Y_i and Y_j.

(d) Among the most used acronyms of statistical parlance is *i.i.d.*, for idependent and identically distributed. Explain in the present setup that $Y = Z_1 + \cdots + Z_n$, where $Z_1, \ldots, Z_n$ are i.i.d., with Z_i taking values $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ with probabilities $p_1, \ldots, p_k$. Derive again the formulae for means, variances, covariances, starting with this representation.

i.i.d.

Ex. 1.6 *Histograms.* Suppose data $Y_1, \ldots, Y_n$ are i.i.d. from some continuous density f over some interval $[a, b]$. Create disjoint cells $C_1, \ldots, C_k$, with $C_j = (a_{j-1}, a_j]$, for $a = a_0 < \cdots < a_k = b$. Let then N_j count the number of data points in cell j. The *histogram*, associated with the chosen cells, is then

histogram

$$\widehat{f}(x) = \widehat{p}_j / h_j \quad \text{for } x \in C_j, \text{ with lengths } h_j = a_j - a_{j-1},$$

where $\widehat{p}_j = N_j/n$ estimates $p_j = \Pr(Y_i \in C_j)$. In most automatic histogram algorithms, the width is taken constant across cells. The notation $\widehat{f}(x)$ indicates that beyond being an effective way of showing the essential spread and shape of the data, it is *an estimate* of the underlying $f(x)$.

(a) Carry out some simulations, samling n datapoints from the standard normal, creating histograms with k cells. For small and big n, play with k small, big, and about right. In `R`, you may use `hist(y,breaks=20,prob=T)`, etc.

(b) Show that $(N_1, \ldots, N_k)$ is multinomial. Taking a constant cell width h, for simplicity, explain that we may write $p_j = \int_{C_j} f \, \mathrm{d}y = f(\zeta_j)h$, for a suitable ζ_j inside cell C_j. From this derive that for $x \in C_j$, we have $\mathrm{E}\,\widehat{f}(x) = f(\zeta_j)$ and $\mathrm{Var}\,\widehat{f}(x) = f(\zeta_j)\{1 - f(\zeta_j)h\}/(nh)$. Argue that for the histogram to achieve $\mathrm{E}\,\widehat{f}(x) \to f(x)$ and $\mathrm{Var}\,\widehat{f}(x) \to 0$, as n increases, we need $h \to 0$ and $nh \to \infty$.

Ex. 1.7 *Hazard rates and survival functions.* Here and below we shall partly follow the implied tradition of using say T and $f(t)$ and $F(t)$, for random variables with their densities and c.d.f.s, rather than say Y and $f(y)$ and $F(y)$, when these relate to *time*. – Consider a random variable T on the halfline $[0, \infty)$, with density f and c.d.f. F. Classes of such distributions are sometimes most conveniently or fruitfully defined and discussed in terms of their hazard or cumulative hazard functions, as opposed to their densities and c.d.f.s, as we outline here; see also Ch. 10.

hazard rate function

(a) Show that

$$\Pr(T \in [t, t+\varepsilon] \,|\, T \geq t) = h(t)\varepsilon + O(\varepsilon^2), \tag{1.3}$$

in terms of the so-called *hazard rate* function $h(t) = f(t)/\{1 - F(t)\}$. With T interpreted as the time to a certain event, the function $h(t)$ describes the chance of this event taking place in the next instance, among those having survived up to t.

(b) So we may deduce hazard rate from the density. Starting instead with $h(t)$, define first *the cumulative hazard* $H(t) = \int_0^t h(s)\,\mathrm{d}s$, and show that $F(t) = 1 - \exp\{-H(t)\}$. The function $S(t) = \Pr(T \geq t) = \exp\{-H(t)\}$ is important in its own right, and is called *the survival function*.

(c) Suppose an individual has survived up to time t_0. Show that

$$\Pr(T \geq t \,|\, T \geq t_0) = \frac{S(t)}{S(t_0)} = \exp[-\{H(t) - H(t_0)\}] \quad \text{for } t \geq t_0.$$

Show that the median lifetime, for such an invidual having lived up to t_0, is $t^* = H^{-1}(H(t_0) + \log 2)$.

Ex. 1.8 *The exponential distribution.* The *exponential distribution* is a simple but important one, in probability theory and statistics, which with positive parameter θ has the density $f(t, \theta) = \theta \exp(-\theta t)$ for $t > 0$. We write $T \sim \mathrm{Expo}(\theta)$ to indicate this.

(a) Show that the cumulative becomes $F(t, \theta) = 1 - \exp(-\theta t)$, and find the median. Show also that we may write $T = T_0/\theta$, where T_0 has the unit exponential distribution with density $\exp(-t_0)$. Show that T has mean and variance $1/\theta$ and $1/\theta^2$.

(b) Using Ex. 1.7, show that the hazard rate is constant, $h(t) = \theta$, and that the cumulative hazard rate is $H(t) = \theta t$. Show also that the exponential distribution is the only one where the hazard rate is constant.

(c) Show that the median survival time is $(\log 2)/\theta$. If an individual has survived up to time t_0, what is the median survival time?

the memoryless property

(d) Assume certain light bulbs have a longevity distribution with the property that $\Pr(T \geq t_0 + t \,|\, T \geq t_0)$ does not depend on t_0. Argue that such light bulbs may be sold as if they were brand new, as long as they are still alive. Show that their distribution must be exponential.

Ex. 1.9 *The Gamma distribution.* The gamma function is important in various branches in mathematics, probability theory, and statistics, and is defined as $\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x)\,\mathrm{d}x$ for a positive. We may hence define a family of probability densities via $g_0(t,a) = \Gamma(a)^{-1} t^{a-1} \exp(-t)$ for $t > 0$. This is called the *Gamma distribution* with shape parameter a.

the gamma function

(a) With T_0 having this density, and b a positive scale parameter, show that $T = T_0/b$ has density

$$g(t,a,b) = \{b^a/\Gamma(a)\} t^{a-1} \exp(-bt) \quad \text{for } t > 0.$$

This is the two-parameter $\mathrm{Gam}(a,b)$ distribution. Verify that $\Gamma(1) = 1$, that $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$, and that $\Gamma(m) = (m-1)!$ for $m = 1, 2, \ldots$.

the Gamma distribution

(b) When T has the $\mathrm{Gam}(a,b)$ distribution, show that the mean and variance are a/b and a/b^2. Find also that $\mathrm{E}\, T^p = \{\Gamma(a+p)/\Gamma(a)\}/b^p$, valid for any p, as long as $p > -a$. Use this to show that the skewness and kurtosis become equal to skew $= 2/a^{1/2}$ and kurt $= 6/a$. Finally, regarding moments, find that the inverse gamma distributed variable $1/T$ has mean $b/(a-1)$ and finite variance $b^2/\{(a-1)^2(a-2)\}$, as long as $a > 2$.

(c) Verify that for $a = 1$ we have the exponential distribution, with density $b\exp(-bt)$ and cumulative $1 - \exp(-bt)$. Show that $a = 2$ gives density $b^2 t \exp(-bt)$ and cumulative $1 - \exp(-ty)(1+bt)$. More generally, show that the cumulative is

$$\int_0^t g(s,a,b)\,\mathrm{d}s = 1 - \exp(-bt)\Big\{1 + bt + \frac{(bt)^2}{2!} + \cdots + \frac{(bt)^{a-1}}{(a-1)!}\Big\}$$

for the case of a being an integer.

(d) For a an integer, give an explicit expression for the hazard function $h(t,a,b)$, as per (1.3), and show that it converges to b as time increases. Show that this is the case also for any a, i.e. not only for integers; it increases from zero to b, if $a > 1$, and decreases from infinity to b, if $a < 1$.

(e) Let T_1, T_2 be independent and exponential with the same θ. Show that $T_1 + T_2 \sim \mathrm{Gam}(2,\theta)$. With $T_1, \ldots, T_k$ seen as the independent waiting times between events, show that the time to event k is a $\mathrm{Gam}(k,\theta)$.

(f) With $T_1 \sim \mathrm{Gam}(a_1,b)$ and $T_2 \sim \mathrm{Gam}(a_2,b)$ independent, show that $T_1 + T_2 \sim \mathrm{Gam}(a_1+a_2,b)$. Generalise. This may indeed be accomplished via the convolution formulae from Ex. A.17, but as for other instances it becomes easier to show such statements via m.g.f.s; see Ex. 1.30–1.31.

Ex. 1.10 *Mixing the exponential.* Sometimes waiting time type data do not follow an exact exponential distribution, but rather one characterised as a mixture of such; T given θ has the Expo(θ) distribution, but the values of θ vary from occasion to occasion, or from individual to individual.

(a) Suppose indeed that $T \,|\, \theta \sim \text{Expo}(\theta)$ but that θ has some density $g(\theta)$. Show that the density of T then becomes $f(t) = \int_0^\infty \theta \exp(-\theta t) g(\theta)\, \mathrm{d}\theta$.

(b) Suppose the distribution of θ is such that $1/\theta$ has mean value $1/\theta_0$ and a positive standard deviation τ. Show, starting with $\mathrm{E}\,(T \,|\, \theta) = 1/\theta$ and $\mathrm{Var}\,(T \,|\, \theta) = 1/\theta^2$, that $\mathrm{E}\, T = 1/\theta_0$ and $\mathrm{Var}\, T = 1/\theta_0^2 + 2\tau^2$; see (1.2). The case of a very tight distribution for the θ corresponds to τ small, which again means the case of a constant rate θ_0 for all.

(c) A convenient class of distributions for θ is the Gamma, with parameters (a, b), from Ex. 1.9. Its mean and variance are a/b and a/b^2; now find also the mean and variance of $1/\theta$. Show that the density of T can be written

$$f(t, a, b) = \int_0^\infty \theta \exp(-\theta t) \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta)\, \mathrm{d}\theta = \frac{a b^a}{(b+t)^{a+1}},$$

and also that its cumulative distribution function is

$$F(t, a, b) = 1 - \Big(\frac{b}{b+t}\Big)^a = 1 - \frac{1}{(1 + t/b)^a}.$$

(d) (xx a bit more. the hazard rate function $h(t) = f(t)/\{1 - F(t)\}$ is decreasing. expressions for quantiles, $t_0(p, a, b) = b\{1/(1-p)^{1/a} - 1\}$, solution to $F(t) = p$, to be used for Story iii.4 for fitting the distribution to the 95 between-war-times. xx)

(e) Find an expression for the hazard rate function $h(t, a, b) = f(t, a, b)/\{1 - F(t, a, b)\}$, and comment on its form, compared to the exponential case.

(f) Find the mean and variance of T, for the $g(t, a, b)$ ditribution. This might be used to estimate (a, b) from data. (xx could point to Story iii.4, perhaps with better calibration. xx)

Ex. 1.11 *Gamma-mixing the gamma.* A given parametric distribution may sometimes be fruitfully extended by placing a separate distribution on one of its parameters. The following is an illustration.

(a) Consider a distribution which for given individuals is a gamma, but where the scale parameter varies between individuals. Specifically, suppose $Y \,|\, b \sim \text{Gam}(a_0, b)$ and that b has a distribution with $\mathrm{E}\, 1/b = 1/b_0$ and $\mathrm{Var}\, 1/b = \tau^2$. Show that Y has mean a_0/b_0 and variance $a_0/b_0^2 + (a_0 + a_0^2)\tau^2$.

(b) For the special case of $b \sim \text{Gam}(c, d)$, thus leading to a 3-parameter model, find the density $f(y, a, c, d)$ for Y. (xx work a bit with parametrisation here; big (c, d) correspond to old gamma. the following to be cleaned and sent to solutions. xx)

$$\bar f(y) = \int_0^\infty \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by) \frac{d^c}{\Gamma(c)} b^{c-1} \exp(-db) \mathrm{d}b = \frac{d^c}{\Gamma(c)} \frac{\Gamma(a+c)}{\Gamma(a)} \frac{y^{a-1}}{(d+y)^{a+c}}.$$

Transformations, uniform, Pareto, Cauchy, Beta, Dirichlet

Ex. 1.12 *Transformation from X to Y.* We often encounter transformations, from one variable X to another Y, also in the vector case. We need formulae for how the density $g(y)$ of the Y can be found in terms of the density $f(x)$ for X.

(a) In the one-dimensional case, suppose $X = h(Y)$, equivalently $Y = h^{-1}(X)$, where h is smooth and increasing. Show that $\Pr(Y \leq y) = \Pr(X \leq h(Y))$, with density formula

$$g(y) = f(h(y))h'(y).$$

Show also that if $x = h(y)$ is continuous and decreasing, the formula becomes $g(y) = f(h(y))|h'(y)|$. Write down density formulae for the variables $Y_1 = \exp(X)$, $Y_2 = 3.33 - 2.22\,X$, $Y_3 = \log X$ (assuming for that case that X is positive).

(b) Show that if X is normal, then a linearly transformed $Y = a + bX$ is also normal. Show that if $X \sim \text{Gam}(a, b)$, with density proportional to $x^{a-1}\exp(-bx)$, then $Y = bX \sim \text{Gam}(a, 1)$.

(c) Suppose then that $X = (X_1, \ldots, X_p)^{\text{t}}$ and $Y = (Y_1, \ldots, Y_p)^{\text{t}}$ are vectors, with transformations binding them together,

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} h_1(Y_1, \ldots, Y_p) \\ \vdots \\ h_p(Y_1, \ldots, Y_p) \end{pmatrix}, \quad \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} h_1^{-1}(X_1, \ldots, X_p) \\ \vdots \\ h_p^{-1}(X_1, \ldots, X_p) \end{pmatrix}.$$

We write this as $X = h(Y)$ and $Y = h^{-1}(X)$, for short. It is assumed that these systems of equations have unique solutions, and that the transformations are smooth, with continuous partial derivatives. In particular, the so-called Jacobi matrix

$$J(y) = \frac{\partial h(y)}{\partial y} = \frac{\partial h(y_1, \ldots, u_p)}{\partial y_1 \cdots \partial y_p},$$

having $\partial h_i(y)/\partial y_j$ as its (i, j) component, exists, and is continuous, with a non-zero determinant $\det(J(y))$ (xx point to real analysis reference xx). – Now, if X has density $f(x)$, show that

$$\Pr(Y \in B) = \int_{h(B)} f(x)\,\mathrm{d}x = \int_B f(h(y))|\det(J(y))|\,\mathrm{d}y.$$

This shows that Y has density $g(y) = f(h(y))|J(y)|$. This is essentially the multidimensional 'integration by substitution' formula of calculus.

(d) For an application, suppose X and Y are independent and standard normal, and transform to polar coordinates, $X = R\cos A$ and $Y = R\sin A$. Find the density $g(r, a)$ for (R, A), with R positive and $A \in [0, 2\pi]$. Show in particular that length R and angle A become independent, with A having a uniform distribution on $[0, 2\pi]$ (i.e. the flat density $1/(2\pi)$ over that interval). Find also the distribution of $Z = Y/X = \tan A$; see also Ex. 1.16.

(e) Let X, Y be independent standard exponentials, and consider $(U, V) = (X+Y, X/Y)$. Show that these become independent, having densities $u\exp(-u)$ and $1/(v+1)^2$.

Ex. 1.13 *Ordering exponentials.* Let Y_1, Y_2, Y_3 be independent unit exponentials, with density $f(y)\exp(-y)$ for y positive, and order them, to $Y_{(1)} < Y_{(2)} < Y_{(3)}$. Then define the so-called spacings between them, $Z_1 = Y_{(1)}$, $Z_2 = Y_{(2)} - Y_{(1)}$, $Z_3 = Y_{(3)} - Y_{(2)}$.

(a) Show first that the joint density of $(Y_{(1)}, Y_{(2)}, Y_{(3)})$ is $3!\, f(y_{(1)})f(y_{(2)})f(y_{(3)})$ on the set $y_{(1)} < y_{(2)} < y_{(3)}$. Find then the joint density for (Z_1, Z_2, Z_3), and show that they are independent.

(b) Then generalise, considering i.i.d. unit exponentials $Y_1, \ldots, Y_n$, ordered into $Y_{(1)} < \cdots < Y_{(n)}$. Work with the scaled spacings $D_1 = nY_{(1)}$, $D_2 = (n-1)(Y_{(2)} - Y_{(1)})$, up to $D_{n-1} = 2(Y_{(n-1)} - Y_{(n-2)})$, $D_n = Y_{(n)} - Y_{(n-1)}$. Show that

$$Y_{(1)} = \frac{V_1}{n}, Y_{(2)} = \frac{V_1}{n} + \frac{V_2}{n-1}, \ldots, Y_{(n)} = \frac{V_1}{n} + \frac{V_2}{n-1} + \cdots + \frac{V_{n-1}}{2} + \frac{V_n}{1},$$

and then show that in fact $V_1, \ldots, V_n$ are i.i.d. unit exponentials.

(c) The Euler constant $\gamma_e = 0.5772...$ is defined as the limit of $1 + 1/2 + \cdots + 1/n - \log n$. Use the above to show that $Y_{(n)} = \max_{i \le n} Y_i$ has mean close to $\log n + \gamma_e$, and variance converging to $\pi^2/6$.

Ex. 1.14 *The Pareto distribution.* (xx pointer to Story iii.5. xx) Consider a variable T defined on the range $T \ge t_0$, for some positive t_0, with c.d.f. $F(t) = 1 - (t_0/t)^\theta$ for $t \ge t_0$, for some positive parameter θ. This is the Pareto distribution, and we may write $T \sim \text{pareto}(t_0, \theta)$ to indicate this.

the Pareto distribution

(a) Use the $\mathrm{E}\, X = \int_0^\infty \{1 - F(x)\}\, \mathrm{d}x$ formula for means of nonnegative variables, see Ex. A.29, to show that the mean of a Pareto is $t_0\theta/(\theta-1)$, for $\theta > 1$. Show furthermore that the variance is $t_0^2\theta/\{(\theta-1)^2(\theta-2)\}$, for $\theta > 2$, and that the median is $\text{med}(T) = t_0 2^{1/\theta}$.

(b) Show that the density is $f(t, \theta) = \theta t_0^\theta / t^{\theta+1}$ for $t \ge t_0$; you may use this to find the mean formula again. Furthermore, explain that the hazard rate is $h(t) = \theta/t$.

(c) Show that $\Pr(T \ge t \,|\, T \ge t_1) = (t_1/t)^\theta$, for $t \ge t_1$. So $T \,|\, (T \ge t_1)$ is Pareto (t_1, θ). Deduce from this that

$$\text{med}(T - t \,|\, T \ge t) = ct = (2^{1/\theta} - 1)t \quad \text{for all } t.$$

This is linked to the so-called Lindy Effect: the longer the life is observed to be (of a company, an idea, a party), the longer is the remaining lifetime expected to be (note that $c > 1$). Show in fact that the Pareto distribution is the only one with the property that $\text{med}(T - t \,|\, T \ge t) = ct$ for all t.

(d) Show that $Y = \log(T/t_0)$ is exponential with parameter θ. So a representation of the Pareto is $T = t_0 \exp(Y)$, with $Y \sim \text{Expo}(\theta)$. You may use this to find the mean formula once more.

Ex. 1.15 *Maxima of i.i.d. samples.* To illustrate one of the many ways in which suitable start models may be generalised and extended, consider $Y_1, \ldots, Y_n$ i.i.d. from some distribution with c.d.f. F and density f, and from these define $M_n = \max_{i \le n} Y_i$. (xx brief pointer to order statistics and sample quantiles in Chs 2, 3. xx)

(a) Show that M_n has c.d.f. $G_n(z) = F(z)^m$ with density $g_n(z) = nF(z)^{n-1} f(z)$. Compute and draw these, for say $n = 1, \ldots, 10$, for the case of the Y_i being standard normal.

(b) Before pursuing some maxima, consider a situation with X_1 and X_2 are independent, find c.d.f.s H_1, H_2 and densities h_1, h_2. Show that

$$\Pr(X_1 \le X_2) = \int H_1(x_2) h_2(x_2)\, \mathrm{d}x_2 = \int \{1 - H_2(x_1)\} h_1(x_1) \mathrm{d}x_1.$$

(c) Long jumpers A and B have jumps being $\mathrm{N}(7.90, 0.10^2)$ and $\mathrm{N}(7.80, 0.15^2)$, respectively (on the metres scale). What is the probability the A jumps longer than B, with one jump each? Consider also p_m, the probability that B after m jumps is better than A after $m = 1, 2, 3, 4, 5, 6$ jumps (in the sense of 'best jump so far'), assuming that all jumps are independent. Show that $p_m = \int F_1(z)^m m F_2(z)^{m-1} f_2(z)\, \mathrm{d}z$, in terms of the two distributions. Compute p_m both via numerical integration and via simulations.

Ex. 1.16 *Ratios and the Cauchy.* If (X, Y) has a certain distribution, what happens to the ratio $V = Y/X$?

(a) Suppose that X and Y are independent with the same density f on $(0, \infty)$. Show that $V = Y/X$ has density $g(v) = \int_0^\infty x f(x) f(vx)\, \mathrm{d}x$. With X and Y independent from the same exponential distribution, show that $g(v) = 1/(1+v)^2$.

(b) With X and Y independent from the same Gamma (a, b), show that $V = Y/X$ has density $\{\Gamma(2a)/\Gamma(a)^2\}\, v^{a-1}/(1+v)^{2a}$.

(c) Suppose now that X and Y are independent from the same density f, symmetric around zero. Show that V has density $g(v) = 2\int_0^\infty x f(x) f(vx)\, \mathrm{d}x$. For the special case of a ratio of two independent standard normals, show that the Cauchy

$$g(v) = (1/\pi)/(1+v^2) \quad \text{with c.d.f.} \quad G(v) = \tfrac{1}{2} + (1/\pi) \arctan v.$$

This is the Cauchy distribution (in its standard form). Show that it has no mean. Find its interquartile range.

(d) There is an infinitely long straight line ahead of you, with distance r from you to the nearest point. You kick a ball towards the line, with some angle $A \in (-\pi/2, \pi/2)$. Show that it crosses the line at position $X = r \tan(A)$. Show that with A having density g and c.d.f. G, over $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$, then X has c.d.f. $F(x) = G(\arctan(x/r))$. When A is uniform over that interval, show that X/r is standard Cauchy. Investigate what happens if A is uniform over a tighter interval, like $(-\pi/4, \pi/4)$.

(e) What distribution does the random angle A need to have, in order for X to be normal? For $r = 1$ unit away from the line, show that the curious density $g(a) = \phi(\tan(a))/\cos^2(a)$ leads to a standard normal X. Graph g to see that it is symmetric and bimodal, with modes at $\pm$ 45 degrees.

Ex. 1.17 *Transformations to the uniform.* We have already touched the uniform distribution in a few points above. Say in general that a variable U is uniform on the interval $[a, b]$ if its density is constant over that interval, i.e. $1/(b-a)$, and zero outside. In particular, we write $U \sim \text{unif}(0,1)$ to indicate a variable with the uniform distribution on the unit interval.

(a) For such a $U \sim \text{unif}(0,1)$, find its mean and variance. Find the probabilities that U lands in $[0.03, 0.04]$, or in $[0.77, 0.78]$.

(b) Let F be a continuous and increasing c.d.f. for a variable X. Show that $U = F(X)$ is uniform on the unit interval. Conversely, we may start with $U \sim \text{unif}(0,1)$ and map to $X' = F^{-1}(U)$. Show that X' has distribution F.

(c) Consider the c.d.f. $F(x) = (x/10)^{3.33}$ on $[0, 10]$. Simulate 10^4 independent X_i from this distribution, by transforming uniforms. Make a fine histogram, with the density $f(x)$ plotted alongside.

the Beta distribution

Ex. 1.18 *The Beta distribution.* An important class of distributions, over the unit interval $(0,1)$, is the *Beta distribution*, with two positive parameters. We write $p \sim \text{Beta}(a,b)$ if its density is

$$\text{be}(p,a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \quad \text{for } p \in (0,1).$$

(a) As an introductory exercise, of separate interest, suppose X and Y are independent Gamma variables with parameters $(a,1)$ and $(b,1)$. Construct from these the sum $Z = X+Y$ and ratio $P = X/(X+Y)$. Finding the joint density for (P,Z), demonstrate that Z and P are independent, with $Z \sim \text{Gam}(a+b,1)$ and P having precisely the $\text{be}(p,a,b)$ density. This in particular shows that the integration constant is the correct one, i.e. that $\int_0^1 p^{a-1}(1-p)^{b-1}\,\mathrm{d}p = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

(b) Compute and display a few of these densities, for (a,b) of your choice. Note that the uniform is the special case of $(a,b) = (1,1)$.

(c) Show that $\text{E}\,p = p_0 = a/(a+b)$ and that $\text{Var}\,p = p_0(1-p_0)/(a+b+1)$.

(d) Find a formula for $\text{E}\,p^m$, for $m = 1, 2, \ldots$, in terms of (a,b), and in terms of the reparametrisation $(cp_0, c(1-p_0))$. Use this to find

$$\text{E}\,(p-p_0)^3 = \frac{2p_0(1-p_0)(1-2p_0)}{(c+1)(c+2)},$$

with a consequent formula for the skewness. For fixed mean p_0, show that the skewness tends to zero with increasing c.

(e) Examine the particular $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ distribution. Show that its density and c.d.f. become $f(p) = (1/\pi)/\{p(1-p)\}^{1/2}$ and $F(p) = (2/\pi)\arcsin(\sqrt{p})$. Find is quantile $F^{-1}(q)$. Plot these functions.

Ex. 1.19 *The Dirichlet distribution.* Let $G_1, \ldots, G_k$ be independent and Gamma distributed, with parameters $(a_1, 1), \ldots, (a_k, 1)$. With $G = G_1 + \cdots + G_k$ their sum, consider the random ratios

$$(X_1, \ldots, X_{k-1}) = (G_1/G, \ldots, G_{k-1}/G).$$

It inherits a distribution, with density $h(x_1, \ldots, x_{k-1})$, worked with below, in the simplex where each $x_i \geq 0$ and $x_1 + \cdots + x_{k-1} < 1$. Taking also $X_k = G_k/G = 1 - (X_1 + \cdots + X_{k-1})$ on board, we have a vector $(X_1, \ldots, X_k)$ of random probabilities summing to 1 over its k categories. Its distribution has a name: it's the Dirichlet distribution, with k categories, and parameters $(a_1, \ldots, a_k)$, which we write as $X \sim \text{Dir}(a_1, \ldots, a_k)$.

the Dirichlet distribution

(a) Suppose $(X_1, X_2, X_3, X_4, X_5, X_6) \sim \text{Dir}(a_1, a_2, a_3, a_4, a_5, a_6)$. Show that $(X_1 + X_4 + X_6, X_2, X_3 + X_5) \sim \text{Dir}(a_1 + a_4 + a_6, a_2, a_3 + a_5)$. Generalise and formalise this summing-over-cells property of the Dirichlet distribution.

(b) With $X \sim \text{Dir}(a_1, \ldots, a_k)$, show that each $X_i \sim \text{Dir}(a_i, a - a_i)$, with $a = a_1 + \cdots + a_k$, and that this is the same as a $\text{Beta}(a_i, a - a_i)$. Show from this that

$$\text{E}\, D_i = \xi_i = a_i/a, \quad \text{Var}\, D_i = \xi_i(1 - \xi_i)/(a + 1),$$

Show also the $\text{cov}(D_i, D_j) = -\xi_i \xi_j/(a + 1)$ for $i \neq j$.

(c) We have been able to derive certain basic properties above, without really needing an expression for the density of a Dirichlet vector. We tend to this now, using the transformation machinery of Ex. 1.12. Show in fact, starting with $(G_1, \ldots, G_k)$ and transforming to $(X_1, \ldots, X_{k-1}, G)$, (i) that $(X_1, \ldots, X_{k-1})$ has the density

$$h(x_1, \ldots, x_{k-1}) = \frac{\Gamma(a)}{\Gamma(a_1) \cdots \Gamma(a_k)} x_1^{a_1 - 1} \cdots x_{k-1}^{a_{k-1} - 1} (1 - x_1 - \cdots - x_{k-1})^{a_k - 1}$$

over the simplex; (ii) that $G \sim \text{Gam}(a, 1)$; (iii) that these are independent; and (iv) that this also verifies the implied formula for integrating $x_1^{a_1 - 1} \cdots x_{k-1}^{a_{k-1} - 1} (1 - x_1 - \cdots - x_{k-1})^{a_k - 1}$ over the simplex. Note that these efforts and results generalise the findings of Ex. 1.18(a).

Ex. 1.20 *Dirichlets inside Dirichlets.* The summing-over-cells property of the Dirichlet, see Ex. 1.19, has other consequences and angles, and we shall learn here that long Dirichlet vectors might be split into Dirichlet parts via Dirichlet cuts.

(a) Start with A, B, C independent gammas with parameters $(a, 1)$, $(b, 1)$, $(c, 1)$. Form from these $X = A/(A + B)$, $Y = (A + B)/(A + B + C)$, $Z = A + B + C$. Explain that we already know that $X \sim \text{Beta}(a, b)$, $Y \sim \text{Beta}(a + b, c)$, $Z \sim \text{Gam}(a + b + c, 1)$. Show in fact that X, Y, Z also are independent. This means working out their joint distribution; establish first that the inverse transform is $A = XYZ$, $B = (1 - X)YZ$, $C = (1 - Y)Z$, and that the associated Jacobi determinant becomes yz^2. Explain that all of this leads to the product Beta representation

$$\frac{A}{A + B + C} = \frac{A}{A + B} \frac{A + B}{A + B + C} = \text{Beta}(a, b)\, \text{Beta}(a + b, c) = \text{Beta}(a, b + c).$$

(b) (xx nils rant, to be edited. link to pinned-down Dirichlet processes. xx) Suppose $(X_1,\ldots,X_k,Y_1,\ldots,Y_\ell)$ is $\mathrm{Dir}(a_1,\ldots,a_k,b_1,\ldots,b_\ell)$, which we may represent as $X_i = G_i/(S+T)$, $Y_j = H_j/(S+T)$, with the G_i and H_j being independent gamma variables with the appropriate parameters, and sums $S=\sum_{i=1}^k G_i$ and $T=\sum_{j=1}^\ell H_j$. Explain that $W=\sum_{i=1}^k X_i = S/(S+T)$ is a $\mathrm{Beta}(a,b)$, with $a=\sum_{i=1}^k a_i$ and $b=\sum_{j=1}^\ell b_j$, and show that

$$X_i = \frac{G_i}{S+T} = W\frac{G_i}{S} = WX_i' \quad\text{and}\quad Y_j = \frac{H_j}{S+T} = (1-W)\frac{H_j}{T} = (1-W)Y_j',$$

where $X'=(X_1',\ldots,X_k') \sim \mathrm{Dir}(a_1,\ldots,a_k)$ and $Y'=(Y_1',\ldots,Y_\ell') \sim \mathrm{Dir}(b_1,\ldots,b_\ell)$, independent of W. We learn that the long Dirichlet vector (X,Y) may be split into two separate Dirichlet vectors $X'=X/w$ and $Y'=Y/(1-w)$, by conditioning on $W=w$.

(c) Now generalise to longer vectors. Let $X=(X_1,\ldots,X_k)$ be a $\mathrm{Dir}(a_1,\ldots,a_k)$, where each X_i is a vector $(X_{i,1},\ldots,X_{i,m_i})$, with corresponding $a_i=(a_{i,1},\ldots,a_{i,m_i})$. Write $W_i=\sum_{j=1}^{m_i} X_{i,j}$ for the sum over X_i. Explain first that $(W_1,\ldots,W_k)\sim\mathrm{Dir}(b_1,\ldots,b_k)$, where $b_i=\sum_{j=1}^{m_i} a_{i,j}$ is the sum over a_i. Show next that with $X_i'=X_i/W_i$, then $X_i'\sim\mathrm{Dir}(a_i)$, and that $X_1',\ldots,X_k'$ are independent of $(W_1,\ldots,W_k)$.

(d) The previous point implies that a long Dirichlet vector may be split into independent Dirichlet components, in several ways. In the setting above, starting with the long $(X_1,\ldots,X_k)$, suppose we condition on $(W_1,\ldots,W_k)=(w_1,\ldots,w_k)$, a given probability vector. With $X_i'=X_i/w_i$, show that $X_1',\ldots,X_k'$ are independent, with $X_i'\sim\mathrm{Dir}(a_i)$. This is the sorcerer's apprentice property of the Dirichlet distribution; pin it down, into subsets with given sums, and see that each part is a scaled Dirichlet again.

Ex. 1.21 *The Beta-binomial distribution.* Sometimes comparable binomial experiments may be modelled and analysed jointly, but where the success probability is not the same across studies. The $p=\Pr(\text{girl})$ may e.g. vary from family to family, see Story i.2.

(a) Suppose in general terms that $Y\,|\,p\sim\mathrm{binom}(n,p)$, and that p has a distribution with mean p_0 and standard deviation τ_0. Using the double expectation rule (1.2), show that

$$\mathrm{E}\,Y = np_0 \quad\text{and}\quad \mathrm{Var}\,Y = np_0(1-p_0)+n(n-1)\tau_0^2.$$

Hence the extra-binomial component of the variance, the $n(n-1)\tau_0^2$, becomes more noticeable with increasing n. The case of $\tau_0=0$ corresponds to the usual binomial.

(b) Suppose $Y\,|\,p\sim\mathrm{binom}(n,p)$ and that $p\sim\mathrm{Beta}(a,b)$. Show that this leads to the distribution

$$\begin{aligned}\Pr(Y=y) &= \int_0^1 \binom{n}{y} p^y(1-p)^{n-y} g(p,a,b)\,\mathrm{d}p \\ &= \binom{n}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(n+a+b)} \quad \text{for } y=0,1,\ldots,n.\end{aligned}$$

Give formulae for the mean and variance of Y. For the special case of the uniform for p, show that all outcomes for Y are equally likely.

Ex. 1.22 *The Dirichlet-multinomial distribution.* Here we deal with the natural extension of the Beta-binomial setup of Ex. 1.21, from the case of two categories to more than two.

(a) Let $Y = (Y_1, \ldots, Y_k)$, for given probability vector $p = (p_1, \ldots, p_k)$, have a multinomial $(n, p_1, \ldots, p_k)$ model, as per Ex. 1.5. Assume then that the p is not fixed, but with p_i variances $\tau_{0,i}^2$ around mean $p_{0,i}$. Show that Y_i, marginally, has mean $np_{0,i}$ and variance $np_{0,i}(1-p_{0,i}) + n(n-1)\tau_{0,i}^2$.

(b) Let in particular $p \sim \mathrm{Dir}(cp_0)$, with parameters $cp_0 = (cp_{0,i}, \ldots, cp_{0,k})$. Show that

$$\mathrm{Var}\, Y_i = \{n + n(n-1)/(c+1)\}p_{0,i}(1-p_{0,i}) = \frac{c+n}{c+1} np_{0,i}(1-p_{0,i}),$$

with a clear overdispersion factor with respect to multinomial variation.

(c) Show that the marginal distribution of $(Y_1, \ldots, Y_k)$, now overdispersed compared to the multinomial, becomes

$$\begin{aligned} \bar{f}(y_1, \ldots, y_k) &= \int \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_{k-1}^{y_{k-1}} (1 - p_1 - \cdots - p_{k-1})^{y_k} \\ &\qquad g(p_1, \ldots, p_{k-1})\, \mathrm{d}p_1 \cdots \mathrm{d}p_{k-1} \\ &= \frac{n!}{y_1! \cdots y_k!} \frac{\Gamma(c)}{\Gamma(cp_{0,1}) \cdots \Gamma(cp_{0,k})} \frac{\Gamma(cp_{0,1} + y_1) \cdots \Gamma(cp_{0,k} + y_k)}{\Gamma(c+n)}. \end{aligned}$$

(d) For the case of Dirichlet parameters $cp_0 = (1, \ldots, 1)$, show that all outcomes $(y_1, \ldots, y_k)$ have the same probability, and find a formula for how many different outcomes there can be.

Laws of small numbers, Poisson, geometric, negative binomial

Ex. 1.23 *The Poisson distribution.* Counting a high number of events with small probabilities leads to the Poisson distribution, which we define here. A count variable Y is said to have the Poisson distribution with parameter $\theta > 0$ if

$$\Pr(Y = y) = \exp(-\theta)\theta^y / y! \quad \text{for } y = 0, 1, 2, \ldots.$$

We write $Y \sim \mathrm{Pois}(\theta)$ to indicate this.

(a) Show that the probabilities indeed sum to 1. Verify next that $\mathrm{E}\, Y = \theta$, $\mathrm{E}\, Y(Y-1) = \theta^2$ and show from this that the variance is equal to the mean. Show further that $\mathrm{E}\, Y(Y-1)(Y-2) = \theta^3$, $\mathrm{E}\, Y(Y-1)(Y-2)(Y-3) = \theta^4$, and with further algebra that for $W = (Y-\theta)/\sqrt{\theta}$, we have skew $= \mathrm{E}\, W^3 = 1/\sqrt{\theta}$, kurt $= \mathrm{E}\, W^4 - 3 = 1/\theta$. Show also that $\mathrm{Var}\,(Y-\theta)^2 = 2\theta^2 + \theta$.

(b) With $Y \sim \mathrm{Pois}(\theta)$, what is the most probable outcome? What is the probability that Y is odd?

(c) Show that the sum of two independent Poisson variables is Poisson, with parameter equal to the sum of the two parameters. Generalise.

(d) Consider $Y \sim \text{binom}(n, p)$, and assume that n grows, while p becomes small, in the fashion of $np \to \theta$. Show that Y then tends to the Pois(θ) distribution, in the sense that the point probabilities converge. See also Ex. 2.8 for a fuller picture.

(e) In some event counting applications there are more zeros than predicted by the Poisson, leading naturally to a more general model with

$$\Pr(Y = 0) = p_0, \quad \Pr(Y = y) = (1 - p_0) \exp(-\theta)\theta^y / y! \,/\{1 - \exp(-\theta)\} \quad \text{for } y \geq 1.$$

Verify that these probabilities sum to 1, and find expressions for the mean and variance. This model is sometimes called the zero-inflated Poisson, since situations with $p_0 > \exp(-\theta)$ are prevalent, but also cases with $p_0 < \exp(-\theta)$ are allowed. Simulate say 1000 datapoints from the model with $\theta = 3.00$ and $p_0 = 0.25$, and check the histogram.

Ex. 1.24 *The geometric distribution.* Suppose Y has the distribution with point probabilities $f(y) = (1 - p)^{y-1}p$ for $y = 1, 2, \ldots$. This is the geometric distribution, and we write $Y \sim \text{geom}(p)$ to indicate this.

(a) Show that the probabilities $f(y)$ indeed sum to 1. Suppose independent experiments are carried out, each time with probability p that a certain event A takes place. With Y the first time A happens, show that $Y \sim \text{geom}(p)$.

(b) Show that Y has mean $1/p$ and variance $(1 - p)/p^2$, via direct summation of $\sum_{y=1}^{\infty} yf(y)$ etc. If Y is the number of times you need to roll a six-sided die until it shows a '6', find the mean and the standard deviation.

(c) Another way of finding the mean and variance is as follows. With probability p, $Y = 1$; with complementary probability $1 - p$, $Y = 1 + Y'$, with Y' having the same distribution as Y. Show that this leads to $\text{E}\, Y = p + (1-p)(1+\text{E}\, Y)$ and solve. Use this representation to also find the variance. Show also that $\text{E}\,(Y - 1/p)^3 = (1-p)(2-p)/p^3$.

(d) Show that $\Pr(Y \geq y) = (1 - p)^{y-1}$, derive a formula for $\Pr(Y \geq y_0 + y \,|\, Y \geq y_0)$, and comment.

(e) A simple related distribution is when one starts counting at 0, not at 1, so to speak. Show that with $Y \sim \text{geom}(p)$, as defined above, the variable $Y_0 = Y - 1$ has point probabilities $\Pr(Y_0 = y) = q^y p$ for $y = 0, 1, \ldots$, writing $q = 1 - p$. Show that Y_0 has mean $(1 - p)/p$ and variance $(1 - p)/p^2$.

(f) Suppose Y given p has geometric probabilities $(1 - p)^{y-1}p$ for $y = 1, 2, \ldots$, but that p itself stems from a uniform distribution. Find the distribution for Y; not that this implies $1/2 + 1/6 + 1/12 + 1/20 + \cdots = 1$.

Ex. 1.25 *Time to last event.* Suppose independent geometric experiments are carried out by m individuals, say players throwing dice until they get the '6'. How long time does it take until the event in question has taken place, for all individuals?

(a) Let $Y_1, \ldots, Y_m$ be the time needed for the m individuals to see the event in question. Show that $F(y) = \Pr(Y_i \leq y) = 1 - (1-p)^y$ for $y = 1, 2, \ldots$. For $Z_m = \max(Y_1, \ldots, Y_m)$,

show that its c.d.f. is $H_m(z) = F(z)^m$, and use this to find a formula for the median time to final event: $z^*_m = \log(1 - (\frac{1}{2})^{1/m})/\log(1-p)$ (which may be rounded to the nearest integer).

(b) For $p = 1/6$, compute also the mean $\mathrm{E}\, Z_m$, for say $m = 1, \ldots, 100$. Plot this, along with the median times z^*_m.

Ex. 1.26 *Mixing the Poisson.* Suppose observations come from Poisson mechanisms, but with different parameters, forming their own distribution. There are several versions and uses of such Poisson overdispersion models. (xx pointer to Poisson regression with overdispersion, perhaps in Ch5. xx)

(a) Suppose $Y \mid \theta \sim \mathrm{Pois}(\theta)$ but that θ has a distribution with mean θ_0 and variance τ_0^2. Show that Y has mean θ_0 and variance $\theta_0 + \tau_0^2$.

(b) Specialise to the case of $\theta \sim \mathrm{Gam}(a, b)$, see Ex. 1.9. Show that $\mathrm{E}\, Y = \theta_0 = a/b$ and that $\mathrm{Var}\, Y = \theta_0(1+1/b)$. Argue that with a large b we come back to pure Poisson. Show also that the marginal distribution of Y becomes

$$f(y,a,b) = \frac{\Gamma(a+y)}{\Gamma(a)\, y!}\frac{b^a}{(b+1)^{a+y}} = \frac{\Gamma(a+y)}{\Gamma(a)\, y!}\Big(\frac{b}{b+1}\Big)^a\Big(\frac{1}{b+1}\Big)^y \quad \text{for } y = 0, 1, 2, \ldots.$$

We are discovering the general *negative binomial* distribution in the process, of the form negative binomial

$$g(y,a,p) = \frac{\Gamma(a+y)}{\Gamma(a)\, y!}(1-p)^y p^a \quad \text{for } y = 0, 1, \ldots, \tag{1.4}$$

for parameters $a > 0$, $p \in (0,1)$; see Ex. 1.27 for more details.

(c) (xx one more thing here, perhaps even mixture of a small and a larger θ value. also point to Consider a regression context, with observed pairs (x_i, Y_i), where $Y_i \,|\, x_i$ has a distribution determined by $Y_i \,|\, \mu_i \sim \mathrm{Pois}(\mu_i)$ but $\mu_i \sim \mathrm{Gam}(\exp(x_i^{\mathrm{t}}\beta)/c, 1/c)$. show that $Y_i \,|\, x_i$ has mean $\exp(x_i^{\mathrm{t}}\beta)$ and inflated variance $\exp(x_i^{\mathrm{t}}\beta)(1+c)$.

Ex. 1.27 *The negative binomial.* We met the negative binomial distribution in Ex. 1.26 and now point to other features and constructions.

(a) Let X_1, X_2 be independent from the geometric distribution $q^x p$ for $x = 0, 1, \ldots$, with $q = 1-p$. Show that $Y = X_1 + X_2$ has distribution $\Pr(Y = y) = (y+1)q^y p^2$, for $y = 0, 1, \ldots$. For $Y = X_1 + X_2 + X_3$ a sum of three such independent geometric variables, show that $\Pr(Y = y) = \binom{y+2}{2} q^y p^3$ for $y = 0, 1, \ldots$.

(b) Generalise to the case of $Y = X_1 + \cdots + X_a$, the sum of a independent geometric variables, each with $q^x p$ for $x = 0, 1, \ldots$. Show that

$$\Pr(Y = y) = \binom{y+a-1}{a-1} q^y p^a = \frac{\Gamma(y+a)}{\Gamma(a)\, y!}(1-p)^y p^a \quad \text{for } y = 0, 1, \ldots,$$

i.e. the negative binomial with parameters (a, p). Deduce that the number of ways in which one may find nonnegative numbers $x_1, \ldots, x_a$ with a given sum y is $\binom{y+a-1}{a-1} = (y+a-1)!/\{(a-1)!\, y!\}$. In how many ways may one find 5 nonnegative numbers with sum 100? And with 10 nonnegative numbers with sum 100?

(c) How do we know that the negative binomial probabilities (1.4) sum to one, also when the a is a non-integer? Deduce from this that

$$\sum_{y=0}^{\infty} \frac{\Gamma(y+a)}{\Gamma(a)} \frac{u^y}{a!} = \frac{1}{(1-u)^a} \quad \text{for } u \in (0,1).$$

Show that $\mathrm{E}\, Y = aq/p$, $\mathrm{Var}\, Y = aq/p^2$.

(d) (xx the step from $Y = X_1 + \cdots + X_r$, counting from zero, to $Y' = X'_1 + \cdots + X'_a$, counting each from one, so that $Y' \geq a$. and just a bit more. reason for the negative binomial term. xx)

(e) In one of the episodes of the television series *Siffer* (NRK, 2011), programme leader Jo Røislien announced he would flip his coin and land 'krone' ten times in a row – which he then proceeded to do. He looked a bit tired, though; he had just kept on doing this, complete with his opening statement, until he had achieved the ten krone in a row event, and then showed only this crowning minute on tv. About how many times did he need to flip his coin, in total, before he (and his camera man) could show that final string of crowns? Simulate the process, and give a histogram of say 1,000 realisations.

Ex. 1.28 *Conditioning on Poisson sums and a generalised binomial.* We start with Poisson sums and see a connection to the binomial. Using generalised Poissons then lead to generalisations of the binomial.

(a) Let X and Y be independent Poissons with parameters θ_1, θ_2. Show that X given $X+Y = n$ is a binomal (n, p), with $p = \theta_1/(\theta_1+\theta_2)$. Generalise to the case of $X_1, \ldots, X_m$ being independent Poissons, with parameters $\theta_1, \ldots, \theta_m$. Show that their conditional distribution given $X_1 + \cdots + X_m = n$ is a multinomial with count n and probabilities $(p_1, \ldots, p_m)$, where $p_j = \theta_j/(\theta_1 + \cdots + \theta_m)$.

(b) To generalise the above, and in its turn also the binomial distribution, consider the gamma-mixed Poissons of 1.26. Specifically, let $X \,|\, \theta_1 \sim \mathrm{Pois}(\theta_1)$ and $Y \,|\, \theta_2 \sim \mathrm{Pois}(\theta_2)$, with gamma distributions $(c\theta_{0,1}, c)$ and $(c\theta_{0,2}, c)$ for θ_1 and θ_2; when c is large, this means tight concentration around $\theta_{1,0}, \theta_{0,2}$, and we're back to Poisson. Show that the conditional distribution of X given $X + Y = n$ may be written

$$f(x \,|\, n) \propto \binom{n}{x} \Gamma(c\theta_{0,1} + x)\Gamma(c\theta_{0,2} + n - x) \quad \text{for } x = 0, 1, \ldots, n,$$

where '$\propto$' means 'proportional to'. For a particular case, explain that with $c\theta_{0,1} = c\theta_{0,2} = 1$, X has the uniform distribution on $0, 1, \ldots, n$.

(c) It may not be easy to sum the terms above directly, to find the normalisation constant, but show via expressions in Ex. 1.21 that we in fact must have

$$f(x \,|\, n) = \binom{n}{x} \frac{\Gamma(c\theta_{0,1} + c\theta_{0,2})}{\Gamma(c\theta_{0,1})\Gamma(c\theta_{0,2})} \frac{\Gamma(c\theta_{0,1} + x)\Gamma(c\theta_{0,2} + n - x)}{\Gamma(c\theta_{0,1} + c\theta_{0,2} + n)}$$

for $x = 0, 1, \ldots, n$. In yet other words, we have reinvented the Beta-binomial distribution with gamma-mixing of Poisson parameters. Also, for large c, we are back to plain binomial (n, p), with $p = \theta_{0,1}/(\theta_{0,1} + \theta_{0,2})$.

(d) To invent other generalisations of the binomial, therefore, we might attempt other extensions of the Poisson. One such is to let X and Y have point probabilities proportional to $\theta_1^x/(x!)^\gamma$ and $\theta_2^x/(y!)^\gamma$, for some $\gamma \neq 1$; we work with such models in Ex. 4.34 and Story iv.6. Show here that X given $X+Y=n$ has the distribution

$$f(x \,|\, n) \propto \binom{n}{x}^\gamma p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n,$$

where $p = \theta_1/(\theta_1+\theta_2)$. Compute these probabilities, for say $n = 50$, $p = 0.33$, and some values of γ around 1. Check then numerically $np_0 = \mathrm{E}\, X$ and the dispersion ratio $\rho = \mathrm{Var}\, X/\{np_0(1-p_0)\}$, to learn that there is overdispersion and underdispersion, compared to the binomial case, for $\gamma < 1$ and $\gamma > 1$, respectively.

Moments, moment-generating functions, characteristic functions

Ex. 1.29 *Moments.* Consider a random variable X with c.d.f. F. Its mean is $\mathrm{E}\, X = \int x \,\mathrm{d}F(x)$, and we may of course define higher moments.

(a) Use results of Ex. A.15 to show that $\mathrm{E}\, X^k$, first seen as $\int y \,\mathrm{d}G_k(y)$, the mean of the variable $Y = X^k$ with distribution G_k inherited from F, is also the same as $\int x^k \,\mathrm{d}F(x)$; thus there is no ambiguity there.

(b) For $r < s$, show that $(\mathrm{E}\,|X|^r)^{1/r} \leq (\mathrm{E}\,|X|^s)^{1/s}$, i.e. $h(r) = (\mathrm{E}\,|X|^r)^{1/r}$ is a non-decreasing function in r. You may use the Jensen inequality (see e.g. Ex. 8.8). In particular, note that if $|X|$ has a finite s-moment, then all moments of smaller order are also finite. Illustrate by computing and graphing $h(r)$ and $\log h(r)$ for the case of $X \sim \mathrm{Expo}(1)$.

(c) For X a standard normal, show that $\mathrm{E}\,|X|^p = 2^{p/2}\Gamma(\frac{1}{2}(p+1))/\sqrt{\pi}$ for $p \geq 0$, a formula that also can be written $(\frac{1}{2})^{p/2}\Gamma(p+1)/\Gamma(\frac{1}{2}p+1)$. Compute and graph the function $h(r)$ for this case. For p an even integer, the formula simplifies to $(\frac{1}{2})^{p/2}p!/(\frac{1}{2}p)!$.

(d) For a random variable X with finite fourth moment, we have defined its skewness skew and kurtosis kurt in Ex. 1.3. Give expressions skew $= h_3(\mu_1, \mu_2, \mu_3)$ and kurt $= h_4(\mu_1, \mu_2, \mu_3, \mu_4)$ in terms of the moments $\mu_j = \mathrm{E}\, X^j$, and also expressions skew $= h_3^*(\mu_2^*, \mu_3^*)$ and kurt $= h_4^*(\mu_2^*, \mu_3^*, \mu_4^*)$ in terms of the centralised moments $\mu_j^* = \mathrm{E}\,(X - \mu_1)^j$.

Ex. 1.30 *Moment-generating functions.* (xx nils has lifted this from App, need post-polish there and here. xx) For a random variable Y, with distribution P, its m.g.f. is

$$M(t) = \mathrm{E}\, \exp(tY) = \int \exp(ty) \,\mathrm{d}P(y),$$

defined for each t at which the expectation exists. The moment-generating function is useful for finding and characterising distributions, for finding their moments, for handling the distributions of sums of variables, and in connection with distributional limits. When Y has a density $f(y)$ (with respect to Lebesgue measure), we have $M(t) = \int \exp(ty) f(y) \,\mathrm{d}y$, and if it is discrete with pointmasses $f(y)$ for sample space S, say,

then $M(t) = \sum_{y \in S} \exp(ty) f(y)$. The expectation operator is more general, however, and $M(t)$ is perfectly defined also for intermediate cases where Y can have both discrete and continuous parts; see Ex. A.15.

(a) For a standard normal $Y \sim \mathrm{N}(0,1)$, show that $M(t) = \exp(\frac{1}{2}t^2)$. When $Y \sim \mathrm{N}(\mu, \sigma^2)$, derive $M(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.

(b) A moment-generating function has that name since it generates moments; we indeed have $M'(0) = \mathrm{E}\, Y$, $M''(0) = \mathrm{E}\, Y^2$, etc., under mild conditions, see Ex. A.31 for details. Use this to find the first four moments of the standard normal.

(c) For $Y \sim \mathrm{Expo}(\theta)$, show that $M(t) = 1/(1 - t/\theta)$, for $t < \theta$.

(d) For $Y \sim \mathrm{Gam}(a, b)$, with density $\{b^a/\Gamma(a)\} y^{a-1} \exp(-by)$, show that $M(t) = \{b/(b - t)\}^a$, for $t < b$. In particular, $M(t) = 1/(1-t)^a$ for $\mathrm{Gam}(a, 1)$.

(e) Suppose Y is equal to zero with probability 0.90, but a standard normal with probability 0.10. Find the $M(t)$, and generalise.

(f) For the binomial (n, p), show that $M(t) = \{1 - p + p \exp(t)\}^n$.

(g) For $Y \sim \mathrm{Pois}(\theta)$, find $M(t) = \exp\{\theta(e^t - 1)\}$. Use this, with Ex. 1.26, to find $M(t)$ also for the negative binomial (a, p). (xx hm, should give the formula here. xx)

(h) Let $Y = \pm 1$ with probabilities $\frac{1}{2}, \frac{1}{2}$. Show that

$$M(t) = \cosh(t) = \tfrac{1}{2}(e^t + e^{-t}) = 1 + (1/2)t^2 + (1/4!)t^4 + (1/6!)t^6 + \cdots .$$

(i) For the uniform distribution on the unit interval, show that $M(t) = \{\exp(t) - 1\}/t$, for $t \neq 0$, and with $M(0) = 1$. For Y having the uniform distribution on the $[-1, 1]$ interval, show that

$$M(t) = \frac{\exp(t) - \exp(-t)}{2t} = \frac{\sinh t}{t},$$

and that this function may be written as the infinite sum $1 + (1/3!)t^2 + (1/5!)t^4 + \cdots$.

Ex. 1.31 *Distribution of sums via moment-generating functions.* Importantly, the m.g.f. $M(t) = \mathrm{E} \exp(tY)$, if it exists in a neighbourhood around zero, characterises the distribution; variables whose m.g.f.s are identical in such a neighbourhood have identical distributions. See details in Ex. A.31.

(a) Suppose $M(t) = (\frac{1}{2})^{25}\{1 + \exp(t)\}^{25}$. What is the underlying distribution? Similarly, if $M(t) = 0.99 + 0.01 \exp(\frac{1}{2}t^2)$, what is the distribution?

(b) With X and Y being independent, with m.g.f.s M_1 and M_2, show that the sum $Z = X + Y$ has m.g.f. $M_X(t) M_Y(t)$.

(c) With X and Y independent and standard normal, show that $X + Y$ is $\mathrm{N}(0, 2)$. Redo the questions of Ex. 1.2.

(d) Redo a part of Ex. 1.23, showing that sums of independent Poissons are Poisson.

(e) If $X_1, \ldots, X_n$ are independent gamma variables, with parameters $(a_1, b), \ldots, (a_n, b)$, find the distribution of $X_1 + \cdots + X_n$.

the Laplace distribution

Ex. 1.32 *The Laplace distribution.* The Laplace or double exponential distribution, in its simplest form, has density $f_0(y) = \frac{1}{2}\exp(-|y|)$, on the real line; note the cusp at its centre point zero.

(a) Let V_1 and V_2 be independent standard exponentials. Show that $Y = V_1 - V_2$ has this density $f_0(y)$. Deduce from this that its m.g.f. is $M_0(t) = 1/(1-t^2)$, for $|t| < 1$.

(b) More generally, consider $Y = V_1 - V_2$ where these two are independent and Expo(θ). Show that Y has density $f(y) = \frac{1}{2}\theta\exp(-\theta\,|y|)$, with zero mean and variance $2/\theta^2$. Also, show that its m.g.f. is $M(t) = 1/\{1-(t/\theta)^2\}$ for $|t| < \theta$. The Laplace with variance 1 is hence that with $\theta = \sqrt{2}$.

(c) Suppose X for given σ is a $\mathrm{N}(0, \sigma^2)$, but that the variance $V = \sigma^2$ has some distribution. Show that the m.g.f. for such a normal scale mixture becomes $M(t) = \mathrm{E}\exp(tX) = M_V(\frac{1}{2}t^2)$, where $M_V(s)$ is the m.g.f. for V. In particular, show that if $X \,|\, \sigma \sim \mathrm{N}(0, \sigma^2)$ and $\sigma^2 \sim \mathrm{Expo}(1)$, then X has the Laplace distribution with variance 1.

(d) If $X \,|\, V \sim \mathrm{N}(0, V)$, and V has density $g(v)$, show that X has density $f(x) = \int_0^\infty \phi(x/v^{1/2})(1/v^{1/2})g(v)\,\mathrm{d}v$. Translate the result above to the interesting formula

$$\begin{aligned}\int_0^\infty \frac{1}{(2\pi)^{1/2}}\exp\Big(-\frac{1}{2}\frac{x^2}{v}\Big)\frac{\exp(-v)}{v^{1/2}}\,\mathrm{d}v &= 2\int_0^\infty \frac{1}{(2\pi)^{1/2}}\exp\{-(\tfrac{1}{2}x^2/w^2 + w^2)\}\,\mathrm{d}w \\ &= \tfrac{1}{2}\sqrt{2}\exp(-\sqrt{2}|x|).\end{aligned}$$

Use this to find a formula for the integral $\int_0^\infty \exp\{-(av^2 + b/v^2)\}\,\mathrm{d}v$.

Ex. 1.33 *Characteristic functions.* Above we have found multiple uses for moment-generating functions, as in Ex. 1.30, and also their close cousin the Laplace transform, see Ex. 1.35, along with generating functions for distributions on the integers. Yet another very useful transform is the characteristic function, worked with in App. A, see Ex. A.34, which is also a crucial technical tool in the development of the large-sample theory of Ch. 2. For a variable with distribution F, its definition is $\varphi(t) = \mathrm{E}\exp(itX) = \int\{\cos(tx) + i\sin(tx)\}\,\mathrm{d}F(x)$, with $\mathrm{d}F(x)$ to be read as $f(x)\,\mathrm{d}x$ when F has a density f. Notably, this $\varphi(t)$ always exists, also for distributions without means, etc. We note here the inversion theorem from Ex. **??**, that if $\int|\varphi(t)|\,\mathrm{d}t$ is finite, then there is a density $f(x) = (2\pi)^{-1}\int\exp(-itx)\varphi(t)\,\mathrm{d}t$, see Ex. **??**, which is also seen to be continuous.

(a) For the standard normal, show that $\varphi(t) = \exp(-\frac{1}{2}t^2)$, and that $\varphi(t) = \exp(-\frac{1}{2}\sigma^2t^2)$ for the $\mathrm{N}(0, \sigma^2)$.

(b) Show that $\varphi(t)$ is real, and then equal to $\mathrm{E}\cos(tX)$, if and only if the distribution is symmetric around zero.

(c) For U a unit exponential, show that $\varphi(t) = 1/(1-it)$. Deduce that $X = U - V$, with its Laplace distribution, has $\varphi(t) = 1/(1+t^2)$, and that

$$\tfrac{1}{2}\exp(-|x|) = (2\pi)^{-1}\int\cos(tx)\frac{1}{1+t^2}\,\mathrm{d}t.$$

(d) Show from this, essentially by changing from (x,t) to (t,x), that the standard Cauchy has $\varphi(t) = \exp(-|t|)$.

(e) In Ex. 1.16(d) we saw that if a ball is kicked from distance r to an infinitely long straight line, with random angle A being uniform on $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$, then the position where the ball crosses the line is $X = rX_0$, where X_0 is Cauchy. Now assume r is not fixed, but comes from a unit exponential distribution. Show that X, a scale mixture of Cauchys, has characteristic function $\phi(t) = 1/(1+|t|)$.

(f) For V a uniform on $[-1,1]$, show $\varphi(t) = (\sin t)/t$. Prove from this that even though $\int (\sin t)/t \, \mathrm{d}t = \pi$, the integral $\int |(\sin t)/t| \, \mathrm{d}t$ must be infinite. Show that

$$f(x) = (2\pi)^{-1} \int \cos(tx) \Big(\frac{\sin t}{t} \Big)^2 \, \mathrm{d}t$$

is equal to the triangular density on $[-2,2]$.

(g) Just as for m.g.f.s, sums of independent components are associated with products of characteristic functions. Show that if X, Y, Z are independent, with $\varphi_1(t), \varphi_2(t), \varphi_3(t)$, then $X+Y+Z$ has $\varphi_1(t)\varphi_2(t)\varphi_3(t)$ as its characteristic function. Whem $X_1, \ldots, X_n$ are i.i.d. with characteristic function $\varphi(t)$, show that $(X_1 + \cdots + X_n)/\sqrt{n}$ has $\varphi_n(t) = \varphi(t/\sqrt{n})^n$. We learn in Ch. 2 that if the X_i have mean zero and finite variance σ^2, then $\varphi_n(t) \to \exp(-\frac{1}{2}\sigma^2 t^2)$. In this connection, you may numerically compute

$$f_n(x) = (2\pi)^{-1} \int \cos(tx) \Big(\frac{\sin t/\sqrt{n}}{t/\sqrt{n}} \Big)^n \, \mathrm{d}t$$

and verify that it tends to the $\mathrm{N}(0, 1/12)$ density; cf. Ex. **??**.

(h) Show that there cannot be two i.i.d. variables with sum having the uniform distribution. (xx but pointer to story with $U = V_1 + V_2$, two independent but with different distributions. xx)

(i) (xx be repaired. xx) If X, Y are independent with the same distribution, and $(X+Y)/\sqrt{2} \sim X$, show that X must be a zero-mean normal. (xx pointer to CLT. xx)

(j) (xx to be repaired. xx) For another characterisation lemma, suppose that $(X+Y)/2 \sim X$, and show by induction that this implies the curious property that with $X_1, \ldots, X_n$ i.i.d. from such a model, then the sample mean $\bar{X}_n$ has the same distribution for every n. Show that X must be Cauchy. Why is this not contradicting the LLN?

Ex. 1.34 *Cumulants and the cumulant-generating function.* For a variable X with m.g.f. $M(t)$, assumed finite in an interval around zero, it is sometimes fruitful to work with the *cumulant-generating function* $K(t) = \log M(t)$. When expanded in a power series around zero, with

cumulants

$$K(t) = K'(0)t + \tfrac{1}{2}K''(0)t^2 + \cdots = \sum_{j=1}^{\infty} \frac{\kappa_j}{j!} t^j,$$

we call the coefficients $\kappa_j = K^{(j)}(0)$ the *cumulants* of the distribution.

(a) For the $\mathrm{N}(\xi, \sigma^2)$ distribution, show that $K(t) = \xi t + \frac{1}{2}\sigma^2 t^2$. For the unit exponential distribution, show $K(t) = t + t^2/2 + t^3/3 + \cdots$, with $\kappa_j = (j-1)!$ for $j \geq 1$.

(b) If X has mean ξ, write $X = \xi + X_0$, Show that $K(t) = \xi t + K_0(t)$, where $K_0(t) = \log M_0(t)$, with M_0 the m.g.f. of the zero-mean variable X_0. Hence the cumulants κ_j for X are the same as the cumulants $\kappa_{j,0}$ for X_0, for $j \geq 2$.

(c) Via successive derivatives of $M(t) = \exp\{K(t)\}$, show that

$$\begin{aligned} M' &= MK', \\ M'' &= M\{K'' + (K')^2\}, \\ M''' &= M\{K''' + 3K''K' + (K')^3\}, \\ M'''' &= M\{K'''' + 4K'''K' + 6K''(K')^2 + 3(K'')^2 + (K')^4\}. \end{aligned}$$

Write ξ and σ^2 for the mean and variance of X. From the equations, show first that $K'(0) = \kappa_1 = \xi$ and that $K''(0) = \kappa_2 = \sigma^2$, and next that

$$\mathrm{E}\, X^3 = \kappa_3 + 3\xi\kappa_2 + \xi^3, \quad \mathrm{E}\, X^4 = \kappa_4 + 4\xi\kappa_3 + 6\xi^2\kappa_2 + 3\kappa_2^2 + \xi^4.$$

(d) Skewness and kurtosis were defined in Ex. 1.3. We may now find useful expressions for these in terms of the cumulants. Show first via formulae above that

$$\mathrm{E}\,(X-\xi)^3 = \kappa_3, \quad \mathrm{E}\,(X-\xi)^4 = \kappa_4 + 3\sigma^4.$$

Explain that this leads to $\mathrm{skew}(X) = \kappa_3/\sigma^3$, $\mathrm{kurt}(X) = \kappa_4/\sigma^4$.

Ex. 1.35 *Generating functions.* Moment-generating functions, studying distributions via the transformation $M(t) = \mathrm{E}\exp(tX)$, have several close relatives, which might be more convenient for certain classes of distributions. It is e.g. common to use *Laplace transformations* $L(s) = \mathrm{E}\exp(-sX)$ for distributions on $[0, \infty)$, then studied for $s \geq 0$. Here we work through the basic properties of *generating functions*, primarily used for distributions on the nonnegative integers. If $\Pr(Y = j) = p_j$, for $j = 0, 1, 2, \ldots$, define $G(s) = \mathrm{E}\, s^Y = \sum_{j=0}^{\infty} p_j s^j = p_0 + p_1 s + p_2 s^2 + \cdots$, called the generating function for that distribution, or for variables having that distribution.

Laplace transforms

generating functions

(a) Show that $G(s) = M(\log s)$, for s such that the latter exists. Demonstrate that $G(s)$ is finite, for $|s| < 1$, and also for $s = 1$. Find the generating functions for (i) the binomial (n, p); (ii) the Poisson with parameter θ; (iii) the geometric with $\Pr(Y = j) = q^{j-1}p$ for $j \geq 1$, with $q = 1 - p$; answers are

$$G_1(s) = (1 - p + ps)^n, \quad G_2(s) = \exp\{-\theta(1-s)\}, \quad G_3(s) = \frac{ps}{1-qs},$$

the latter valid for $|s| < 1/q$.

(b) Returning to the general case, give an expression for $G'(s)$, show that $G'(1) = \mathrm{E}\, Y$, and that $G''(1) = \mathrm{E}\, Y(Y-1)$. Find the mean and variance for the Poisson using generating functions.

(c) Suppose X and Y are random variables taking on values in $\{0, 1, 2, \ldots\}$, and that their generating functions are equal, on an interval around zero. Show that X and Y must have identical distributions.

(d) Show that if X, Y, Z are independent, with generating functions G_1, G_2, G_3, then the generating function for $X+Y+Z$ is $G_1(s)G_2(s)G_3(s)$. Show from this, and the previous point, yet again, that a sum of independent Poissons is a Poisson.

Ex. 1.36 *Getting Y from two copies of X.* Let X_1, X_2 be independently drawn from some distribution on the nonnegative integers, and from these draw Y from the binomial $(X_1 + X_2, \frac{1}{2})$.

(a) Writing ξ and σ^2 for the mean and variance for the X distribution, set up formulae for the conditional mean and variance of Y, given X_1, X_2. From these show that $\mathrm{E}\, Y = \xi$, $\mathrm{Var}\, Y = \frac{1}{2}\xi + \frac{1}{2}\sigma^2$.

(b) Write $G_0(s) = \mathrm{E}\, s^X$ for the generating function of the X_i. Show that the generating function for Y may be written $G(s) = G_0(\frac{1}{2} + \frac{1}{2}s)^2$. When $X_i \sim \mathrm{binom}(m, p)$, show from this that $Y \sim \mathrm{binom}(2m, \frac{1}{2}p)$. Then show that if the X_i are Poisson, then Y reproduces the same distribution.

Ex. 1.37 *Sums of random lengths and the compound Poisson.* Let $X_1, X_2, \ldots$ be i.i.d., from a distribution with mean ξ, variance σ^2, and m.g.f. $M_0(t)$. Consider then a random sum of these random elements; $Z = \sum_{i=1}^N X_i$, where N has some distribution with mean λ, variance τ^2, and generating function $G(s) = \mathrm{E}\, s^N$. We define Z as zero if $N = 0$.

(a) Show that Z has m.g.f. $M(t) = G(M_0(t))$. Show that Z has mean $\lambda\xi$ and variance $\lambda\sigma^2 + \xi^2\tau^2$.

compound Pòisson

(b) Consider the so-called compound Poisson variable $Z = \sum_{i=1}^N X_i$, where the X_i are i.i.d. with m.g.f. $M_0(s)$ and $N \sim \mathrm{Pois}(\lambda)$. Show that the m.g.f. of Z may be written

$$\mathrm{E}\, \exp(tZ) = \mathrm{E}\, M_0(s)^N = \exp[\lambda\{M_0(t) - 1\}],$$

and that this leads to mean $\lambda\xi$ and variance $\lambda(\xi^2 + \sigma^2)$. Find an expression for $\mathrm{E}\, \exp(tZ)$ for the particular case of $X_i \sim \mathrm{Gam}(a, b)$.

(c) Consider now $Z = \sum_{i=1}^N X_i$, where the X_i are i.i.d. unit exponentials. For given N, the Z is a $\mathrm{Gam}(N, 1)$. With $N \sim \mathrm{geom}(p)$, show that $Z \sim \mathrm{Expo}(p)$. Generalise.

(d) Using terminology and results from Ex. 1.34, show that the cumulant-generating function for the compound Poisson is $K(t) = \lambda\{M_0(t) - 1\}$, with cumulants $\kappa_j = \lambda\, \mathrm{E}\, X^j$. Show from this that $\mathrm{skew}(Z) = (1/\lambda)^{1/2}\gamma_3$ and $\mathrm{kurt}(Z) = (1/\lambda)\gamma_4$, with $\gamma_3 = \mathrm{E}\, X^3/(\mathrm{E}\, X^2)^{3/2}$ and $\gamma_4 = \mathrm{E}\, X^4/(\mathrm{E}\, X^2)^2$. For a Poisson sum $Z = \sum_{i=1}^N X_i$ of i.i.d. standard normals, so that $Z \,|\, (N = n) \sim \mathrm{N}(0, n)$, find the variance and kurtosis.

Ex. 1.38 *The logarithmic distribution.* Consider a variable X with point probabilities $\Pr(X = x) = c(p)^{-1}p^x/x$ for $x = 1, 2, \ldots$, with p a parameter in $(0, 1)$. This distribution is sometimes called the logarithmic distribution.

(a) Show that we must have $c(p) = -\log(1-p)$. Find expressions for the m.g.f. $M(t)$, its mean, and its variance. Comment on the cases where p is close to zero, or close to one. Show also for its generating function that $G(s) = \mathrm{E}\, s^X = c(ps)/c(p)$.

(b) Consider $Z = \sum_{i=1}^N X_i$, with the X_i being i.i.d. with this logarithmic distribution, and N is Poisson, with parameter expressed as $\lambda c(p)$. Find the mean and variance of Z, and show that its distribution is a negative binomial. We learn that the negative binomial is inside the class of compound Poissons.

The multinormal, the t, the chi-squared, the F

Ex. 1.39 *Mean and variance matrix for a random vector.* (xx calibrate with Story vii.1. xx) Consider a random vector of length p, say $Y = (Y_1, \ldots, Y_p)^{\mathrm{t}}$. Its mean vector is defined as the vector of means, i.e. $\xi = \mathrm{E}\, Y = (\mathrm{E}\, Y_1, \ldots, \mathrm{E}\, Y_p)^{\mathrm{t}}$, and its variance matrix Σ (also often called the covariance matrix), of dimension $p \times p$, has the variances on the diagonal and the covariances outside.

(a) Show that the elements of Σ are $\mathrm{cov}(Y_i, Y_j)$, for $i, j = 1, \ldots, p$. If we transform Y to $Z = AY + b$, with A a $m \times p$ matrix and b a vector of length m, show that $\mathrm{E}\, Z = A\xi + b$ and that $\mathrm{Var}\, Z = A\Sigma A^{\mathrm{t}}$. Explain that this generalises the usual rule $\mathrm{Var}\,(aY) = a^2 \mathrm{Var}\, Y$ for one-dimensional variables.

(b) Let X_1, X_2, X_3, X_4 be i.i.d. standard normals. Set up the variance matrix for $(X_1, X_1 + X_2, X_1 + X_2 + X_3, X_1 + X_2 + X_3 + X_4)^{\mathrm{t}}$.

(c) For the multinomial model studied in Ex. 1.5, with $Y = (Y_1, \ldots, Y_k)^{\mathrm{t}}$ counting the number of n events that fall in categories $1, \ldots, k$, with probabilities $p_1, \ldots, p_k$, show that $\mathrm{E}\, Y = np$, where p is the vector of $p_1, \ldots, p_k$. Show also that its variance matrix can be written $\Sigma = D - pp^{\mathrm{t}}$, where D is diagonal with elements p.

(d) For the multinomial model, with $\mathbf{1}$ the vector $(1, \ldots, 1)^{\mathrm{t}}$, show that $\mathbf{1}^{\mathrm{t}} D \mathbf{1} = 0$. Explain that this is related to the linear relationships $\sum_{j=1}^k p_j = 1$ and $\sum_{j=1}^k Y_j = n$. The matrix is hence not of full rank, and not invertible. Work however with the shorter vector $Y_0 = (Y_1, \ldots, Y_{k-1})^{\mathrm{t}}$, and show that its variance matrix may be written $\Sigma_0 = D_0 - p_0 p_0^{\mathrm{t}}$, where $p_0 = (p_1, \ldots, p_{k-1})^{\mathrm{t}}$ and D_0 is diagonal with these elements. Show that $\Sigma_0^{-1} = D_0^{-1} + (1/p_k)\mathbf{1}_0 \mathbf{1}_0^{\mathrm{t}}$, where $\mathbf{1}_0$ is the vector of $k-1$ 1s. For any vector $x = (x_1, \ldots, x_k)$ with sum zero, and with x_0 being the shortened version $(x_1, \ldots, x_{k-1})$, demonstrate that $x_0^{\mathrm{t}} \Sigma_0^{-1} x_0 = \sum_{j=1}^k x_j^2 / p_j$.

Ex. 1.40 *The multinormal distribution.* Let $Y = (Y_1, \ldots, Y_p)^{\mathrm{t}}$ be a random vector of length p. We say that it is multinormally distributed, with mean vector ξ and variance matrix Σ, which needs to be positive definite, provided its joint density can be written

$$f(y) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\tfrac{1}{2}(y-\xi)^{\mathrm{t}} \Sigma^{-1} (y-\xi)\},$$

where the domain for y is all of $\mathbb{R}^p$. We write $Y \sim \mathrm{N}_p(\xi, \Sigma)$ to indicate this distributon.

(a) Show that $\int y f(y)\, \mathrm{d}y$ indeed is equal to ξ, so calling it the mean vector is appropriate. Show also that $\mathrm{E}\,(Y-\xi)(Y-\xi)^{\mathrm{t}}$, calculated from the density, is equal to Σ.

(b) Show that if $Y \sim \mathrm{N}_p(\xi, \Sigma)$, then $Y - \xi \sim \mathrm{N}_p(0, \Sigma)$.

(c) Assume now that A is an invertible $p \times p$ matrix, and consider the transformation $Z = AY$. Show that if $Y \sim \mathrm{N}_p(\xi, \Sigma)$, then $Z = AY \sim \mathrm{N}_p(A\xi, A\Sigma A^{\mathrm{t}})$.

(d) By the spectral decomposition theorem of linear algebra, there is an orthonormal matrix P, with $PP^{\mathrm{t}} = I = P^{\mathrm{t}}P$, such that $P\Sigma P^{\mathrm{t}} = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, with these values being the eigenvalues of Σ. Show that $Z = P(Y - \xi)$ has components $Z_1, \ldots, Z_p$ which are independent, with $Z_j \sim \mathrm{N}(0, \lambda_j)$.

(e) Show that a vector Y is multinormal if and only if all linear combinations are normal. In particular, if $Y \sim \mathrm{N}_p(\xi, \Sigma)$, then $V = c^{\mathrm{t}}Y = c_1Y_1 + \cdots + c_pY_p$ is normal $\mathrm{N}(c^{\mathrm{t}}\xi, c^{\mathrm{t}}\Sigma c)$. [xx need to say something careful about allowing constants to be seen as normal, with zero variance. xx]

(f) Generalise point (c) to state that with any matrix A, of size say $q \times p$, the transformed $Z = AY$ is a multinormal $\mathrm{N}_q(A\xi, A\Sigma A^{\mathrm{t}})$.

(g) For the binormal case, with means ξ_1, ξ_2, standard deviations σ_1, σ_2, and correlation ρ, show that the density may be written

$$f(x,y) = \frac{1}{2\pi}\frac{1}{\sigma_1\sigma_2(1-\rho^2)^{1/2}}\exp\Big[-\tfrac{1}{2}\frac{1}{1-\rho^2}\Big\{\Big(\frac{x-\xi_1}{\sigma_1}\Big)^2 + \Big(\frac{y-\xi_2}{\sigma_2}\Big)^2 - 2\rho\Big(\frac{x-\xi_1}{\sigma_1}\Big)\Big(\frac{x-\xi_1}{\sigma_1}\Big)\Big\}\Big].$$

Show that X and Y are independent if and only if the correlation is zero.

(h) We learn that the situation is easy and clean for the multinormal case, where independence is equivalent to zero correlation. This is in general more complicated. (i) Let $X \sim \mathrm{N}(0,1)$ and $Y = DX$, with D a random sign with equal probabilities for $-1, 1$. Show that Y also is standard normal, that the correlation is zero, but that they are dependent. (ii) For another example, again with $X \sim \mathrm{N}(0,1)$, let $Y = X$ if $|X| \le a$ but $Y = -X$ if $|X| > a$. Show that Y is standard normal. Show that the correlation $\rho(a)$ goes from -1 to 1 as a goes from zero to infinity; thus there is an a_0 for which the correlation is zero. Also, find this a_0 numerically (answer: 1.537). In these zero-correlation constructions, with normal marginals, the point is that there is not joint binormality.

Ex. 1.41 *The multinormal and conditional distributions.* Consider a multinormally distributed vector, of length $p+q$, blocked into subvectors of sizes p and q. Let us write this as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}_{p+q}\Big(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\Big).$$

(a) By carrying out a linear transformation, and using results from Ex. 1.40, show that

$$Z = Y_1 - \Sigma_{12}\Sigma_{22}^{-1}Y_2 \sim \mathrm{N}_p(\xi_1 - \Sigma_{12}\Sigma_{22}^{-1}\xi_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}),$$

and that this Z is independent of Y_2.

(b) Show that the distribution of Y_1 given $Y_2 = y_2$ must be multinormal. Derive the important formulae for the conditional mean and variance,

$$\mathrm{E}\,(Y_1 \,|\, y_2) = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \xi_2), \qquad \mathrm{Var}\,(Y_1 \,|\, y_2) \,=\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Note that the conditional mean is a linear function in y_2 and that the conditional variance matrix is constant, not depending on y_2.

(c) Now study the simplest two-dimensional prototype case, with

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, & \rho \\ \rho, & 1 \end{pmatrix}\Big).$$

Show that $Y \,|\, x \sim \mathrm{N}(\rho x, 1-\rho^2)$ and that $X \,|\, y \sim \mathrm{N}(\rho y, 1-\rho^2)$. Discuss implications for situations where an easy to measure X might be a proxy for a harder to come by Y. How can you estimate Y from X, and with what precision?

(d) In generalisation of this special binormal case, consider a binormal

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} \xi \\ \xi_0 \end{pmatrix}, \begin{pmatrix} \sigma^2, & \rho\sigma\sigma_0 \\ \rho\sigma\sigma_0, & \sigma_0^2 \end{pmatrix}\Big).$$

In particular, the X is seen as stemming from a $\mathrm{N}(\xi_0, \sigma_0^2)$ distribution. Show that $Y \,|\, (X = x)$ is normal, with constant variance $\sigma^2(1-\rho^2)$ and linear mean function $\mathrm{E}\,(Y \,|\, x) = \xi + \rho(\sigma/\sigma_0)x$. This is essentially the linear regression model, for $Y \,|\, x$, here derived as a consequence of binormality. Without pretensions of having precise numbers, guess the binormal parameters in your population for (X, Y) being height and weight, and deduce from this a model predicting a person's weight from his or her height.

(e) Generalise this to situations with a $(p+1)$-dimensional normal distribution for $(Y, X_1, \ldots, X_p)$, with Y seen as the main outcome, influenced by covariates $X_1, \ldots, X_p$. Show that Y given these covariates is normal, with a constant variance, and find the linear conditional mean function $\mathrm{E}\,(Y \,|\, X_1, \ldots, X_p)$. This is essentially the linear multiple regression model; see Ex. 3.31.

(f) (xx a bit here on regression towards the mean. xx)

Ex. 1.42 *How tall is Nils?* Assume that the heights of Norwegian men above the age of twenty follow the normal distribution $\mathrm{N}(\xi, \sigma^2)$ with $\xi = 180$ cm and $\sigma = 9$ cm.

(a) Given this information only, what is your point estimate of his height, and what is your 95 percent prediction interval?

(b) Assume now that you learn that his four brothers are actually 195 cm, 207 cm, 196 cm, 200 cm tall, and furthermore that correlations between brothers' heights in the population of Norwegian men is equal to $\rho = 0.80$. Use this information about his four brothers to revise your initial point estimate of his height, and provide the updated 95 percent prediction interval. Is Nils a statistical outlier in his family?

(c) Suppose that Nils has n brothers and that you learn their heights. Give formulae for the updated normal parameters ξ_n and σ_n, in the conditional distribution of his height given these extra pieces of information. Use this to clarify the following statistical point: Even if you get to know all facts concerning 99 brothers, there should be a limit to your confidence in what you may infer about Nils.

Ex. 1.43 *The chi-squared.* This exercise goes through some basic properties of the chi-squared; see also Ex. 1.47 for its eccentric cousin, the noncentral or eccentric chi-squared.

(a) We say that a nonnegative variable X has the chi-squared distribution, with degrees of freedom m, and write $X \sim \chi^2_m$ for this, when its density takes the form

$$g_m(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} \exp(-\tfrac{1}{2}x) \quad \text{for } x > 0.$$

Show that its m.g.f. becomes $M(t) = (1-2t)^{-m/2}$, for $t < \frac{1}{2}$. Show further that $\mathrm{E}\, X = m$, $\mathrm{Var}\, X = 2m$, and that the skewness, i.e. $\mathrm{E}\, W^3$ with $W = (X - \mathrm{E}\, X)/(\mathrm{Var}\, X)^{1/2} = (X-m)/(2m)^{1/2}$, is $(8/m)^{1/2}$. From the density, show $\mathrm{E}\,(\chi^2_m)^p = 2^p\Gamma(m/2+p)/\Gamma(m/2)$.

(b) Via the m.g.f., show the simple and basic convolution property for the chi-squared, that if $X_1, \ldots, X_n$ are independent and chi-squared distributed with degrees of freedom $m_1, \ldots, m_n$, then the sum $Z = \sum_{i=1}^n X_i$ is chi-squared too, with degrees of freedom $\sum_{i=1}^n m_i$. A generalisation is given in Ex. 1.47.

(c) If N is standard normal, show that $X = N^2 \sim \chi^2_1$. Establish that if $X \sim \chi^2_m$, with m a natural number, then it may be represented as $X = N_1^2 + \cdots + N_m^2$, in terms of independent standard normals $N_1, \ldots, N_m$. Note, however, that the χ^2_m with the density $g_m(x)$ above may be used also when m is not a natural number.

(d) There are connections between the chi-squared and the Gamma distribution (see Ex. 1.9). Show that the $\mathrm{Gam}(\frac{1}{2}m, \frac{1}{2})$ is the χ^2_m; that if $Y \sim \mathrm{Gam}(a, b)$, then $X = 2bY \sim \chi^2_{a/2}$; and that if $Z \sim \chi^2_m$, then $\frac{1}{2}Z \sim \mathrm{Gam}(\frac{1}{2}m, 1)$.

(e) Consider independent $X \sim \chi^2_a$ and $Y \sim \chi^2_b$. Show, perhaps via Gamma distribution ratios in Ex. 1.12, that $R = X/(X+Y) \sim \mathrm{Beta}(\frac{1}{2}a, \frac{1}{2}b)$.

(f) When $X \sim \chi^2_m$, show that $\mathrm{E}\log X = \log 2 + \psi(\frac{1}{2}m)$, where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function.

(g) Consider $Z = XY$, a product of independent standard normals. Use (1.2) to show that its m.g.f. is $M(t) = 1/(1-t^2)^{1/2}$, for $|t| < 1$. Deduce from this that Z has the same distribution as $\frac{1}{2}(K - L)$, where K and L are independent with χ^2_1 distributions.

(h) Consider the second degree equation $x^2 + Bx + C = 0$ from school. If there are many such equations, with B and C being independent standard normal, how many of these equations will have both roots real?

Ex. 1.44 *Tell me about X and $X + Y$.* In exercises above we have seen that sums of independent normals, Poissons, chi-squares are respectively normal, Poisson, chi-square. Sometimes we meet questions going the other way.

(a) Suppose X and Y are independent, and that you learn the distributions of X and $Z = X + Y$. Show that Y must have m.g.f. $M_Y(t) = M_Z(t)/M_X(t)$. As an illustration, suppose $X \sim \mathrm{N}(0,1)$ and $X + Y \sim \mathrm{N}(0,2)$. Show that $Y \sim \mathrm{N}(0,1)$.

(b) Explain similarly that if X and Y are independent, with X and $X+Y$ being Poisson, then Y is also Poisson.

(c) Suppose X and Y are independent, that $X \sim \chi^2_a$ and that $Z = X + Y \sim \chi^2_{a+b}$. Show that the only possibility is then the expected one, that $Y \sim \chi^2_b$.

(d) Consider $X_1, \ldots, X_n$ i.i.d. standard normal. Writing as usual $\bar{X}$ for the sample mean, we know that $\sqrt{n}\bar{X}$ is also standard normal. Show that the vector of $X_i - \bar{X}$ is independent of $\bar{X}$. Writing $Q = \sum_{i=1}^n (X_i - \bar{X})^2$, show from $\sum_{i=1}^n X_i^2 = Q + n\bar{X}^2$ that $Q \sim \chi^2_{n-1}$. (xx calibrate with other things; this is giving a simpler proof of the lemma below, with no orthogonal transformations. xx)

(e) (xx for later, perhaps in Ch3: also nice for p-dimensional things with $Q(\theta) = \sum_{j=1}^k (y_j - \theta)^{\mathrm{t}} \Sigma_j^{-1} (y_j - \theta)$, where $Q_{\min} = Q(\widehat{\theta}) \sim \chi^2_{(k-1)p}$, independent of $\widehat{\theta} = A^{-1} \sum_{j=1}^k \Sigma_j^{-1} y_j$. xx)

Ex. 1.45 *Orthonormal transformations.* [xx check and calibrate with chi-squared things, to get the order right. restructure text. xx] We have seen in Ex. 1.40 that a multinormal vector can be sent via a linear transformation to independent one-dimensional normal components, and vice versa. This also leads to useful characterisation and representation theorems involving independence. In the present exercise we shall e.g. find a proof that the sample mean $\bar{Y}$ and the sample variance statistic $S = \sum_{i=1}^n (Y_i - \bar{Y})^2$ are independent; this fact, which does not hold outside the normal family, was actively used in Ex. 3.4, and will also be utilised (xx in other exercises, like Ex. 3.6 xx).

(a) Suppose $X = (X_1, \ldots, X_n)$ is a vector with i.i.d. and standard normal components, and let A be an orthonormal matrix, which means $AA^{\mathrm{t}} = I = A^{\mathrm{t}}A$. In yet other words, each row of A and each column of A has length 1, rows are ortogonal, as well as columns. Show that $Y = AX$ must have components $Y_1, \ldots, Y_n$ which are also i.i.d. and standard normal. – Here you may also use the general transformation formula of Ex. 1.12.

(b) To exemplify the above, show that if X_1, X_2 are independent and standard normal, then also Y_1, Y_2, where

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2}, & 1/\sqrt{2} \\ 1/\sqrt{2}, & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} (X_1 + X_2)/\sqrt{2} \\ (X_1 - X_2)/\sqrt{2} \end{pmatrix},$$

must be independent and standard normal.

(c) When A is orthonormal, show that it preserves length, so $\|Au\| = \|u\|$, for any vector u; here $\|u\|$ is Euclidean length, so $\|u\|^2 = u_1^2 + \cdots + u_n^2$.

(d) Let again $X_1, \ldots, X_n$ be i.i.d. standard normals. Construct an orthogonal matrix A by letting its first row be $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$, and define $Y = AX$. Then show that $Y_1 = \sqrt{n}\bar{X} = \sum_{i=1}^n X_i/\sqrt{n}$, and that

$$Z = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=2}^n Y_i^2.$$

(e) Conclude from this that (i) $\sqrt{n}\bar{X} \sim \mathrm{N}(0,1)$, (ii) $Z \sim \chi^2_{n-1}$, and (iii) $\bar{X}$ and Z are idependent. This was proven more directly in Ex. 1.44 above, via decompositions of chi-squares.

(f) Show that this implies the following classical and important properties, starting with an independent sample $Y_1, \ldots, Y_n$ from the $\mathrm{N}(\mu, \sigma^2)$: The statistics $\bar{Y}$ and $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2$ are independent, with $\bar{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$ and $Z \sim \sigma^2 \chi^2_{n-1}$. Show also from this that the classical empirical variance

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \tag{1.5}$$

is unbiased for the population variance, i.e. $\mathrm{E}\,\widehat{\sigma}^2 = \sigma^2$. Construct an unbiased estimator for σ, of the type $c_n \widehat{\sigma}$. [xx point to generalisations for the linear regression model. xx]

(g) Consider the general multinormal distribution $Y \sim \mathrm{N}_p(\xi, \Sigma)$, with invertible Σ. Show that $K = (Y - \xi)^{\mathrm{t}} \Sigma^{-1} (Y - \xi) \sim \chi^2_p$. Suppose $Y = y_{\mathrm{obs}}$ is observed, that Σ is known, but ξ unknown. Give a confidence region R such that $\xi \in R$ with probability 90 percent. How does this region shrink, if you observe 100 vectors from the multinormal, rather than merely 1?

Ex. 1.46 *The t distribution.* Consider independent variables $X \sim \mathrm{N}(0,1)$ and $K \sim \chi^2_m$. The ratio $t = X/(K/m)^{1/2}$ is then said to have the t distribution, with m degrees of freedom. We write $t \sim t_m$ to indicate this.

(a) Find the mean and variance of t.

(b) Show that its density can be written

$$g_m(x) = \frac{\Gamma((m+1)/2)}{\Gamma(m/2)} \frac{1}{\sqrt{m\pi}} \frac{1}{(1 + x^2/m)^{(m+1)/2}}.$$

Show that $g_m(x)$ tends to the standard normal density $\phi(x)$ as m increases, and explain why this is to be expected. Show also that the Cauchy distribution, see Ex. 1.16, is the t_1 distribution. (xx find the little fun fact from Hjort (1994). xx)

(c) Find also the skewness and kurtosis for the t_m distribution. In particular, show that the latter is $\mathrm{kurt} = 6/(m-4)$ for $m > 4$.

(d) Assume $Y_1, \ldots, Y_n$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$. With $\widehat{\sigma}$ the empirical standard deviation, from (1.5), show that

$$t = (\bar{Y} - \mu)/(\widehat{\sigma}/\sqrt{n})$$

has the t-distribution with $n-1$ degrees of freedom. This is the classic t-statistic dating all the way back to Student (1908),

Ex. 1.47 *The noncentral chi-squared.* Consider also the so-called noncentral chi-squared distribution, say $K \sim \chi^2_m(\lambda)$, with λ the excentre or eccentricity parameter; the case of $\lambda = 0$ corresponds to the ordinary $K \sim \chi^2_m$. It is the distribution of $Y_1^2 + \cdots + Y_m^2$, where the Y_i are independent normals, with $Y_i \sim \mathrm{N}(\mu_i, 1)$, and $\lambda = \sum_{i=1}^m \mu_i^2$.

(a) Show that the m.g.f. of $K \sim \chi^2_m(\lambda)$ may be written

$$M(t) = \mathrm{E}\,\exp(tK) = \frac{\exp\{\lambda t/(1-2t)\}}{(1-2t)^{m/2}} \quad \text{for } t < \tfrac{1}{2}.$$

(b) Also the noncentral chi-squared distributions have convolution properties, generalising those of Ex. 1.43. If $K_i \sim \chi^2_{m_i}(\lambda_i)$, and these are independent, for $i = 1, \ldots, n$, show that $\sum_{i=1}^n K_i$ is another noncentral chi-squared, with degrees of freedom $\sum_{i=1}^n m_i$ and excentre parameter $\sum_{i=1}^n \lambda_i$.

(c) Another property which can be established, remarkbly without yet having seen the density of a $\chi^2_m(\lambda)$, is the following, using Ex. 1.44: if X and Y and independent, with $X \sim \chi^2_{m_1}(\lambda_1)$ and $X + Y \sim \chi^2_{m_1+m_2}(\lambda_1 + \lambda_2)$, then by necessity $Y \sim \chi^2_{m_2}(\lambda_2)$.

(d) Its density can be expressed in several ways; show that this is one such valid formula:

$$f(k, m, \lambda) = \sum_{j=0}^{\infty} \{\exp(-\tfrac{1}{2}\lambda)(\tfrac{1}{2}\lambda)^j/j!\}\, g_{m+2j}(k),$$

where $g_{m+2j}(k)$ is the χ^2_{m+2j} density. In other words, the noncentral chi-squared is a Poisson mixture of central chi-squared distributions. Show that this entails the representation $K \,|\, (J = j) \sim \chi^2_{m+2j}$, where $J \sim \mathrm{Pois}(\frac{1}{2}\lambda)$. Also non-integer values of m are allowed here.

(e) Establish that for $K \sim \chi^2_m(\lambda)$, we have $\mathrm{E}\,K = m + \lambda$ and $\mathrm{Var}\,K = 2m + 4\lambda$. Show also that the skewness of K becomes $2^{3/2}(m + 3\lambda)/(m + 2\lambda)^{3/2}$. What is required in order for this skewness to tend to zero?

(f) Let $K = (\lambda^{1/2} + N)^2$, which has the $\chi^2_1(\lambda)$ distribution. Consider the normalised variable

$$\frac{K - (1+\lambda)}{(2+4\lambda)^{1/2}} = \frac{N^2 + 2\lambda^{1/2}N - 1}{(2+4\lambda)^{1/2}}.$$

Work out its m.g.f. and show that it tends to $\exp(\frac{1}{2}t^2)$ for growing λ.

(g) More generally, with $K \sim \chi^2_m(\lambda)$, work out a formula for the m.g.f. $M(t)$ for $Z = \{K - (m + \lambda)\}/(2m + 4\lambda)^{1/2}$. For any fixed m, show that $M(t) \to \exp(\frac{1}{2}t^2)$ as λ grows, and comment on this finding.

Ex. 1.48 *Noncentral chi-squared for empirical variances.* We saw in Ex. 1.45 that if $X_1, \ldots, X_n$ are i.i.d. $\mathrm{N}(a, 1)$, with common a, then $Z = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$, with consequences for the empirical variance estimator. Here are some fruitful generalisations.

(a) Let the X_i have non-identical means, $X_i \sim \mathrm{N}(a_i, 1)$. Show that $Z \sim \chi^2_{n-1}(\lambda)$, with noncentrality parameter $\lambda = \sum_{i=1}^n (a_i - \bar{a})^2$.

(b) Assume now that $X_i \sim \mathrm{N}(a_i, 1/m_i)$ for $i = 1, \ldots, n$, perhaps reflecing sample sizes m_i for different groups, and with $M = \sum_{i=1}^n m_i$. Consider $Z = \sum_{i=1}^n m_i(X_i - \widetilde{X})^2$, with $\widetilde{X} = \sum_{i=1}^n (m_i/M)X_i$. Show that $Z = \sum_{i=1}^n m_i X_i^2 - M\widetilde{X}^2$, and that its distribution is a $\chi^2_{n-1}(\lambda)$, with $\lambda = \sum_{i=1}^n m_i(a_i - \widetilde{a})^2$, where $\widetilde{a} = \sum_{i=1}^n (m_i/M)a_i$.

(c) Let $X \sim \mathrm{N}_p(\xi, \Sigma)$. Show that $Z = X^{\mathrm{t}}\Sigma^{-1}X \sim \chi^2_p(\xi^{\mathrm{t}}\Sigma^{-1}\xi)$.

Ex. 1.49 *The F distribution.* As we have seen Ex. 1.45, with a normal sample $X_1, \ldots, X_n$, the distribution of the classical empirical variance estimator $\widehat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is governed by $\widehat{\sigma}^2 \sim \sigma^2\chi^2_m/m$, where $m = n-1$ is the degrees of freedom. Suppose there are two independent samples, from normal distributions $\mathrm{N}(\xi_1, \sigma_1^2)$ and $\mathrm{N}(\xi_2, \sigma_2^2)$, of sample sizes n_1 and n_2, with estimators $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$.

(a) Let $\rho = \sigma_1/\sigma_2$, the ratio of standard deviations. For the ratio of the two empirical variances, show that

$$R^2 = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} = \rho^2 F, \quad \text{where } F \sim \frac{\chi^2_{m_1}/m_1}{\chi^2_{m_2}/m_2},$$

with degrees of freedom $m_1 = n_1 - 1$ and $m_2 = n_2 - 1$, and with the two chi-squareds being independent. We say that F has the F distribution, or Fisher distribution, with degrees of freedom (m_1, m_2), and write $F \sim F(m_1, m_2)$.

(b) Show that when $F \sim F(m_1, m_2)$, then $1/F \sim F(m_2, m_1)$. Show furthermore that

$$\mathrm{E}\, F = \frac{m_2}{m_2 - 2}, \quad \mathrm{E}\, F^2 = \frac{m_1 + 2}{m_1}\frac{m_2^2}{(m_2 - 2)(m_2 - 4)},$$

these expressions being finite when $m_2 > 2$ and $m_2 > 4$, respectively. Find also an expression for the variance. Verify that both $\mathrm{E}\, F$ and $\mathrm{E}\, F^2$ tend to 1 as the degrees of freedom increase.

(c) The main aspects of the F distribution have been worked out, above, from the constructive definition given in ((a)), without actually needing any formula for its density; also, probabilities are found using software packages, like `pf(x,m1,m2)` in `R`. Once in a while one needs the density function, however. Show first that the cumulative function can be written

$$\Pr(F \le x) = \Pr(\chi^2_{m_1}/m_1 \le x\chi^2_{m_2}/m_2) = \int_0^\infty G(xy(m_1/m_2), m_1)g(y, m_2)\,\mathrm{d}y,$$

in terms of the cumulative $G(\cdot, m)$ and densiy $g(\cdot, m)$ of the χ^2_m. Then take the derivative to get

$$h(x, m_1, m_2) = \int_0^\infty g(xy(m_1/m_2), m_1)y(m_1/m_2)g(y, m_2)\,\mathrm{d}y.$$

Complete the math to land at

$$h(x, m_1, m_2) = \frac{\Gamma(\frac{1}{2}(m_1 + m_2))}{\Gamma(\frac{1}{2}m_1)\Gamma(\frac{1}{2}m_2)}(m_1/m_2)^{m_1/2}\frac{x^{m_1/2-1}}{\{1 + (m_1/m_2)x\}^{(m_1+m_2)/2}}.$$

The exponential family class

Ex. 1.50 *The exponential family class.* Many parametric models fall under the wide umbrella of the exponential family class, which we treat in this and the folllowing exercises. This will be properly generalised and extended down the road, but we start with this definition: Suppose Y has model density of the form

exponential family

$$\begin{aligned} f(y,\theta) &= \exp\{\theta_1 T_1(y) + \cdots \theta_p T_p(y) - k(\theta_1,\ldots,\theta_p)\}h(y) \\ &= \exp\{\theta^{\mathrm{t}} T(y) - k(\theta)\}h(y), \end{aligned} \tag{1.6}$$

for appropriate functions $T_1(y),\ldots,T_p(y)$ and $h(y)$; the $k(\theta)$ function is there to secure integration to one. We require that there is no fixed linear relationship among $T_1(y),\ldots,T_p(y)$, and also that the support of the distribution, the smallest closed set having probability 1, is the same, for all parameter values. We then say Y is of the exponential family class, with *data functions* $T(y) = (T_1(y),\ldots,T_p(y))^{\mathrm{t}}$ and *natural parameters* $\theta = (\theta_1,\ldots,\theta_p)^{\mathrm{t}}$.

(a) Before we start developing the general theory for the full class, we verify that a few classic models are under its umbrella. For the following models, write the model density in a form matching (1.6). (i) $Y \sim \text{binom}(n,p)$. (ii) $Y \sim \text{Pois}(\theta)$. (iii) $Y \sim \text{Beta}(a,b)$. (iv) $Y \sim \text{Gam}(a,b)$. (v) $Y \sim \text{N}(\xi,\sigma^2)$, with known σ (and see Ex. 1.51). (vi) Let (X,Y,Z) be trinomial (n,p,q,r), with $r = 1-p-q$. First work with $f(x,y,p,q)$, and then with the submodel where $p = a^2$, $q = 2a(1-a)$, $r = (1-a)^2$.

(b) Show that we must have

$$k(\theta) = \log\Big(\int \exp\{\theta^{\mathrm{t}} T(y)\}h(y)\,\mathrm{d}y\Big),$$

assumed to be finite for at least some θ. Let in fact Ω be the set of θ such that $k(\theta)$ is finite, called the natural parameter region. Show that Ω is a convex set.

natural parameter region

(c) The score function of a model is $u(y,\theta) = \partial \log f(y,\theta)/\partial\theta$; see Ex. 5.14 for more on this. For the class of models studied here, show that $u(y,\theta) = T(y) - \xi(\theta)$, where

$$\xi(\theta) = \frac{\partial k(\theta)}{\partial\theta} = \frac{\int T(y)\exp\{\theta^{\mathrm{t}} T(y)\}h(y)\,\mathrm{d}y}{\int \exp\{\theta^{\mathrm{t}} T(y)\}h(y)\,\mathrm{d}y}.$$

Show that the score function must have mean zero, which here means $\mathrm{E}_\theta\, T(Y) = \xi(\theta)$. Show also that $\mathrm{Var}_\theta\, T(Y) = \partial^2 k(\theta)/\partial\theta\partial\theta^{\mathrm{t}}$, giving variances and covariances of the $T_j(Y)$ in one matrix formula.

(d) We come back to general likelihood theory in Ch. 5, but for now show that if $Y_1,\ldots,Y_n$ are i.i.d. from the exponential family density, then the logarithm of the joint density can be written

$$\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i,\theta) = n\{\theta^{\mathrm{t}}\bar{T} - k(\theta)\},$$

with $\bar{T} = (1/n)\sum_{i=1}^n T(Y_i)$ the vector of averages $\bar{T}_j = (1/n)\sum_{i=1}^n T_j(y_i)$. If this is a family with say $p = 3$ parameters, and $n = 10000$, then the full relevant information required for computing the ℓ_n function is captured in the 3 averages $\bar{T}_1,\bar{T}_2,\bar{T}_3$. This is related to the sufficiency concept, returned to in Ch. 5. Show also that this $\ell_n(\theta)$ is a concave function.

Ex. 1.51 *The normal and binormal exponential family members.* The normal and binormal (and multinormal) classes of distributions belong under the exponential family umbrella.

(a) Consider first the normal (ξ, σ^2) model. Explain that its density may be written

$$f(y) = \frac{1}{(2\pi)^{1/2}} \exp\Big(-\tfrac{1}{2}\frac{y^2}{\sigma^2} + \frac{\xi y}{\sigma^2} - \tfrac{1}{2}\frac{\xi^2}{\sigma^2}\Big),$$

and argue from this that the normal is of exponential family type, with natural parameters $1/\sigma^2$ and ξ/σ^2, and data functions equivalent to (y, y^2).

(b) Consider then (X, Y) binormal, with its five parameters $\xi_1, \xi_2, \sigma_1, \sigma_2, \rho$, see Ex. 1.40. Show that the distribution is of the exponential family type, with natural parameters

$$\frac{1}{(1-\rho^2)\sigma_1^2}, \frac{\xi_1}{(1-\rho^2)\sigma_1^2}, \frac{1}{(1-\rho^2)\sigma_2^2}, \frac{\xi_2}{(1-\rho^2)\sigma_2^2}, \frac{\rho}{(1-\rho^2)\sigma_1\sigma_2}.$$

In the terminology of exponential families, identify also the associated $T_1(x, y), \ldots, T_5(x, y)$.

(c) Show that $(\xi_1, \xi_2, \sigma_1, \sigma_2, \rho)$ is in one-to-one correspondence with the five exponential class natural parameters above.

(d) Suppose now that $\xi_1 = \xi_2$ and $\sigma_1 = \sigma_2$, so this distribution for (X, Y) has now three parameters. Show that it is inside the exponential family, and identify its natural parameters.

Ex. 1.52 *A log-linear density model on the unit interval.* The machinery of the exponential family class makes it easy to construct new models, starting with relevant data functions T_j, in the language of Ex. 1.50.

(a) For densities on the unit interval, start with data function $T(y) = y - \frac{1}{2}$, to define $f(y, \theta) = \exp\{\theta T(y) - k(\theta)\}$ for $y \in [0, 1]$. Find the $k(\theta)$, a formula for the c.d.f. $F(y, \theta)$, and the mean and variance for Y having this density.

(b) Then go on to

$$f(y, \theta_1, \theta_2) = \exp\{\theta_1(y - \tfrac{1}{2}) + \theta_2(y - \tfrac{1}{2})^2 - k(\theta_1, \theta_2)\},$$

i.e. using $T_j(y) = (y - \frac{1}{2})^j$ as data functions, for $j = 1, 2$. Set up an integral function for $k(\theta_1, \theta_2)$ and some algorithms for determining the means, variances, and covariance, for $T_1(Y), T_2(Y)$, for any given (θ_1, θ_2).

(c) Spell out the necessary details for the third order log-linear density model, which uses $T_j(y) = (y - \frac{1}{2})^j$ for $j = 1, 2, 3$. Again, set up algorithms so that you may compute the means, variances, covariances for T_1, T_2, T_3, for any $(\theta_1, \theta_2, \theta_3)$.

Yet other models: log-normal, Weibull, Gompertz, Gumbel, et al.

the log-normal distribution

Ex. 1.53 *The log-normal distribution.* Starting with $X \sim \mathrm{N}(\xi, \sigma^2)$, the variable $Y = \exp(X)$ is said to be a *log-normal*, and we write $Y \sim \log\mathrm{N}(\xi, \sigma^2)$ to indicate this.

(a) Consider the view that the distribution should or could have been named the exponormal instead – would you agree? Show that with $Y \sim \log\mathrm{N}(\xi, \sigma^2)$, its mean and variance are $\exp(\xi + \frac{1}{2}\sigma^2)$ and $\{\exp(2\sigma^2) - \exp(\sigma^2)\}\exp(2\xi)$.

(b) Show that its density may be written $\phi_\sigma(\log y - \xi)/y = \sigma^{-1}\phi(\sigma^{-1}(\log y - \xi))/y$, for $y > 0$. Find its mode.

(c) Assume that $Y \,|\, \xi \sim \text{logN}(\xi, \sigma^2)$, and that $\xi \sim \text{N}(\xi_0, \tau^2)$. Show that marginally, $Y \sim \text{logN}(\xi_0, \sigma^2 + \tau^2)$. Make explicit the connection to Ex. 1.59.

(d) Show that a product of independent log-normals is log-normal. Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the $\text{logN}(\xi, \sigma^2)$ distribution. Explain what happens to their harmonic mean, $Z_n = (Y_1 \cdots Y_n)^{1/n}$.

(e) Assume a random time variable T has the $\text{logN}(0, 1)$ distribution. Find a formula for its hazard rate $h(t)$, and show that $h(t) \doteq (\log t)/t$ for growing t. Plot the exact hazard rate, along with its approximation, and comment.

Ex. 1.54 *The Weibull distribution.* The Weibull distribution, with positive parameters (a, b), has c.d.f. $F(t) = 1 - \exp\{-(t/a)^b\}$ for $t \geq 0$. The b is called the shape parameter, with a a scale parameter. The Weibull generalises the exponential distribution, which is the special case of $b = 1$. Other parametrisations are sometimes convenient, as with $1 - \exp(-ct^b)$.

(a) Find a formula for the median, and more generally for the q-quantile $F^{-1}(q)$. Show that the density can be written $f(t) = \exp\{-(t/a)^b\}bt^{b-1}/a^b$ for $t > 0$. Find also a formula for the hazard rate, and draw this in a diagram, for $b = 0.9, 1.0, 1.1$, say for $a = 1$.

(b) Use the general mean formula of Ex. 1.14(a) to work through the details of

$$\text{E}\, T^p = \int_0^\infty \Pr(T^p \geq u)\, \text{d}u = \int_0^\infty \exp\{-(u^{1/p}/a)^b\}\, \text{d}u = a^p\, \Gamma(1 + p/b).$$

Show that this leads to mean $a\,\Gamma(1 + 1/b)$ and variance $a^2\,\{\Gamma(1 + 2/b) - \Gamma(1 + 1/b)\}^2$. With T from the Weibull (a, b), plot the function $\text{sd}(T)/\text{E}\, T$ as a function of b. (xx write out. also reparametrisation, with $1 - \exp(-ct^b)$. point to story. xx)

(c) Show that $V = (T/a)^b \sim \text{Expo}(1)$, and use this to give a recipe for simulating outcomes from any Weibull.

Ex. 1.55 *The Gompertz distribution.* The Gompertz distribution, with positiver parameters (a, b), has hazard rate $h(t) = a \exp(bt)$.

(a) Find the cumulative hazard rate, the c.d.f., and the density. Find also a formula for the median, and more generally for the q quantile, expressed via (a, b). (xx pointer to Story ii.1. more; round off. xx)

(b) Suppose an individual has survived up to time t_0. Show that her cumulative hazard rate, for the remaining lifetime, is $H(t) - H(t_0) = (a/b)\{\exp(bt) - \exp(bt_0)\}$. Give a formula for $t^*(t_0)$, her median survival time. (xx then brief application of this, for Norwegian women, using perhaps rough estimates of (a, b), via data from Human Mortality Index. give $(t_0, t^*(t_0))$ as a graph, for women born in perhaps 1900, 1960, 2020. can also be a Story. xx)

Ex. 1.56 *The Gumbel distribution.* Here we work with the Gumbel distribution, useful e.g. in models for extreme values.

the Gumbel distribution

(a) Let $X_1, \ldots, X_n$ be i.i.d. from the standard exponential distribution. Show that their maximum value M_n has c.d.f. $\{1 - \exp(-m)\}^n$. Deduce that $M_n - \log n$ has c.d.f. $G_n(u) = \{1 - (1/n)\exp(-u)\}^n$, for all $u \geq -\log n$.

(b) Show that the limit c.d.f. for $M_n - \log n$ becomes $G(u) = \exp\{-\exp(-u)\}$, and that this defines a c.d.f. on the full line. This is called the *Gumbel distribution*. Find its median and interquartile range.

(c) Find its density $g(u) = \exp\{-u - \exp(-u)\}$, and draw it in a diagram, along with the densities g_n for say $n = 10, 20, 30$, for $M_n - \log n$.

(d) With U having the Gumbel distribution, show that its m.g.f. becomes $M(t) = \Gamma(1-t)$, for $t < 1$. To find the moments, show by taking derivatives of $\Gamma(a) = \int_0^\infty v^{a-1}\exp(-v)\,\mathrm{d}v$ at position $a = 1$ that

$$\mathrm{E}\,U^p = (-1)^p\Gamma^{(p)}(1) = (-1)^p\int_0^\infty (\log v)^p \exp(-v)\,\mathrm{d}v.$$

These may be found numerically. You may also work with the cumulant-generating function, where a certain connection can be made to the Riemann zeta function, namely

$$K(t) = \log\Gamma(1-t) = \gamma_e t + \sum_{j=2}^\infty \frac{\zeta(j)}{j}t^j$$

for $|t| < 1$. Here $\gamma_e = 0.5772...$ is the Euler constant, see Ex. 1.13, and $\zeta(j) = \sum_{n=1}^\infty 1/n^j$. In the terminology of Ex. 1.34, show that $\kappa_1 = \gamma_e$, $\kappa_2 = \zeta(2)$, $\kappa_3 = 2\zeta(3)$, $\kappa_4 = 6\zeta(4)$. With $\zeta(2) = \pi^2/6$, $\zeta(3) = \pi^3/25.7943$, $\zeta(4) = \pi^4/90$, establish that $\mathrm{E}\,U = \gamma_e$, $\mathrm{Var}\,U = \sigma^2 = \pi^2/6$, $\mathrm{skew}(U) = 2\zeta(3)/\sigma^3 \doteq 1.1395$, $\mathrm{kurt}(U) = 12/5 = 2.4$.

(e) To appreciate the perhaps strange-looking connection from Gumbel moments to the zeta function, show from efforts of Ex. 1.13 that

$$M_n - \log n = \sum_{i=1}^n W_i/i + a_n$$

where $W_i = V_i - 1$, for i.i.d. unit exponentials V_i, and $a_n \to \gamma_e$. Find formulae for the mean, variance, skewness, kurtosis of $M_n - \log n$ based on this, and take their limits; these will agree with formulae found above.

Ex. 1.57 *The logistic distribution.* Consider the logistic distribution (in its standard form), with c.d.f. $H(x) = \exp(x)/\{1 + \exp(x)\}$, over the real line.

the logistic distribution

(a) Show that H indeed is a proper c.d.f., and that its density is $h(x) = \exp(x)/\{1 + \exp(x)\}^2 = H(x)\{1 - H(x)\}$, symmetric around zero. Find its interquartile range.

(b) With X having the logistic distribution, show that its m.g.f. $M(t)$ becomes

$$\int \frac{\exp\{(t+1)x\}}{\{1+\exp(x)\}^2}\,\mathrm{d}x = \int_0^\infty \frac{u^t}{(1+u)^2}\,\mathrm{d}u = \int_0^1 v^t(1-v)^{1-t}\,\mathrm{d}t = \Gamma(1+t)\Gamma(1-t),$$

for $|t|<1$. Show from this that actually $X = U - V$, in terms of independent Gumbel distributed U and V, see Ex. 1.56. Demonstrate this fact also directly, via convolution. Use this representation to show that $\mathrm{Var}\,X = \pi^2/3$ and that its kurtosis is $\mathrm{kurt} = 6/5 = 1.2$.

(c) Here we managed to find moments of the logistic via the Gumbel difference representation. We may also use the intriguing formula $\Gamma(1+t)\Gamma(1-t) = \pi t/\sin(\pi t)$, a consequence of Euler's reflection formula. Take derivatives of this function to find the variance and kurtosis formulae.

(d) A scaled version of the logistic has c.d.f. $H(x,\tau) = H(x/\tau) = \exp(x/\tau)/\{1+\exp(x/\tau)\}^2$, for a suitable positive τ. Show that the variance is $\tau^2\pi^2/3$, which is 1 for $\tau = \sqrt{3}/\pi$. Draw a figure with the density, alongside the standard normal, and comment.

Ex. 1.58 *The logistic-normal distribution on the unit interval.* The most prominent model for distributions on the unit interval is the Beta, also useful in various guises and roles outside its direct use for fitting datasets; see Ex. 1.18. Here we work with a different class, useful for versions of logistic regression, which we meet in Ch. 5. Let $H(u) = \exp(u)/\{1+\exp(u)\}$ the logistic transform, worked with in Ex. 1.57, taking any real u to the unit interval.

(a) Solve $H(u) = v$ to find the inverse transform $H^{-1}(v) = \log\{v/(1-v)\}$. Start with X a standard normal and define $V = H(X)$. Show that its c.d.f. is $G(v) = \Pr(X \le H^{-1}(v)) = \Phi((H^{-1}(v)))$, with density

$$g(v) = \phi(H^{-1}(v))\frac{1}{v(1-v)} = \frac{1}{(2\pi)^{1/2}}\exp[-\tfrac{1}{2}\{\log v - \log(1-v)\}^2]\frac{1}{v(1-v)}.$$

(b) Generalise to the case where $X \sim \mathrm{N}(\xi,\sigma^2)$, where the density becomes

$$g(v,\xi,\sigma) = \phi((H^{-1}(v)-\xi)/\sigma)\frac{1}{\sigma}\frac{1}{v(1-v)}.$$

We call this the logistic-normal with parameters (ξ,σ). Draw c.d.f.s and densities of this type, for some combinations of parameters. Explain that if $X_1,\ldots,X_p$ has a joint multinormal distribution, then the random probability

$$p(X_1,\ldots,X_p) = H(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)$$

has the logistic-normal distribution, and identify its parameters. In this sense the logistic-normal class is closed under linear combinations of the underlying $H^{-1}(V)$, whereas the Beta class does not have similar properties.

(c) Suppose $Y \,|\, p$ is binom(n,p), and that p is logistic-normal. There are no closed-form expressions for the resulting probabilities

$$\bar f(y) = \binom{n}{y} \int_0^1 p^y (1-p)^{n-y} g(p,\xi,\sigma)\,\mathrm{d}p \quad \text{for } y = 0,1,\ldots,n,$$

but it may be worked with numerically. Display a few of these $\bar f(y)$, for say $n = 100$ and a few values of (ξ,σ).

(d) We may also work the other way here: suppose $V = H(X)$ is Beta$(cp_0, c(1-p_0))$; what is then the distribution for X?

Ex. 1.59 *A normal with a normal mean is normal.* (xx preliminary version; need just a few editorial decisions regarding where to place it, and how. xx) The normal distribution has a convenient coherence type property: if X given its mean parameter is normal, and this mean parameter itself is normal, then X, marginally, is again normal. This is also related to what is found in Ex. 1.41.

(a) Consider first independent X_1 and X_2 with densities f_1 and f_2. Show that $\int f_1 f_2\,\mathrm{d}x$ is the density of $X_1 - X_2$, evaluated at zero. Write then $\phi_\sigma(x-\xi) = \sigma^{-1}\phi(\sigma^{-1}(x-a))$ for the N(ξ,σ^2) density. Show that

$$\int \phi_{\sigma_1}(x-\xi_1)\phi_{\sigma_2}(x-\xi_2)\,\mathrm{d}x = \phi_{(\sigma_1^2+\sigma_2^2)^{1/2}}(\xi_1-\xi_2).$$

(b) Assume that X given ξ has mean ξ and variance σ^2, and further that ξ stems from its own distribution, with mean ξ_0 and variance τ^2. Show that X, marginally, has mean ξ_0 and variance $\sigma^2+\tau^2$. Then specialise to the normal case, with $X \,|\, \xi$ and ξ having normal distributions. Show that X indeed also is normal, (i) by integrating out the ξ, with respect to its distribution, and also (ii) by arguing via $X = \xi + \varepsilon = \xi_0 + \delta + \varepsilon$, where ε and δ are zero-mean normals with variances σ^2 and ε^2. (xx point to connected thing for logN. xx)

(c) Suppose $Y \,|\, (x_1,x_2)$ is normal N$(a + b_1x_1 + b_2x_2, \sigma^2)$, as in linear regression models we will study in later chapters; see e.g. Ex. 3.31. Assume then that (x_1,x_2) themselves have a distribution, in its space of covariate pairs, and that this distribution is binormal. Show that Y, marginally, is normal, and give formulae for its mean and variance.

Ex. 1.60 *Mixing the normal scale.* (xx at the moment nils thinks this exercise will go away, partly with material in Story v.2 and partly elsewhere. point back to and calibrate with Ex. 1.32. xx) Suppose $X \sim \mathrm{N}(\xi,\sigma^2)$ for given parameters (ξ,σ), but that there are background mechanisms producing these (ξ,σ). In various settings this leads to good 'mixtures of normals' models for actually observed data.

(a) Suppose a given individual has his ξ and that his associated X is a N(ξ,σ^2). Assume next that in a population of such X, there is a distribution $\xi \sim \mathrm{N}(\xi_0, \sigma^2_{\mathrm{extra}})$ of their means. Show that an X sampled from that population is a N$(\xi_0, \sigma^2 + \sigma^2_{\mathrm{extra}})$. From a statistical modelling viewpoint we have simply 'put in more in the σ', perhaps stretched

its interpretation a little, without inventing or having to invent a new model for the observed X, per se. Also, without knowing more, or perhaps having a separate experiment, we cannot identify the components of the observed variance.

(b) Now turn attention to the scale. With the mean ξ kept fixed, but σ having some density $\pi(\sigma)$, show that X has density $\bar{f}(x) = \int (1/\sigma)\phi((x-\xi)/\sigma)\pi(\sigma)\,\mathrm{d}\sigma$, and the variance of X is the mean of the distribution of σ^2. Assume for simplicity of presentation that $\xi = 0$, and work with the case where the distribution of σ is such that $1/\sigma^2 \sim \mathrm{Gam}(a,b)$ (it is common to express this by saying that σ^2 has an inverse gamma distribution). Work out that

$$\bar{f}(x) = \frac{1}{(2\pi)^{1/2}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+\frac12)}{(b+\frac12 x^2)^{a+1/2}}.$$

It is useful to transform this to a member of the well-known distributions, to facilitate computations of probabilities etc. Show therefore that

$$V = (\tfrac12 X^2/b)/(1+\tfrac12 X^2/b) \sim \mathrm{Beta}(\tfrac12, a),$$

and express the c.d.f. of X in terms of the c.d.f. of this Beta distribution: $\Pr(|X| \le x) = \mathrm{Be}((\frac12 x^2/b)/(1+\frac12 x^2/b), \frac12, a)$. Check these details via simulations. (xx nils jotting down the details here, to land in solutions. xx) solving the V equation for X, this inverse transformation is $X = (2b)^{1/2}\{v/(1-v)\}^{1/2}$. Since X is symmetric around zero, the density of V must be

$$\bar{h}(v) = 2\bar{f}((2b)^{1/2}\{v/(1-v)\}^{1/2})\,(2b)^{1/2}\tfrac12\{(1-v)/v\}^{1/2}/(1-v)^2.$$

Sorting out terms with v and $1-v$ gives the desired Beta distribution density.

Ex. 1.61 *Normal mixtures.* [xx to come. mean, variance, skewness. xx] Suppose Y is such that with probability p_j, it is a normal (μ_j, σ_j^2), with probabilities $p_1, \ldots, p_k$ summing to 1. Its density may be written $f(y) = \sum_{j=1}^k p_j \phi_{\sigma_j}(y-\mu_j)$, where $\phi_\sigma(u) = \sigma^{-1}\phi(\sigma^{-1}u)$ is the density of a $\mathrm{N}(0,\sigma^2)$. Such distributions are called normal mixtures.

(a) With J taking values $1, \ldots, k$, with probabilities $p_1, \ldots, p_k$, let $Y \mid (J=j) \sim \mathrm{N}(\mu_j, \sigma_j^2)$. Show that this Y has the density above; this amounts to a way of representing and interpreting a normal mixture.

(b) From $\mathrm{E}\,(Y \mid J) = \mu_J$ and $\mathrm{Var}\,(Y \mid J) = \sigma_J^2$, show that

$$\mathrm{E}\,Y = \bar{\mu} = \sum_{j=1}^k p_j \mu_j, \quad \mathrm{Var}\,Y = \mathrm{E}\,(Y-\bar{\mu})^2 = \sum_{j=1}^k p_j \sigma_j^2 + \sum_{j=1}^k p_j (\mu_j - \bar{\mu})^2.$$

(c) (xx someting; display a few. xx)

Ex. 1.62 *The hypergeometric distribution.* You draw a sample of n items from a bag of N, which has A of Type One and $B = N - A$ of type Two. Consider X, the number among the sampled n which are of Type One.

(a) Show that X has distribution

$$f(x) = \Pr(X = x) = \binom{A}{x}\binom{B}{n-x}\Big/\binom{N}{n}.$$

For which x is this positive? Explain the identity $\sum_{x=0}^{A}\binom{A}{x}\binom{B}{n-x} = \binom{A+B}{n}$.

(b) Show that $\mathrm{E}\,X = nA/N = np$, with $p = A/N$ the proportion of Type One in the bag. This may be done work with $\sum_{x=0}^{n} xf(x)$, or by writing $X = J_1 + \cdots + J_n$, with J_i and indicator for selected item i being a Type One or not.

(c) Explain that if one samples one item at the time, followed by replacing the item, then the $J_1, \ldots, J_n$ above are independent Bernoulli variables with probability $p = A/N$, leading in that case to binomial variance $np(1-p)$. For the present hypergeometric setting, where n items sampled in one go without replacement, the J_i are dependent; show that $\mathrm{cov}(J_1, J_2) = p(A-1)/(N-1) - p^2 = -p(1-p)/(N-1)$. Deduce that the variance formula becomes $\mathrm{Var}\,X = c_n np(1-p)$, with c_n being the shrinking factor $(N-n)/(N-1)$. This may be accomplished working algebraically with $\mathrm{E}\,X(X-1)$, or via the representation above.

(d) (xx just a bit more. comparing with binomial. xx)

Ex. 1.63 *Building bivariate dependence models.* (xx something here. need to have something beside the multinormal for dependence. xx) Let X and Y have densities $f_1(x)$ and $f_2(y)$, with c.d.f.s $F_1(x)$ and $F_2(y)$. To model dependence between them, the idea pursued here is mapping them to the normal scale, then using the binormal model, and mapping back.

(a) Write $X = F_1^{-1}(\Phi(U))$ and $Y = F_2^{-1}(\Phi(V))$, where (U, V) is binormal with zero means, unit variances, and correlation ρ. Show that X and Y indeed have densities f_1 and f_2. Writing $g_\rho(u, v)$ for the binormal density, show that the joint density can be written

$$f(x, y) = f_1(x)f_2(y)R_\rho(\Phi^{-1}(F_1(x)), \Phi^{-1}(\Phi(y))), \quad \text{with } R_\rho(u, v) = \frac{g_\rho(u, v)}{\phi(u)\phi(v)}.$$

with the dependence factor R_ρ of course being 1 when $\rho = 0$. Show in fact that

$$R_\rho(u, v) = \frac{1}{(1-\rho^2)^{1/2}} \exp\Big\{-\tfrac{1}{2}\frac{1}{1-\rho^2}(\rho^2u^2 + \rho^2v^2 - 2\rho uv)\Big\}.$$

(b) Now consider $f_1(x) = \exp(-x)$ and $f_2(y) = \exp(-y)$, i.e. two unit exponentials. Construct a bivariare pair (X, Y) via the recipe above, with a density $f_\rho(x, y)$ having unit exponential marginals. Compute the correlations $\mathrm{corr}(X, Y)$ as a function of ρ, which might be easiest via simulations. (xx note very different cases, $\rho = -0.99$ and $\rho = 0.99$. xx)

(c) (xx one more example. for the uniform case, we have $\mathrm{corr}(X, Y)$ quite close to ρ, but they remain different. xx)

Ex. 1.64 *Alice and Bob correlate their binomials.* Here we show how correlated binomials may be constructed, leading also to correlated random walks.

(a) Alice flips her fair coin n times, with i.i.d. 0-1 outcomes $A_1, \ldots, A_n$. Bob has two coins, and mixes between them depending on Alice's outcomes: if $A_i = 1$, he uses the plus-coin with probability $\frac{1}{2} + a$ for heads; and if $A_i = 0$ he uses his minus-coin with $\frac{1}{2} - a$ for heads. With B_i his outcome, show that $\Pr(B_i = 1) = \frac{1}{2}$, and argue therefore that both X_n and Y_n are binomial $(n, \frac{1}{2})$ variables, where $X_n = \sum_{i=1}^n A_i$ and $Y_n = \sum_{i=1}^n B_i$ are the number of heads for Alice and for Bob. Show that the correlation between these two binomials is $2a$.

(b) In the little story-telling above, Bob observes Alice's outcomes, one by one, which then influence his choice between two coins; Alice doesn't even need to be aware of Bob's existence. Explain however that we from observed pairs of coin flips (A_i, B_i) never can see the difference between that scenario and the alternative one, that Bob is the one flipping his fair coin, without caring for Alice, before she chooses between two biased ones. This is arguably an instance of what Breiman (2001) alludes to as the Rashomon Effect (from a Japanese movie in which different persons report very differently about something they have all observed): data alone cannot help us uncover which of the chains of action have been at work. Show indeed that as long as Alice and Bob have a joint scheme of producing outcomes $(0,0), (0,1), (1,0), (1,1)$, with probabilities respectively $\frac{1}{4}(1+a), \frac{1}{4}(1-a), \frac{1}{4}(1-a), \frac{1}{4}(1+a)$, then (X_n, Y_n) have the correlated binomial distribution.

the Rashomon Effect: different models may offer equally good explanations

(c) Find a way to compute $f(x,y) = \Pr(X_n = x, Y_n = y)$, for $x, y = 0, 1, \ldots, n$.

(d) Leaving the Rashomon aspects to the side, generalise the first setup to the case of two correlated binomials (n, p), where p is not necessarily $\frac{1}{2}$. Take indeed $\Pr(A_i = 1) = p$ and then $\Pr(B_i = 1 \,|\, A_i = 1) = p + a$, $\Pr(B_i = 0 \,|\, A_i = 0) = 1 - p + ap/(1-p)$, for $a < \min(p, 1-p)$, and show that this works properly. What is the correlation between X_n and Y_n?

(e) Show that $\sqrt{n}(X_n/n - p, Y_n/n - p)$ tends in distribution to a binormal zero-mean (X, Y), with variances $p(1-p)$ and covariance ap.

(f) (xx brief pointer to two correlated random walks, Ch9, with two correlated Brownian motions. also good ML exercise, finding $\widehat{a}$ based on having observed Alice and Bob random walks, easy $\sqrt{n}(\widehat{a} - a)$, test for $a = 0$, etc.)

Notes and pointers

(xx to come. a bit old literature, but crisply, and not systematic. brief genesis of the normal, a few sentences on the chi-squared, Pearson (1900), the t, Student (1908), the F, the Dirichlet, more. we also point to essential things in later chapters. point out that the normal is famous and useful also because of a host of approximation methods. xx)

(xx where do we have a precise theorem on $M_X = M_Y$ implying $F = G$? inversion formula? xx)

The chi-squared is on the list over deservedly famous distributions in probability theory and statistics, and stems from Karl Pearson's famous 1900 paper, 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling'. [xx a bit more: he establishes the chi-square distribution, the test carrying the chi-squared name, and sets up a rigorous concepttual framework for hypothesis testing. xx]

(xx point to uses of expofamily in both Chs. 4 and 5. xx)

I.2

Large-sample theory

The broad themes of this chapter are the concepts, details, methods, results, applications pertaining to three modes of convergence for random variables: convergence in probability, convergence almost surely, convergence in distribution. The first two have to do with random variables X_n coming close to some limit X, with increasing n, typically indexed by sample size; often the limit is merely a constant. The chief result here, with various extensions and uses, is the Law of Large Numbers, than the empirical average of a sequence of observations tends to the expected value of the underlying distribution. The third mode of convergence rather involves the distribution of X_n coming close to the distribution of some limit X, with the Central Limit Theorem being a prime statement. These machineries also lead to practically useful approximations; the idea is that a complicated distribution may be approximated by something much simpler. This is in particular helped by a collection of approximation methods called the delta method. The theory is developed first for functions of i.i.d. sequences, involving tools of moment-generating functions and characteristic functions, along with various probability inequalities. It is then extended to cover cases of independent variables from non-equal distributions, culminating in the Lindeberg theorem, giving precise conditions under which a sum of independent components approaches normality. Methods and results from this chapter are crucial for developing the likelihood theory of Ch. 5, and also for several later chapters.

Key words: central limit theorems, characteristic functions, delta method, large-sample approximations, laws of large numbers, Lindeberg conditions, modes of convergence

In this chapter we study convergence of sequences $(X_n)_{n\geq 1}$ of random elements. The index n typically refers to the sample size, and X_n is some function of the n data points available. The modes of convergence we study take place as n grows without bounds; hence the name large-sample theory.

A random element X is a function defined on a probability space $(\Omega, \mathcal{F}, \Pr)$, taking its values in some space $\mathcal{X}$, equipped with an appropriate σ-algebra. When $\mathcal{X}$ is a subset of the real line, X a random variable; when $\mathcal{X}$ is a subset of $\mathbb{R}^k$ for some $k \geq 1$, X is a random vector (so a random vector of dimension one is a random variable); and when $\mathcal{X}$

is a function space (i.e., a set of functions), X is called a random or stochastic process. In this chapter we concentrate mainly on convergence of random variables and vectors, with the more involved themes of convergence of stochastic processes studied in Ch. 9. Many of the results of the present chapter are, however, valid for stochastic process, which ones will be clear from the context.

Applications of large-sample theory are plentiful in probability and statistics, partly to understand crucial phenomena better, and partly to provide fruitful and practical approximations; an estimator or a statistic might have a very complicated exact distribution, but have a simple to use and sometimes accurate large-sample approximation. The key convergence concepts, with ensuing approximations and applications, are as follows.

First, if $X_n = (X_{n,1}, \ldots, X_{n,k})$ is a sequence of are random vectors in $\mathbb{R}^k$, and $a = (a_1, \ldots, a_k)$ is some point in $\mathbb{R}^k$, we say that X_n *converges to* a *in probability*, written $X_n \to_{\rm pr} a$, if for each positive ε,

$$\Pr(\|X_n - a\| \geq \varepsilon) \to 0, \quad \text{as } n \text{ tends to infinity.} \tag{2.1}$$

Here $\|x\| = (\sum_{j=1}^k x_j)^{1/2}$ is simple Euclidean distance, and hence ordinary distance in the one-dimensional case. If $X_n = \widehat{\theta}_n$ is an estimator, for some parameter θ, we say that the estimator is *consistent* if $\widehat{\theta}_n \to_{\rm pr} \theta_0$, where θ_0 is the true parameter value. consistency of an estimator

Second, a stronger version of convergence is X_n *converges to* a *almost surely*, that is, the convergence occurs with probability one. This means that the event

$$N = \{\omega \in \Omega \colon \lim_n X_n(\omega) \neq a\} \quad \text{has probability zero,} \tag{2.2}$$

and we write $X_n \to_{\rm a.s.} a$. We say that an estimator $\widehat{\theta}_n$ is strongly consistent for θ_0 if $\widehat{\theta}_n \to_{\rm a.s.} \theta_0$. A strong achievement indeed is the *strong Law of Large Numbers* (LLN), the LLN which says that

$$\text{if } X_1, X_2, \ldots \text{ are i.i.d. with finite mean } \xi, \text{ then } \bar{X}_n = n^{-1} \sum_{i=1}^n X_i \to_{\rm a.s.} \xi, \tag{2.3}$$

with no further assumptions required. One may readily prove weaker versions of the LLN, as in Ex. 2.11(a) and Ex. 2.16, but with the development of sharper tools, and separate valuable results along the way, we reach the strong LLN in Ex. 2.52–2.53. It immediately has many applications and uses, as we shall see.

We note that limits in probability and almost surely can easily be defined also when the limit is a random variable X, just by replacing a with the X in (2.1)–(2.2). Most often, though, these two convergence concepts are used for cases when the limit is a constant.

The third and statistically speaking most important concept is that of *convergence in distribution*. If the random vectors X_n and X of dimension $k \geq 1$ have c.d.f.s F_n and F, we say that X_n *converges in distribution to* X, or, equivalently, that F_n converges in distribution to F, if

$$F_n(x) \to F(x), \quad \text{for all continuity points } x = (x_1, \ldots, x_k) \text{ of } F. \tag{2.4}$$

the CLT

We write $X_n \to_d X$ or $F_n \Rightarrow F$ to indicate this, allowing for simplicity also statements like $X_n \to_d \mathrm{N}(0,1)$. The bigger sibling of the LLN is the Central Limit Theorem (CLT): if $X_1, X_2, \ldots$ are i.i.d. random variables with variance σ^2, then

$$\sqrt{n}(\bar{X}_n - \mathrm{E}\, X_1) \to_d \mathrm{N}(0, \sigma^2), \tag{2.5}$$

again with no further assumptions needed beyond a finite variance. Below we go considerably further, however, in detail, in extensions, in applications. In particular, with additional tools and efforts we reach the Lindeberg theorem, with precise necessary conditions for a sum of independent variables from different distributions to approach normality. Such results are, for example, used to establish approximate normality for estimators in regression models.

Reaching the CLTs, with variations, requires hard mathematical work, with various technical details to sort out. When the main theorems have been established, however, along with further tools, their actual use for statistical applications might be relatively straightforward. In particular, functions of approximately normal variables are also approximately normal, as we learn in the subsection on delta methods.

Modes of convergence

Ex. 2.1 *Almost surely implies* pr *implies d.* Of the three modes of convergence, convergence almost surely is the strongest, convergence in distribution the weakest, with convergence in probability lying somewhere in the middle. In the following, X_n and X are random vectors. In statistical applications, $(X_n)_{n\geq 1}$ will often be a sequence of estimators, and the limit a constant.

(a) Show that if $X_n \to_{\text{a.s.}} X$, then $X_n \to_{\text{pr}} X$. To show this, consider the sets $A_n = \{|X_n - X| \geq \varepsilon\}$ for some $\varepsilon > 0$, and establish that $\cap_{n\geq 1} \cup_{m\geq n} A_m \subset A$, where A is the set where X_n does not converge to X, and take it from there.

(b) Show that if $X_n \to_{\text{pr}} X$ then $X_n \to_d X$.

Markov's inequality

(c) Let Y be a nonnegative random variable, and show Markov's inequality: for any $a > 0$, $\Pr(Y \geq a) \leq \mathrm{E}\,(Y)/a$. Use this to show that if $\mathrm{E}\,\|X_n - X\|^p$ for some $p \geq 1$, then X_n converges in probability to X. In applications, we often have $p = 2$ and X constant, equal to a parameter θ, say. Show that $\mathrm{E}\,\|X_n - \theta\|^2 \to 0$ if and only if $\mathrm{E}\, X_n \to \theta$ and $\mathrm{Var}(X_n) \to 0$.

Ex. 2.2 *The Borel–Cantelli lemma and convergence almost surely.* Let $A_1, A_2, \ldots$ be a sequence of events. Consider $A_{\text{i.o.}} = \cap_{n\geq 1} \cup_{m\geq n} A_m$, the full-sequence event corresponding to the A_n occurring infinitely often.

(a) Let $A_1, A_2, \ldots$ be a sequence of events, and let N be the total number of occurrences of the A_i. Show that $\mathrm{E}\, N = \sum_{i=1}^{\infty} \Pr(A_i)$.

The Borel–Cantelli lemma

(b) Show that if $\sum_{n=1}^{\infty} \Pr(A_n)$ is convergent, then $\Pr(A_{\text{i.o.}}) = 0$. So sooner or later there will be a finite (but random n), such that none of the A_m will ever occur, for $m > n$.

(c) Assume in addition that the $A_1, A_2, \ldots$ are independent events. Show that if $\sum_{i=1}^{\infty} p_i$ is divergent, then $\Pr(A_{\text{i.o.}}) = 1$. To show this, prove and use the inequality $1+x \leq \exp(x)$, valid for all $x \in \mathbb{R}$. In particular, for the case of independent events, there can't be say a 50 percent chance that there will be infinitely many occurrences.

(d) Let $X_1, X_2, \ldots$ be i.i.d. from the unit exponential distribution. Will there be infinitely many cases with $X_i \geq 0.99 \log i$, with $X_i \geq \log i$, with $X_i \geq 1.01 \log i$?

(e) Let $X_1, X_2, \ldots$ be i.i.d. standard normal. Show first that

$$\Pr(X_i \geq a) = 1 - \Phi(a) = \frac{\phi(a)}{a}(1 + o(1)), \quad \text{as } a \to \infty.$$

Show that there will be infinitely many cases with $|X_i| \geq (2 \log i)^{1/2}$.

(f) (xx one or two more. new records, $\Pr(R_n = 1) = 1/n$. xx)

(g) Show that $\|X_n - X\| \to_{\text{a.s.}} 0$ is equivalent to $|X_{n,j} - X_j| \to_{\text{a.s.}} 0$ for each j. The same is true for convergence in probability. Consider the random vectors $X_n = (X_{n,1}, X_{n,2})$ and $X = (X_1, X_2)$. Show that $(X_{n,1}, X_{n,2}) \to_{\text{pr}} (X_1, X_2)$, by the definition of (2.1), is equivalent to $X_{n,1} \to_{\text{pr}} X_1$ and $X_{n,2} \to_{\text{pr}} X_2$, that is, ordinary one-dimensional convergence for each component. Generalise to higher dimensions.

Ex. 2.3 *The converses do not hold.* This exercise shows that the implications arrows in Ex. 2.1 only point in one direction, that is, the converses do not hold.

(a) Let $X_1, X_2, \ldots$ be Bernoulli random variables with success probabilities $p_1, p_2, \ldots$. Show that $X_n \to_{\text{pr}} 0$ as long as $p_n \to 0$, but that $X_n \to_{\text{a.s.}} 0$ takes more. What is the probability for having infinitely many $X_n = 1$, for $p_n = 1/n^{0.99}$, for $p_n = 1/n$, for $p_n = 1/n^{1.01}$? Conclude that we can have convergence in probability, but not almost surely.

(b) Let U be unif $(0, 1)$, and divide the unit interval in $2, 3, \ldots$ pieces: $A_1 = [0, 1/2]$, $A_2 = (1/2, 1]$, $A_3 = [0, 1/4]$, $A_4 = (1/4, 1/2]$, $A_5 = (1/2, 3/4]$, and so on. Define the random variables $X_n = I(U \in A_n)$, and show that $X_n \to_{\text{pr}} 0$, but not almost surely.

(c) Find an example where $X_n \to_d X$, but there is not convergence in probability.

(d) Let X_n be a sequence of binary random variable with distributions $\Pr(X_n = a_n) = p_n$ and $\Pr(X_n = 0) = 1 - p_n$. Construct the a_n and p_n sequences such that $X_n \to_{\text{pr}} 0$, but $\mathrm{E}\,|X_n|$ does not converge to zero. In the same vein, Let $X_n = \theta + \varepsilon_n$, for $\theta \in \mathbb{R}$, where $\varepsilon_1, \varepsilon_2, \ldots$ are independent random variable with distribution $\Pr(\varepsilon_n = 2^n) = \Pr(\varepsilon_n = -2^n) = 1/(3n)$ and $\Pr(\varepsilon_n = 0) = 1 - 1/(3n)$. Show that $X_n \to_{\text{pr}} \theta$, but that $\mathrm{E}\,|X_n - \theta|$ does not converge to zero.

Ex. 2.4 *Partial converses.* As we have just seen, the implication arrows of Ex. 2.1 only point in one direction. We do, however, have certain partial converses.

(a) As as partial converse to Ex. 2.1(a), show that if $X_n \to_{\text{pr}} X$, then there is a subsequence $(n_k)_{k \geq 1}$ such that $X_{n_k} \to_{\text{a.s.}} X$. To construct such a subsequence, you may

use the Borel–Cantelli lemma. Since any subsequence of a sequence converging in probability also converges in probability, you have just proved that if $X_n \to_{\mathrm{pr}} X$, then every subsequence has a further subsequence converging almost surely.

(b) As a partial converse to Ex. 2.1(b), show that if $X_n \to_d a$ for some constant $a \in \mathbb{R}$, then $X_n \to_{\mathrm{pr}} a$. Generalise to higher dimensions.

(c) As a partial converse to Ex. 2.1(c), we have that if $X_n \to_{\mathrm{pr}} X$ and $\|X_n\|$ is bounded by a random variable Y such that $|Y|^p$ is integrable, then $\mathrm{E}\,\|X_n - X\|^p \to 0$, for some $p \geq 1$. Notice that this is a version of Lebesgue's dominated convergence theorem (see Ex. A.11(d), and you may consult that exercise before solving this one).

(d) Another partial converse to Ex. 2.1(c) is provided by (a version of) Scheffé's lemma, which states that if X_n is a sequence of nonnegative random vectors such that $X_n \to_{\mathrm{pr}} X$ and $\mathrm{E}\,X_n \to \mathrm{E}\,X$, with X integrable, then $\mathrm{E}\,\|X_n - X\| \to 0$. Prove it first in one dimension, where you may use that $|y| = y + 2\max(-y, 0)$, then generalise.

Ex. 2.5 *Uniform integrability.* Yet another partial converse to Ex. 2.1(c) is obtained by introducing the concept of *uniform integrability*. A sequence $X_1, X_2, \ldots$ of random variables is uniformly integrable if

$$\lim_{M\to\infty} \sup_{n\geq 1} \mathrm{E}\,|X_n| I(|X_n| > M) = 0. \tag{2.6}$$

As we show in this exercise, if $(X_n)_{n\geq 1}$ is uniformly integrable and $X_n \to_{\mathrm{pr}} X$, then $\mathrm{E}\,|X_n - X| \to 0$. Thus, compared to Ex. 2.4(c) (which is essentially Lebesgue's dominated convergence theorem), the uniform integrability assumption takes the place of the boundedness assumption imposed in the dominated convergence theorem. Moreover, as we show in (c) and (i), uniformly integrability is a bit weaker than boundedness, and turns out to be a necessary for convergence in mean.

(a) Show that $\mathrm{E}\,|X| < \infty$ is equivalent to $\lim_{M\to\infty} \mathrm{E}\,|X| I(|X| > M) = 0$. This explains why the property in (2.6) is called uniform integrability.

(b) Show that if $(X_n)_{n\geq 1}$ is uniformly integrable, then $\sup_n \mathrm{E}\,|X_n| < \infty$. Exhibit a sequence of random variables for which the converse does not hold.

(c) Show that if $(X_n)_{n\geq 1}$ is dominated by an integrable Y, then it is uniformly integrable. Construct a sequence of random variables that is uniformly integrable, but that is not dominated by an integrable random variable.

(d) Let $(X_n)_{n\geq 1}$ be a uniformly integrable sequence of random variables such that $X_n \to_{\mathrm{a.s.}} X$. Show that the limit X must be integrable. To show this you may first consult Fatou's lemma, see Ex. A.11(b) in the appendix.

(e) Let $(X_n)_{n\geq 1}$ be as in (d). Show that $\mathrm{E}\,|X_n - X| \to 0$. The trick is to truncate $Y_n = |X_n - X|$ and take it from there.

(f) Show that in (e), it is sufficient that X_n converges in probability to X.

(g) Show that if $(X_n)_{n\geq 1}$ is such that $\mathrm{E}\,|X_n|^{1+\delta} < \infty$ for all n, for some $\delta > 0$, then it is uniformly integrable.

(h) Show that $(X_n)_{n\geq 1}$ is uniformly integrable if and only if it is asymptotically uniformly integrable, that is, $\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{E}\,|X_n| I(|X_n| > M) = 0$.

(i) Finally, (e) has a converse. Show that if $(X_n)_{n\geq 1}$ and X are nonnegative and integrable, $X_n \to_{\text{a.s.}} X$, and $\mathrm{E}\,X_n \to \mathrm{E}\,X$, then $(X_n)_{n\geq 1}$ is uniformly integrable. Deduce from this that if $(X_n)_{n\geq 1}$ and X are integrable, $X_n \to_{\text{a.s.}} X$, and $\mathrm{E}\,|X_n - X| \to 0$, then $(X_n)_{n\geq 1}$ is uniformly integrable.

Ex. 2.6 *Scheffé's lemma.* If we replace the random vectors in Ex. 2.4(d) with the densities $(f_n)_{n\geq 1}$ and f, then the limit is automatically integrable, and we obtain the lemma typically known by the name of Scheffé.

(a) Let Y_n and Y be random vectors with densities f_n and f, respectively. Show that if $f_n(y) \to f(y)$ for all y, then $\int |f_n(y) - f(y)|\,\mathrm{d}y \to 0$. Conclude from this that $Y_n \to_d Y$. Scheffé's lemma

(b) Let $X_1, X_2, \ldots$ be i.i.d. unif $(0,1)$, and set $M_n = \max_{i\leq n} X_i$. Use Scheffé's lemma to show that $n(1 - M_n)$ converges in distribution to a unit exponential. [xx this is not a great examples xx]

(c) [xx a couple of simple examples here, where $f_n \to f$. xx]

(d) Show that under the conditions of (a) we have $\sup_B |\Pr(Y_n \in B) - \Pr(Y \in B)| \to 0$. Since $Y_n \to_d Y$ only requires $\Pr(Y_n \in B) \to \Pr(Y \in B)$ for sets of the form $B = (-\infty, y]$ (or, more generally for all continuity sets B, see Ex. 2.19), convergence $\Pr(Y_n \in B) \to \Pr(Y \in B)$ uniformly in B is stronger than what is required for convergence in distribution: For a classical example thereof, consider the sequence $Y_n = 1/n$. Clearly, $Y_n \to_d 0$, but show that $\sup_B |\Pr(Y_n \in B) - \Pr(Y \in B)| = 1$. More examples are given in Ex. 2.7.

(e) If Y_n and Y have densities f_n and f, and $Y_n \to_d Y$, we should expect $f_n \to f$. This is not alway happening, however. Consider the case of $F_n(y) = y + (n\pi)^{-1} \sin(n\pi y)$. Plot the F_n and its density f_n, for some n. Show that $Y_n \to_d$ unif, but that $f_n(y)$ does not converge to 1 for all y. It is also instructive to transform to the approximate normal scale, via $X = \Phi^{-1}(Y)$. Show that X then has density $g_n(x) = \phi(x)\{1 + \cos(n\pi\Phi(x))\}$, with very notable oscillations, but where the c.d.f. $G_n(x)$ nevertheless tends to the standard normal.

Ex. 2.7 *From discrete to continuous.* Often enough discrete distributions have continuous limits.

(a) Let X_n have distribution $\Pr(X_n = j/n) = 1/(n+1)$ for $j = 0, 1, \ldots, n$. Show that $X_n \to_d X$, where X has the uniform distribution on the unit interval. Show also that $\sup_B |\Pr(X_n \in B) - \Pr(X \in B)|$ does *not* converge to zero (cf. Ex. 2.6(d)).

(b) With perhaps similar techniques as in (a), consider X_n with distribution $\Pr(X_n = j/n) = j/\{n(n+1)/2\}$ for $j = 1, \ldots, n$. Find its limit distribution.

(c) Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with success probability p. Only using (2.4), show that $\sqrt{n}(\bar{X}_n - p)\} \to_d X$, with X a $\mathrm{N}(0, p(1-p))$ distribution. You may use Stirling's formula $n! \sim \sqrt{2\pi n}(n/e)^n$ here; see Ex. 2.39.

(d) Suppose $X \sim \mathrm{Pois}(\lambda)$ and that λ grows. Again, using only the definition in (2.4), show that $(X - \lambda)/\lambda^{1/2} \to_d \mathrm{N}(0, 1)$.

Ex. 2.8 *Many small probabilities give a Poisson.* The Law of Small Numbers, der Gesetz der kleinen Zahlen, says that if we sum a high number of 0-1 variables, with each having a small probability of 1, then we're close to a Poisson.

(a) Suppose Y_n is binomial (n, p_n), with p_n becoming small with growing n in a way which has $np_n \to \lambda$. Show that $Y_n \to_d \mathrm{Pois}(\lambda)$.

(b) More generally, suppose $X_1, \ldots, X_n$ are independent 0-1 Bernoulli variables with $p_i = \Pr(X_i = 1)$. Show that if $\max_{i \le n} p_i \to 0$ and $\sum_{i=1}^n p_i \to \lambda$, then $\sum_{i=1}^n X_i \to_d \mathrm{Pois}(\lambda)$.

(c) Suppose $X_1, X_2, \ldots$ are independent Bernoulli with $p_i = i/n$, and consider $Y_n(t) = \sum_{i \le t\sqrt{n}} X_i$. Show that $Y_n(t) \to_d \mathrm{Pois}(\frac{1}{2}t^2)$. The limit is actually a full Poisson process in t, with independent increments (see Ch. 9).

(d) Suppose (X_n, Y_n) has the trinomial distribution, with parameters (n, p_n, q_n), see Ex. 1.4. Assume now that p_n, q_n become small with n, such that $np_n \to \lambda_1$, $nq_n \to \lambda_2$. Show that the correlation betweem X_n and Y_n tends to zero, and that $(X_n, Y_n) \to_d (X, Y)$, where X and Y are independent and Poisson with parameters λ_1, λ_2. Generalise to a situation extending that of point (b); use the multinomial model of Ex. 1.5.

Ex. 2.9 *Maximum of uniforms.* Let $X_1, \ldots, X_n$ be i.i.d. random variables with the uniform distribution on $(0, 1)$, and set $M_n = \max_{i \le n} X_i$.

(a) Show that $M_n \to_{\mathrm{pr}} \theta$, that is, the maximum of the observations is consistent for the unknown endpoint.

(b) Find the limit distribution of $V_n = n(\theta - M_n)$, and use this result to find an approximate 90 percent confidence interval for θ, i.e., a random interval $[L_n(M_n), U_n(M_n)]$ such that $\Pr(L_n \le \theta \le U_n)$ is approximately 0.95. on confidence intervals.

Ex. 2.10 *Stochastic O_{pr} and o_{pr} symbols.* [xx can we make one or two nice exercises? And where to place it? xx]

Convergence in probability and tail bounds

Ex. 2.11 *Markov, Chebyshov, and the Law of Large Numbers.* In view of the definition of convergence in probability in (2.1), it is certainly useful to find mathematically manageable bounds for so-called tail probabilities, i.e., $\Pr(|X| \ge a)$ for a random variable X. There are several such, as we learn here, with more to come in Ex. 2.17. Their uses include assessing how likely it might be that an estimator is some distance off its target.

(a) Markov's inequality was proven in Ex. 2.1(c): If X is nonnegative, then $\Pr(X \geq a) \leq \mathrm{E}\,(X)/a$ for each $a > 0$. More generally, if $h(x)$ nonnegative and nondecreasing, with $h(a)$ positive, show that $\Pr(X \geq a) \leq \mathrm{E}\,h(X)/h(a)$. From this, deduce the Chebyshov inequality: if X has finite variance (but is not necessarily nonnegative), then

the Chebyshov inequality

$$\Pr(|X - \mathrm{E}\,X| \geq \varepsilon) \leq \mathrm{Var}\,(X)/\varepsilon^2, \quad \text{for } \varepsilon > 0.$$

(b) With $X_1, \ldots, X_n$ being i.i.d. from the same distribution as X, with mean ξ and standard deviation σ, show that for the empirical mean $\bar{X}_n$ that $\Pr(|\bar{X}_n - \xi| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2)$. Since the right hand side tends to zero as n tends to infinity, you have now proven the version of the *Law of Large Numbers* (LLN) that says that the empirical mean of n i.i.d. random variables with finite second moments converges in probability to its expectation as n tends to infinity.

the LLN via Chebyshov

(c) Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with expectation θ and finite variance σ^2. Consider the weighted average $\widehat{\theta} = \sum_{i=1}^n w_i Y_i / \sum_{i=1}^n w_i$, for nonnegative and fixed weights w_i. Give a condition for consistency of the estimator, in terms of these weights. What happens for $w_i = 1/i$, and for $w_i = 1/i^{1.5}$?

(d) If X has mean ξ, and a finite fourth moment, show that $\Pr(|X - \xi| \geq \varepsilon) \leq \mathrm{E}\,|X - \xi|^4/\varepsilon^4$. For $X \sim \mathrm{N}(\xi, \sigma^2)$, deduce that $\Pr(|X - \xi| \geq \varepsilon) \leq 3\sigma^4/\varepsilon^4$.

(e) With $X_1, \ldots, X_n$ i.i.d. from a distribution with finite fourth moment, write $\gamma_4 = \mathrm{E}\,\{(X_i - \xi)/\sigma\}^4 - 3$ for its kurtosis. Show that

$$\mathrm{E}\,|\bar{X}_n - \xi|^4 = \frac{\sigma^4}{n^4}\{n\gamma_4 + 3n(n-1)\} = \frac{\sigma^4}{n^2}\{3 + (1/n)(\gamma_4 - 3)\}.$$

Show hence that $\Pr(|\bar{X}_n - \xi| \geq \varepsilon) \leq 3.01\,\sigma^4/(n^2\varepsilon^2)$, for all large enough n. When is this a sharper result than that of the Chebyshov inequality?

Ex. 2.12 *Continuous mapping for convergence in probability.* The theorem known as the continuous mapping theorem (which will be proved in its entirety during the course of this chapter) says that the three modes of convergence introduced above are preserved under a continuous mapping. In this exercise we look at the convergence in probability part of this theorem. Below you may take X_n and X as random vectors or variables, the proofs are much the same.

(a) Suppose $X_n \to_{\mathrm{pr}} a$, with a being a constant. Show that if $g\colon \mathbb{R}^k \to \mathbb{R}^m$ (so $k = m = 1$ in the one dimensional case) is a function continuous at point $x = a$, then indeed $g(X_n) \to_{\mathrm{pr}} g(a)$.

(b) Suppose more generally that $X_n \to_{\mathrm{pr}} X$, with the limit being a random variable or vector such that $\Pr(X \in C) = 1$, for $C \subset \mathbb{R}$ or $C \subset \mathbb{R}^k$, respectively. Show that if $g\colon \mathbb{R}^k \to \mathbb{R}^m$ is uniformly continuous on C, then $g(X_n) \to_{\mathrm{pr}} g(X)$.

(c) Show that in (b) it is sufficient that g is continuous. [xx might include a hint or two here xx].

Ex. 2.13 *The binomial and the empirical distribution function.* For i.i.d. observations $Y_1, \ldots, Y_n$, we form *the empirical cumulative distribution function*, the e.c.d.f., with $F_n(t) = n^{-1} \sum_{i=1}^n I(Y_i \leq t)$. Plotting F_n for a given dataset is informative, also for comparing with given distributions.

(a) Before returning to F_n, consider $X_n \sim \text{binom}(n, p)$ and the familiar ratio $\widehat{p}_n = X_n/n$. Show that indeed $\widehat{p}_n \to_{\text{pr}} p$. Historically speaking, this is the earliest clear LLN statement. Explain that this also provides an interpretation of what probability means.

(b) Coming back to the empirical c.d.f., explain that $F_n(t)$ for given t is just a binomial proportion, and that $F_n(t) \to_{\text{pr}} F(t)$ for each t. In particular, the events $F_n(t) \leq F(t) - \varepsilon$ and $F_n(t) \geq F(t) + \varepsilon$ both have probabilities going to zero, for positive ε.

(c) We come back to finer analysis of F_n in Ex. 2.55 and 9.22, but find the variance of $F_n(t)$ and the covariance betweem $F_n(t_1)$ and $F_n(t_2)$.

Ex. 2.14 *Quantiles.* Here we learn that the empirical quantiles are consistent for their population counterparts. Further informative analyses are in Ex. 3.18, including convergence in distribution, but here we limit attention to direct convergence in probability.

(a) Suppose first that $U_1, \ldots, U_n$ are i.i.d. from the uniform distribution on the unit interval, and let M_n be the empirical median. With the empirical c.d.f. F_n from Ex. 2.13, use the fact that $F_n(t) \to_{\text{pr}} t$, for each t, to show that $M_n \to_{\text{pr}} \frac{1}{2}$. Then generalise to the case of observations $Y_1, \ldots, Y_n$ from a distribution with continuous and strictly increasing c.d.f. F, showing for the empirical median that $M_n \to_{\text{pr}} F^{-1}(\frac{1}{2})$.

(b) As in (a), let $U_1, \ldots, U_n$ are i.i.d. from the uniform distribution on the unit interval, and M_n the empirical median. Use Scheffé's lemma (Ex. 2.6) to show that $\sqrt{n}(M_n - \frac{1}{2}) \to_d \text{N}(0, 1/4)$.

(c) More generally, for a $q \in (0,1)$, let $Q_n = Q_n(q)$ be the empirical q-quantile, defined here to be $Y_{(\lfloor nq \rfloor)}$, the $\lfloor nq \rfloor$ order statistic, where $\lfloor x \rfloor = \max\{m \in \mathbb{Z} \colon m \leq x\}$. Show that Q_n is consistent for the population quantile $F^{-1}(q)$. [xx comment briefly on different definitions of quantile, but this does not matter here, differences are small. xx]

(d) Show that the interquartile range, the 0.75 quantile minus the 0.25 quantile, is consistent for the population interquartile range $F^{-1}(0.75) - F^{-1}(0.25)$. Show more generally that if $\theta = g(F^{-1}(q_1), \ldots, F^{-1}(q_r))$ is any continuous function of a finite number of quantiles, then $\widehat{\theta} = g(Q_n(q_1), \ldots, Q_n(q_r))$ is consistent for θ.

Ex. 2.15 *Smooth functions of means and quantiles.* (xx write down. all the easy consequences. continuous functions of means and quantiles are consistent. more to come. xx)

Ex. 2.16 *Improving on the weak LLN.* We have seen in Ex. 2.11(a) that the LLN holds if the distribution has a finite variance. Here we get rid of the finite variance condition. For the almost sure version of the LLN see Ex. 2.52.

(a) Let $X_1, X_2, \ldots$ be i.i.d. with finite mean ξ. Show $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i \to_{\rm pr} \xi$ by truncating the random variables involved, i.e., write $X_i = X_i I(|X_i| > M) + X_i I(|X_i| \le M)$ for some $M > 0$, then show that

$$\Pr(|\bar{X}_n - \xi| \ge \varepsilon) \le \frac{8M}{n\varepsilon^2}\mathrm{E}\,|X| + \frac{2}{\varepsilon}\mathrm{E}\,X I_{X>M},$$

for any $\varepsilon > 0$; and conclude. You will need Jensen's inequality, see Ex. A.15(f), at some point in this argument.

(b) Let $X_1, X_2, \ldots$ be i.i.d. random variables with fininte variance σ^2. Show that the empirical variance $n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is consistent for σ^2.

Ex. 2.17 *Further tail bound inequalities.* In Ex. 2.11 we learned about the Markov and Chebyshov inequalities; here we work out further tail bounds for our toolboxes.

(a) Suppose X has a finite moment-generating function (m.g.f.) $M(t) = \mathrm{E}\,\exp(tX)$, as per Ex. 1.30. Show that

$$\Pr(X \ge a) \le q(a) = \min\{t\colon\ \exp(-ta)M(t)\}.$$

Writing $M(t) = \exp\{K(t)\}$, show that this leads to $q(a) = \exp\{K(t_a) - at_a\}$, where t_a is the solution to $K'(t_a) = a$. Apply this to $X \sim \mathrm{N}(0,1)$ and a positive, and show that $\Pr(X \ge a) \le \exp(-\frac{1}{2}a^2)$. Show that this is indeed sharper than the tail bound $1/a^2$, from the simpler Chebyshov inequality, for all $a > 0$.

(b) For $X \sim \mathrm{N}(0,1)$ and a positive, show that

$$\Pr(|X| \ge a) \text{ is smaller than each of } \frac{1}{a^2}, \frac{3}{a^4}, \frac{15}{a^6}, 2\exp(-\tfrac{1}{2}a^2).$$

(xx a bit more here, rounding it off. more inequalities in Ch. 2. xx)

(c) (xx here or later, perhaps after the mgf things. we also do the expo, which is simpler than the χ^2. xx) Let $X \sim \chi^2_m$, which has mean and variance m and $2m$. Consider $p_m(a) = \Pr(X \ge m + am^{1/2})$. Show that

$$p_m(a) \le \min\{t\colon\ \frac{(1-2t)^{-m/2}}{\exp\{t(m+am^{1/2})\}}\} = (1 + a/m^{1/2})^{m/2}\exp(-\tfrac{1}{2}am^{1/2}).$$

Compare this bound both with bounds from the Markov inequality, and with the exact limit of $p_m(a)$, as m grows.

(d) Let $X_1, X_2, \ldots$ be i.i.d. with mean zero and variance one, so that $\sqrt{n}\bar{X}_n \to_d \mathrm{N}(0,1)$, but you don't need to use that here. Assume the m.g.f. $M(t) = \mathrm{E}\,\exp(tX_1) = \exp\{K(t)\}$ is finite. Show that

$$\Pr(\sqrt{n}\bar{X}_n \ge a) \le \frac{M(t/\sqrt{n})^n}{\exp(ta)} = \exp\{nK(t/\sqrt{n}) - ta\},$$

for each t. (xx then a bit more. tail inequality. not too far from good bound $\exp(-\frac{1}{2}a^2)$. briefly mention and point to large deviations theory. xx)

Ex. 2.18 *Bernshteĭn and Weierstraß.* [xx Nils, can you fix the notation in this exercise. I propose that we drop the $\doteq$ notation, and use $o_p(1)$, etc. instead xx] In c. 1885, Karl Weierstraß proved one of the fundamental and insightful results of approximation theory, that any given continuous function can be approximated uniformly well, on any finite interval, by polynomials; see also Hveberg (2019). A generation or so later, such results had been generalised to so-called Stone–Weierstraß theorems, stating, in various forms, that certain classes of functions are rich enough to deliver uniform approximations to bigger classes of functions. This is useful also in branches of probability theory. Here we give a constructive and relatively straightforward proof of the Weierstraß theorem, involving so-called Bernshteĭn polynomials. Let $g\colon [0,1] \to \mathbb{R}$ be continuous, and construct

$$B_n(p) = \mathrm{E}_p\, g(X_n/n) = \sum_{j=0}^{n} g(j/n) \binom{n}{j} p^j (1-p)^{n-j} \quad \text{for } p \in [0,1],$$

where $X_n \sim \mathrm{binom}(n,p)$. Note that $B_n(p)$ is a polynomial of degree n.

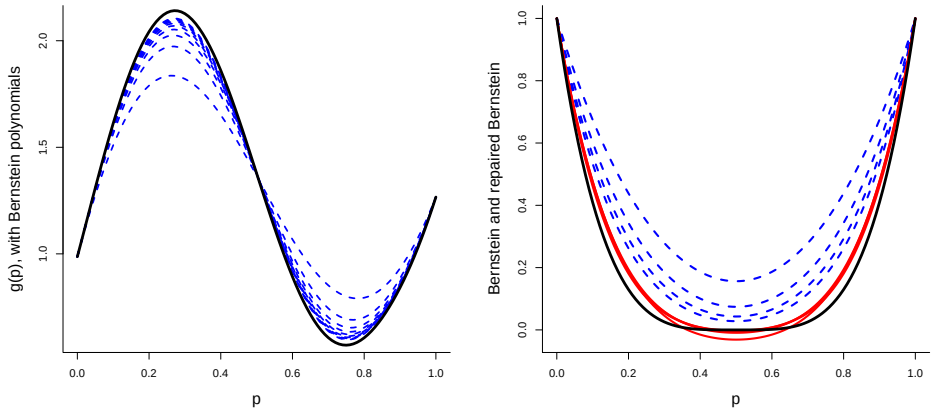


Figure 2.1: Left panel: The given non-polynomial function $g(p)$ (full black curve), along with approximating Bernshteĭn polynomials of order $10, 20, \ldots, 100$. Right panel: For the function $g_2(p) = (2p-1)^4$, the Bernshteĭn polynomials B_n, along with the bias-modified ones B_n^*, or order $4, 6, 8, 10$.

(a) Show that $B_n(p) \to_{\mathrm{pr}} g(p)$, for each p. Then show that the convergence is actually uniform. In some detail, for $\varepsilon > 0$, find $\delta > 0$ such that $|x-y| < \delta$ implies $|g(x)-g(y)| < \varepsilon$ (which is possible, as a continuous function on a compact interval is always uniformly continuous). Then show

$$|B_n(p) - g(p)| \le \varepsilon + 2M \Pr(|X_n/n - p| \ge \delta),$$

with M a bound on $|g(p)|$. Show from this that indeed $\max_p |B_n(p) - g(p)| \to 0$.

(b) Consider the marvellous function

$$g_1(p) = \sin(2\pi p) + \exp(1.234 \sin^3 \sqrt{p}) - \exp(-4.321 \cos^5 p^2)$$

on the unit interval. Compute the Bernshteĭn polynomials of various orders, and display these in a diagram, alongside the curve of g. Construct a version of Figure 2.1, left panel, which does this for $n = 10, 20, \ldots, 90, 100$. How high n is needed for the maximum absolute difference to creep below 0.01?

(c) Assuming smoothness of the g function, and writing $\widehat{p} = X_n/n$ for the binomial fraction, use Taylor expansion $g(\widehat{p}) \doteq g(p) + g'(p)(\widehat{p}-p) + \frac{1}{2}g''(p)(\widehat{p}-p)^2$ to show that $B_n(p) \doteq g(p) + \frac{1}{2}g''(p)p(1-p)/n$. This invites a repaired Bernshteĭn polynomial, of the form $B_n^*(p) = B_n(p) - \frac{1}{2}C_n(p)p(1-p)/n$, with $C_n(p) = \mathrm{E}_p g''(X_n/n)$. Explain that this is still a polynomial, now of order $n+2$, and show in examples that it often succeeds better than B_n in coming close to the target $g(p)$. Make a version of Figure 2.1, right panel, which shows the B_n and the B_n^*, of order $4, 6, 8, 10$, for the case of $g_2(p) = (2p-1)^4$.

(d) Assuming g has a derivative, construct a confidence band of the type $B_n(p) \pm 1.96\,\widehat{\sigma}(p)/\sqrt{n}$, for a certain $\widehat{\sigma}(p)$, with the property that it for each given p covers the underlying $g(p)$ with probability tending to 0.95.

(e) Let now $g(x,y)$ be an arbitrary function on the unit simplex, $\{(x,y)\colon x \geq 0, y \geq 0, x+y \leq 1\}$. Construct a mixed polynomial $B_n(x,y)$ of degree n such that it converges uniformly to g on the simplex.

Convergence in distribution

Ex. 2.19 *The Portmanteau theorem.* So far we have taken as our definition of $X_n \to_d X$ (see (2.4)) that $F_n(x) \to F(x)$ for all continuity points x of F, where F_n and F are the c.d.f.s of X_n and X, respectively. A limitation of this definition is that c.d.f.s are only defined for random vectors, and since we soon enough want to study convergence in distribution in spaces other than $\mathbb{R}^k$, we need a more general definition. Let P_n and P be probability measures on some measurable space $(\mathcal{X}, \mathcal{B})$. We say that P_n converges weakly to P, denoted $P_n \Rightarrow P$, if

weak convergence $P_n \Rightarrow P$

$$\int g\,\mathrm{d}P_n \to \int g\,\mathrm{d}P, \quad \text{for all } g \in C_b(\mathcal{X}), \tag{2.7}$$

where $C_b(\mathcal{X})$ is the collection of all continuous and bounded functions $g\colon \mathcal{X} \to \mathbb{R}$. If P_n and P are the distributions of random vectors X_n and X, then (2.7) is equivalent to $X_n \to_d X$ as defined in (2.4) (a fact we prove in (g)). As will become clear as we proceed, however, the definition in (2.7) is vastly more general, in particular, it works when $\mathcal{X}$ is a function space, see Ch. 9. It is also often much easier to work with. In this exercise we first prove the 'bare bones' Portmanteau theorem, valid for all types of metric spaces, then proceed to proving its equivalence with $X_n \to_d X$. Ex. 2.20 presents additional equivalent statements. The Portmanteau theorem says the following four statements are equivalent:

$C_b(\mathcal{X})$

(i) $P_n \Rightarrow P$;
(ii) $\limsup_n P_n(F) \leq P(F)$ for every closed set F;
(iii) $\liminf_n P_n(B) \geq P(B)$ for every open set B;

(iv) $P_n(C) \to P(C)$ for every set C that is P-continuous, in the sense that $P(\partial C) = 0$, where $\partial C = \bar{C} \setminus C^\circ$ is the boundary of C (the closure minus the interior);

Notice that in the case that X_n and X are random vectors, (i) can be written $\mathrm{E}\, f(X_n) \to \mathrm{E}\, f(X)$ for all $f \in C_b(\mathbb{R}^k)$; and (ii) $\limsup_n \Pr(X_n \in F) \le \Pr(X \in F)$ for every closed set $F \subset \mathbb{R}^k$; and so on. We follow the classical text Billingsley (1968) and prove the Portmanteau theorem through a string of subexercises.

(a) Let d be a metric, and for any set A define $d(x, A) = \inf\{d(x, y)\colon y \in A\}$, i.e., the distance from x to A. For $\varepsilon > 0$ and a set A, define the function $f_{A,\varepsilon}(x) = \eta(d(x, A)/\varepsilon)$ where

$$\eta(t) = \begin{cases} 1, & \text{if } t \le 0, \\ 1 - t, & \text{if } 0 \le t \le 1, \\ 0, & \text{if } t \ge 1. \end{cases}$$

The function $f_{A,\varepsilon}(x)$ will be used to approximate the indicator function I_A. Let $F = [a, b]$ be a closed interval on the real line, and make a sketch of $f_{F,\varepsilon}(x)$ for some $\varepsilon > 0$. For arbitrary closed sets F and $\varepsilon > 0$, show that $f_{F,\varepsilon}$ is continuous (it is clearly bounded), that $f_{F,\varepsilon} = 1$ for $x \in F$; and that $f_{F,\varepsilon}(x) = 0$ when $d(x, F) \ge \varepsilon$.

(b) Show that (i) implies (ii): Let F be a closed set. Given $\delta > 0$ we can find $\varepsilon > 0$ such that the open set $B = \{x\colon d(x, F) < \varepsilon\}$ is such that $P(B) < P(F) + \delta$. Notice that $f_{F,\varepsilon}(x) = 1$ for all $x \in B$. Use these facts to show that $\limsup_n P_n(F) < P(F) + \delta$, and conclude.

(c) Show that (ii) and (iii) are equivalent.

(d) Show that (ii) implies (i): Let g be a bounded and continuous function, $a \le g(x) \le b$ for all x, say. Define $f(x) = \{g(x) - a\}/(b - a)$, so that $0 \le f(x) \le 1$. Why does it suffice to prove the implication for f? Since f is continuous, the sets $F_i = \{x\colon g(x) \ge i/k\} = f^{-1}([i/k, 1])$ for $i = 0, 1, \ldots, k$ are closed. Define the functions

$$f_{k,\mathrm{low}}(x) = \sum_{i=1}^{k} \frac{i-1}{k} I_{F_{i-1}/F_i}(x), \quad \text{and} \quad f_{k,\mathrm{up}}(x) = \sum_{i=1}^{k} \frac{i-1}{k} I_{F_{i-1}/F_i}(x).$$

Show that $f_{k,\mathrm{low}}(x) \le f(x) \le f_{k,\mathrm{up}}(x)$ for all x, and that

$$\int f_{k,\mathrm{low}}\, \mathrm{d}P = \frac{1}{k} \sum_{i=1}^{k} P(F_i), \quad \text{and} \quad \int f_{k,\mathrm{up}}\, \mathrm{d}P = \frac{1}{k} + \frac{1}{k} \sum_{i=1}^{k} P(F_i),$$

and use these facts to prove the implication.

(e) Show that (ii) implies (iv).

(f) Show that (iv) implies (ii): For a closed set F and an arbitrary $\delta > 0$, the boundary $\partial\{x\colon d(x, F) \le \delta\} \subset \{x\colon d(x, F) = \delta\}$. For any sequence of positive δ_k tending to zero as $k \to \infty$, $F_k = \{x\colon d(x, F) \le \delta\}$ decreases to F as $k \to \infty$. Explain why we can choose these δ_k so that the F_k are continuity sets, and use this fact to prove the implication.

(g) It is time to prove that the definition of $X_n \to_d X$, as we defined it in (2.4), is equivalent to (i)–(iv) of the Portmanteau theorem. In this setting, P_n and P are the distributions of X_n and X, respectively, e.g., $P_n = \Pr X_n^{-1}$, where Pr is the probability measure on the space on which X_n is defined; and $F_n(x) = P_n(-\infty, x]$ and $F(x) = P(-\infty, x]$ are the respective c.d.f.s. The reader is of course welcome to prove the equivalence with any of the four statements of that theorem. Here we give a gentle push towards proving that (2.4) implies (iii) for the case of random variables: Any open set $B \subset \mathbb{R}$ can be written as a countable union of disjoint open intervals, $B = \cup_{j=1}^{\infty}(a_j, b_j)$. Explain why we, for any $\varepsilon > 0$ and each of these intervals, can find continuity points a_j' and b_j' of F such that $(a_j', b_j'] \subset (a_j, b_j)$ but $P(a_j', b_j'] \geq P(a_j, b_j) - \varepsilon/2^j$ for $j = 1, 2, \ldots$. Take it from here, and try to generalise this proof to the setting where X_n and X are random vectors.

Ex. 2.20 *More in the Portmanteau.* In the exercise we add some equivalent statements to the Portmanteau theorem, and look at some consequences. Throughout, P_n and P are measures on $\mathbb{R}$ equipped with the Borel σ-algebra. Each of the equivalent statements hold for general metric spaces, but the proofs pointed to in these exercises may then need to be modified.

(a) Let $f\colon \mathbb{R} \to \mathbb{R}$ be a bounded and continuous function. Let $X \sim F$ and $Z \sim \mathrm{N}(0, 1)$, be independent, and $\sigma > 0$. Show that $\mathrm{E}\, f(X + \sigma Z) = \mathrm{E}\, f_\sigma(X)$ where

$$f_\sigma(x) = \int f(u)\phi((u - x)/\sigma)/\sigma \, \mathrm{d}u,$$

and ϕ is the standard normal density; see Ex. A.35.Show that $f_\sigma(x)$ is bounded and continuous, and has continuous derivatives of all orders.

(b) Let Z be a standard normal random variable independent of $(X_n)_{n\geq 1}$. Show that $X_n + \sigma Z \to_d X + \sigma Z$ for all $\sigma > 0$ if and only if $X_n \to_d X$.

(c) From (a) and (b), to conclude that $X_n \to_d X$ if and only if $\mathrm{E}\, f(X_n) \to \mathrm{E}\, f(X)$ for all bounded and continuous functions, having continuous derivatives of all orders.

(d) Show that $X_n \to_d X$ if and only if $\mathrm{E}\, f(X_n) \to \mathrm{E}\, f(X)$ for all continuous functions that vanish outside of a compact set, i.e., for all continuous functions that are nonzero only on a compact set.

(e) In fact, $X_n \to_d X$ if and only if $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$ for infinitely smooth functions that vanish outside of a compact set. Let f be an infinitely smooth density with support $[-1, 1]$ (see Ex. A.17 for the existence of such a density), and let g be a continuous function that vanishes outside of a compact set C. For some $0 < \delta < 1$, define $f_\delta(x) = f(x/\delta)/\delta$. From Ex. A.17(e) we know that the convolution $(g * f_\delta)(z) = \int g(z - x) f_\delta(x)\, \mathrm{d}x = \int f_\delta(z - y) g(y)\, \mathrm{d}y$ is also infinitely smooth and vanishes outside of a compact set. Find a compact interval $[-a, a]$ containing C, such that

$$\sup_{z\in[-a,a]} |(g * f_\delta)(z) - g(z)| \to 0, \quad \text{as } \delta \to 0.$$

In words, the continuous function that vanish outside of a compact set can be uniformly approximated by infinitely smooth functions that vanish outside of a compact set. These latter functions then have bounded derivatives of all orders, a fact we exploit in Ex. 2.28.

(f) Finally, prove the if-and-only-if statement at the start of (d).

(g) If f is a bounded function such that $\Pr(X \in C_f) = 1$, where C_f is the set of continuity points of f, then, for any $\varepsilon > 0$, there exists bounded and continuous functions f_{low} and f_{up} such that $f_{\text{low}} \leq f \leq f_{\text{up}}$ and $\mathrm{E}\, f_{\text{up}}(X) - f_{\text{low}}(X) < \varepsilon$. Consider the functions $f_{\text{low},k}(x) = \inf_y\{f(y) + k|x - y|\}$ and $f_{\text{up},k}(x) = \sup_y\{f(y) - k|x - y|\}$ for $k = 1, 2, \ldots$, and show the existence of functions f_{low} and f_{up} as described. [xx cite Ferguson 1996 for this one xx].

(h) Show that $X_n \to_d X$ if and only if $\mathrm{E}\, f(X_n) \to \mathrm{E}\, f(X)$ for all bounded and measurable functions such that $\Pr(X \in C_f) = 1$, where $C_f = \{x\colon f \text{ is continuous at } x\}$.

(i) Let $X, X_1, X_2, \ldots$ and $Y, Y_1, Y_2, \ldots$ be random variables. Show that $(X_n, Y_n) \to_d (X, Y)$ is equivalent to both (1) $\mathrm{E}\, f(X_n)g(Y_n) \to \mathrm{E}\, f(X)g(Y)$ for all bounded and continuous f and g; and (2) $\mathrm{E}\, f(X_n)g(Y_n) \to \mathrm{E}\, f(X)g(Y)$ for all bounded and continuous f, and all bounded g such that $\Pr(Y \in C_g) = 1$, where $C_g = \{y\colon g \text{ is continuous at } y\}$.

Continuous mapping

Ex. 2.21 *Continuous mapping for convergence in distribution.* Let $X_1, X_2, \ldots$ and X be k dimensional random vectors, and $g\colon \mathbb{R}^k \to \mathbb{R}^m$ a function that is continuous on a set $C \subset \mathbb{R}^k$ such that $\Pr(X \in C) = 1$. Suppose that $X_n \to_d X$.

(a) Show that $g(X_n) \to_d g(X)$ when g is continuous on all of $\mathbb{R}^k$, that is, when $C = \mathbb{R}^k$.

(b) Now suppose that C is a subset of $\mathbb{R}^k$, and show that $h(X_n) \to_d h(X)$ in this more general case as well. You may first show that for F a closed set, the closure of $h^{-1}(F)$ is included in $h^{-1}(F) \cup C^c$, then use the Portmanteau theorem. [xx try to find another proof. Can use Portmanteau theorem (8) directly? xx].

Ex. 2.22 *Skorokhod's theorem* [xx I'm not sure if this is the right place to introduce the probability transform, nor am I sure Ex. 7.1 is (which repeats (a)). we can have it in the appendix, and there use it to prove the existence of infinite sequences of independent random variables xx]

probability transform

(a) Let $U \sim \text{unif}(0, 1)$. For any c.d.f. F, define $X(u) = \inf\{x\colon F(x) \geq u\}$ for $0 < u < 1$, so $X(u) = F^{-1}(u)$ whenever F is continuous. Show that $X(U) \sim F$..

(b) Suppose that $F_n \Rightarrow F$ and let $([0, 1], \mathcal{B}, \lambda)$ be the unit interval equipped with its Borel-σ-algebra and Lebesgue measure. On this space, construct the random variables $Y_n(\omega) = \inf\{x\colon F_n(x) \geq \omega\}$ and $Y(\omega) = \inf\{x\colon F(x) \geq \omega\}$. Show that $Y_n \to_{\text{a.s.}} Y$.

(c) Let $(X_n)_{n\geq 1}$ and X be random variables. Show that if $X_n \to_d X$, then $\mathrm{E}|X| \leq \liminf_n \mathrm{E}|X_n|$; and that if $(X_n)_{n\geq 1}$ is uniformly integrable, then X is integrable and $\mathrm{E}\, X_n \to \mathrm{E}\, X$.

(d) Show that $X_n \to_d X$ if and only if $f(X_n) \to_d f(X)$ for all $f \in C_b(\mathbb{R})$.

Ex. 2.23 *The Cramér–Slutsky rules.* The utility of the three results in (c) below, together known as the Cramér–Slutksy rules, will become abundantly clear as we progress.

(a) Show that if $X_n \to_d X$, and $Y_n - X_n \to_{\text{pr}} 0$, then also $Y_n \to_d X$. This says that variables which are essentially close, for growing n, have identical limit distributions.

(b) Show that if X_n and Y_n are sequences of random vectors such that $X_n \to_d X$ and $Y_n \to a$, for a random variable X and a constant a, then $(X_n, Y_n) \to_d (X, a)$.

(c) Show that if $X_n \to_d X$ and $Y_n \to_{\text{pr}} a$, as above, then Cramér–Slutsky

(i) $X_n + Y_n \to_d X + a$;
(ii) $X_n Y_n \to_d Xa$;
(iii) $X_n / Y_n \to_d X/a$, provided $a \neq 0$.

Explain why rules (ii) and (iii) also hold when Y_n and a are matrices.

(d) Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p). Look back at Ex. 2.7(c) and Ex. 2.16(b) and show that $\sqrt{n}(\bar{X}_n - p)/\widehat{\sigma}_n^2 \to_d \text{N}(0,1)$, where $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

Ex. 2.24 *Showing convergence in two steps.* (xx needs xref and calibration, depending on how it is presented and where applications follow. it is valid for any metric space with distance $d(x,y)$, not merely $\mathbb{R}^k$. xx) Suppose one wishes to prove that $X_n \to_d X$, but that technical issues make it easier to first prove that an approximation to X_n converges to an approximation to X. With a suitable extra condition this might suffice. (xx may point briefly to Story vii.7 to see this in use. xx)

(a) For the approximations $A_{n,k}$ to X_n and A_k to X, suppose (i) that $A_{n,k} \to_d A_k$, for each k; (ii) that $A_k \to_d X$ as $k \to \infty$; and also (iii) that

$$\lim_{k\to\infty} \limsup_{n\to\infty} \Pr(d(X_n, A_{n,k}) \geq \varepsilon) = 0 \quad \text{for each } \varepsilon > 0.$$

Show that $X_n \to_d X$.

(b) Suppose somebody clever has managed to prove the CLT for bounded i.i.d. variables (for example using m.g.f.s): if $A_1, A_2, \ldots$ are i.i.d. with mean zero with variance one, and $|A_i| \leq k$, then $Z_n = \sqrt{n}\bar{A}_n \to_d \text{N}(0,1)$. How can you then prove that the CLT is valid also for unbounded i.i.d. random variables, as long at only their variance is finite?

Ex. 2.25 *Tightness, Helly, and Prokhorov.* Consider the sequence $(X_n)_{n\geq 1}$ of random variables with distribution $\Pr(X_n = x) = 1/3$ for $x = 0, 1/2, n$. You may verify that the sequence $F_n(x) = \Pr(X_n \leq x)$ of distribution functions converges pointwise the limiting function,

$$G(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{3}, & 0 \leq x < \frac{1}{2}, \\ \frac{2}{3}, & x \geq \frac{1}{2}. \end{cases}$$

The function $G(x)$ is a limit of distributions functions, but it is *not* itself a distribution function: the problem is that $G(\infty) = 2/3$, meaning that one third of the probability mass of $F_n(x)$ has escaped to inifinity! Tightness is a condition ensuring that a limit of distribution functions is itself a distribution function. Here is the definition: A sequence $Y_1, Y_2, \ldots$ of random variables is tight if for any $\varepsilon > 0$ there exists a constant K so that Tightness

$$\Pr(|Y_n| > K) < \varepsilon, \quad \text{for all } n.$$

You may verify that the sequence X_n defined above is not tight. A tight sequence of random variables is also said to be bounded in probability (see Ex. 2.10).

We start with an exercise on other characterisations of tightness, before we proceed to Prokohorov's theorem, and finally Helly's theorem which is key to proving Prokohorov's. In Ex. 2.40 the notion of tightness is extended to random vectors, and Prokohorov's theorem is proven for random vectors.

(a) Show first that any random variable is tight. Next, let $(Y_n)_{n\geq 1}$ be a sequence of random variables with c.d.f.s F_n, and show that the following three statements are equivalent:

(i) $(Y_n)_{n\geq 1}$ is tight;
(ii) For any $\varepsilon > 0$ there is a $K > 0$ so that $\limsup_{n\to\infty} \Pr(|Y_n| > K) < \varepsilon$;
(iii) For any $\varepsilon > 0$ there are a and b such that $F_n(a) < \varepsilon$ and $F_n(b) > 1 - \varepsilon$ for all n.
(iv) For any sequence of constants $a_1, a_2, \ldots \geq 0$ tending to zero, $a_n Y_n \to_p 0$.

Finally, show that (v) if for some $\delta > 0$ there is an $M > 0$ and an $n_0 \geq 1$ so that $\mathrm{E}\,|Y_n|^\delta \leq M$ for all $n \geq n_0$, then $(Y_n)_{n\geq 1}$ is tight; and that (v) implies (i) if there is a $\delta > 0$ so that $|Y_n|^\delta$ is integrable.

(b) Prokhorov's theorem says that

Prokhorov's theorem

(i) If $X_n \to_d X$ for some random variable X, then $(X_n)_{n\geq 1}$ is tight;
(ii) If $(X_n)_{n\geq 1}$ is tight, then there is a subsequence $(n_k)_{k\geq 1}$ such that $X_{n_k} \to_d X$ for some random variable X.

Use the Portmanteau theorem (Ex. 2.19) to prove the first part of Prokhorov's theorem; and Helly's theorem, which we prove in (c), to prove the second part.

Notice that since any subsequence of a tight sequence must itself be tight, (ii) is the same as saying that if $(X_n)_{n\geq 1}$ is tight, then any subsequence $(n_k)_{k\geq 1}$ has a further subsequence $(n_{k_j})_{j\geq 1}$ such that $X_{n_{k_j}}$ converges in distribution.

(c) Let $(F_n)_{n\geq 1}$ be a sequence of distribution functions on the real line. Helly's theorem says that for any such sequence there is a right continuous and nondecreasing function F with range contained in $[0, 1]$ and a subsequence $(n_k)_{k\geq 1}$ such that $F_{n_k}(x) \to F(x)$ at every continuity point of F. Thus F has two of the defining properties of c.d.f.s (see A.14(a)). The property that may be lacking is that $F(x)$ may not tend to 0 or 1 as $x \to -\infty$ or $x \to \infty$, respectively. To prove Helly's theorem, let $\mathbb{Q} = \{q_1, q_2, \ldots\}$ be the rational numbers, and consider the infinite array

Helly's theorem

$$\begin{array}{ccccc} F_1(q_1) & F_2(q_1) & F_3(q_1) & \ldots \\ F_1(q_2) & F_2(q_2) & F_3(q_2) & \ldots \\ F_1(q_3) & F_2(q_3) & F_3(q_3) & \ldots \\ \vdots & \vdots & \vdots & \end{array}$$

Since $F_n(q_k)$ lies between zero and one for all n and k, each row of this array is bounded, and, as we know from the Bolzano–Weierstrass theorem, every bounded sequence has a convergent subsequence. In particular, there is a subsequence $n_{1,1}, n_{1,2}, \ldots$ so that

$F_{n_{1,k}}(q_1)$ has a limit as k tends to infinity. Call this limit $G(q_1)$. Extract a further subsequence $n_{2,1}, n_{2,2}$ from $n_{1,1}, n_{1,2}, \ldots$ along which $F_{n_{2,j}}$ converges to a limit, say $G(q_2)$, as j tends to infinity. Continue like this and argue that the diagonal sequence $n_k = n_{k,k}$ of the array

the diagonal method

$$\begin{matrix} n_{1,1} & n_{1,2} & n_{1,3} & \ldots \\ n_{2,1} & n_{2,2} & n_{2,3} & \ldots \\ n_{3,1} & n_{3,2} & n_{3,3} & \ldots \\ \vdots & \vdots & \vdots & \end{matrix}$$

is such that $F_{n_k}(q_j) \to G(q_j)$ for $j = 1, 2 \ldots$ as k tends to infinity. Define the function

$$F(x) = \inf\{G(q) \colon q > x\},$$

and use that the rationals are dense in the reals to show that $F_{n_k}(x)$ converges to $F(x)$ as $k \to \infty$ for every continuity point x of F. Show that F necessarily has two of the three defining properties of a c.d.f., but not necessarily the third, as described above.

(d) The following lemma is often useful for proving convergence in distribution, see Ex. 2.27(c) and Ex. 2.29(c) for applications. Suppose that $(X_n)_{n\geq 1}$ is tight, and that every subsequence of X_{n_k} that converges weakly at all, converges to the same random variable X. Show that then $X_n \to_d X$. To prove this, assume that $(X_n)_{n\geq 1}$ does *not* converge in distribution to X, and use Prokohorov's theorem to derive a contradiction.

subsequence lemma

Ex. 2.26 *Characteristic functions converging to a characteristic function.* Let $(X_n)_{n\geq 1}$ be a sequence of random variables with characteristic functions $\varphi_1(t), \varphi_2(t), \ldots$. Suppose that $\varphi_n(t) \to \varphi(t)$, and that we recognise $\varphi(t)$ as the characteristic function of some random variable X. In such cases we can conclude that $X_n \to_d X$ without further ado.

(a) Suppose that X_n and X have characteristic functions $\varphi_n(t)$ and $\varphi(t)$, respectively, and that $\varphi_n(t) \to \varphi(t)$. As an auxiliary assumption, suppose that $\varphi_n(t)$ is dominated by a function $g(t)$ such that $\int g(t)\,\mathrm{d}t < \infty$, so that by Ex. A.35(h) the X_n have densities

$$f_n(x) = \frac{1}{2\pi} \int \exp(-itx)\varphi_n(t)\,\mathrm{d}t.$$

Show that under these assumptions $X_n \to_d X$.

(b) Next, we reduce the general case (i.e., the case of X_n that do not necessarily have integrable characteristic functions) to the special case studied in (a), by using the result from Ex. 2.20(b). Suppose that X_n and X have characteristic functions $\varphi_n(t)$ and $\varphi(t)$, respectively, and that $\varphi_n(t) \to \varphi(t)$. Let Z be a standard normal random variable, and σ some positive constant. Show that $X_n + \sigma Z \to_d X + \sigma Z$, and conclude from **??** that $\varphi_n(t) \to \varphi(t)$ implies $X_n \to_d X$.

Ex. 2.27 *Lévy's continuity theorem and some more.* In Ex. 2.26 we made two crucial assumptions: that $\varphi_n(t) \to \varphi(t)$, *and* that the limiting function $\varphi(t)$ is a characteristic function. What if we drop the second assumption? That is, suppose that X_n is such that its characteristic functions φ_n converge to *some* function $\varphi(t)$, but we do not immediately

know that this limit is a characteristic function. Lévy's continuity theorem says that if this limiting function is continuous at zero, then it is a characteristic function, for some appropriate random variable X, and $X_n \to_d X$. The key to this theorem is that the continuity of $\varphi(t)$ at zero entails that X_n is tight. In fact, given that $\varphi_n(t) \to \varphi(t)$ for some function $\varphi(t)$, there is equivalence between the following three statements:

(i) $\varphi(t)$ is continuous at zero;
(ii) The X_n sequence is tight;
(iii) $\varphi(t)$ is a characteristic function.

In (a)-(b) we prove the implication (i)⇒(ii), which is Lévy's continuity theorem; the implication (ii)⇒(iii) is proven in (c); while the implication (iii)⇒(i) is immediate from the uniform continuity of characteristic functions, see Ex. A.34(a).

(a) Start by using Fubini's theorem (see Ex. A.16) to show that if X has characteristic function φ and cumulative distribution function F, then

$$\int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\}\, \mathrm{d}t = 2\varepsilon \int \big(1 - \frac{\sin(x\varepsilon)}{x\varepsilon}\big)\, \mathrm{d}F(x).$$

In particular, the integral of $\varphi(t)$ on a symmetric interval around zero is really a real number, that is, the complex component disappears. Deduce that

$$\frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\}\, \mathrm{d}t \geq 2 \int_{|x\varepsilon|\geq c} \big(1 - \frac{\sin(x\varepsilon)}{x\varepsilon}\big)\, \mathrm{d}F(x) \geq 2(1 - 1/c)\Pr\{|X| \geq c/\varepsilon\},$$

with the value $c = 2$ yielding the tail inequality

$$\Pr\{|X| \geq 2/\varepsilon\} \leq \frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\}\, \mathrm{d}t. \tag{2.8}$$

(b) If we now have a collection of random variables, where their characteristic functions have approximately the same level of smoothness around zero, then we should get tightness from (2.8). Assume that $X_1, X_2, \ldots$ have characteristic functions $\varphi_1, \varphi_2, \ldots$, and that $\varphi_n(t)$ converges to *some* function $\varphi(t)$ that is continuous at zero. For a given $\varepsilon' > 0$, find $\varepsilon > 0$ such that $|1 - \varphi(t)| \leq \varepsilon'$ for $|t| \leq \varepsilon$. Show that

$$\limsup_{n\to\infty} \Pr\{|X_n| \geq 2/\varepsilon\} \leq \frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\}\, \mathrm{d}t \leq 2\varepsilon'.$$

We've hence found a broad interval, namely $[-2/\varepsilon, 2/\varepsilon]$, inside which each X_n lies, with high enough probability, which means that the X_n sequence is tight, and thus establishes the implication (i)⇒(ii).

(c) Now we prove the implication (ii)⇒(iii), i.e., that if $(X_n)_{n\geq 1}$ is tight and $\varphi_n(t)$ converges to *some* function $\varphi(t)$, then $\varphi(t)$ must be the characteristic function of some random variable. To prove this, assume that $(X_n)_{n\geq 1}$ is tight and $\varphi_n(t) \to \varphi(t)$, but that $X_n \to_d X$ is false; then use tightness to extract a subsequence X_{n_k} converging to a random variable X, and the lemma in Ex. 2.25(d) to extract another subsequence $X_{n'_k}$ converging to some other random variable Y; and derive a contradiction.

(d) Since we're at it with tightness and characteristic functions: Show that if $X_1, X_2, \ldots$ is a tight sequence of random variables with characteristic functions $\varphi_1, \varphi_2, \ldots$, then $(\varphi_n)_{n\geq 1}$ is uniformly equicontinuous. That is, for each $\varepsilon > 0$ there is a $\delta > 0$ such that $|t - s| < \delta$ implies $|\varphi_n(t) - \varphi_n(s)| < \varepsilon$ for all n.

Central limit theorems

Ex. 2.28 *The central limit theorem, Lindeberg's proof.* Let $X_1, X_2 \ldots$ be i.i.d. random variables with mean zero and unit variance, and define $X_{n,i} = X_i/\sqrt{n}$. The goal is to show that $\mathrm{E}\, g(\sum_{i=1}^n X_{n,i}) \to \mathrm{E}\, g(Z)$, for all infinitely smooth functions g with compact support, where Z is a standard normal random variable, which is equivalent to $\sum_{i=1}^n X_{n,i} \to_d Z$, see the Portmanteau theorem, Ex. 2.19. Introduce $Z_{n,i} = Z_i/\sqrt{n}$, where the $Z_1, Z_2, \ldots$ are i.i.d. standard normals, so that $\sum_{i=1}^n Z_{n,i} \sim Z$ by Ex. A.35(g).

(a) What is in essence Lindeberg's idea was to show that $\mathrm{E}\, g(\sum_{i=1}^n X_{n,i}) - \mathrm{E}\, g(\sum_{i=1}^n Z_{n,i})$ tends to zero by replacing the summands $X_{n,i}$ by the Gaussiand summands $Z_{n,i}$, one by one. Convince yourself that by so doing the difference $g(\sum_{i=1}^n X_{n,i}) - g(\sum_{i=1}^n Z_{n,i})$ is equal to the telescoping sum on the right, that is,

$$g(\sum_{i=1}^n X_{n,i}) - g(\sum_{i=1}^n Z_{n,i}) = \sum_{k=1}^n \{g(\sum_{i=1}^k X_{n,i} + \sum_{i=k+1}^n Z_{n,i}) - g(\sum_{i=1}^{k-1} X_{n,i} + \sum_{i=k}^n Z_{n,i})\}.$$

(b) Since g is infinitely smooth with compact support, we have from Taylor's theorem that there exists a $K < \infty$ and $\delta > 0$ such that for any $\varepsilon > 0$,

$$|g(x+y) - g(x) - g'(x)y - \tfrac{1}{2}g''(x)y^2| \leq \varepsilon y^2, \quad \text{when } |y| \leq \delta, \text{ and,}$$
$$|g(x+y) - g(x) - g'(x)y - \tfrac{1}{2}g''(x)y^2| \leq K y^2, \quad \text{when } |y| > \delta.$$

Show that for each $k = 1, \ldots, n$,

$$\mathrm{E}\,\{g(\sum_{i=1}^k X_{n,i} + \sum_{i=k+1}^n Z_{n,i}) - g(\sum_{i=1}^{k-1} X_{n,i} + \sum_{i=k}^n Z_{n,i})\} = \mathrm{E}\, r_n(X_k) + \mathrm{E}\, r_n(Z_k),$$

where

$$r_n(x) \leq \frac{\varepsilon}{n}x^2 + \frac{K}{n}x^2 I(|x| \geq \sqrt{n}\delta).$$

Please conclude, and you will have shown the CLT for i.i.d. random variables.

(c) Suppose $Y_1, Y_2, \ldots$ are i.i.d. with finite mean ξ and standard deviation σ. Show that the random sum, normalised to have mean zero and variance one, i.e.,

$$\Big(\sum_{i=1}^n Y_i - n\xi\Big)/\sqrt{n\sigma^2} = \sqrt{n}(\bar{Y}_n - \xi)/\sigma,$$

tends to the $\mathrm{N}(0,1)$ in distribution.

(d) We now extend the central limit theorem to independent random variables that do not necessarily have the same distribution. Let $X_1, X_2, \ldots$ be independent mean zero random

variables with variances $\sigma_1^2, \sigma_2^2 \ldots$, and form $X_{n,j} = X_j/B_n$, where $B_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$. In view of the remainder term in (b), the Lindeberg condition is natural: Assume that for every $\varepsilon > 0$, it is the case that

$$\sum_{j=1}^{n} X_{n,j}^2 I\{|X_{n,j}| \geq \varepsilon\} \to 0, \quad \text{as } n \to \infty.$$

Show that $\sum_{j=1}^n X_{n,j} \to_d \mathrm{N}(0,1)$ by making the appropriate modifications to the proof of the i.i.d. case worked with in (a) and (b).

Ex. 2.29 *Proving the CLT with moment-generating functions.* In this exercise we prove a central limit theorem for i.i.d. random variables $X_1, X_2, \ldots$ whose moment generating functions $M(t) = \mathrm{E}\exp(tX_1)$ exist on an interval around zero. See Ex. A.31 for what this assumption entails about the random variables. Throughout, we assume that $\mathrm{E}\,X_1 = 0$ and $\mathrm{Var}(X_1) = \sigma^2$. [xx we should perhaps move 1.30, jamfør fortellerproblem Nils mentions xx]

(a) Show that $M(t) = 1 + \frac{1}{2}\sigma^2 t^2 + o(t^2)$ as $t \to 0$.

(b) Show that $\sqrt{n}\bar{X}_n = n^{-1/2}\sum_{i=1}^n X_i$ has moment generating function of the form

$$M_n(t) = M(t/\sqrt{n})^n = \{1 + \tfrac{1}{2}\sigma^2 t^2/n + o(t^2/n)\}^n,$$

and conclude that $M_n(t) \to \exp(\frac{1}{2}t^2\sigma^2)$ as $n \to \infty$, where we recognise the limit as the m.g.f. of the $\mathrm{N}(0,\sigma^2)$ distribution (see Ex. 1.30). Exercise (c) gives us what we need to conclude that this indeed implies that $\sqrt{n}\bar{X}_n$ converges in distribution to a $\mathrm{N}(0,\sigma^2)$ distributed random variable.

(c) The result of this exercise can be seen as an m.g.f. analogue of Lévy's continuity theorem. Let $(X_n)_{n\geq 1}$ have m.g.f.s $\{M_n(t)\}_{n\geq 1}$ and suppose that all these m.g.f.s exist on the same interval $[-a,a]$. Expand on Ex. 2.17(a) to show that

$$\Pr(|X_n| \geq K) \leq \exp(-Ka)\{M_n(a) + M_n(-a)\}, \quad \text{for all } n.$$

Show that if $M_n(t)$ converges to *some* function $M(t)$ for all t in some interval around zero that contains $[-a,a]$, then $(X_n)_{n\geq 1}$ is tight; and $X_n \to_d X$ where X is a random variable with m.g.f. $M(t)$. You may use Ex. 2.27(d) and Ex. A.38 to show this.

Ex. 2.30 *Two useful lemmas.* In proving the central limit theorems below, the following two lemmas will be useful.

(a) Demonstrate that with $z_{n,1}, z_{n,2}, \ldots$ a sequence of real numbers coming closer to zero, we have $\prod_{i=1}^n (1 + z_{n,i}) \to \exp(z)$, provided (i) $\sum_{i=1}^n z_{n,i} \to z$; (ii) $\max_{i\leq n} |z_{n,i}| \to 0$; and (iii) $\sum_{i=1}^n |z_{n,i}|$ stays bounded. It may be helpful to show first that

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \cdots = x + K(x)x^2,$$

with $K(x)$ a continuous function such that $|K(x)| \leq 1$ for $|x| \leq \frac{1}{2}$, and $K(x) \to -\frac{1}{2}$ when $x \to 0$. These statements are valid also when the $z_{n,i}$ and the x are complex numbers inside the unit ball, in which case the logarithm is the natural complex extension of the real logarithm.

(b) Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be complex numbers such that $|x_j|, |y_j| \leq 1$ for all j. Show that $|x_1 x_2 - y_1 y_2| \leq |x_1 - y_1| + |x_2 - y_2|$, and proceed by induction to show that

$$|\prod_{j=1}^{n} x_j - \prod_{j=1}^{n} y_j| \leq \sum_{j=1}^{n} |x_j - y_j|, \quad \text{for all } n.$$

Ex. 2.31 *The Lindeberg CLT via truncations.* In this exercise we present a proof of the CLT for independent random variables using m.g.f.s., but without assuming that the m.g.f.s. of the random variables involved necessarily exist (this is a slight generalisation of a proof presented in Inlow (2010)). Since the m.g.f. of a bounded random variable always exists (prove it), the trick is to truncate the random variables involved: Let $X_1, X_2, \ldots$ be independent random variables with mean zero and variances $\sigma_1^2, \sigma_2^2, \ldots$. Write $B_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$ and $X_{n,j} = X_j / B_n$ for $j = 1, \ldots, n$. We are to show that $\sum_{j=1}^{n} X_{n,j}$ converges in distribution to a standard normal random variable, provided the Lindeberg condition is satisfied, $L_n(\varepsilon) = \sum_{j=1}^{n} \mathrm{E}\, X_{n,j}^2 I(|X_{n,j}| \geq \varepsilon) \to 0$, for each $\varepsilon > 0$, which we assume throughout.

(a) For some $\varepsilon > 0$, define $\mu_{n,j} = \mathrm{E}\, X_{n,j} I(|X_{n,j}| < \varepsilon)$, as well as the random variables $Y_{n,j} = X_{n,j} I(|X_{n,j}| < \varepsilon) - \mu_{n,j}$ and $v_{n,j} = X_{n,j} I(|X_{n,j}| \geq \varepsilon) + \mu_{n,j}$. Verify that

$$X_{n,j} = Y_{n,j} + v_{n,j}.$$

Verify that $|Y_{n,j}| \leq 2\varepsilon$, which implies that the m.g.f. of $Y_{n,j}$ exist for all j and n. Show that

$$\mathrm{E}\, \exp(tY_{n,j}) = 1 + \tfrac{1}{2} t^2 \mathrm{E}\, Y_{n,j}^2 + r_{n,j}(t),$$

where the remainders are so that $|\sum_{j=1}^{n} r_{n,j}(t)| \leq \varepsilon \exp(2\varepsilon)$ for $|t| \leq 1$.

(b) By the first lemma of Ex. 2.30, showing that $\prod_{j=1}^{n} \mathrm{E}\, \exp(tY_{n,j}) \to \exp(\frac{1}{2} t^2)$ now comes down to verifying that (i) $\sum_{j=1}^{n} \mathrm{E}\, Y_{n,j}^2 \to 1$; (ii) that $\max_{j \leq n} \mathrm{E}\, Y_{n,j}^2 \to 0$ and $\max_{j \leq n} |r_{n,j}| \to 0$; and (iii) that $\limsup_{n \to \infty} |\mathrm{E}\, \exp(tY_{n,j}) - 1|$ is finite. Please verify these conditions (remember that it is sufficient to consider $|t| \leq 1$).

(c) Finally, show that $\sum_{j=1}^{n} v_{n,j} \to_{\mathrm{pr}} 0$ and conclude that $\sum_{j=1}^{n} X_{n,j} \to_d \mathrm{N}(0,1)$.

Ex. 2.32 *Proving the CLT with characteristic functions.* It can be argued that the most elegant, unified, and general proof of the CLT is obtained by the use of characteristic functions. This is because, contrary to the m.g.f., the characteristic function of a random variable always exists.

(a) Show that if X has finite mean ξ, then its characteristic function satisfies $\varphi(t) = 1 + i\xi t + o(t)$ as $t \to 0$. Also, its derivative exists, with $\varphi'(t) = \mathrm{E}\, iX \exp(itX)$, and in particular $\varphi'(0) = i\xi$ (see Ex. A.34 for more details).

(b) Show similarly that if X has finite variance σ^2, then

$$\varphi(t) = 1 + i\xi t - \tfrac{1}{2}(\xi^2 + \sigma^2) t^2 + o(t^2) \quad \text{as } t \to 0.$$

(c) If $X_1, X_2, \ldots$ are i.i.d. with mean zero and finite variance σ^2, then show that $Z_n = \sqrt{n}\bar{X}_n = n^{-1/2}\sum_{i=1}^n X_i$ has characteristic function of the form

$$\varphi_n(t) = \{1 - \tfrac{1}{2}\sigma^2 t^2/n + o(1/n)\}^n.$$

Prove the CLT from this.

Ex. 2.33 *The Liapunov and Lindeberg theorems: main story.* Let $X_1, X_2, \ldots$ be independent zero-mean variables with at the outset different distributions $F_1, F_2, \ldots$ and hence different standard deviations $\sigma_1, \sigma_2, \ldots$. Below we also need their characteristic functions $\varphi_1, \varphi_2, \cdots$. The question is when we can rest assured that the normalised sum

Liapunov and Lindeberg theorems

$$Z_n = (X_1 + \cdots + X_n)/B_n = \sum_{i=1}^n X_i \Big/ \Big(\sum_{i=1}^n \sigma_i^2\Big)^{1/2},$$

really tends to the standard normal, as n increases.

(a) Show that Z_n has characteristic function

$$\kappa_n(t) = \mathrm{E}\exp(itZ_n) = \varphi_1(t/B_n)\cdots\varphi_n(t/B_n).$$

(b) From Ex. A.34 we have that $|\exp(ix) - 1 - ix - \frac{1}{2}(ix)^2| \le \frac{1}{6}|x|^3$ and that $\varphi_i(s) = 1 - \frac{1}{2}\sigma_i^2 s^2 + o(s)$ for small s, so the essential idea is to write

$$\kappa_n(t) = \prod_{i=1}^n \{1 - \tfrac{1}{2}\sigma_i^2 t^2/B_n^2 + \varepsilon_{n,i}(t)\},$$

and prevail until one has found conditions that secure convergence to the desired $\exp(-\frac{1}{2}t^2)$. In view of the first lemma in Ex. 2.30, this essentially takes (i) $\sum_{i=1}^n \varepsilon_{n,i}(t) \to 0$; (ii) $\max_{i\le n}\sigma_i^2/B_n^2 \to 0$ and $\max_{i\le n}|\varepsilon_{n,i}(t)| \to 0$; and (iii) $\sum_{i=1}^n |1 - \varphi_i(t/B_n)|$ staying bounded. Show that

$$\begin{aligned}|\varphi_i(s) - (1 - \tfrac{1}{2}\sigma_i^2 s^2)| &= \Big|\int \{\exp(isx) - 1 - isx - \tfrac{1}{2}(isx)^2\}\,\mathrm{d}F_i(x)\Big| \\ &\le \int |\exp(isx) - 1 - isx - \tfrac{1}{2}(isx)^2|\,\mathrm{d}F_i(x) \le \tfrac{1}{6}|s|^3\,\mathrm{E}\,|X_i|^3.\end{aligned}$$

(c) This leads to the условие Ляпунова version of the Lindeberg theorem: show that if the variables all have finite third order moments, with $B_n \to \infty$ and

$$\sum_{i=1}^n \mathrm{E}\Big|\frac{X_i}{B_n}\Big|^3 \to 0,$$

then $\kappa_n(t) \to \exp(-\frac{1}{2}t^2)$, which from Ex. 2.26 we know is equivalent to the desired $Z_n \to_d \mathrm{N}(0,1)$. This is (already) a highly significant extension of the CLT. If the X_i are uniformly bounded, for example, with B_n^2/n having a positive limit, which would rather often be the case, then the Lyapunov condition holds. It is also possible to refine arguments and methods to show that

$$\sum_{i=1}^n \mathrm{E}\Big|\frac{X_i}{B_n}\Big|^{2+\delta} \to 0, \quad \text{for some } \delta > 0,$$

is sufficient for limiting normality.

(d) The issue waits however for even milder and actually minimal conditions, and that is, precisely, the Lindeberg condition:

$$L_n(\varepsilon) := \sum_{i=1}^n \mathrm{E}\Big|\frac{X_i}{B_n}\Big|^2 I\Big\{\Big|\frac{X_i}{B_n}\Big| \geq \varepsilon\Big\} \to 0 \quad \text{for all } \varepsilon > 0. \tag{2.9}$$

Show that if the Lyapunov condition is in force, then the Lindeberg condition holds (so Lindeberg assumes less than Lyapunov).

Ex. 2.34 *The Lindeberg theorems: nitty-gritty details.* The essential story, regarding Lyapunov and Lindeberg, has been told in the previous exercise. Here we tend to the smaller-level but nevertheless crucial remaining details, in order for the ball to be shoven across the finishing line after all the preliminary work. Again, let $X_1, X_2, \ldots$ be independent, with distributions $F_1, F_2, \ldots$, zero means, standard deviations $\sigma_1, \sigma_2, \ldots$, and characteristic functions $\varphi_1, \varphi_2, \ldots$. The crucial random variable studied is

$$Z_n = \frac{X_1 + \cdots + X_n}{(\sigma_1^2 + \cdots + \sigma_n^2)^{1/2}} = \sum_{i=1}^n \frac{X_i}{B_n},$$

with $B_n^2 = \sum_{i=1}^n \sigma_i^2$. We assume that the Lindeberg condition holds, i.e. (2.9) is true.

(a) Show that $\max_{i\leq n}(\sigma_i^2/B_n^2) \to 0$, from the Lindeberg condition. Show further that

$$\begin{aligned} |\varphi_i(t/B_n) - 1| &\leq \int |\exp(itx/B_n) - 1 - itx/B_n|\,\mathrm{d}F_i(x) \\ &\leq \tfrac12 t^2 \int (x/B_n)^2\,\mathrm{d}F_i(x) \leq \tfrac12 t^2 \max_{i\leq n}(\sigma_i^2/B_n^2), \end{aligned}$$

so all $\varphi_i(t/B_n)$ are eventually inside radius say $\frac12$ of 1. We are hence in a position to take the logarithm of $\kappa_n(t) = \mathrm{E}\exp(itZ_n)$, and work with $\log\kappa_n(t) = \sum_{i=1}^n \log\varphi_i(t/B_n)$, etc.; see the first lemma in Ex. 2.30.

(b) In continuation and refinement of arguments above, show that

$$r_n(t) = \varphi_i(t/B_n) - (1 - \tfrac12\sigma_i^2 t^2/B_n^2),$$

can be bounded, as follows:

$$\begin{aligned} |r_n(t)| &= \Big|\int \{\exp(itx/B_n) - 1 - itx/B_n - \tfrac12(itx/B_n)^2\}\,\mathrm{d}F_i(x)\Big| \\ &\leq \int |\exp(itx/B_n) - 1 - itx/B_n - \tfrac12(itx/B_n)^2|\,\mathrm{d}F_i(x) \\ &\leq \int_{|x|/B_n\leq\varepsilon} \tfrac16 \frac{|t|^3|x|^3}{B_n^3}\,\mathrm{d}F_i(x) + \int_{|x|/B_n>\varepsilon} \Big(\tfrac12\frac{|t|^2|x|^2}{B_n^2} + \tfrac12\frac{|t|^2|x|^2}{B_n^2}\Big)\,\mathrm{d}F_i(x) \\ &\leq \tfrac16|t|^3\varepsilon\frac{\sigma_i^2}{B_n^2} + t^2\,\mathrm{E}\Big|\frac{X_i}{B_n}\Big|^2 I\Big\{\Big|\frac{X_i}{B_n}\Big| \geq \varepsilon\Big\}. \end{aligned}$$

(c) Show that this leads to

$$\sum_{i=1}^n \big|\varphi_i(t/B_n) - (1 - \tfrac12\sigma_i^2 t^2/B_n^2)\big| \leq \tfrac16|t|^3\varepsilon + t^2\,L_n(\varepsilon),$$

with $L_n(\varepsilon)$ as defined in (2.9) (the Lindeberg condition), and by way of the first lemma in Ex. 2.30, that this secures what we were after, namely that

$$\prod_{i=1}^{n} \varphi_i(t/B_n) \to \exp(-\tfrac{1}{2}t^2),$$

and hence triumphantly $Z_n \to_d \mathrm{N}(0,1)$, under the Lindeberg condition only.

(d) Suppose that $\alpha > 0$ is so that $B_n/n^\alpha \to \sigma^2 > 0$, and that the Lindeberg condition

$$\frac{1}{n^\alpha} \sum_{i=1}^{n} \mathrm{E}\, X_i^2 I(|X_i| \geq n^{\alpha/2}\varepsilon) \to 0, \quad \text{for each } \varepsilon > 0,$$

holds. Show that then $n^{-\alpha/2} \sum_{i=1}^{n} X_i \to_d \mathrm{N}(0, \sigma^2)$.

Ex. 2.35 *Lindeberg, Liapunov, etc..* Let us summarise the implications related to the various conditions mentioned in the preceding couple of exercises, and some more. Let $X_1, X_2, \ldots$ be independent random variables with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. Let $B_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$, and set $X_{n,i} = (X_i - \mu_i)/B_n$. In Ex. 2.33(d) it was established that the Liapunov condition implies the Lindeberg condition, i.e., if $\sum_{i=1}^{n} |X_{n,i}|^{2+\delta} \to 0$ for some $\delta > 0$, then $\sum_{i=1}^{n} \mathrm{E}\,|X_{n,i}|^2 I(|X_{n,i}| \geq \varepsilon) \to 0$ for each $\varepsilon > 0$. Also, in Ex. 2.34(a) it was show that if the Lindeberg condition holds, then $\max_{i \leq n} \sigma_i^2/B_n^2 \to 0$. In this exercise we explore some other implications related to the Lindeberg condition.

(a) Show that if $|X_i| < M$ for all i, and $B_n \to \infty$, then the Lindeberg condition holds.

(b) Show that $\mathrm{E}\,|X_i|^{2+\delta} < M$ for all i, and the sequence of variances $\sigma_1^2, \sigma_2^2, \ldots$ is either (i) bounded below; or (ii) such that $B_n/n \to \sigma^2 > 0$; or (iii) $\sigma_i^2 \to \sigma^2 > 0$, then the Lindeberg condition holds.

(c) Suppose that there is a sequence of constants $K_1, K_2, \ldots$ such that $|X_i| < K_i$ almost surely for each i, and that $K_n/B_n \to 0$. Show that the $\sum_{i=1}^{n} X_{n,i} \to_d \mathrm{N}(0,1)$.

(d) Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with finite second moments. Show that $\max(Y_1, \ldots, Y_n)/\sqrt{n}$ tends to zero in probability.

(e) Show that if $Y_1, Y_2, \ldots$ are i.i.d. with mean μ and variance σ^2, then the Lindeberg condition holds. Thus, the Lindeberg CLT contains the CLT for i.i.d. random variables.

Ex. 2.36 *The Lindeberg CLT, yet again.* Let $(X_{n,i})_{i \leq n, n \geq 1}$ be a triangular array of mean zero random variables with variances $(\sigma_{n,i}^2)_{i \leq n, n \geq 1}$ such that $X_{n,1}, \ldots, X_{n,n}$ are independent for each n. We say that the triangular array is rowwise independent. If $\sum_{i=1}^{n} \sigma_{n,i}^2 \to 1$, and the Lindeberg condition holds, i.e., $\sum_{i=1}^{n} X_{n,i}^2 I(|X_{n,i}| \geq \varepsilon) \to 0$ for each $\varepsilon > 0$, then $\sum_{i=1}^{n} X_{n,i} \to_d \mathrm{N}(0,1)$. In this exercise we prove this fact using slightly different means than in Ex. 2.33–2.34.

(a) Let $(Y_{n,i})_{i\le n, n\ge 1}$ be a triangular array of rowwise independent mean zero normally distributed random variables with variances $\mathrm{Var}(Y_{n,i}) = \sigma_{n,i}^2$ for all i and n. Using the second lemma of Ex. 2.30, show that for any $\varepsilon > 0$

$$|\mathrm{E}\,\exp(it\sum_{i=1}^n X_{n,i}) - \mathrm{E}\,\exp(it\sum_{i=1}^n Y_{n,i})| \le \tfrac{1}{3}|t|^3\varepsilon\sum_{i=1}^n \sigma_{n,i}^2 + L_n(\varepsilon) + L_n'(\varepsilon),$$

where $L_n(\varepsilon) = \sum_{i=1}^n X_{n,i}^2 I(|X_{n,i}| \ge \varepsilon)$ and $L_n'(\varepsilon) = \sum_{i=1}^n Y_{n,i}^2 I(|Y_{n,i}| \ge \varepsilon)$. Deduce that $\sum_{i=1}^n X_{n,i} \to_d \mathrm{N}(0,1)$.

(b) (xx we invent one more point to make the exercise have both (a) and (b). xx)

Ex. 2.37 *Limiting normality of linear combinations of i.i.d. variables.* Let $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. from some distribution with mean zero and finite variance σ^2. For a sequence of multiplicative constants $a_1, a_2, \ldots$, consider

$$Z_n = \frac{\sum_{i=1}^n a_i\varepsilon_i}{B_n} = \sum_{i=1}^n (a_i/B_n)\varepsilon_i, \quad \text{with } B_n^2 = \sum_{i=1}^n a_i^2,$$

which has mean zero and variance 1. The question is what should be demanded of the a_i sequence, to ensure that $Z_n \to_d \mathrm{N}(0,1)$ (even if the ε_i distribution might be looking say skewed and multimodal and strange).

(a) Let $D_n = \max_{i\le n} |a_i|/B_n$. Writing G for the distribution of ε_i, show that

$$\sum_{i=1}^n \mathrm{E}\,\Big|\frac{a_i\varepsilon_i}{B_n}\Big|^2 I\Big(\Big|\frac{a_i\varepsilon_i}{B_n}\Big| \ge \delta\Big) \le \sum_{i=1}^n \frac{a_i^2}{B_n^2}\mathrm{E}\,\varepsilon_i^2\, I(D_n|\varepsilon_i| \ge \delta) \le \int_{|u|\ge\delta/D_n} u^2\,\mathrm{d}G(u),$$

and conclude that $Z_n \to_d \mathrm{N}(0,1)$ provided $D_n \to 0$.

(b) Under a variety of setups, one actually has $D_n \to 0$, which is hence not at all a strict condition. Verify that the condition holds, and hence limiting normality, in the following cases: (i) $a_i = 1$ (which corresponds to the plain CLT); (ii) all $|a_i|$ inside some positive $[b,c]$ interval; (iii) $a_i = i$; (iv) $a_i = i^2$ (and generalise); (v) $a_i = 1/\sqrt{i}$. Show however that the condition does not hold for $a_i = 1/i$.

(c) A grasshopper sits at zero and then starts jumping, to the right and left with equal probability, and with jump sizes $1, 2, 3, \ldots$. With S_n her position after n jumps, show that $S_n/n^{3/2}$ has a normal limit. Though she keeps on passing zero, she will not be in that vicinity; show that $\Pr(|S_n| \le cn^{1.49}) \to 0$, for each c.

(d) [xx fikse xx]Another important case to understand well is when the a_i can be considered an i.i.d. sequence, drawn from their own distribution. Show that $D_n \to_{\mathrm{pr}} 0$ if the a_i distribution has finite variance. (xx nils thinks this is if and only if, actually. what happens with Z_n if the a_i are drawn from say the $1/|x|^2$ distribution, for $|x| \ge 1$? xx)

Ex. 2.38 *Higher-order expansions of m.g.f.s.* (xx to be polished. the aim is to give more info for CLT, of the type $|F_n(t) - \Phi(t)| \le c/\sqrt{n}$ for some c. xx) [xx intro text here, perhaps by Nils. (xx nils pushes an earlier thing from Ch2 to this place; then we edit and prune and clean. xx)]

(a) Consider a variable Y, with m.g.f. $M(t) = \mathrm{E}\exp(tY)$, assumed to be finite in at least a neighbourhood around zero. We have seen in Ex. A.31 that $\mathrm{E}\,Y^r = M^{(r)}(0)$. Write ξ and σ^2 for the mean and variance of Y. Show that $M(t) = 1 + \xi t + o(h)$, for $|t|$ small. Taking a Taylor expansion to the next step, show that $M(t) = 1+\xi t+\frac{1}{2}(\xi^2+\sigma^2)t^2+o(t^2)$. Deduce also that $\log M(t) = \xi t + \frac{1}{2}\sigma^2 t^2 + o(t^2)$.

(b) We may also take the expansion to the third order, but it is simpler and more insightful to proceed from $Y = \xi + Y_0$, with Y_0 having mean zero. Show that

$$M(t) = \exp(t\xi)\,\mathrm{E}\exp(tY_0) = \exp(t\xi)\{1 + \tfrac{1}{2}\sigma^2 t^2 + \tfrac{1}{6}\gamma_3 t^3 + o(|t|^3)\},$$

where $\gamma_3 = \mathrm{E}\,(Y-\xi)^3$.

(c) Consider $Y_1, \ldots, Y_n$ i.i.d. from a distribution with mean zero and m.g.f. $M(t)$ being finite around zero. Show that $Z_n = \sqrt{n}\bar{Y}$ has

$$\begin{aligned} M_n(t) &= \mathrm{E}\exp(tZ_n) = M(t/\sqrt{n})^n \\ &= \{1 + \tfrac{1}{2}\sigma^2 t^2/n + \tfrac{1}{6}\gamma_3 t^3/n^{3/2} + o(|t|^3/n^{3/2})\}^n. \end{aligned}$$

Show from this that under the assumptions given, $\log M_n(t) = \frac{1}{2}\sigma^2 t^2 + \frac{1}{6}\gamma_3 t^3/\sqrt{n} + o(1/\sqrt{n})$. Explain why this is a proof of the CLT (via criteria given in Ex. **??**, with attention to certain further details in Ch 3 xx).

(d) (xx round off, point to CLT, identify remainder term with skewness. xx)

the Stirling approximation

Ex. 2.39 *Proving the Stirling formula.* The approximation formula

$$n! \doteq n^n e^{-n}\sqrt{2\pi n}, \quad \text{in the sense of} \quad \lim_{n\to\infty} \frac{n!}{n^n \exp(-n)(2\pi n)^{1/2}} = 1,$$

is a famous one, named after J. Stirling (1692–1770) (xx though stated earlier by A. de Moivre xx). Here we shall prove this formula via the CLT for Poisson variables.

(a) If $X_n \sim \mathrm{Pois}(n)$, show that $Z_n = (X_n - n)/\sqrt{n} \to_d Z$, a standard normal. Show that $\exp(-n)(1 + n/1 + n^2/2! + \cdots + n^n/n!) \to \frac{1}{2}$.

(b) Show that with ε small,

$$\sum_{n\le j\le n+\varepsilon\sqrt{n}} \frac{j-n}{\sqrt{n}} \exp(-n)\frac{n^j}{j!} \doteq \frac{1}{(2\pi)^{1/2}}\varepsilon,$$

and attempt to prove Stirling from this. Show also that

$$\mathrm{E}\,\max(0, Z_n) = \sum_{j\ge n} \frac{j-n}{\sqrt{n}} \exp(-n)\frac{n^j}{j!} \to \mathrm{E}\,\max(0, Z),$$

that the left hand side may be written $\sqrt{n}\exp(-n)n^n/n!$, and that the right hand side is $1/(2\pi)^{1/2}$. Deduce Stirling from this. As part of your solution, show that $\sum_{j\ge n}(j-n)p(j,n) = np(n,n)$.

Joint convergence in distribution

Ex. 2.40 *Continuity theorem for vector variables.* With $X = (X_1, \ldots, X_p)^{\mathrm{t}}$ a random vector, in dimension p, we define its characteristic functions as

$$\varphi(t_1, \ldots, t_p) = \mathrm{E}\, \exp(it^{\mathrm{t}}X) = \mathrm{E}\, \exp\{i(t_1X_1 + \cdots + t_kX_p)\}$$

for $t = (t_1, \ldots, t_p)^{\mathrm{t}}$. See Ex. A.36 in the appendix for multi-dimensional inversion formulae, and other properties of $\varphi(t_1, \ldots, t_p)$ that you will need in solving the this exercise. In this exercise we concentrate on extending the results of Ex. 2.26 and Ex. 2.27 to the multi-dimensional setting.

(a) Let $X_n = (X_{n,1}, \ldots, X_{n,p})^{\mathrm{t}}$ be a sequence of random vectors with characteristic functions $\varphi_n(t_1, \ldots, t_p)$. Suppose that $\varphi_n(t_1, \ldots, t_p) \to \varphi(t_1, \ldots, t_p)$, where φ is the characteristic function of some random vector $X = (X_1, \ldots, X_p)^{\mathrm{t}}$. Mimic the steps taken in Ex. 2.26 to show that then $X_n \to_d X$.

(b) Show that $X_n \to_d X$ if and only if $a^{\mathrm{t}}X_n \to_d a^{\mathrm{t}}X$ for each a, i.e. if and only if all linear combinations converge. This is the Cramér–Wold theorem. Cramér–Wold theorem

(c) Show that a random pair (X_n, Y_n) converges in distribution to the binormal $\mathrm{N}_2(0, \Sigma)$, with Σ having diagonal elements 1, 1, and correlation ρ, if and only if $aX_n + bY_n \to_d \mathrm{N}(0, a^2 + b^2 + 2\rho ab)$ for each (a, b).

(d) A sequence of random vectors $X_n = (X_{n,1}, \ldots, X_{n,p})^{\mathrm{t}}$ is tight if for any $\varepsilon > 0$ there is a K such that tightness in dimension k

$$\Pr(\|X_n\| > K) < \varepsilon \quad \text{for all } n,$$

where $\|x\| = (x_1 + \cdots + x_k)^{1/2}$. Show that all the $X_{n,1}, \ldots, X_{n,p}$ are tight if and only if X_n is tight. Show also that X_n is tight if and only if any linear combination

$$a^{\mathrm{t}}X_n = a_1X_{n,1} + \cdots + a_pX_{n,p},$$

is tight, where $a = (a_1, \ldots, a_p)^{\mathrm{t}}$.

(e) Show that if the sequence $(X_n)_{n\geq 1}$ of random vectors is tight, then the corresponding sequence of characteristic functions $\{\varphi_n(t_1, \ldots, t_p)\}_{n\geq 1}$ is uniformly equicontinuous.

(f) We must extend Prokohorov's theorem, proven for the one-dimensional case in Ex. 2.25(b), to the multi-dimensional case: Demonstrate that (i) if $X_n \to_d X$, then X_n is tight; and, (ii) that if $(X_n)_{n\geq 1}$ is tight, then there is a subsequence $(n_k)_{k\geq 1}$ such that $(X_{n_k})_{k\geq 1}$ converges in distribution. Prokohorov's theorem in $\mathbb{R}^p$

(g) For each n, let ϕ_n be the characteristic function of the random vector $X_n = (X_{n,1}, \ldots, X_{n,p})$. Amend the derivations of Ex. 2.27 to show that for any $\varepsilon > 0$, Lévy's continuity theorem in dim. p

$$\frac{1}{\varepsilon^p}\int_{-\varepsilon}^{\varepsilon}\cdots\int_{-\varepsilon}^{\varepsilon}\{1 - \varphi_n(t_1, \ldots, t_p)\}\,\mathrm{d}t_1 \cdots \mathrm{d}t_p \geq 2^p(1 - 1/c^p)\Pr(|X_{n,j}| \geq c/\varepsilon \text{ for all } j).$$

Deduce from this that if $\phi_n(t_1, \ldots, t_k)$ converges to a limit function that is continuous at zero, say $\varphi(t_1, \ldots, t_p)$, then $X_n \to_d X$, where X is a random vector with characteristic function $\varphi(t_1, \ldots, t_p)$. This is thus Lévy's continuity theorem in dimension p.

Ex. 2.41 *The multi-dimensional CLT.* In this exercise we prove the two central limit theorems for random vectors taking values in $\mathbb{R}^p$.

multivariate CLT

(a) Suppose $X_1, X_2, \ldots$ are i.i.d. random vectors with mean ξ and variance matrix Σ. Invoke the Cramér–Wold device to show that $\sqrt{n}(\bar{X}_n - \xi) \to_d \mathrm{N}_p(0, \Sigma)$.

multivariate Lindeberg theorem

(b) Let $X_1, X_2, \ldots$ be independent random vectors in dimension p, with means $\xi_1, \xi_2, \ldots$, and with finite positive definite variance matrices $\Sigma_1, \Sigma_2, \ldots$. Let $B_n = (\Sigma_1 + \cdots + \Sigma_n)^{1/2}$, and write $Z_{n,i} = B_n^{-1}(X_i - \xi_i)$. Show that

$$Z_n = (\Sigma_1 + \cdots + \Sigma_n)^{-1/2} \sum_{i=1}^n (X_i - \xi_i) = \sum_{i=1}^n Z_{n,i} \to_d \mathrm{N}_p(0, I_p),$$

provided the multivariate Lindeberg condition holds, i.e.,

$$\sum_{i=1}^n \mathrm{E}\, \|Z_{n,i}\|^2 I(\|Z_{n,i}\| \geq \varepsilon) \to 0, \quad \text{for each } \varepsilon > 0. \tag{2.10}$$

(c) Show that the multidimensional Liapunov condition $\sum_{i=1}^n \mathrm{E}\, \|Z_{n,i}\|^{2+\delta} \to 0$ for some $\delta > 0$, implies the Lindeberg condition in (2.10).

(d) In analogy with Ex. 2.34(d), suppose that B_n^2/n converges to some positive definite Σ, and that the Lindeberg condition

$$\frac{1}{n} \sum_{i=1}^n \mathrm{E}\, \|X_i - \xi_i\|^2 I(\|X_i - \xi_i\| \geq \sqrt{n}\varepsilon) \to 0, \quad \text{for each } \varepsilon > 0,$$

holds. Show that $n^{-1/2} \sum_{i=1}^n (X_i - \xi_i) \to_d \mathrm{N}_p(0, \Sigma)$.

(e) Suppose that B_n^2/n converges to some positive definite Σ and that the third moment of each component of $X_i - \xi_i$ is bounded for all i, i.e., there is a K so that $\mathrm{E}\, |X_{i,j} - \xi_{i,j}|^3 \leq K$ for $j = 1, \ldots, p$ and all $i \geq 1$. Show that $n^{-1/2} \sum_{i=1}^n (X_i - \xi_i) \to_d \mathrm{N}_p(0, \Sigma)$.

(f) Suppose $X_1, X_2, \ldots$ are i.i.d. with mean ξ and standard deviation σ. We assume that also the skewness and kurtosis are finite, $\gamma_3 = \mathrm{E}\, (X_i - \xi)^3/\sigma^3$ amd $\gamma_4 = \mathrm{E}\, (X_i - \xi)^4/\sigma^4 - 3$. Show from the two-dimensional CLT that

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - \xi) \\ \sqrt{n}(\widehat{\sigma}_0^2 - \sigma^2) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2, & \gamma_3\sigma^3 \\ \gamma_3\sigma^3, & \sigma^4(2 + \gamma_4) \end{pmatrix} \Big),$$

where $\widehat{\sigma}_0^2 = n^{-1} \sum_{i=1}^n (X_i - \xi)^2$.

(g) Then show that with $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$, which is a 'real estimator', as opposed to $\widehat{\sigma}_0^2$, which uses ξ, then we have $\sqrt{n}(\widehat{\sigma}^2 - \widehat{\sigma}_0^2) \to_{\mathrm{pr}} 0$; see Ex. 2.23. Conclude that the two-dimensional limit distribution result above continues to hold with $\sqrt{n}(\widehat{\sigma}^2 - \sigma^2)$ replacing $\sqrt{n}(\widehat{\sigma}_0^2 - \sigma^2)$.

(h) Let in particular the distribution of the X_i be normal, so $X_i \sim \mathrm{N}(\xi, \sigma^2)$. Show that γ_3 and γ_4 are equal to zero, so the general result above simplifies to

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - \xi) \\ \sqrt{n}(\widehat{\sigma}^2 - \sigma^2) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2, & 0 \\ 0, & 2\sigma^4 \end{pmatrix} \Big).$$

For another instructive application, involving the Gamma distribution, see Ex. 3.25.

The delta method

Ex. 2.42 *The delta method: the basics.* There are important and often reasonably simple to use approximation methods, in probability theory and statistics, going by the name of *the delta method*. It is related to functions of variables often being approximately linear, if the variables in question are not too spread out, with consequences for approximate normality.

(a) We start out in dimension one. Suppose in generic terms that $\sqrt{n}(A_n - a) \to_d Z$, with random variables A_n and a constant a. The limits are taken with respect to the index n tending to infinity, where n typically but not always is the size of an underlying sample. that $A_n \to_{\rm pr} a$. Now consider a function g, defined in at least a neighbourhood around a, and assumed to have a continuous derivative there. Show via the mean value theorem that $g(A_n) - g(a) = g'(B_n)(A_n - a)$, for some B_n between A_n and a, and use this to show that $\sqrt{n}\{g(A_n) - g(a)\} \to_d g'(a)Z$.

(b) For the typical case where the limit is a zero-mean normal, say $Z \sim {\rm N}(0, \tau^2)$, explain that $\sqrt{n}\{g(A_n) - g(a)\} \to_d {\rm N}(0, g'(a)^2\tau^2)$.

(c) For the vector case, assume $\sqrt{n}(A_n - a) \to_d Z$ in dimension p, and consider a smooth function $g(a) = g(a_1, \ldots, a_p)$, assumed to have continuous first order derivatives in a neighbourhood around a. Explain that $A_n \to_{\rm pr} a$, so A_n will be inside that neighbourhood with probability tending to one. With $c = g^*(a)$ the vector of derivatives $\partial g(a)/\partial a_1, \ldots, \partial g(a)/\partial a_p$, evaluated at a, show that $\sqrt{n}\{g(A_n) - g(a)\} \to_d c^{\rm t}Z = c_1Z_1 + \cdots + c_pZ_p$. In particular, if $Z \sim {\rm N}_p(0, \Sigma)$, explain that the limit $c^{\rm t}Z$ is a ${\rm N}(0, c^{\rm t}\Sigma c)$. The main point is that the convergence in distribution, from the previous points, holds also jointly.

(d) Sometimes one also needs the vector-to-vector extension of results above, leading to the joint limit distribution of several such g functions. Let $g = (g_1, \ldots, g_q)^{\rm t} \colon \mathbb{R}^p \to \mathbb{R}^q$ be such a function, with component functions having first order derivatives at position a, yielding a Jacobi matrix $J = \partial g(a)/\partial a$ of dimension $q \times p$. Show that $\sqrt{n}\{g(A_n) - g(a)\} \to_d JZ$. In particular, if $Z \sim {\rm N}_p(0, \Sigma)$, then the limit is ${\rm N}_q(0, J\Sigma J^{\rm t})$.

(e) The delta method has been formulated here in terms of limit distributions, implying approximate normality for function of approximately normal variables. Explain that in the framework above, with $\sqrt{n}(A_n - a) \to_d {\rm N}_p(0, \Sigma)$, then $g(A_n)$ is approximately normal, with mean $g(a)$ and variance $c^{\rm t}\Sigma c/n$. The delta method hence gives approximations also to means, variances, and covariances of smooth functions of other variables.

(f) The $\sqrt{n}$ factor above comes from the most typical uses of these methods, where variances of estimators go to zero with speed $1/n$, in terms of sample size. Show however that the mathematics goes through for any increasing sequence; if $d_n(A_n - a) \to_d Z$, with d_n tending to infinity, then $d_n\{g(A_n) - g(a)\} \to_d g'(a)Z$. There are indeed situations where the rate might be $n^{2/3}$ or $n^{1/3}$.

Ex. 2.43 *Applying the delta method.* Here we exercise our delta method muscles, to see how the general recipes of Ex. 2.42 may be applied in a few situations. As is clear, once

we have established limiting normality for certain quantities, e.g. via the CLT, there is a host of easy-to-harvest applications for functions of these start quantities.

(a) If $\sqrt{n}(A_n-a) \to_d \mathrm{N}(0,1)$, find out what happens to $\sqrt{n}(A_n^3-a^3)$, and to $\sqrt{n}\{\exp(kA_n)-\exp(ka)\}$, where k is a constant.

(b) For Y a binomial (n,p), we have of course $\mathrm{Var}\,\widehat{p}=p(1-p)/n$ for the classic estimator $\widehat{p}=Y/n$. Use the delta method to find approximations to the means, variances, and distributions of (i) the estimated odds ratio $\widehat{p}/(1-\widehat{p})$; (ii) the estimated log-odds-ratio $\log\widehat{p}-\log(1-\widehat{p})$; (iii) the transformed estimator $\widehat{\gamma}=2\arcsin(\widehat{p}^{1/2})$. In particular, for the latter, show $\sqrt{n}(\widehat{\gamma}-\gamma)\to_d \mathrm{N}(0,1)$, with $\gamma=2\arcsin(p^{1/2})$.

(c) Suppose $\widehat{p}_1=Y_1/n$ and $\widehat{p}_2=Y_2/n$ are two binomial estimates, with the same sample size n. Then $\sqrt{n}(\widehat{p}_1-p_1)\to_d Z_1$ and $\sqrt{n}(\widehat{p}_2-p_2)\to_d Z_2$, where $Z_j\sim\mathrm{N}(0,p_j(1-p_j))$. Find the approximate normal distribution of $\widehat{p}_1/\widehat{p}_2$, viewed as an estimator of p_1/p_2. Modify arguments appropriately to find a good approximation to the variance of $\widehat{p}_1/\widehat{p}_2$, and its approximate normal distribution, also in the case of unequal sample sizes, say n_1 and n_2. [xx pointer to Story i.1. xx]

(d) Suppose $Y_1,\ldots,Y_n$ are independent from the geometric distribution with $\Pr(Y_i=y)=(1-p)^{y-1}p$ for $y=1,2,\ldots$. We learned in Ex. 1.24 that the mean and variance are $1/p$ and $(1-p)/p^2$. Find first the limiting distribution of $\sqrt{n}(\bar{Y}-1/p)$ and then that of $\sqrt{n}(\widehat{p}-p)$, where $\widehat{p}=1/\bar{Y}$.

(e) Suppose (a,b) is a certain position on the map, where one only has estimates, say A_n and B_n, for its x- and y-coordinates. Assume these are independent, approximately unbiased, and approximate normal, after n measurements. We formalise a version of this as $\sqrt{n}(A_n-a)\to_d N_1$ and $\sqrt{n}(B_n-b)\to_d N_2$, the limit variables N_1,N_2 being independent and standard normal. Having observed A_n and B_n, explain how you can put up 90 percent confidence intervals for a and b separately. Construct also a 90 percent confidence circle for (a,b).

(f) Let us pass from Cartesian to polar coordinates, letting

$$R_n=\|(A_n,B_n)\|=(A_n^2+B_n^2)^{1/2}\quad\text{and}\quad\widehat{\alpha}_n=\arctan(B_n/A_n),$$

seen as estimators of the length $r=\|(a,b)\|$ and angle $\alpha=\arctan(b/a)$. Find the limit distributions for $\sqrt{n}(R_n-r)$ and $\sqrt{n}(\widehat{\alpha}_n-\alpha)$, and show that these are independent in the limit.

(g) Suppose one observes $(A_n,B_n)=(4.44,2.22)$, with $n=100$. Construct and display an approximate 90 percent confidence circle for (a,b), and then approximate 90 percent confidence intervals for the length r and angle α. How can you construct confidence intervals for r and α jointly, say $I_{r,n}$ and $I_{\alpha,n}$, such that the probability that $(r\in I_{r,n})\cap(\alpha\in I_{\alpha,n})$ converges to 0.90?

Ex. 2.44 *Limiting normality for multinomials.* Consider the multinomial setup of Ex. 1.5, with $(Y_1,\ldots,Y_k)$ counting the number of events of type $1,\ldots,k$ in n independent

experiments, each time with probabilities $p = (p_1, \ldots, p_k)^{\rm t}$. Here we sort out the basic large-sample behaviour of the relative frequencies $\widehat{p}_j = Y_j/n$. This is used e.g. in the Karl Pearson 1900 Story vii.1.

(a) Show that these $\widehat{p}_j = Y_j/n$ are consistent, with $\sqrt{n}(\widehat{p}_j - p_j) \to_d {\rm N}(0, p_j(1-p_j))$. Show more generally that there is full joint convergence in distribution here; $X_n = \sqrt{n}(\widehat{p} - p) \to_d Z \sim {\rm N}_k(0, \Sigma)$, where Σ is the matrix with elements $\sigma_{j,\ell} = p_j\delta_{j,\ell} - p_j p_\ell$. It may be written $\Sigma = D - pp^{\rm t}$ with D diagonal with elements p_j. Verify that this is consistent with $\sum_{j=1}^k Z_j = 0$.

(b) For $\gamma = g(p_1, \ldots, p_k)$ any smooth function of the relative frequencies, with natural estimator $\widehat{\gamma} = g(\widehat{p}_1, \ldots, \widehat{p}_k)$, show that $\sqrt{n}(\widehat{\gamma} - \gamma) \to_d {\rm N}(0, \tau^2)$, with $\tau^2 = c^{\rm t}\Sigma c = c^{\rm t}Dc - (c^{\rm t}p)^2$, where $c = \partial g(p)/\partial p$. Check what this says, for the case of $\gamma = p_1 + \cdots + p_k$.

(c) Consider (X, Y, Z) being trinomial (n, p, q, r). With $\widehat{p} = X/n$, $\widehat{q} = Y/n$, $\widehat{r} = Z/n$, find the limit distribution for $\widehat{\gamma} = \widehat{p}/(\widehat{q}\widehat{r})^{1/2}$, as well as for $\widehat{\delta} = 2\arcsin(\widehat{p}^{1/2}) - 2\arcsin(\widehat{q}^{1/2})$.

Ex. 2.45 *Delta method calculus for the normal case.* Let $Y_1, \ldots, Y_n$ be i.i.d. from the normal ${\rm N}(\xi, \sigma^2)$, with standard estimators $\widehat{\xi} = \bar{Y}$ and $\widehat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. In various exercises in Ch. 1 we have worked with *exact finite-sample* calculus, for certain basic parameters, like the mean, variance, the quantile $\gamma_q = \xi + z_p\sigma$. Here we show how the delta method, starting with the basic limit distributions for the two parameters, can be used to put up large-sample normal approximations for *any* functions of the parameters, in cases where it might be too hard to carry out exact finite-sample calculus.

(a) Since the skewness and the kurtosis for the normal are zero, show that the general result from Ex. 2.41 implies that $(\sqrt{n}(\bar{Y} - \xi), \sqrt{n}(\widehat{\sigma} - \sigma))^{\rm t}$ tends to say $(A, B)^{\rm t}$, with these being independent and zero-mean normals with variances σ^2 and $\frac{1}{2}\sigma^2$. Show this directly, from the normality assumptions, as opposed to deriving it as a special case of the general statement. Note also that the $\sqrt{n}(\bar{Y} - \xi) \sim {\rm N}(0, \sigma^2)$ holds exactly, for each finite n.

(b) With $\alpha = g(\xi, \sigma)$, for any smooth function of the two parameters, the natural estimator is $\widehat{\alpha} = g(\bar{Y}, \widehat{\sigma})$. Show that

$$\sqrt{n}(\widehat{\alpha} - \alpha) \to_d cA + dB \sim {\rm N}(0, (c^2 + \tfrac{1}{2}d^2)\sigma^2),$$

where c and d are the partial derivatives of g, evaluated at the position (ξ, σ). Show how this leads to construction of confidence intervals for the α parameter.

(c) Consider the probability $p = \Pr(Y \geq y_0) = 1 - \Phi((y_0 - \mu)/\sigma)$, for some given threshold y_0, and the associated estimator $\widehat{p} = 1 - \Phi((y_0 - \bar{Y})/\widehat{\sigma})$. Finr the limit distribution of $\sqrt{n}(\widehat{p} - p)$, and use this to put up a confidence interval for p, with coverage level converging to 0.90. Compare with the simpler estimator $p^* = n^{-1}\sum_{i=1}^n I(Y_i \geq y_0)$, the binomial proportion, which bypasses the normal assumption.

(d) Then consider the parameter $\kappa = \xi/\sigma$, the normalised mean (so its value is unchanged when one passes from say millimetres to metres). Find the limit distribution for $\widehat{\kappa} = \bar{Y}/\widehat{\sigma}$, and construct an approximate 90 percent confidence interval. [xx also try exact inference for this parameter, and compare. xx]

(e) Assume ξ is positive, so that the so-called coefficient of variation $\gamma = \sigma/\xi$ has a natural interpretation. With $\widehat{\gamma} = \widehat{\sigma}/\widehat{\xi}$ the plug-in estimator, show that $\sqrt{n}(\widehat{\gamma} - \gamma)$ tends to a normal with variance $\tau^2 = \frac{1}{2}\gamma^2(1+2\gamma^2)$. Find also a variance stabilising transformation, from γ to $\gamma^* = h(\gamma)$, with the property that $\sqrt{n}(\widehat{\gamma}^* - \gamma^*) \to_d \mathrm{N}(0,1)$: one such is $h(\gamma) = \log\gamma - \log(1 + (1+2\gamma^2)^{1/2})$. Explain how confidence intervals can be set for γ via this.

Ex. 2.46 *Estimating mean and standard deviation outside normality.* Let $Y_1, \ldots, Y_n$ be i.i.d. from a distribution with finite fourth moment, and consider the usual mean $\bar{Y}_n$ and empirical standard deviation S_n. Under normality we have precise finite-sample results regarding their distributions, see Ex. 1.45, but here we investigate behaviour outside normality.

(a) Let as on previous occasions γ_3 and γ_4 be the skewness and kurtotis of the distribution. Use the delta method, with previous results from Ex. 2.41, to show that

$$\begin{pmatrix} \sqrt{n}(\bar{Y}_n - \xi) \\ \sqrt{n}(\widehat{\sigma}_n - \sigma) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1, & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3, & \frac{1}{2} + \frac{1}{4}\gamma_4 \end{pmatrix}\Big).$$

Explain that this implies $\Pr(|\bar{Y}_n - \xi| \le 1.96\,\widehat{\sigma}) \to 0.95$, even outside normality; the normality-based standard recipes regarding inference for the mean parameter is hence not muich hampered by that modelling assumption.

(b) The situation is different when it comes to inference for the standard deviation, however. Show that $\sqrt{n}(\widehat{\sigma} - \sigma)/\widehat{\kappa} \to_d \mathrm{N}(0,1)$, if $\widehat{\kappa}$ is a consistent estimator for $\kappa = (\frac{1}{2} + \frac{1}{4}\gamma_4)^{1/2}$. Construct indeed such an estimator; see also Ex. 3.12. For an application of this, with sample size up to a million, see Story vii.2.

Ex. 2.47 *Approximate variances and the delta method.* We have built a well-working apparatus around the delta method and seen various applications. Statements reached are in terms of precise limit distributions, though without going into the quality of the resulting approximations. The present exercise goes into the details of variances and covariances for functions of sample averages.

(a) Suppose $X_1, \ldots, X_n$ are i.i.d. with mean zero, variance σ^2, and finite skewness $\gamma_3 = \mathrm{E}\,(X_i/\sigma)^3$ and kurtosis $\gamma_4 = \mathrm{E}\,(X_i/\sigma)^4 - 3$. With $\bar{X}_n$ as usual being the average, show then that

$$\begin{aligned} \mathrm{E}\,\bar{X}_n^2 &= \sigma^2/n, \\ \mathrm{E}\,\bar{X}_n^3 &= (\sigma^3/n^2)\gamma_3, \\ \mathrm{E}\,\bar{X}_n^4 &= (\sigma^4/n^2)(3 + \gamma_4/n), \\ \mathrm{Var}\,\bar{X}_n^2 &= (\sigma^4/n^2)(2 + \gamma_4/n). \end{aligned}$$

(b) Let then $Y_1, \ldots, Y_n$ be i.i.d., with finite mean ξ, standard deviation σ, skewness γ_3, kurtosis γ_4. With $\bar{Y}_n$ the sample average, consider then the variable

$$Z_n = a_0 + a_1(\bar{Y}_n - \xi) + \tfrac{1}{2}a_2(\bar{Y}_n - \xi)^2.$$

Show that $\mathrm{E}\, Z_n = a_0 + \frac{1}{2}a_2\sigma^2/n$, and that

$$\mathrm{Var}\, Z_n = a_1^2\sigma^2/n + (1/n^2)\{\tfrac{1}{4}a_2^2\sigma^4(2+\gamma_4/n) + a_1a_2\sigma^3\gamma_3\}.$$

(c) Consider any smooth function $Z_n = g(\bar{Y}_n)$. Since $\bar{Y}_n$ is close to ξ with high probability (see Ex. 2.11), it makes sense to carry out a Taylor expansion,

$$Z_n = g(\xi) + g'(\xi)(\bar{Y}-\xi) + \tfrac{1}{2}g''(\xi)(\bar{Y}-\xi)^2 + \delta_n,$$

where δ_n is a smaller-sized remainder term – you may prove that $n^{3/2}\delta_n$ is bounded in probability, provided g has three derivatives in a neighbourhood around ξ. Show from the above that

$$\mathrm{Var}\, Z_n = g'(\xi)^2\sigma^2/n + (1/n^2)\{\tfrac{1}{4}g''(\xi)^2\sigma^4(2+\gamma_4/n) + g'(\xi)g''(\xi)\sigma^3\gamma_3\} + o(1/n^2).$$

There is hence a clear leading $O(1/n)$ term for the variance, with other terms being of size $O(1/n^2)$. Explain how this relates to the basics of the delta method from Ex. 2.42.

(d) Suppose A is a random variable with mean a and finite variance, and that $g(y)$ is smooth in a neighbourhood around a. Use the Taylor approximation

$$g(y) = g(a) + g'(a)(y-a) + \tfrac{1}{2}g''(a)(y-a)^2 + O(|y-a|^3),$$

valid for y close to a, to show that

$$\mathrm{E}\, g(A) \doteq g(a) + \tfrac{1}{2}g''(a)\mathrm{Var}\, Y, \quad \mathrm{Var}\, g(Y) \doteq g'(a)^2\,\mathrm{Var}\, Y,$$

and indicate the sizes of the error terms involved.

Ex. 2.48 *The empirical correlation coefficient under binormality.* For i.i.d. data pairs (X_i, Y_i), the classical empirical correlation coefficient is

$$R_n = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\{\sum_{i=1}^n (X_i-\bar{X})^2\}^{1/2}\{\sum_{i=1}^n (Y_i-\bar{Y})^2\}^{1/2}} = n^{-1}\sum_{i=1}^n \frac{X_i-\bar{X}}{\widehat{\sigma}_1}\frac{Y_i-\bar{Y}}{\widehat{\sigma}_2}, \quad (2.11)$$

with $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ the empirical variances $n^{-1}\sum_{i=1}^n (X_i-\bar{X})^2$ and $n^{-1}\sum_{i=1}^n (Y_i-\bar{Y})^2$. Here we find the the limit distribution of R_n under binormality.

(a) Assume first that the (X_i, Y_i) pairs are from a zero-mean binormal with variances 1 and correlation $\rho \in (-1,1)$; see Ex. 1.40. Use results from Ex. 1.41, including $Y_i \,|\, X_i \sim \mathrm{N}(\rho X_i, 1-\rho^2)$, to derive expressions for $\mathrm{E}\, X_i^2Y_i^2$, $\mathrm{E}\, X_i^3Y_i$, $\mathrm{E}\, X_iY_i^3$. Use these to show that

$$\Sigma = \mathrm{Var}\begin{pmatrix} X_i^2 \\ Y_i^2 \\ X_iY_i \end{pmatrix} = \begin{pmatrix} 2, & 2\rho^2, & 2\rho \\ 2\rho^2, & 2, & 2\rho \\ 2\rho, & 2\rho, & 1+\rho^2 \end{pmatrix}.$$

Use the CLT to argue that

$$\begin{pmatrix} A_n \\ B_n \\ C_n \end{pmatrix} = \sqrt{n}\begin{pmatrix} n^{-1}\sum_{i=1}^n X_i^2 - 1 \\ n^{-1}\sum_{i=1}^n Y_i^2 - 1 \\ n^{-1}\sum_{i=1}^n X_iY_i - \rho \end{pmatrix} \to_d \begin{pmatrix} A \\ B \\ C \end{pmatrix} \sim \mathrm{N}_3(0,\Sigma).$$

With $R_{n,0} = C_n/(A_nB_n)^{1/2}$, use the delta method to show that $\sqrt{n}(R_{n,0}-\rho) \to_d Z = -\frac{1}{2}\rho A - \frac{1}{2}\rho B + C$, and that in fact $Z \sim \mathrm{N}(0,(1-\rho^2)^2)$.

(b) Then generalise to the situation where the (X_i, Y_i) pairs are i.i.d. from a zero-mean binormal, with standard deviations σ_1, σ_2 and correlation ρ. Show that we still have $\sqrt{n}(R_{n,0} - \rho) \to_d \mathrm{N}(0, (1-\rho^2)^2)$.

(c) Then go one step further, to the full five-parameter binormal situation, with unknown means ξ_1, ξ_2, standard deviations σ_1, σ_2, and correlation ρ. Argue first that we must have $\sqrt{n}(R_{n,1} - \rho) \to_d \mathrm{N}(0, (1-\rho^2))$, where $R_{n,1}$ is as in (2.11) but using ξ_1, ξ_2 instead of $(\bar{X}, \bar{Y})$. Then, finally, show what we really are after, that $\sqrt{n}(R_n - \rho)$ must have the same limit distribution.

variance stabilising transformation

(d) Explain that if $h(\rho)$ is a smooth function, then $\sqrt{n}\{h(R_n) - h(\rho)\} \to_d h'(\rho)(1-\rho^2)\,\mathrm{N}(0,1)$. Show that with the particular choice $\zeta = \frac{1}{2}\log\{(1+\rho)/(1-\rho)\}$ the variance is being stabilised, and $\sqrt{n}(\widehat{\zeta} - \zeta) \to_d \mathrm{N}(0,1)$, where $\widehat{\zeta} = \frac{1}{2}\log\{(1+R_n)/(1-R_n)\}$. This is called Fisher's zeta. Show that $\widehat{\zeta} \pm 1.645/\sqrt{n}$ becomes an approximate 90 percent confidence interval for ζ, and transform this to an approximate 90 percent confidence interval for ρ.

Fisher's zeta

Ex. 2.49 *The empirical correlation coefficient, general case.* Here we use some of the arguments of Ex. 2.48 to find the limit distribution of the empirical correlation R_n of (2.11) also outside binormality. Assume $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. pairs from a distribution with means ξ_1, ξ_2, standard deviations σ_1, σ_2, and correlation ρ. Write $a_{j,k} = \mathrm{E}\, U_i^j V_j^k$ for cross moments of the standardised $U_i = (X_i - \xi_1)/\sigma_1$ and $V_i = (Y_i - \xi_2)/\sigma_2$, where it is assumed that fourth order moments $a_{4,0}$ and $a_{0,4}$ are finite.

(a) Show that $\sqrt{n}(R_n - R_{n,0}) \to_{\mathrm{pr}} 0$, where $R_{n,0}$ is as R_n, but using the real ξ_1, ξ_2 instead of their estimators $\bar{X}, \bar{Y}$. Show also that the distribution of R_n and $R_{n,0}$ must depend on ρ but not on $\xi_1, \xi_2, \sigma_1, \sigma_2$. We may hence carry out our large-sample investigation with the standardised (U_i, V_i) rather than the (X_i, Y_i). Work with

$$\begin{pmatrix} A_n \\ B_n \\ C_n \end{pmatrix} = \begin{pmatrix} \sqrt{n}(n^{-1}\sum_{i=1}^n U_i^2 - 1) \\ \sqrt{n}(n^{-1}\sum_{i=1}^n V_i^2 - 1) \\ \sqrt{n}(n^{-1}\sum_{i=1}^n U_i V_i - \rho) \end{pmatrix},$$

and show that $(A_n, B_n, C_n)^{\mathrm{t}} \to_d (A, B, C)^{\mathrm{t}} \sim \mathrm{N}_3(0, \Sigma)$, for the variance matrix Σ of $(U_i^2, V_i^2, U_i V_i)^{\mathrm{t}}$. Spell out the elements of this matrix, using the $a_{j,k}$. Check that this agrees with the Σ of Ex. 2.48 under binormality.

(b) Then show that $\sqrt{n}(R_n - \rho) \to_d Z = -\frac{1}{2}\rho A - \frac{1}{2}\rho B + C$, and give an expression for the limit distribution variance τ^2. Explain how τ may be estimated from the data, and how this leads to confidence intervals of the type $R_n \pm 1.96\,\widehat{\tau}/\sqrt{n}$ for ρ.

(c) For a concrete illustration, consider the joint density $f(x,y) = 1 + a(x - \frac{1}{2})(y - \frac{1}{2})$ for (x,y) in the unit square. Find the allowed parameter range for a, and a formula for the correlation coefficient ρ in terms of a. Then apply the above to find the limit distribution of $\sqrt{n}(R_n - \rho)$.

Ex. 2.50 *The delta method outside root-n terrain.* (xx to come. not always $\sqrt{n}(X_n - a) \to_d \mathrm{N}(0, \tau^2)$ terrain. different limits, different speeds. xx)

Ex. 2.51 *Stretching the delta method.* (xx to be filled in. with $\sqrt{n}(X_n - a) \to_d V$, we have $Z_n = \sqrt{n}\{g(X_n) - g(a)\} \to_d g'(a)V$ for a fixed $g(x)$. here we consider $Z_n = \sqrt{n}\{g_n(X_n) - g_n(a)\}$. with $g_n''(a) = o(\sqrt{n})$ we may still have the right approximation. example: $Z_n = \sqrt{n}\{\exp(c_n X_n) - 1\}$. xx)

The strong law of large numbers

Ex. 2.52 *The Strong Law of Large Numbers: the basics.* (xx to be cleaned. xx) Suppose $X_1, X_2, \ldots$ are i.i.d. from a distribution with finite $\mathrm{E}\,|X_i|$. Then the mean $\xi = \mathrm{E}\,X_i$ exists, and we are aiming to prove the strong LLN of (2.3), that the event

$$A = \{\bar{X}_n \to \xi\} = \cap_{\varepsilon>0} \cup_{n_0 \geq 1} \cap_{n \geq n_0} \{|\bar{X}_n| \leq \varepsilon\}$$

has probability equal to one hundred percent. We may for simplicity and without loss of generality take $\xi = 0$ below.

(a) Show that A is the same as $\cap_{N \geq 1} \cup_{n_0 \geq 1} \cap_{n \geq n_0} \{|\bar{X}_n| \leq 1/N\}$, and deduce in particular from this that A is actually measurable – so it does make well-defined sense to work with its probability.

(b) Show that if $\Pr(A_N) = 1$ for all N, then $\Pr(\cap_{N \geq 1} A_N) = 1$ – if you're fully certain about a countable number of events, then you're also fully certain about all of them, jointly. This is actually not true with a bigger index set: if $X \sim \mathrm{N}(0,1)$, then you're 100 percent sure that $B_x = \{X \text{ is not } x\}$ takes place, for each single x, but from this does it *not* follow that you should be sure about $\cap_{\text{all } x} B_x$. Explain why.

(c) Show that $\Pr(A) = 1$ if and only if $\Pr(B_{n_0}) \to 0$, for each $\varepsilon > 0$, where $B_{n_0} = \cup_{n \geq n_0} \{|\bar{X}_n| \geq \varepsilon\}$. In words: for a given ε, the probability should be very low that there is *any* $n \geq n_0$ with $|\bar{X}_n| \geq \varepsilon$.

(d) A simple bound is of course $\Pr(B_{n_0}) \leq \sum_{n \geq n_0} \Pr\{|\bar{X}_n| \geq \varepsilon\}$, so it suffices to show, if possible, under appropriate conditions, that $\sum_{n \geq 1} \Pr\{|\bar{X}_n| \geq \varepsilon\}$ is a convergent series. With finite variance σ^2, show that the classic simple Chebyshov bound, see Ex. 2.11, does *not* solve any problem here.

(e) (xx calibrate better with Ex. 2.11. xx) Show, however, that if the fourth moment is finite, then

$$\Pr\{|\bar{X}_n| \geq \varepsilon\} \leq \frac{1}{\varepsilon^4} \mathrm{E}\,|\bar{X}_n|^4 \leq \frac{c}{\varepsilon^4} \frac{1}{n^2},$$

for a suitable c. So under this condition, which is moderately hard, we've proven the strong LLN.

(f) One may squeeze more out of the chain of arguments below, which we indicate here, without full details. Assume $\mathrm{E}\,|X_i|^r$ is finite, for some $r > 2$, like $r = 2.02$. Then one may show, via arguments in von Bahr (1965), that the sequence $\mathrm{E}\,|\sqrt{n}\bar{X}_n|^r$ is bounded. This leads to the bound

$$\Pr\{|\bar{X}_n| \geq \varepsilon\} \leq \frac{1}{(\sqrt{n}\varepsilon)^r} \mathrm{E}\,|\sqrt{n}\bar{X}_n|^r,$$

and these form a convergent series. We have hence proven (modulo the von Bahr thing) that the strong LLN holds for finite $\mathrm{E}\,|X_i|^{2+\varepsilon}$, an improvement over the finite $\mathrm{E}\,|X_i|^4$ condition. – To get further, trimming away on the conditions until we are at the Kolmogorovian position of only requiring finite mean, we need more technicalities; see the following Ex. 2.53.

Ex. 2.53 *The Strong Law of Large Numbers: nitty-gritty details.* This exercise goes through the required extra technical details, along with a few intermediate lemmas, to secure a full proof of the full LLN theorem: as long as $\mathrm{E}\,|X_i|$ is finite, the infinite sequence of sample means $\bar{X}_n$ will with probability equal to a hundred percent converge to $\xi = \mathrm{E}\,X_i$.

(a) We start with Kolmogorov's inequality: Consider independent zero-mean variables $X_1, \ldots, X_n$ with variances $\sigma_1^2, \ldots, \sigma_n^2$, and with partial sums $S_i = X_1 + \cdots + X_i$. Then

$$\Pr\{\max_{i\le n} |S_i| \ge \varepsilon\} \le \frac{\mathrm{Var}\,S_n}{\varepsilon^2} = \frac{1}{\varepsilon^2}\sum_{i=1}^{n}\sigma_i^2.$$

Note that this is a much stronger result than the special case of caring only about $|S_n|$, with $\Pr\{|S_n| \ge \varepsilon\} \le \mathrm{Var}\,S_n/\varepsilon^2$, which is the Chebyshov inequality. To prove it, work with the disjoint decomposition

$$A_i = \{|S_1| < \varepsilon, \ldots, |S_{i-1}| < \varepsilon, |S_i| \ge \varepsilon\} \quad \text{and} \quad A = \cup_{i=1}^n A_i = \{\max_{i\le n} |S_i| \ge \varepsilon\}.$$

Show that $\mathrm{E}\,S_n^2 \ge \mathrm{E}\,S_n^2 I(A) = \sum_{i=1}^n \mathrm{E}\,S_n^2 I(A_i)$, that

$$\mathrm{E}\,S_n^2 I(A_i) = \mathrm{E}\,(S_i + S_n - S_i)^2 I(A_i) \ge \varepsilon^2 \Pr(A_i),$$

and that this leads to the inequality asked for.

(b) Consider a sequence of independent $X_1, X_2, \ldots$ with means zero and variances $\sigma_1^2, \sigma_2^2, \ldots$. Show that if $\sum_{i=1}^\infty \sigma_i^2$ is convergent, then $\sum_{i=1}^\infty X_i$ is convergent with probability 1. – It suffices to show that the sequence of partial sums $S_n = X_1 + \cdots + X_n$ is Cauchy with probability 1. Show that this is the same as

$$\lim_{n\to\infty} \Pr\big[\cup_{i,j\ge n}\{|S_i - S_j| \ge \varepsilon\}\big] = 0 \quad \text{for each } \varepsilon > 0.$$

Use the Kolmogorov inequality to show this.

(c) A quick example to illustrate this result is as follows. Consider $X = X_1/10 + X_2/100 + X_3/1000 + \cdots$, a random number in the unit interval, with the X_i independent, and with no further assumptions. Show that X exists with probability 1.

(d) Prove that if $\sum_{i=1}^\infty a_i/i$ converges, then $\bar{a}_n = (1/n)\sum_{i=1}^n a_i \to 0$. To show this, consider $b_n = \sum_{i=1}^n a_i/i$, so that $b_n \to b$ for some b. Show $a_n = n(b_n - n_{n-1})$, valid also for $n = 1$ if we set $b_0 = 0$, and which leads to $\sum_{i=1}^n a_i = nb_n - b_0 - b_1 - \cdots - b_{n-1}$.

(e) From the above, deduce that if $X_1, X_2, \ldots$ are independent with means $\xi_1, \xi_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$, and $\sum_{i=1}^\infty \sigma_i^2/i^2$ converges, then $\bar{X}_n - \bar{\xi}_n \to_{\text{a.s.}} 0$. Here $\bar{\xi}_n = (1/n)\sum_{i=1}^n \xi_i$.

(f) Use the above to show that if $X_1, X_2, \ldots$ are independent with zero means, and all variances are bounded, then indeed $\bar{X}_n \to_{\text{a.s.}} 0$. Note that this is a solid generalisation of what we managed to show in (xx calibrate xx) – first, the distributions are allowed to be different (not identical); second, we have landed at a.s. convergence with the mild assumption of finite and bounded variances, whereas we there needed the harsher conditions of finite fourth moments.

(g) We're close to the Pole. For i.i.d. zero mean variables $X_1, X_2, \ldots$, split them up with the little trick

$$X_i = Y_i + Z_i, \quad \text{with} \quad Y_i = X_i I(|X_i| < i), \quad Z_i = X_i I(|X_i| \geq i).$$

We have $\bar{X}_n = \bar{Y}_n + \bar{Z}_n$, so it suffices to demonstrate that $\bar{Y}_n \to_{\text{a.s.}} 0$ and $\bar{Z}_n \to_{\text{a.s.}} 0$ (since an intersection of two sure events is sure). Use the Borel–Cantelli lemma, in concert with $\mathrm{E}\,|X_i| = \int_0^\infty \Pr(|X_i| \geq x)\,\mathrm{d}x$, to show that only finitely many Z_i are non-zero. Then use previous results to demonstrate $\bar{Y}_n - \bar{\xi}_n \to_{\text{a.s.}} \to 0$ and $\bar{\xi}_n \to 0$, where $\bar{\xi}_n$ is the average of $\xi_i = \mathrm{E}\,Y_i$.

(h) So we've managed to prove the Strong LLN; good. Attempt also to prove the interesting converse that if $\mathrm{E}\,|X_i| = \infty$, then the sequence of sample means is pretty erratic indeed:

$$\Pr\{\limsup_{n\to\infty} \bar{X}_n = \infty\} = 1, \quad \Pr\{\liminf_{n\to\infty} \bar{X}_n = -\infty\} = 1.$$

Simulate a million realisations from the density $f(x) = 1/x^2$, for $x \geq 1$, in your nearest computer, display the sequence of $\bar{X}_n$ on your screen, and comment.

Ex. 2.54 *Yes, we converge with probability 1.* We've proven that the sequence of empirical means converges almost surely to the population mean, under the sole condition that this mean is finite. This half-automatically secures almost sure convergence of various other natural quantities, almost without further efforts.

(a) Suppose $X_1, X_2, \ldots$ are i.i.d. with finite variance σ^2. Show that the classical empirical standard deviation $\widehat{\sigma} = \{\sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1)\}^{1/2}$ converges a.s. to σ. Note again that nothing more is required than a finite second moment.

(b) Suppose the third moment is finite, such that the skewness $\gamma_3 = \mathrm{E}\,\{(X-\xi)/\sigma\}^3$ is finite. Show that $\widehat{\gamma}_{3,n} = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)^3/\widehat{\sigma}^3$ is strongly consistent for γ_3.

(c) Then suppose the fourth moment is finite, such that the kurtosis $\gamma_4 = \mathrm{E}\,\{(X - \xi)/\sigma\}^4 - 3$ is finite. Construct a strongly consistent estimator for this kurtosis.

(d) Assume that $(X_1, Y_1), (X_2, Y_2), \ldots$ is an i.i.d. sequence of random pairs, with finite variances, and define the population correlation coefficient in the usual fashion, as $\rho = \mathrm{cov}(X, Y)/(\sigma_1\sigma_2)$. Show that the usual empirical correlation coefficient

$$R_n = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\{\sum_{i=1}^n (X_i - \bar{X}_n)^2\}^{1/2}\{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2\}^{1/2}}$$

converges with probability one hundred percent to ρ.

(e) Formulate and prove a suitable statement regarding almost sure convergence of smooth functions of means.

(f) Let $X_1, X_2, \ldots$ be an i.i.d. sequence of nonnegative variables such that $\mathrm{E} \log X_i = \xi$ is finite. Show that the harmonic means $H_n = (X_1 \cdots X_n)^{1/n}$ converge with probability 1 to $\exp(\xi)$.

Ex. 2.55 *Glivenko–Cantelli theorem.* For i.i.d. observations $Y_1, \ldots, Y_n$, we form the empirical c.d.f., as in Ex. 2.13, with $F_n(t) = n^{-1} \sum_{i=1}^n I(Y_i \leq t)$. We have seen that since $F_n(t)$ is just a binomial ratio, $F_n(t) \to_{\text{a.s.}} F(t)$, for each t. It is a remarkable fact that this convergence also takes place uniformly, with probability 1. This is the Glivenko–Cantelli theorem: with $D_n = \max_t |F_n(t) - F(t)|$, the max taken over all t in the domain in question, we have $\Pr(D_n \to 0) = 1$. This means that regardless of any strange or complicated aspects of the distribution F, with enough data one will be able to learn these. See also Ex. 3.9 and 9.22 for more information, regarding the speed with which $D_n \to 0$.

Glivenko–Cantelli

(a) Choose $t_1 < \cdots < t_m$, creating a finite number of cells $[t_j, t_{j+1})$, where we take $t_0 = -\infty$ and $t_{m+1} = \infty$. With $A_{m,j}$ the event that $F_n(t_j) \to F(t_j)$, argue that $\Pr(\cap_{j=1}^m A_{m,j}) = 1$.

(b) Consider any t in the cell $[t_j, t_{j+1})$. Writing $D_n(t) = F_n(t) - F(t)$, use monotinicity of F_n and F to show that

$$D_n(t_j) - \{F(t_{j+1}) - F(t_j)\} \leq F_n(t) - F(t) \leq D_n(t_{j+1}) + F(t_{j+1} - F(t_j).$$

Deduce that

$$\max_{t_j \leq t < t_{j+1}} |D_n(t)| \leq B_m + C_m,$$

where $B_m = \max_{1 \leq j \leq m} |D_n(t_j)|$ and $C_m = \max_{1 \leq j \leq m} \{F(t_{j+1}) - F(t_j)\}$.

(c) Show that $\Pr(\limsup D_n \leq C_m) = 1$.

(d) For each $\varepsilon > 0$, show that a partition into cells can be arranged, with high m if required, so that $C_m \leq \varepsilon$. Conclude that $\Pr(D_n \to 0) = 1$.

(e) Choose some moderately complicated normal mixture, of the type $f = \sum_{j=1}^k p_j \mathrm{N}(\mu_j, \sigma_j^2)$; see Ex. 1.61. Then simulate a high number n of data from this distribution, and read off $D_n = \max_t |F_n(t) - F(t)|$. Check out how high n needs to be to have $D_n \leq 0.01$, say, with high probability, in a few situations.

Stable and conditional convergence, and nonnormal limits

Ex. 2.56 *Stable convergence.* This exercise introduces the notion of stable convergence of random variables, which is a form of convergence lying between convergence in probability and convergence in distribution (to paraphrase Jacod and Mémin (1981), an early article on the topic). On a probability space $(\Omega, \mathcal{F}, \Pr)$, let $(X_n)_{n \geq 1}$ be a sequence of random

variables, and let $\mathcal{G} \subset \mathcal{F}$. The sequence X_n converges $\mathcal{G}$-stably to the random variable X, where X is defined on an extension $(\Omega \times \mathbb{R}, \mathcal{G} \otimes \mathcal{B}(\mathbb{R}), \Pr')$, if

$$\mathrm{E}\, \xi f(X_n) \to \mathrm{E}'\, \xi f(X), \tag{2.12}$$

for every bounded $\mathcal{G}$-measurable random variable ξ, and every bounded and continuous function f. See Ex. A.27 for extensions of probability spaces. Here $\mathrm{E}'(\cdot)$ denotes the expectation with respect to $\Pr'$. We write $X_n \to_{\mathcal{G}\text{-st.}} X$ to indicate this form of convergence.

(a) Show that if $X_n \to_{\mathcal{G}\text{-st.}} X$ then $X_n \to_d X$.

(b) To see that the converse of (a) is not true, consider the sequence $X_n = I(n \text{ odd})Y + I(n \text{ even})Y'$, where Y and Y' are i.i.d. random variables with common distribution F. Show that $X_n \to_d F$, but that X_n fails to converge $\mathcal{G} = \sigma(Y)$ stably, for example. This example is from Aldous and Eagleson (1978).

(c) Suppose that $(X_n, Y) \to_d (X, Y)$ for every $\mathcal{G}$-measurable Y. We'll soon see that this is equivalent to $X_n \to_{\mathcal{G}\text{-st.}} X$. For now, assume that the limit X is also $\mathcal{G}$-measurable, and show that then $X_n \to_p X$. – This illustrates that to have $\mathcal{G}$-stable convergence of X_n to X without also having convergence in probability, X must be realised in a fashion that does not render it $\mathcal{G}$-measurable, hence the extension of the original probability space.

(d) Show that the following are equivalent:

(i) $X_n \to_{\mathcal{G}\text{-st.}} X$;
(ii) $(X_n, Y) \to_d (X, Y)$ for every $\mathcal{G}$-measurable random vector Y;
(iii) $(X_n, Y) \to_{\mathcal{G}\text{-st.}} (X, Y)$ for every $\mathcal{G}$-measurable random vector Y;
(iv) $(X_n, Y_n) \to_d (X, Y)$ for every sequence $(Y_n)_{n\geq 1}$ of random variables and every $\mathcal{G}$-measurable Y such that $Y_n \to_p Y$;
(v) $\mathrm{E}\, I_G f(X_n) \to \mathrm{E}'\, I_G f(X)$ for every $G \in \mathcal{G}$ and every bounded and continuous f;
(vi) $\mathrm{E}\,\{f(X_n) \,|\, G\} \to \mathrm{E}'\,\{f(X) \,|\, G\}$, for every $G \in \mathcal{G}$ with $\Pr(G) > 0$, and every bounded and continuous f;
(vii) $\mathrm{E}\, I_G \exp(itX_n) \to \mathrm{E}'\, I_G \exp(itX)$ for every $G \in \mathcal{G}$.

To prove this, Ex. **????** and Ex. A.28(c) might be of help.

(e) [xx rewrite xx] Let Q_n and Q be versions of the conditional distributions of X_n and X given $\mathcal{G}$, respectively, and let $Q_n f = \int_{\mathbb{R}} f(x) Q_n(\cdot, \mathrm{d}x)$, with Qf similarly defined. Show that $X_n \to_{\mathcal{G}\text{-st.}} X$ is equivalent to $\mathrm{E}\, \xi Q_n f \to \mathrm{E}\, \xi Q f$, for every ξ and f as above.

(f) [xx cramer slutsky for stable convergence here xx]

Ex. 2.57 *Conditional convergence.*[xx introtext here xx]

(a) Suppose that $(X_n, Y_n) \to_d (X, Y)$. Show that we then also have marginal convergence, that is $X_n \to_d X$ and $Y_n \to_d Y$.

(b) Show that $(X_n, Y_n) \to_d (X, Y)$ is equivalent to

$$\mathrm{E}\{f(X_n) \,|\, Y_n \in B\} \to \mathrm{E}\{f(X_n) \,|\, Y_n \in B\},$$

for all $f \in C_b(\mathbb{R})$ and all sets $B \in \mathcal{B}(\mathbb{R})$ such that $\Pr(Y \in B) > 0$ and $\Pr(Y_n \in B) > 0$ for all n, and $\Pr(Y \in \partial B) = 0$, i.e., B is a continuity set of the distribution of Y.

(c) Suppose $(X_n, Y_n) \to_d (X, Y)$, a binormal zero-mean limit, say $\mathrm{N}_2(0, \Sigma)$, with

$$\Sigma = \begin{pmatrix} \sigma_X^2, & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y, & \sigma_Y^2 \end{pmatrix}.$$

Deduce from (b) that $X_n \,|\, (|Y_n| \le \varepsilon) \to_d X \,|\, (|Y| \le \varepsilon)$ for each $\varepsilon > 0$, i.e., that the conditional distribution of X_n given $|Y_n| \ge \varepsilon$ converges in distribution to that of X given $|Y| \ge \varepsilon$. Show that $X \,|\, (|Y| \le \varepsilon) \to_d X \,|\, (Y = 0) \sim \mathrm{N}(0, (1-\rho^2)\sigma_X^2)$ as $\varepsilon \to 0$.

(d) Consider a distribution for X_i with mean zero and variance σ^2, where we indeed know that $\sqrt{n}\bar{X}_n \to_d \mathrm{N}(0, \sigma^2)$. Suppose X_i also has an integrable characteristic function $\varphi(t)$, implying by Ex. A.35 the existence of a smooth density f for X_i and also a density f_n for $\sqrt{n}\bar{X}_n$. Show that

$$f_n(z) = 1/(2\pi) \int \exp(-itz)\varphi(t/\sqrt{n})^n \,\mathrm{d}t \to \sigma^{-1}\phi(\sigma^{-1}z),$$

i.e., that there is convergence not merely for the cumulatives, but for the densities too. To do so, you may split the domain of integration in two parts $|t| > \varepsilon\sqrt{n}$ and $|t| \le \varepsilon\sqrt{n}$, for some $\varepsilon > 0$, and use Ex. A.37.

(e) Suppose next that X_i is discrete, without a continuous density; for a concrete example, consider $X_i = \pm 1$ with equal probabilities, and for which $\varphi(t) = \cos t$. Then $Z_n = \sqrt{n}\bar{X}_n$ does not have a density, but we may add a little Gaussian noise, to form $Z_n^* = \sqrt{n}\bar{X}_n + \xi_n$, with $\xi_n \sim \mathrm{N}(0, \varepsilon_n^2)$. Show that Z_n^* has density

$$f_n^*(z^*) = 1/(2\pi) \int \exp(-itz^*)\varphi(t/\sqrt{n})^n \exp(-\tfrac{1}{2}\varepsilon_n^2 t^2) \,\mathrm{d}t,$$

and that this again converges to the normal density $\sigma^{-1}\phi(\sigma^{-1}z^*)$ provided merely that $\varepsilon_n \to 0$.

(f) We may use the same trick also in the vector case. Specifically, with a p-dimensional X_i having zero mean and covariance matrix Σ, show (i) that if X_i has an integrable characteristic function, then the density for $Z_n = \sqrt{n}\bar{X}_n$ tends to the $\mathrm{N}_p(0, \Sigma)$ density; and (ii) that if X_i is discrete, without a density, then $Z_n^* = \sqrt{n}\bar{X}_n + \xi_n$, with some small Gaussian added noise $\xi_n \sim \mathrm{N}_p(0, \varepsilon_n^2\Sigma)$ with $\varepsilon_n \to 0$, has a density $f_n^*(z^*)$ which tends to the same $\mathrm{N}_p(0, \Sigma)$ density.

(g) Suppose $(X_n, Y_n)^{\mathrm{t}} \to_d (X, Y)^{\mathrm{t}}$, a zero-mean binormal. If there is also density convergence, with (X_n, Y_n) having density $f_n(x, y)$ tending to the appropriate binormal density $f(x, y)$, show that $X_n \,|\, (|Y_n| \le \delta_n)$ tends to $X \,|\, (Y = 0)$, as long as $\delta_n \to 0$. Show that the same limiting distribution statement holds also when (X_n, Y_n) has a discrete distribution, using the 'adding small Gaussian noise to get densities' trick.

Ex. 2.58 *Limiting normality of rank sums statistics.* (xx to be edited and polished; nils rant so far. we put it in if it looks smooth enough, and with a brief pointer to Story v.5. point to Swensen (1983). xx) In a population of n individuals, followed on some continuous scale, a subgroup of interest, of size m, has ranks $X_1, \ldots, X_m$. These form a randomly selected subset of size m from $\{1, \ldots, n\}$, with all such $\binom{n}{m}$ subsets equally likely. The rank sum $Z_n = X_1 + \cdots + X_m$ is the Wilcoxon statistic.

(a) Explain that one may write $Z_n = \sum_{i=1}^n iJ_i$, where the 0-1 variables J_i are such that precisely m of them are 1, and with all $\binom{n}{m}$ subsets of such 1s being equally likely. Find $\mathrm{E}\, J_i$, $\mathrm{Var}\, J_i$, $\mathrm{cov}(J_i, J_j)$ for $j \neq i$. Writing $p = m/n$ for the sample ratio, show using either of the representations $\sum_{i=1}^m X_i$ or $\sum_{i=1}^n iJ_i$ that

$$\mathrm{E}\, Z_n = \tfrac{1}{2}m(n+1) \doteq \tfrac{1}{2}n^2 p, \quad \mathrm{Var}\, Z_n = (1/12)(n+1)m(n-m) \doteq (1/12)n^3 p(1-p).$$

(b) We aim indeed at showing limiting normality of Z_n here, with both n and m becoming larger, with $m/n \to p$. Explain that this must mean

$$(Z_n - \tfrac{1}{2}n^2 p)/n^{3/2} \to_d \mathrm{N}(0, (1/12)p(1-p)).$$

We cannot use CLT or Lindeberg for studying Z_n, since the X_i are dependent, as are the J_i. Consider however a different parallel setup, involving independent Bernoulli variables $J_1^*, \ldots, J_n^*$ with $\Pr(J_i^* = 1) = p = m/n$. Explain that the distribution of Z_n is the same as the distribution of $Z_n^* = \sum_{i=1}^n iJ_i^*$ given $\sum_{i=1}^n J_i^* = m$. Show now that

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = \begin{pmatrix} (1/\sqrt{n})\sum_{i=1}^n (i/n)(J_i^* - p) \\ (1/\sqrt{n})\sum_{i=1}^n (J_i^* - p) \end{pmatrix} \to_d \begin{pmatrix} A \\ B \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/3, & 1/2 \\ 1/2, & 1 \end{pmatrix}\Big).$$

(c) Find the distribution of $A \,|\, (B = b)$, and show in particular that $A \,|\, (B = 0) \sim \mathrm{N}(0, 1/12)$. This gives a clear limiting normality statement for Z_n^*, conditional on $\sum_{i=1}^n J_i^* = m$, and nicely solves our Wilcoxon problem; show that it corresponds precisely to the $(Z_n - \frac{1}{2}n^2 p)/n^{3/2}$ limiting statement above. (xx some extra care needed. but an easy and instructive way to show normality for Wilcoxon, and also for other related variables. to illustrate, find limiting normality for $X_1^{1/2} + \cdots + X_m^{1/2} = \sum_{i=1}^n i^{1/2} J_i$. xx)

(d) (xx if we manage: also a link via the uniform order statistic process, and to integrals of salt-and-pepper processes, with $W_n = \int_0^1 ([ns]/n)\, \mathrm{d}C_n(s)$, with the $\mathrm{d}C_n(s)$ being $\mathrm{d}s$ or 0 with probabilities p and $1-p$, but conditional on the random region $\int_0^1 \mathrm{d}C_n(s)$ being $p = m/n$. if we're lucky there is a limit expressible as integral or Brownian bridge, for Ch9. nils will attempt to fix this and at least make the idea more precise. xx)

Ex. 2.59 *Nonnormal limits.* (xx polish this. point to process version, with more results for hitting times, etc., in Ch. 9. xx) Normally limits are normal, but not always. Here we shall indeed work with variables with mean zero and variance one, where the sample averages have nonnormal limits. The basic construction is as follows. Let $U_1, U_2, \ldots$ be i.i.d., with mean zero and variance one, and with m.g.f. $M_0(s) = \mathrm{E}\exp(sU_i)$ finite in a neighbourhood around zero; in particular, all moments for the U_i are finite. Let

independently of these $J_1, J_2, \ldots$ be independent Bernoulli variables with $\Pr(J_i = 1) = 1/i$, $\Pr(J_i = 0) = 1 - 1/i$. Then form

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} J_i \sqrt{i}\, U_i = \sum_{i=1}^{n} J_i \sqrt{i/n}\, U_i.$$

A picture to have in mind is that most of the terms will be zero, with non-zero contributions becoming both more rare and more big as time proceeds.

(a) Show that there will with probability one be infinitely many $J_i = 1$, i.e. non-zero terms in the Z_n sum as n grows.

(b) Show that the terms $J_i \sqrt{i} U_i$ have mean zero and variance one; hence also the normalised sample average Z_n has mean zero and variance one. Find also an expression for the kurtosis $\kappa_n = \mathrm{E}\, Z_n^4 - 3$ of Z_n, and show that $\kappa_n \to \frac{1}{2} a_4$, where $a_4 = \mathrm{E}\, U_i^4$. Compare this to what we are 'used to' from the Lindeberg theorem.

(c) We already know that if Z_n has a limit distribution, it can't be normal. Working with the m.g.f., show that

$$M_n(t) = \mathrm{E} \exp(tZ_n) = \prod_{i=1}^{n} \Big[1 + \frac{1}{i} \{ M_0(t\sqrt{i/n}) - 1 \} \Big],$$

for all t around zero for which $M_0(t)$ is finite.

(d) Show that

$$\prod_{i=1}^{n} \Big[1 + \frac{1}{i} \{ M_0(t\sqrt{i/n}) - 1 \} \Big] \to \exp\Big\{ \int_0^1 \frac{M_0(t\sqrt{x}) - 1}{x}\, \mathrm{d}x \Big\}.$$

Work first with Special Case One, where we let U_i have the simple symmetric two-point distribution $\Pr(U_i = 1) = \Pr(U_i = -1) = \frac{1}{2}$. Find the limiting kurtosis for Z_n in this case. Show that $M_0(s) = \frac{1}{2} e^s + \frac{1}{2} e^{-s} = 1 + (1/2!) s^2 + (1/4!) s^4 + \cdots$, and use this to find an infinite-sum expression for the limit of $M_n(t)$. Have you now proved that Z_n has a limit distribution?

(e) Then work with Special Case Two, where the U_i have a double exponential distribution, of the form $f(u) = \frac{1}{2} \sqrt{2} \exp(-\sqrt{2}|u|)$ on the real line (the $\sqrt{2}$ factor is there to ensure variance one). Find the m.g.f. $M_0(s)$ for the U_i, and then the m.g.f. $M(t)$ for the limit distribution of Z_n.

(f) For most cases, regarding the distribution for the U_i, it is hard to learn the explicit distribution for Z_n (even in cases where there might be a clear distribution for its limit). For Special Case Two, however, find the explicit distribution for Z_n, for any given n.

CLTs for dependent random variables

Ex. 2.60 *A CLT for 1-dependent variables.* (xx decide later if these few should be pushed to Ch. 12. xx) Consider a stationary sequence $Y_1, Y_2, \ldots$, with mean zero and

variance one, being 1-dependent. Stationarity means $(Y_1, \ldots, Y_r)$ having the same distribution as $(Y_{i+1}, \ldots, Y_{i+r})$, for any i and block lengths r, and 1-dependence means that Y_i, Y_{i+1} may be dependent, but $Y_1, \ldots, Y_i$ is independent of $Y_{i+2}, Y_{i+3}, \ldots$. This exercise reaches a CLT for $\sum_{i=1}^n Y_i$, representing a genuine extension of the usual CLT and Lindeberg theorems from independence.

(a) Writing $\rho = \text{corr}(Y_i, Y_{i+1})$, show that $(1/k)\text{Var}(Y_1 + \cdots + Y_k) = 1 + 2(1-1/k)\rho$ which then goes to $1 + 2\rho$ for increasing k.

(b) For a given block length k, split $Y_1 + \cdots + Y_n$ into $[n/k]$ blocks, and write block j of these as $U_j + V_j$, with U_j as sum of $k-1$ consecutive observations and V_j the last one of that block. Write then

$$Z_n = (1/\sqrt{n}) \sum_{i=1}^n Y_i = (1/\sqrt{n}) \Big(\sum_{j=1}^{[n/k]} U_j + \sum_{j=1}^{[n/k]} V_j + E_n \Big),$$

with E_n any extra left after the $k[n/k]$ variables captured in these first $[n/k]$ blocks.

(c) Explain why $U_1, U_2, \ldots$ are independent, so that the usual CLT applies to these. Show that $(1/\sqrt{n}) \sum_{j=1}^{[n/k]} U_j \to_d \text{N}(0, \tau_k^2)$, with $\tau_k^2 = (1/k)\text{Var}\,(Y_1 + \cdots + Y_{k-1})$.

(d) Then use Ex. 2.24 to prove that $(1/\sqrt{n}) \sum_{i=1}^n Y_i \to_d \text{N}(0, 1 + 2\rho)$, i.e. a CLT for 1-dependent variables.

(e) Assume $X_1, X_2, \ldots$ are i.i.d. with mean zero and variance one. Consider $Z_n = (1/\sqrt{n})(X_1X_2 + X_2X_3 + \cdots + X_{n-1}X_n)$. Show that $Z_n \to_d \text{N}(0,1)$. Show also that

$$Z_n' = (1/\sqrt{n}) \sum_{i=1}^{n-1} (X_i - \bar{X}_n)(X_{i+1} - \bar{X}_n)$$

has the same limit distribution, where $\bar{X}_n$ as usual is the sample mean.

Ex. 2.61 *A CLT for m-dependent variables.* In natural generalisation of Ex. 2.60, consider a stationary m-dependent sequence $Y_1, Y_2, \ldots$, with mean zero and variance σ^2. There is accordingly potential dependence among $Y_1, \ldots, Y_m$, but for any i, $(Y_1, \ldots, Y_i)$ is independent of $(Y_{i+m+1}, \ldots, Y_n)$.

(a) Writing $\text{cov}(Y_i, Y_j) = \sigma^2 \rho(|j-i|)$, with the autocorrelation function $\rho(\cdot)$, show first that in general terms,

$$(1/n)\,\text{Var}\Big(\sum_{i=1}^n Y_i \Big) = \sigma^2 \Big\{ 1 + 2 \sum_{j=1}^n (1 - j/n)\rho(j) \Big\}.$$

Then show that for the case of m-dependence, for any $k \geq m$, we have $(1/k)\,\text{Var}(Y_1 + \cdots + Y_k) \to \sigma^2\{1 + 2\sum_{j=1}^m \rho(j)\}$,

(b) Extend arguments and techniques from Ex. 2.60 to show that $(1/\sqrt{n}) \sum_{i=1}^n Y_i$ tends to a zero-mean normal with variance $\sigma^2\{1 + 2\rho(1) + \cdots + 2\rho(m)\}$.

(c) (xx a bit on how the acf works for an i.i.d. sequence: $\sqrt{n}\widehat{\rho}(j) \to_d \text{N}(0,1)$, for each j. xx)

[xx new title xx]

Ex. 2.62 *Local asymptotics.* The CLT and Lindeberg machineries yield normal limits and hence approximations in situations where independent observations come from given models. It is sometimes useful to extend such results to situations where observations stem from distributions close to, but not equal to, the postulated start models. The standard $\sqrt{n}$ speed of convergence for the CLT and relatives leads naturally to the notion of $O(1/\sqrt{n})$ neighbourhoods. If there is limiting zero-mean normality of variables like $\sqrt{n}(\widehat{\theta}-\theta_0)$, under a relevant null model at θ_0, then such variables typically have limiting non-zero-mean normal limits at such $O(1/\sqrt{n})$ alternatives.

(a) A simple setup illustrating such ideas is the following. Suppose $X_1,\ldots,X_n$ are i.i.d. from a distribution with mean $\xi+\delta/\sqrt{n}$ and variance $\sigma_n^2=\sigma^2+d/n$. Consider then $Z_n=\sqrt{n}(\bar{X}_n-\xi)$. Use the Lindeberg theorem, or a triangular version of the CLT, to demonstrate that $Z_n\to_d \mathrm{N}(\delta,\sigma^2)$.

(b) (xx for Ch. 5, an exercise with $\sqrt{n}(\widehat{\theta}-\theta_0)\to_d \mathrm{N}(b\delta, J^{-1})$, when data stem from $f(y,\theta_0)+\delta/\sqrt{n}h(y)$. natural special case: $f(y,\theta_0,\gamma_0\delta/\sqrt{n})$. xx)

(c) xx

Ex. 2.63 *Approximate normality when combining information sources.* (xx this is a nils rant, so far. it needs intro sentences. the point is partly that yes, Lindeberg gives us limiting normality of sums, but we also need consistent variance estimators. xx) To illustrate the general themes, in a situation exhibiting these general components, consider the following setup. There are Poisson parameters $\theta_1,\ldots,\theta_k$, with associated independent Poisson observations $y_{j,},\ldots,y_{j,m_j}$ for θ_j, leading to $\widehat{\theta}_j=\bar{y}_j=(1/m_j)\sum_{\ell=1}^{m_j}y_{i,\ell}$. The object is to make inference for the linear combination $\phi=a^{\mathrm{t}}\theta=\sum_{j=1}^k a_j\theta_j$, for which we use the estimator $\widehat{\phi}=a^{\mathrm{t}}\widehat{\theta}=\sum_{j=1}^k a_j\widehat{\theta}_j$, with variance

$$B_k^2=\mathrm{Var}\,\widehat{\phi}=\sum_{j=1}^k a_j^2\theta_j/m_j.$$

(a) For $Y\sim\mathrm{Pois}(\theta)$, show that $\mathrm{E}\,(Y-\theta)^3=\theta\ldots$, and that this implies that its skewness is $1/\theta^{1/2}$. Show also that with $Y_1,\ldots,Y_m$ i.i.d. from this distribution, we have $\mathrm{E}\,(\bar{Y}-\theta)^3=\theta/m$, with $\mathrm{skew}(\bar{Y})=1/(m\theta)^{1/2}$. Thus the skewness tends to zero, indicating limiting normality, as long as with θ, or m, or both, grow.

(b) Show furthermore that

$$\mathrm{skew}(\widehat{\phi})=\mathrm{E}\Big(\frac{\widehat{\phi}-\phi}{B_k}\Big)^3=\frac{\sum_{j=1}^k a_j^3\theta_j/m_j}{(\sum_{j=1}^k a_j^2\theta_j/m_j)^{3/2}}.$$

(xx then some Lindeberg things here, understanding when this tends to zero, leading to $Z_{k,0}=(\widehat{\phi}-\phi)/B_k\to_d \mathrm{N}(0,1)$. play a bit with a_j,m_j. xx)

(c) (xx then wish to find a case where variance is not well enough estimated. xx) We estimate the variance using $\widehat{B}_k^2=\sum_{j=1}^k a_j^2\widehat{\theta}_j/m_j$. To make inference for ϕ we need not

merely the result of (b), but also relative consistency of the variance estimate. Show that $V_k = \widehat{B}_k^2/B_k^2$ has mean 1 and variance

$$\operatorname{Var} V_k = \frac{\sum_{j=1}^k a_j^4\theta_j/m_j^2}{(\sum_{j=1}^k a_j^2\theta_j/m_j)^2}.$$

(xx rigging the game so that $Z_{k,0} \to_d \mathrm{N}(0,1)$, but not Z_k. As a special case to consider, take a common $m_j = m_0$ for all sample sizes, $a_j = j$, and assume $\theta_j = 1/j$. What happens to B_k, $\widehat{B}_k$, V_k, and the natural ratio $Z_k = (\widehat{\phi} - \phi)/\widehat{B}_k$? xx)

(d) (xx then find the typical behaviour of V_k, to ensure also $Z_n = (\widehat{\phi}-\phi)/\widehat{B}_k \to_d \mathrm{N}(0,1)$. make connections to chapter 4 stuff on deviance and wilks. the Wilks thing is close to Z_n^2. xx)

Ex. 2.64 *Limiting normality of the sample variance matrix.* (xx can be better placed, inside Ch 3. results are used for Ex. **??**. xx) Consider i.i.d. vectors $Y_1, \ldots, Y_n$ from the multinormal $\mathrm{N}_p(\xi, \Sigma)$, first with known mean vector ξ, which we for convenience then set to zero. The estimated variance matrix is $\widehat{\Sigma} = (1/n)\sum_{i=1}^n Y_i Y_i^{\mathrm{t}}$.

(a) Write $\sigma_{j,k}$ for the elements of the $p \times p$ matrix Σ, and σ_j^2 for $\sigma_{j,j}$, the variance of component j of Y_i. Show that its estimator is $\widehat{\sigma}_j^2 = (1/n)\sum_{i=1}^n Y_{i,j}^2$, and that it has the distribution of $\sigma_j^2\chi_n^2/n$. Show also that $\sqrt{n}(\widehat{\sigma}_{j,j} - \sigma_{j,j}) \to M_{j,j}$, with this limit having the $\mathrm{N}(0, 2\sigma_j^4)$ distribution.

(b) Using first the one-dimensional CLT, show that $\sqrt{n}(\widehat{\sigma}_{j,k} - \sigma_{j,k})$ has a normal limit $M_{j,k}$, and find its variance.

(c) Then show that there is convergence in distribution of the full matrix, say $\sqrt{n}(\widehat{\Sigma} - \Sigma) \to_d M$, with $M = (M_{i,j})_{i,j=1,\ldots,p}$ multinormal with zero means, and that

$$\operatorname{cov}(M_{i,j}, M_{k,l}) = \sigma_{i,k}\sigma_{j,l} + \sigma_{i,l}\sigma_{k,j}.$$

(d) Assume Σ has full rank p. Show that limiting normality for $\widehat{\Sigma}$ implies limiting normality for $\widehat{\Sigma}^{-1}$, and that in fact $\sqrt{n}(\widehat{\Sigma}^{-1} - \Sigma^{-1}) \to_d M^* = -\Sigma^{-1}M\Sigma^{-1}$. Writing $\sigma^{j,k}$ for the elements of Σ^{-1}, show that $\operatorname{cov}(M_{i,j}^*, M_{k,l}^*) = \sigma^{i,k}\sigma^{j,l} + \sigma^{i,l}\sigma^{j,k}$.

(e) In case of an unknown mean vector, one uses the sample variance matrix $\widetilde{\Sigma} = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^{\mathrm{t}}$. Show that $\sqrt{n}(\widetilde{\Sigma} - \widehat{\Sigma}) \to_{\mathrm{pr}} 0$, where $\widehat{\Sigma} = n^{-1}\sum_{i=1}^n (Y_i - \xi)(Y_i - \xi)^{\mathrm{t}}$ uses the ξ. Deduce that $\sqrt{n}(\widetilde{\Sigma} - \Sigma) \to_d M$ and $\sqrt{n}(\widetilde{\Sigma}^{-1} - \Sigma^{-1}) \to_d M^* = -\Sigma^{-1}M\Sigma^{-1}$, i.e. with the same limits as above.

(f) (xx something to round if off. perhaps dimension 2. mention the Wishart distribution, but here we derive limits without knowing or using that. also, not lost at sea outside multinormality, but the covariance structure for the limit M becomes much more complicated. xx)

Ex. 2.65 *Summing geometrically many terms.* Suppose $Y_1, Y_2, \ldots$ are i.i.d. with mean zero, variance σ^2, and m.g.f. $M_0(t)$. With $\Pr(N = n) = (1-p)^{n-1}p$ for $n \geq 1$, i.e. a geometric distribution, consider $Z_p = p^{1/2}(Y_1 + \cdots + Y_N)$, with the Y_i being independent of N.

(a) Show first that the generating function for N is $\mathrm{E}\, s^N = ps/\{1-(1-p)s\}$ for $|s| < 1/(1-p)$; see Ex. 1.35. Show that Z_p has variance σ^2, and that its m.g.f. may be written

$$K_p(t) = \mathrm{E}\, \exp(tZ_p) = pM_0(p^{1/2}t)/\{1-(1-p)M_0(p^{1/2}t)\}.$$

(b) Then use $M_0(t) = 1 + \frac{1}{2}\sigma^2 t^2 + o(|t|^2)$ to demonstrate that as $p \to 0$, with increasing number of terms $\mathrm{E}\, N = 1/p$, we have $K_p(t) \to 1/(1-\frac{1}{2}\sigma^2 t^2)$ for $|t| < \sqrt{2}/\sigma$. This shows that $Z_p \to_d L_\sigma$, the Laplace distribution with standard deviation σ, see Ex. 1.32.

(c) For the particular case of a sum of randomly many normal terms, let the Y_i be i.i.d. standard normal. Show what $Z_p \,|\, N \sim \mathrm{N}(0, pN)$, and that $pN \to_d \mathrm{Expo}(1)$ as $p \to 0$. Explain how this matches Ex. 1.32.(c).

(d) (xx one or two further illustrations, with Laplace limit of $p^{1/2}(Y_1 + \cdots + Y_N)$. with $Y_i = Q_i - 1$, Poisson, we learn $p^{1/2}(V_N - N) \to_d L$, where $V_N \sim \mathrm{Pois}(N)$. similarly with $p^{1/2}(W_N - N)$, where $W_N \,|\, N \sim \chi^2_N$, randomly many degrees of freedom. xx)

(e) (xx something to simulate, to illustrate the cusp behaviour at the centre. with the Y_i having mean ξ and variance 1, we have

$$p^{1/2}(Y_1 - \xi + \cdots + Y_N - \xi) = p^{1/2}N(\bar{Y} - \xi) = (pN)^{1/2}\, N^{1/2}(\bar{Y} - \xi) \to_d L_1.$$

this gives different inference for ξ, and different predictions for $\bar{Y}$, than what we're used to from normal terrain. we're also in scale mixtures of normal terrain, with random variance tending to a unit exponential. i'll look for variations. i do like the $pN \to_d \mathrm{Expo}(1)$, since it gives the cool cusp in the limit for Z_p, but other variations for $pN \to_d V$ are ok too. xx)

(f) (xx just ranting away a bit until things settle. xx) More generally, with $Y_1, Y_2, \ldots$ being i.i.d. with zero mean and unit variance, consider

$$Z_p = p^{1/2}(Y_1 + \cdots + Y_N) = (pN)^{1/2} N^{1/2} \bar{Y}_N,$$

with N having a distribution such that $pN \to_d V$, say, as $p \to 0$. From the CLT, $N^{1/2}\bar{Y}_N \to_d \mathrm{N}(0,1)$, so this amounts to a situation with a normal limit, but a random variance, as in $X \,|\, V \sim \mathrm{N}(0, V)$ but V random. Here $\mathrm{E}\, \exp(tX) = \mathrm{E}\, \exp(\frac{1}{2}t^2 V) = M_0(\frac{1}{2}t^2)$, where $M_0(u) = \mathrm{E}\, \exp(uV)$ is the mgf for V. Also, X has density

$$\bar{f}(x) = \int \phi(x/v^{1/2})(1/v^{1/2})\, \mathrm{d}H(v),$$

with H the distribution of V. – The special case above amounts to $N \sim \mathrm{geom}(1/p)$, where pN tends to the unit exponential, and where X gets the Laplace distribution. For another case, consider $N \,|\, \lambda \sim \mathrm{Pois}(\lambda/p)$, and with λ having its own distribution with mean and variance λ_0 and τ_0^2, say. Show that $\mathrm{E}\, pN = \lambda_0$ and that $\mathrm{Var}\, pN \to \tau_0^2$. Also consider the special case for this setup where $\lambda \sim \mathrm{Gam}(a, b)$. From

$$\mathrm{E}\, \{\exp(-spN) \,|\, \lambda\} = \exp[-(\lambda/p)\{1 - \exp(-sp)\}]$$

deduce

$$\mathrm{E}\,\exp(-spN) = \frac{1}{[1+(1/b)(1/p)\{1-\exp(-sp)\}]^a} \to \frac{1}{(1+s/b)^a},$$

and that $pN \to_d V_{a,b}$, another $\mathrm{Gam}(a,b)$. With the construction above, the normal scale mixture variable X has mgf $1/(1-\frac{1}{2}t^2/b)^a$, and density

$$\bar f(x) = \int \phi(x/v^{1/2})(1/v^{1/2})\frac{b^a}{\Gamma(a)}v^{a-1}\exp(-bv)\,\mathrm{d}v.$$

This is a Laplace, for $a=1$. (xx but different, and interesting, for other a. round this off. xx)

(g) (xx i think we can use these tools to form a full Laplace process, for BNP use or otherwise. we should tune in to a $Z_p(t) = p^{1/2}(Y_1+\cdots+Y_{N(t)})$, with a clever $N(t)$. will look at $N(t)$ being a negative binomial process with mean t/p. nils thinks this works: (i) $\lambda \sim \mathrm{Gam}(1,b)$, with mean $\lambda_0 = 1/n$. (ii) $N \,|\, \lambda \sim \mathrm{Pois}(\lambda/p)$. write down $\mathrm{E}\,(pN \,|\, \lambda)$ and $\mathrm{Var}\,(pN \,|\, \lambda)$, then unconditional mean and variance for pN. (iii) find limit in distribution of pN. (iv) study $Z_p = p^{1/2}\sum_{i=1}^N Y_i$, close to $(pN)^{1/2}$ times a normal, etc. (v) make it into a full Laplace process, by having $\lambda_t \sim \mathrm{Gam}(1,b_t)$. xx)

Ex. 2.66 *Maximum sample value of exponentials and the Gumbel distribution.* (xx nils reorganises some of these exercises which need tidying up. i now start with the gumbel and the maximum of exponentials, before taking up other gumbel related matters. xx) We define the *Gumbel distribution* on the real line by its cumulative distribution function $G_0(u) = \exp\{-\exp(-u)\}$.

the Gumbel distribution

(a) Show that G_0 is indeed a cumulative distribution function, that its density is $g_0(u) = \exp\{-u-\exp(-u)\}$, and that its Laplace transform is $L_0(s) = \mathrm{E}\,\exp(-sU) = \Gamma(1+s)$, in terms of the gamma function.

(b) Use properties of the gamma function to show that the mean and variance of the Gumbel distribution is γ_e and $\pi^2/6$, where $\gamma_e = 0.5772...$ is the Euler constant. The latter has several equivalent definitions, among which is that $H_n - \log n \to \gamma_e$, where $H_n = 1+1/2+\cdots+1/n$ is the partial sum of the divergent harmonic series.

the Euler–Mascheroni constant

(c) With U having the Gumbel distribution, show also that its mode is 0 and that its median is $-\log(\log 2) = 0.367$. Find an expression for the q-quantile $G_0^{-1}(q)$. Show that $\Pr(-1.097 \le U \le 2.970) = 0.90$.

(d) The Gumbel distribution turns up in various contexts concerning extreme values. The simplest such case is as follows: let $X_1,\ldots,X_n$ be i.i.d. from the unit exponential model, with $M_n = \max_{i\le n} X_i$ their maximum. Show that $M_n - \log n \to_d U$, the Gumbel distribution.

(e) We may learn more about the distribution of M_n via first investigating the spacings. With $X_{(1)} < \cdots < X_{(n)}$ being the order statistics, let $D_1 = X_{(1)}$, $D_2 = X_{(2)} - X_{(1)}$, up to $D_n = X_{(n)} - X_{(n-1)}$. We have seen via Ex. 1.12 and 1.13 that the spacings are

independent (for this special case of the exponential), with $D_i \sim \text{Expo}(n-i+1)$. Show that this leads to the representation

$$X_{(n)} = D_1 + \cdots + D_n = V_1/1 + V_2/2 + \cdots + V_n/n,$$

with $V_1, \ldots, V_n$ being i.i.d. and unit exponential.

(f) Show from this that M_n has mean $H_n \doteq \log n + \gamma_e$ and variance $\sum_{i=1}^n 1/i^2$, tending to $\pi^2/6$. This is agreement with the Gumbel limit for $M_n - \log n$.

(g) Show that $M_n - H_n \to_d U - \gamma_e$, the zero-mean version of the Gumbel. Deduce from this that

$$\lim_{n\to\infty} \text{E}\exp\{-s(M_n - H_n)\} = \prod_{i=1}^{\infty} \frac{\exp(s/i)}{1+s/i} = \Gamma(1+s)\exp(\gamma_e s).$$

This infinite-product form of the gamma function is actually equivalent to a famous formula by Weierstraß. Show also that

$$\sum_{i=1}^{\infty} \{s/i - \log(1+s/i)\} = \gamma_e s + \log\Gamma(1+s) = \sum_{j=2}^{\infty} (-1)^j \frac{\zeta(j)}{j} s^j,$$

valid for $|s| < 1$, where $\zeta(j) = 1 + 1/2^j + 1/3^j + \cdots$ is Riemann's zeta function, at j. (xx two more sentences. here we derive these deep mathematical facts from a simple convergence in distribution result; could also go the other way, if we start with gamma function knowledge. xx)

(h) Yet another fruitful perspective on what we've learned above is in terms of an infinite sum of smaller and smaller exponentials. Consider independent exponentials $W_1, W_2, \ldots$, where $W_i \sim \text{Expo}(i)$, i.e. with mean $1/i$. Show that $W = \sum_{i=1}^{\infty}(W_i - 1/i)$ is finite, with probability one, and that its distribution is that of $U - \gamma_e$, the zero-mean Gumbel.

Ex. 2.67 *Weibull and Gamma maxima.* The basic result of Ex. 2.66, concerning the maximum of a sample of exponentials, leads to limit distribution results also for maxima from other distributions.

(a) Suppose that $X_1, \ldots, X_n$ are i.i.d. from the Weibull distribution, with cumulative function $F(x) = 1 - \exp\{-(x/a)^b\}$, for certain parameters (a, b). With M_m the sample maximum, show that $(M_n/a)^b - \log n$ tends to the Gumbel distribution.

(b) In two minutes, simulate $n = 1000$ values from the Weibull with $(a, b) = (1, \frac{1}{2})$. Guess in advance how large M_n will be, using the representation $M_n = a(\log n + U_n)^{1/b}$, where U_n tends to the Gumbel.

(c) Similarly consider the Gamma distribution with parameters $(a, b) = (2, 1)$, where the cumulative can be expressed as $F(x) = 1 - \exp(-x)(1+x)$, see Ex. 1.9. With M_n the maximum of a sample of size n from this distribution, show that $M_n - \log(1+M_n) - \log n$ tends to the Gumbel distribution. What is the approximate median for the M_n?

(d) More generally, suppose $F(x) = 1 - \exp\{-A(x)\}$, with $A(x)$ the cumulative hazard rate, and let again M_n be the maximum value from a sample of size n. Show that $A(M_n) - \log n$ tends to the Gumbel.

Ex. 2.68 *Maximum of independent geometric variables.* Let T have the geometric waiting time distribution with parameter p, i.e. with point probabilities $(1-p)^{t-1}p$ for $t = 1, 2, \ldots$. We write $T \sim \text{geom}(p)$ to indicate this distribution; see Ex. 1.24.

(a) Show that $V = pT$ has mean 1 and variance $1 - p$. Show also that if $p \to 0$, then $V = pT$ tends to the unit exponential in distribution. Give an approximate formula for the median of a geometric distribution with small p.

(b) Now suppose $V_1, \ldots, V_n$ are independent geometric waiting times with parameter $1/n$, hence with mean value n. With $Z_n = \max(V_1, \ldots, V_n)$ the time until all waiting times have been completed, we then have $Z_n/n = \max(V_1/n, \ldots, V_n/n)$, which is close in distribution to $\max(E_1, \ldots, E_n)$, with these E_i being independent unit exponentials. By results of Ex. 2.66 we should there expect $Z_n/n - \log n \to_d U$, the Gumbel. Show that indeed this holds. (xx needs some technicalities and a hint, see nils diehard. xx)

(c) (xx one more thing here. can we make something useful out of this, with mgf for Z_n/n. not easy. xx) we do have

$$\Pr(Z_n/n \le v) = \Pr(V_i \le nv)^n = [1 - (1-1/n)^{nv}]^n.$$

Ex. 2.69 *Collecting cards: how long time?* (xx nils will reorganise this a bit, after the abels taarn things. plan is basic things $T_1 + \cdots + T_n$ here, Gumbel limit, a bit more in next exercise, this being Ch4. then likelihood things in Story iv.5 about estimating n from $V_r = T_1 + \cdots + T_r$, time to having seen r different cards. then story about estimating n from observed V_r. xx) Consider a deck of n cards, with $X_1, X_2, \ldots$ independent draws from these, i.e. uniform on $\{1, \ldots, n\}$. How many such random draws are necessary, before you have seen all n cards? – There are several reformulations of this card collecting problem, and with other metaphors. You may think of a fair die, with n faces, and ask how many times you need to roll it until you've seen all faces.

(a) Show that the time needed, until we have seen all n cards, can be represented as $V_n = T_1 + \cdots + T_n$, where T_i is geometric with parameter $p_i = (n - i + 1)/n$. Hence $\mathrm{E}\, T_i = n/(n - i + 1)$, and the card finding process is easy in the beginning, then steadily harder. We may also re-order the T_i to $V_n = T_1' + \cdots + T_n'$, where $T_i' \sim \text{geom}(i/n)$, which for some purposes is an easier representation.

(b) Let $(N_1, \ldots, N_n)$ be the number of times cards $1, \ldots, n$ have been seen, in the course of z independent random draws from the deck. Show that this is a multiomial with count z and probabilities $(1/n, \ldots, 1/n)$; in particular, $N_i \sim \text{binom}(z, 1/n)$. Show that the correlation between N_i and N_j is $-1/(n-1)$.

(c) Show also that another representation of V_n is as $\max(W_1, \ldots, W_n)$, where W_i is the first time $N_i \ge 1$. Show that $W_i \sim \text{geom}(1/n)$, with mean n. These are however dependent, so the Gumbel limit result of Ex. 2.68 does not immediately apply. Show

that the correlation between W_i and W_j is small, however, namely $-\frac{1}{2}/(n-1)$, indicating that $(V_n - n\log n)/n$ should converge to the Gumbel, even with these waiting times being dependent. (xx polish wording here. xx)

(d) (xx nils will coordinate and calibrate this with what is placed in Ex. 1.24. xx) For T_i, with distribution $(1-p_i)^{t-1}p_i$ for $t = 1, 2, 3, \ldots$, show that

$$\mathrm{E}\, T_i = \frac{1}{p_i}, \quad \mathrm{Var}\, T_i = \frac{1-p_i}{p_i^2}, \quad \mathrm{E}\,(T_i - 1/p_i)^3 = \frac{(1-p_i)(2-p_i)}{p_i^3},$$

so the skewness of T_i is $\mathrm{E}\,(T_i - 1/p_i)^3/\sigma_i^3 = (2-p_i)/(1-p_i)^{1/2}$.

(e) Show that

$$\mathrm{E}\, V_n = n(1 + 1/2 + \cdots + 1/n) = nH_n \doteq n(\gamma + \log n),$$

using Ex. 2.66. Show also that

$$\mathrm{Var}\, V_n = \sum_{i=1}^n \Big(\frac{n^2}{i^2} - \frac{n}{i} \Big) \doteq n^2(\pi^2/6) - n(\gamma + \log n).$$

(f) (xx limit of skewness. not zero. xx) Now consider

$$U_n = \frac{V_n - \mathrm{E}\, V_n}{(\mathrm{Var}\, v_n)^{1/2}}, \quad U_{n,0} = \frac{V_n - n(\gamma + \log n)}{n\pi/\sqrt{6}},$$

and show that $U_n - U_{n,0} \to_{\mathrm{pr}} 0$. Show further that

$$\mathrm{E}\, U_n^3 = \frac{\sum_{i=1}^n \mathrm{E}\,(T_i - p_i)^3}{(\mathrm{Var}\, Z_n)^{3/2}} \doteq \frac{2n^3 \sum_{i=1}^n (1/i^3) + O(n^2)}{n^3\pi^3/6^{3/2}} \to \frac{2 \cdot 1.2021}{\pi^3/6^{3/2}} = 1.1396.$$

(g) So we're outside limiting normality; show indeed that the Lindeberg condition cannot hold here. (xx limit distribution. other things. xx)

(h) (xx check this. xx) With $U_n' = (V_n - n\log n)/n = \bar T_n - \log n$, show that

$$\begin{aligned} \mathrm{E}\, \exp(-sU_n') &= \exp(s\log n) \prod_{i=1}^n \mathrm{E}\, \exp\{-(s/n)T_i\} \\ &= n^s \prod_{i=1}^n \frac{(i/n)\exp(-s/n)}{1 - (1-i/n)\exp(-s/n)} \\ &= n^s \frac{(n!/n^n)\exp(-s)}{\prod_{i=1}^n \{1 - (i/n)\exp(-s/n)\}}. \end{aligned}$$

Then show that $U_n' \to_d U$. (xx hmm, have not landed this properly yet, but can be cool story. and if we prove $U_n' \to_d U$ in some other way, we are automatically deriving the side consequence

$$A_n(s) = \prod_{i=1}^n \{1 - (i/n)\exp(-s/n)\} \doteq \frac{n^s \exp(-n-s)(2\pi n)^{1/2}}{\Gamma(1+s)},$$

or

$$\prod_{i=1}^{n}\{\exp(s/n)-(i/n)\} \doteq \frac{n^s\exp(-n)(2\pi n)^{1/2}}{\Gamma(1+s)},$$

which for $s=0$ is Stirling. need a bit more work. xx)

(i) It is also useful to find the distribution of $G_n(v)=\Pr(V_n\le v)$ explicitly. Argue that $V_n\le v$ is equivalent to $A_1\cap\cdots\cap A_n$, where A_i is the event that i is seen in the course of the first z attempts. With $B_i=A_i^c$ its complement, that i has not been seen during these first v attempts. Use this to deduce that $1-G_n(v)$ can be written

$$\begin{aligned}\Pr(B_1\cup\cdots\cup B_n) &= \binom{n}{1}\Pr(B_1)-\binom{n}{2}\Pr(B_1\cap B_2)+\binom{n}{3}\Pr(B_1\cap B_2\cap B_3)-\ldots\\ &= \sum_{j=1}^{n}(-1)^{j-1}\binom{n}{j}(1-j/n)^v,\end{aligned}$$

for $v\ge n$. Use algebra to also derive

$$g_n(v)=\Pr(V_n=v)=\sum_{j=1}^{n}(-1)^{j-1}\binom{n-1}{j-1}(1-j/n)^{v-1}\quad\text{for } v\ge n.$$

Use $x^{n-1}+x^n+\cdots=x^{n-1}/(1-x)$ for $|x|<1$ to derive the identity

$$\sum_{j=1}^{n-1}(-1)^{j-1}\binom{n}{j}(1-j/n)^{n-1}=1.$$

(j) Use the case $(T_1=1,\ldots,T_n=1)$ to derive

$$\prod_{i=1}^{n}(i/n)=\frac{n!}{n^n}=\sum_{j=1}^{n-1}(-1)^{j-1}\binom{n-1}{j-1}(1-j/n)^{n-1},$$

and argue that these expressions are close to $\exp(-n)(2\pi n)^{1/2}$, by the Stirling approximation. Show via arrangements of this formula that $\sum_{j=0}^{n}(-1)^j\binom{n}{j}j^n=n!$. (xx $-4\cdot1+6\cdot16-4\cdot81+256=24$, etc. xx)

(k) (xx pointer here to a different story, where we estimate n based on how long time it took us to reach level r, i.e. $W_r=T_1+\cdots+T_r$. it might be a CD story with $C_r(n)=\Pr_n(W_r<W_{r,\mathrm{obs}})+\frac12\Pr_n(W_r=W_{r,\mathrm{obs}})$. how many Italians in my neighbourhood? xx)

Ex. 2.70 *The 2nd largest, 3rd largest, etc., for exponentials.* Let as in Ex. 2.66 $X_1,\ldots,X_n$ be i.i.d. from the unit exponential model. For the largest observation we saw there that $X_{(n)}-\log n\to U$, the Gumbel distribution with c.d.f. $\exp\{-\exp(-u)\}$. Here we shall work with the 2nd largest, the 3rd largest, etc.

(a) For a positive, consider W_a, with density

$$g_a(w)=\Gamma(a)^{-1}\exp\{-aw-\exp(-w)\} \tag{2.13}$$

on the real line. Show that $V_a=\exp(-W_a)$ has the gamma distribution with parameters $(a,1)$, and that the Laplace transform becomes $\mathrm{E}\exp(-tW_a)=\Gamma(a+t)/\Gamma(a)$. The Gumbel distribution is the case of $a=1$, so we may consider (2.13) a generalised Gumbel.

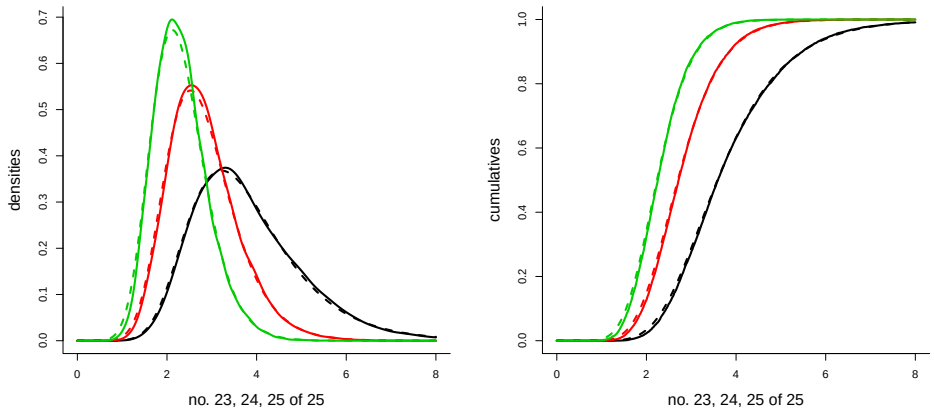


Figure 2.2: *For $n = 25$, the dashed curves are limit-approximation densities (left panel) and cumulatives (right panel) for order statistics 23, 24, 25 for i.i.d. exponentials, computed via the limits $\log n + W_i$ for $i = 1, 2, 3$. The full curves are the real densities and cumulatives, based on having simulated 10^4 outcomes for each.*

(b) Deduce that $W_a = \log(1/V_a)$ has mean $-\psi(a)$ and variance $\psi'(a)$, with $\psi(a) = \partial \log \Gamma(a)/\partial a$ the digamma function. Show that the c.d.f.s for g_1, g_2, g_3 are

$$\begin{aligned} G_1(w) &= \exp\{-w - \exp(-w)\}, \\ G_2(w) &= \exp\{-w - \exp(-w)\}\{1 + \exp(-w)\}, \\ G_3(w) &= \exp\{-w - \exp(-w)\}\{1 + \exp(-w) + \tfrac{1}{2}\exp(-2w)\}. \end{aligned}$$

(c) With order statistics $X_{(1)} < \cdots < X_{(n)}$, consider $W_{n,i} = X_{(n-i+1)} - \log n$, for given i; the case $i = 1$ is $W_{n,1} = X_{(n)} - \log n$ already considered in Ex. 2.66. Show that

$$\Pr(X_{(n-i+1)} - \log n \le w) = \Pr(U_{(n-i+1)} \le 1 - (1/n)\exp(-w)),$$

in terms of the order statistics for the uniform. Use the Beta connection of Ex. 3.17 to deduce that the density of $W_{n,i}$ may be written

$$g_{n,i}(w) = \mathrm{be}(1 - (1/n)\exp(-w), n - i + 1, i)(1/n)\exp(-w)$$

in terms of the Beta density with parameters $(n - i + 1, i)$. Take the limit to prove that $W_{n,i} \to_d W_i$.

(d) Construct a version of Figure 2.2, showing that the approximations based on the limit distributions work well already for $n = 25$. The dashed curves are using the limits, with $x_{(n-i+1)} = \log n + W_i$, wheras the dashed curves are the real densities and cumulatives, obtained by 10^4 simulations from $X_{(n-2)}, X_{(n-1)}, X_{(n)}$. (xx these are kernel estimates so need a pointer to Ch. 13. can we make a story out of this, insurance company cares for these most extreme outcomes. xx)

Ex. 2.71 *How many records are set?* Consider i.i.d. observations $X_1, X_2, \ldots$ from some continuous distribution on the line. Define $R_n = 1$ if X_n is bigger than all previous datapoints, i.e. a new record has been set.

(a) Explain first that the number of records set, in the course of the first n occasions, is $Z_n = \sum_{i=1}^n R_i$. Show that $\Pr(R_n = 1) = 1/n$, so that $\mathrm{E}\, Z_n = H_n = 1+1/2+\cdots+1/n$, the partial sum of the harmonic series. As we have noted earlier, $H_n - \log n \to \gamma_e = 0.5772...$, the Euler constant, so the number of records set is approximately $\log n + \gamma_e$.

(b) Show that $R_1, \ldots, R_n$ are independent, and deduce from this that $\mathrm{Var}\, Z_n = H_n - \sum_{i=1}^n 1/i^2 \doteq \log n + \gamma_e - \pi^2/6$. Use the Lindeberg theorem to show that $(Z_n - H_n)/H_n^{1/2} \to_d \mathrm{N}(0,1)$. Show also that $(Z_n - \log n)/(\log n)^{1/2} \to_d \mathrm{N}(0,1)$, which gives a simpler but somewhat more crude approximation to Z_n probabilities. Use these normal limits to give a prediction band, with $\Pr(a_n \le Z_n \le b_n) \to 0.95$.

(c) Now consider $A_n = Z_{2n} - Z_n$, the number of records set during $n, n+1, \ldots, 2n$. Show that A_n tends to the Poisson with mean $\log 2$. Generalise.

Ex. 2.72 *When the 00 box is hidden.* Consider a 2×2 table setup with counts $N_{0,0}, N_{0,1}, N_{1,0}, N_{1,1}$, corresponding to $N_{i,j}$ counting the cases of $(X = i, Y = j)$, for $i, j = 0, 1$, for two factors X and Y. We take the four counts to be a multinomial vector with probabilities $p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1}$. Assume now that the 00 box is hidden, hence also the total number $N = N_{0,0} + N_{0,1} + N_{1,0} + N_{1,1}$; one has observed counts $N_{0,1}, N_{1,0}, N_{1,1}$, but not the $N_{0,0}$ in question. How can one estimate the hidden $N_{0,0}$, and then in its turn N? Such questions and estimation methods go back to Petersen (1896), who needed to estimate fish populations in the 1890ies. Versions of methods methods developed below is being used in Story iii.13 to estimate the number of killed persons in Srebrenica 1995. See also Ex. **??** for generalisations.

(a) Assume in this exercise that factors X and Y are independent, with $\Pr(X = 1) = p = p_{1,\cdot}$ and $\Pr(Y = 1) = q = p_{\cdot,1}$; we use '$\cdot$' notation to indicate that the index in question is being summed over. Show that

$$p_{0,0} = (1-p)(1-q), \quad p_{0,1} = (1-p)q, \quad p_{1,0} = p(1-q), \quad p_{1,1} = pq.$$

(b) Argue that $N_{1,\cdot}N_{\cdot,1}/N^2$ and $N_{1,1}/N$ are both valid estimates of $p_{1,1}$. Discuss conditions under which $N^* = N_{1,\cdot}N_{\cdot,1}/N_{1,1}$ is a reasonable estimator of N.

(c) The N is unknown, but we may still study the usual ratios $\widehat{p}_{i,j} = N_{i,j}/N$. Show that there is joint convergence in distribution, say $N^{1/2}(\widehat{p}_{i,j} - p_{i,j}) \to_d A_{i,j}$, as N increases, with the $A_{i,j}$ forming a four-dimensional mean zero normal. Give its variance matrix. Under independence, show that $(N^* - N)/N^{1/2} = N^{1/2}(N^*/N - 1)$ has limit distribution

$$\begin{aligned} U &= (1/p)(A_{1,0} + A_{1,1}) + (1/q)(A_{0,1} + A_{1,1}) - \{1/(pq)\}A_{1,1} \\ &= \{pA_{0,1} + qA_{1,0} + (p+q-1)A_{1,1}\}/(pq). \end{aligned}$$

This is a normal $(0, \tau^2)$; show that indeed $\tau^2 = (1-p)(1-q)/(pq)$. How can this be used to form a confidence interval for N? (More general schemes, for estimators and confidence intervals, are developed in See Ex. **??**.)

(d) Show that the N^* leads to the natural estimator $\widehat{p} = N_{1,1}/N_{1,\cdot}$ for p. Find its approximate distribution, and assess how much is lost in precision by not knowing N. (xx check also with the implied $N^*_{0,0} = N^* - (N_{0,0} + N_{0,1} + N_{1,0})$. xx)

(e) The setup and methods above can be used in a variety of setups, for estimating the sizes of populations based on incomplete surveys; the N^* estimator above goes back to Petersen (1896), estimating the number of fish based on capture-recapture surveys. Carry out a few simulation experiments, as follows. There are fish $\{1, \ldots, N\}$ in your pond. Your first catch, with fish being caught as in a binomial setup with probability p_1, gives the index set A_1; your captured fish are marked and released in the pond. Similarly your second catch, with catch probability p_2, gives index set A_2. By counting the numbers $N_{1,0}, N_{0,1}, N_{1,1}$ in the associated Venn diagram, estimate the total number of fish N (and your analysis should work without knowing p_1, p_2). Check if your 95 percent confidence interval captures the real N. To play with these methods, also to understand how the catch probabilities p_1, p_2 influence estimates and precision, there are helpful and easy-to-use algorithms in `R`, namely `intersect`, `union`, `setdiff`.

(f) (xx might bypass this point. but might include the case of three surveys, or leave it to $\ell_{\rm prof}(N)$ analysis in Ch 5. but we can ask for analysis of the estimator

$$n^*_{0,0,0} = \frac{n_{1,0,0}n_{0,1,0} + n_{1,0,0}n_{0,0,1} + n_{0,1,0}n_{0,0,1}}{n_{1,1,0} + n_{1,0,1} + n_{0,1,1}},$$

used in Lum et al. (2013). xx)

Notes and pointers

(xx to come. we point to certain famous things from the past: Kolmogorov (1933b), Lindeberg (1922), Borel and Cantelli. tail bounds. emil's extension of the Inlow (2010) paper, from CLT to Lindeberg. more on Lindeberg and the history of CLT developments in Cramér (1976), also see Schweder (1980, 1999). xx)

[xx the first of the two lemmas: Nils beta 1990. xx]

(xx include also something on Fra Preface i *Life and Times of CLT*: "For those who teach a course in probability whose objective is to prove the Central limit theorem of interest [...] is commentary on the characteristic function approach employed by Lyapunov versus the 'very simple' proof, as Le Cam describes it, given by Lindeberg". also, LeCam (1986). xx)

[xx for Scheffé: see Scheffé (1947), but also Kusolitsch (2010), who explains that the result is a special case of results published by F. Riesz in 1928. see also what Scheffe says in his paper about comments he got from Morse. xx]

(xx push this to Notes. xx) Inlow (2010) has shown how one can prove the usual CLT without the technical use of characteristic and hence complex functions. Essentially, he writes the X_i in question as $Y_i + Z_i$ with $Y_i = X_i\, I\{|X_i| \le \varepsilon\sqrt{n}\}$ and $Z_i = X_i\, \{|X_i| > \varepsilon\sqrt{n}\}$, after which 'ordinary' m.g.f.s may be used for the part involving the Y_i, yielding the normal limit, supplemented with analysis to show that the part involving the Z_i tends to zero in probability. – It is a non-trivial matter to extend Inlow's arguments,

from the CLT to the Lindeberg theorem, but this is precisely what is done in Stoltenberg (2019). Check that note, on the book website, and make sure you understand its main tricks and steps.

(xx When Jarl Waldemar Lindeberg was reproached for not being sufficiently active in his scientific work, he said, 'Well, I am really a farmer'. And if somebody happened to say that his farm was not properly cultivated, his answer was, 'Of course my real job is to be a mathematics professor'. Hundred years ago!, i.e. in 1920, he published his first paper on the CLT, and in 1922 he generalised his findings to the classical Lindeberg Theorem, with the famous Lindeberg Condition, securing limiting normality of a sum of independent but not identically distributed random variables. He did not know about Ляпунов's earlier work, and therefore not about условие Ляпунова, the Lyapunov condition, which we treat below as a simpler-to-reach condition than the more general one of Lindeberg. Other luminaries whose work touch on these themes around the 1920ies and beyond include Paul Lévy, Harald Cramér, William Feller, and, intriguingly, Alan Turing who (allegedly) won the war and invented computers etc. xx)

(xx point to a couple of characterisation theorems books, kagan linnik rao, one more. xx)

(xx for Notes: The little $\log(1+x)$ lemma is stated, proven, and used in Hjort (1990b, Appendix). xx)

(xx the material is from Hjort and Pollard (1993) and Hjort (1986a). xx)

[xx For notes and pointers: Ex. 2.20(a) is from Jacod and Protter (2004, p. 166), they have it from Pollard (1984) xx].

ToDo notes, of 12-August-2024.

Clean and calibrate. Include a couple of classic nonparametric test procedures, like Wilcoxon, the sign test, more, to showcase the use of Lindeberg things to show limiting normality of such statistics too. point to Hjort and Pollard (1993) and Hjort (1986a).

Include some non-normal limits. Can take $\widehat{\mu}^* = \{1-c(D_n)\}\widehat{\mu}_{\rm narr} + c(D_n)\widehat{\mu}_{\rm wide}$ from model selection. And point to $n^{2/5}$ rates for f estimation.

I.3

Parameters, estimators, precision, confidence

With data observed from a statistical model, the theme of this chapter is that of constructing estimators for unknown statistical parameters, along with assessing their precision. This provides ways of comparing competing estimation methods. Basic concepts include the bias, the variance, the mean squared error of estimators. The development also naturally leads to the important notion and basic machinery of confidence intervals. General estimation methods covered here include the method of moments and the method of quantiles; these can also be combined. For regression setups, with response variables influenced by covariates, we go through the method of least squares. To understand and utilise the properties of classes of estimators in general models, we utilise the machinery of large-sample normal approximations, from Ch. 2. This also enables one to assess precision and to compare different competing estimators. Similar remarks apply also for the more versatile method of maximum likelihood, treated in Ch. 5.

Key words: approximations, confidence, estimators, linear regression, model parameters, moment matching, quantile matching, risk

Most statistical models have *parameters*, as we learn from the generous variety of models in Ch. 1. Parameters may then be fine-tuned, or estimated, from data, which is the grand theme of the present chapter. In generic terms, if a model has density $f(y, \theta)$, with $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{t}}$ its parameter vector, we use data $\mathcal{D}$ to construct *an estimator* $\widehat{\theta} = \widehat{\theta}(\mathcal{D})$. Thus $f(y, \widehat{\theta})$ is the fitted model, which we use for interpretation and inference, themes we return to in more detail in later chapters. The data $\mathcal{D}$ can often be in the form of direct independent observations $y_1, \ldots, y_n$ from the model, but can also be different in character, involving censoring mechanisms, or measurement error.

A sensible minimum demand for an estimator is that it should tend to the right value, with increasing data volume. Formally, if $\widehat{\theta} = \widehat{\theta}_n$ is the estimator for some parameter whose true value is θ_0, based on a sample of size n, then we say that the estimator (or, more pedantically, the sequence of estimators) is *consistent* provided $\widehat{\theta}_n$ converges in probability to θ_0, i.e. $\widehat{\theta}_n \to_{\mathrm{pr}} \theta_0$ in the terminology of Ch. 2.

focus parameter

One often needs estimates and inference methods for *focus parameters*, those of particular and context-driven interest, which are one-dimensional functions $\phi = \phi(\theta_1, \ldots, \theta_p)$

of the underlying model parameter vector. If $\widehat{\phi}$ is an estimator for this parameter, we often care about its mean, represented here as

$$\mathrm{E}_\theta\, \widehat{\phi} = \phi + b(\theta). \tag{3.1}$$

The footscript signals that the expectation operator is at work at the parameter position θ. The $b(\theta)$ is termed *the bias*; in various cases it is a function of ϕ only, but in general it depends on the full parameter vector θ. If $\mathrm{E}_\theta\, \widehat{\phi} = \phi$, at all positions θ, we say the estimator is *unbiased*. In addition to wishing for estimators with small bias, we care about its variability, and often about its *mean squared error*

unbiased estimator

mean squared error

$$\mathrm{mse}(\widehat{\phi}, \theta) = \mathrm{E}_\theta\, (\widehat{\phi} - \phi)^2 = \mathrm{Var}_\theta\, \widehat{\phi} + b(\theta)^2, \tag{3.2}$$

the classic variance plus squared bias. This is a function of the unknown parameter, and gives a way of understanding and comparing performance for competing estimation schemes. When we can sort out the mathematics properly, depending on the situation at hand, we then choose estimators with smaller mse than those of competitors.

The $\mathrm{mse}(\widehat{\phi}, \theta)$ of (3.2) is sometimes called the risk, or risk function, and relates to having chosen squared error $(\widehat{\phi} - \phi)^2$ as the underlying measure of quality. Other ways in which to compare and rank performance, involving also different quality functions and risk functions, will be dealt with in Ch. 8.

Fundamental and conspicuous instruments in the statistical toolkits, when summarising and reporting findings of investigations, are *confidence intervals* and *testing of null hypotheses*. The development in the present chapter deals with the former, regarding interpretation, construction, properties, performance, whereas the following chapter handles the latter, along with further connections.

Consider in general terms data y, perhaps a long vector or a data matrix, from a model with parameter $\theta = (\theta_1, \ldots, \theta_p)$, with $\phi = \phi(\theta_1, \ldots, \theta_p)$ a parameter of interest. Then $[L(y), U(y)]$ is a confidence interval, with confidence level α, like 0.90 or 0.95 or an even higher 0.99, provided

confidence interval

$$\Pr_\theta\{L(Y) \le \phi \le U(Y)\} = \alpha \quad \text{for all } \theta. \tag{3.3}$$

Thus $[L(Y), U(Y)]$ is a random interval, and with a high number of repeated situations it will capture the underlying ϕ a fraction α of the times. The reported confidence interval is $[L_{\mathrm{obs}}, U_{\mathrm{obs}}] = [L(y_{\mathrm{obs}}), U(y_{\mathrm{obs}})]$, computed based on the actually observed data y_{obs}. Occasionally one also wants to construct *confidence regions* for a parameter vector, as opposed to confidence intervals for each of its components. Thus is (a, b) is a parameter pair in a model, a random region R, based on data, is a 95 percent confidence region provided $\Pr_\theta((a, b) \in R) = 0.95$ for all model parameters θ.

In various setups one can study distributions, biases, variances, confidence etc. quite accurately, as will be seen in many exercises below. Often enough this might be too complicated, however, and one relies instead on good approximations. There is indeed a host of normal approximations, sometimes with additional tools for finetuning these, as studied in Ch. 2, with yet more to come in Ch. 5. These methods may be understood, appreciated, seen in action, and used for new situations, without necessarily having been through each δ and ε of Ch. 2.

In this chapter we learn certain estimation principles, including those associated with the method of moments and the method of quantiles. There is also room for combining such methods, or for coming up with new estimators in unfamiliar waters. We go on to more advanced models and hence estimation methods in later chapters (and in several of our stories), but included below is the basics of linear regression and the least sum of squares methods. The more versatile and often well-performing method of *maximum likelihood* will be studied with care in Ch. 5.

(xx ToDo nils, not yet fully done as of 12-August-2024: moment method; quantile method; both with transformation and delta method things from Ch3; least squares with a bit more written out for linear regression. xx)

[xx In this brief intro there should be a figure, conveying some basic ideas. We may snikinnføre confidence intervals, but that comes with more weight in Ch. 4, along with testing and power and p-values. we do mention a few key concepts here in intro, like unbiasedness, low variance, etc. xx]

Precise estimation in a few classical models

Ex. 3.1 *Mean squared error.* Suppose data lead to an estimator $\widehat{\phi}$ for a focus parameter $\phi = \phi(\theta)$, in a model with parameter θ.

(a) Verify the mse formula (3.2), and note its Pythagorean character. Make a little right-angle triangle figure with absolute bias and standard deviation for the two short sides and root-mse, rmse $=$ mse$^{1/2}$ on the long side. These root operations bring the risk components down to the original scale of the measurements.

(b) For a simple situation, let $Y \sim \mathrm{N}(\theta, 1)$, with θ to be estimated. Find formulae for the mean squared errors of the three estimators $0.9\,Y$, Y, $1.1\,Y$. Note the interplay between bias and variance.

(c) Generalise to the case of $Y_1, \ldots, Y_n$ being i.i.d. from $\mathrm{N}(\theta, 1)$. Find $\mathrm{mse}(\widehat{\theta}, \theta)$ for the three estimators $0.99\,\bar{Y}$, $\bar{Y}$, $1.01\,\bar{Y}$, with $\bar{Y}$ the sample average. Comment on what you find.

(d) In somewhat more general terms, consider an i.i.d. sample $Y_1, \ldots, Y_n$ from a distribution with unknown mean μ and variance σ^2. Show that $\bar{Y}$ is unbiased with variance σ^2/n. If your estimator for μ is $c_n \bar{Y}$, what is required of c_n, in order for the mean squared error to go to zero with growing n?

Ex. 3.2 *Binomial estimation.* Consider Y being binomial (n, p), as in Ex. 1.3.

(a) To estimate p the canonical choice is $\widehat{p} = Y/n$. Find its mean and variance, and a formula for $\mathrm{mse}(\widehat{p}, p) = \mathrm{E}_p\,(\widehat{p} - p)^2$.

(b) Then compare the simple binomial unbiased proportion with the estimator $\widehat{p}_B = (Y + 1)/(n + 2)$ Find its bias and variance, and a formula for $\mathrm{mse}(\widehat{p}_B, p)$. Draw the two mse functions in a diagram, for say $n = 10$. When is the Bayes estimator better than the Y/n, according to this criterion?

Ex. 3.3 *Estimating the normal mean.* Suppose we have independent observations $Y_1, \ldots, Y_n$ from the normal distribution $\mathrm{N}(\mu, \sigma^2)$.

(a) Prove that the sample average $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ has the $\mathrm{N}(\mu, \sigma^2/n)$ distribution. This $\bar{Y}$ is the canonical estimator for μ. Find also a clear formula for its risk, or mean squared error, namely $\mathrm{mse}(\bar{Y}, \mu, \sigma) = \mathrm{E}_{\mu,\sigma}(\bar{Y} - \mu)^2$. The subscript indicates that the mean operator is with respect to the probability mechanism dictated by (μ, σ).

(b) Then generalise the above somewhat, by finding the mean and variance also for the estimator $\widehat{\mu} = b\bar{Y}$, with b a constant (which might be close to 1). Use this to put up a clear expression for

$$\mathrm{mse}(b\bar{Y}, \mu, \sigma) = \mathrm{E}_{\mu,\sigma}(b\bar{Y} - \mu)^2.$$

Illustrate this, for values $b = 0.98, 1.00, 1.02$, and comment. For what values of the parameters (μ, σ) will the estimator $0.98\,\bar{Y}$ be better than the classic $\bar{Y}$? Are there values of the parameters where $1.02\,\bar{Y}$ is better than the plain $1.00\,\bar{Y}$?

(c) Suppose the starting assumptions about the data at hand are changed to merely saying that the Y_i are i.i.d. with mean μ and standard deviation σ, i.e. we avoid saying that the distribution of the error terms $\varepsilon_i = (Y_i - \mu)/\sigma$ needs to be exactly normal. How does this affect your findings and claims for the previous points?

Ex. 3.4 *Estimating the normal variance.* As in Ex. 3.3, suppose there are i.i.d. data $Y_1, \ldots, Y_n$ from the $\mathrm{N}(\mu, \sigma^2)$. Here we care about the standard deviation parameter σ. As we saw in Ex. 1.45, $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_m$, where $m = n - 1$. Also, the Z is stochastically independent of the sample mean $\bar{Y}$.

(a) Use the statement above to find the mean and variance of $\widehat{\sigma}^2 = cZ$ (where c ought to be about $1/n$). Find the mean squared error $\mathrm{mse}(cZ, \sigma) = \mathrm{E}_\sigma\,(cZ - \sigma^2)^2$. Check in particular the result for $c = 1/(n-1)$, the classical factor to make the estimator unbiased; for $c = 1/n$, which comes out of the maximum likelihood paradigm (see Ch. 5); and for $c = 1/(n+1)$.

(b) Find the best possible constant c for estimators of this type cZ, using the mean squared error on the σ^2 scale as criterion.

(c) Find also the mean and variance of $dZ^{1/2}$, seen as an estimator of σ, i.e. on the standard deviation scale, not that of the variance. Find an expression for $\mathrm{mse}(dZ^{1/2}, \sigma) = \mathrm{E}_\sigma\,(dZ^{1/2} - \sigma)^2$. Find the best d, according to this criterion.

(d) We may also finetune σ estimation on the log-scale. Examine the risk function $\mathrm{mse}(kZ^{1/2}, \sigma) = \mathrm{E}_\sigma\,\{\log(kZ^{1/2}) - \log\sigma\}^2$, and find the best value of k.

(e) A different solution to the issue of determining 'the best constant' when estimating σ, disregarding tradition and mathematical convenience, might be as follows. With $\widehat{\sigma}^2 = Z/(n-1)$ being the traditional sample variance, with $1/(n-1)$ selected to achieve unbiasedness on the σ^2 scale, consider $\sigma^* = c_n\widehat{\sigma}$, with c_n to be fine-tweaked perhaps a little bit away from 1. Find the c_n that makes $\mathrm{risk}(c_n\widehat{\sigma}, \sigma) = \mathrm{E}_\sigma\,|c_n\widehat{\sigma} - \sigma|$ smallest.

This means relying on absolute error as loss function, and the solution needs numerical minimisation of a function which needs numerical integration. Give a table with these optimal c_n for say $n = 10, \ldots, 30$. Show that $c_n \to 1$ as n grows.

Ex. 3.5 *Confidence interval for an exponential rate.* Choose a sample size n, and simulate i.i.d. variables $Y_1, \ldots, Y_n$ from the exponential distribution, see Ex. 1.8, with parameter θ equal to say $\theta_0 = 3.33$.

(a) Construct a 90 percent confidence interval $[L, U]$ for θ. Check if θ_0 is contained in this interval, for the data you generated. Repeat the experiment say 100 times, and record how often the intervals contain θ_0. What is in fact the distribution of N, the number of the 100 intervals which cover the truth?

(b) In addition to checking whether the intervals cover the truth, compute the length $D = U - L$, and give a histogram of its distribution. Find $\mathrm{E}\, D$. Repeat also these experiments with a couple of other sample sizes, and comment.

Ex. 3.6 *Confidence interval for a normal variance.* Let $Y_1, \ldots, Y_n$ be i.i.d. from a normal distribution. How can we set up confidence intervals for the standard deviation σ (or, equivalently, for the variance σ^2)? Writing $m = n - 1$, the sample variance is $\widehat{\sigma}^2 = Z/m$, with $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_m$, from Ex. 1.45.

(a) Start with $[a, b] = [\Gamma_m^{-1}(0.05), \Gamma_m^{-1}(0.95)]$, an interval covering the χ^2_m with probability 0.90. Transform $a \le m\widehat{\sigma}^2/\sigma^2 \le b$ to the confidence interval $\mathrm{ci} = [(m/b)^{1/2}\widehat{\sigma}, (m/a)^{1/2}\widehat{\sigma}]$. Show in detail that indeed $\Pr_\sigma(\sigma \in \mathrm{ci}) = 0.90$.

(b) Most often one wishes to estimate and assess the σ parameter directly, being on the same scale as the measurements, but once in a while it would be more natural to communicate and interpret results on the variance scale. Show in suitable detail that $\Pr_\sigma((m/b)\widehat{\sigma}^2 \le \sigma^2 \le (m/a)\widehat{\sigma}^2) = 0.90$; confidence intervals can in this fashion be easily transformed, say from θ to $g(\theta)$, as here, from σ^2 to σ, or the other way around.

(c) The construction above is 'equitailed', starting with 0.05 probability to the left of a and 0.05 probability to the right of b. One might somewhat more generally use any $[a, b]$ with 0.90 probability for the χ^2_m, needing $\Gamma_m(b) - \Gamma_m(a) = 0.90$. The length of the 90 percent σ interval above is proportional to $1/a^{1/2} - 1/b^{1/2}$. Minimise this function, say for $m = 10, 20, 30, 40$. Compare these length-minimising 0.90 intervals with the simpler ones, and comment.

(d) (xx same exercise for minimising $1/a - 1/b$, for σ^2. moral: it doesn't matter so much, and we're largely happy with the equitailed scheme. xx)

(e) (xx simple illustration with an easy dataset. xx)

Ex. 3.7 *Confidence interval for a normal mean.* Here we go through the basics for constructing confidence intervals for normal means. Since approximations to normality abound in statistical theory and practice, what we learn here quickly finds use also outside the strict normality assumptions.

(a) Start with the simplest prototype setup, a single Y from the $\mathrm{N}(\xi, 1)$ model. Show that the random interval $[Y - 1.96, Y + 1.96]$ captures ξ with probability 0.95.

(b) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the $\mathrm{N}(\xi, \sigma^2)$ model, at the moment with standard deviation parameter σ taken known. With $\bar{Y}$ as usual being the data average, show that $\sqrt{n}(\bar{Y} - \xi)/\sigma$ is standard normal, and deduce from this that $\bar{Y} \pm 1.96\,\sigma/\sqrt{n}$ is a 95 percent confidence interval for ξ.

(c) In most cases also the σ is unknown, however. Let $\widehat{\sigma}$ be the usual empirical standard deviation, see e.g. Ex. 3.4. Show that the natural construction

$$t = \frac{\bar{Y} - \xi}{\widehat{\sigma}/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \xi)}{\widehat{\sigma}} = \frac{\sqrt{n}(\bar{Y} - \xi)/\sigma}{\widehat{\sigma}/\sigma} \tag{3.4}$$

has a distribution not depending on the two parameters, and that this distribution, call it G_n, is symmetric around zero. Deduce also that with $t_{0,n} = G_n^{-1}(0.975)$, the random interval $\bar{Y} \pm t_{0,n}\widehat{\sigma}/\sqrt{n}$ covers ξ with probability 0.95.

(d) It is then 'only' a matter of finding and perhaps tabulating the distribution G_n of t. It is in fact the celebrated t distribution, with $\mathrm{df} = n - 1$ degrees of freedom, see Ex. 1.46. But even without that specific knowledge detail, we could easily have simulated a high number for t, from (3.4), and read off the required quantile. (xx also: $t_{0,n}$ not far from 1.96 with n moderate to big. xx)

(e) In more generality, suppose β is a model parameter for which there is an estimator $\widehat{\beta}$ with distribution $\mathrm{N}(\beta, c_n^2\sigma^2)$, say, with a known factor c_n. Suppose also that there is a statistically independent estimator $\widehat{\sigma}$ for σ, with the property that $\widehat{\sigma}^2/\sigma^2 \sim \chi^2_m/m$, for a known m. Then show that $t = (\widehat{\beta} - \beta)/(c_n\widehat{\sigma}) \sim t_m$. Put up a 0.99 confidence interval for β based on this.

(f) Suppose $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are random samples from two normal groups with equal variance, say $\mathrm{N}(\xi_1, \sigma^2)$ and $\mathrm{N}(\xi_2, \sigma^2)$. Find the distribution of the difference of sample means $D_n = \bar{X} - \bar{Y}$, and construct a confidence interval for $\delta = \xi_1 - \xi_2$. (xx check for other cases in the book, testing, CDs. xx)

Ex. 3.8 *Normal quantiles: estimation and confidence.* Consider again the setup of Ex. 3.3, with a sample of Y_i from the normal $\mathrm{N}(\mu, \sigma^2)$ model. In the present exercise we care about quantiles, as opposed to 'only' the mean or the standard deviation.

(a) Show that the c.d.f. of Y_i may be written $F(y) = \Pr(Y_i \le y) = \Phi((y - \mu)/\sigma)$, with $\Phi(x)$ the c.d.f. for the standard normal (i.e. the `pnorm` function in `R`). Show that the q quantile $F^{-1}(q)$ is equal to $\gamma_q = \mu + z_q\sigma$, with $z_q = \Phi^{-1}(q)$. Thus the 0.95 quantile is $\gamma_{0.95} = \mu + 1.645\,\sigma$, etc.

(b) Find the mean and variance of the natural estimator $\widehat{\gamma}_q = \bar{Y} + z_q\,\widehat{\sigma}$, where $\widehat{\sigma} = (Z/m)^{1/2}$, with Z as in Ex. 3.4 and $m = n - 1$.

(c) Show that we may write

$$\begin{aligned} \widehat{\gamma}_q - \gamma_q &\sim \mu + (\sigma/\sqrt{n})N + z_q\sigma(K_m/m)^{1/2} - \mu - z_q\sigma \\ &= \sigma[(1/\sqrt{n})N + z_q\{(K_m/m)^{1/2} - 1\}], \end{aligned}$$

in terms of $N \sim \mathrm{N}(0,1)$ and $K_m \sim \chi^2_m$, with these being independent. Verify that the pivot

$$W_{n,q} = \frac{\sqrt{n}(\widehat{\gamma}_q - \gamma_q)}{\widehat{\sigma}} \sim \frac{N + z_q\sqrt{n}\{(K_m/m)^{1/2} - 1\}}{(K_m/m)^{1/2}}$$

has a distribution not depending on μ, σ. It can be simulated, for any given sample size n and quantile level q. With a_n such that $\Pr(-a_n \le W_{n,q} \le a_n) = 0.95$, convert this to a 95 percent confidence interval for γ_q. Note that for $q = 0.50$, the median case, the distribution of the $W_{n,q}$ is the t_m, the t with degrees of freedom $m = n - 1$.

(d) (xx nils calibrate better, in view of large-sample things coming below. xx) Use the delta method of Ex. 2.47 to show that $\sqrt{n}\{(K_m/m)^{1/2} - 1\} \to_d \mathrm{N}(0, \frac{1}{2})$, and explain that $W_{n,q} \to_d \mathrm{N}(0, 1 + \frac{1}{2}z_q^2)$. Construct an approximate 95 percent confidence interval for γ_q based on this. [xx simple data illustration, perhaps inside story xx]

(e) Consider more generally *any* smooth function $\gamma = g(\mu, \sigma)$. With $\widehat{\gamma} = g(\widehat{\mu}, \widehat{\sigma})$, use the delta method to find the limit distribution of $\sqrt{n}(\widehat{\gamma} - \gamma)$. Use this to set up a confidence interval for γ.

Ex. 3.9 *The empirical distribution function.* (xx nils will perestroik this in view of exercises in Ch2. xx) Assume there is an i.i.d. dataset $Y_1, \ldots, Y_n$ from an unknown distribution, with c.d.f. $F(t) = \Pr(Y_i \le t)$. The *empirical distribution function* is $F_n(t) = n^{-1}\sum_{i=1}^n I(Y_i \le t)$, the simple binomial proportion of points falling in $(-\infty, t]$. We saw in Ex. 2.55 that F_n with probability 1 tends uniformly to F; here we learn about how fast this happens, via convergence in distribution.

the empirical distribution function

(a) Explain that the empirical distribution function is the cumulative of the probability measure that puts probability mass $1/n$ at each data point. This is the natural nonparametric estimator of the unknown F.

(b) Construct a version of Figure 3.1, left panel, where $n = 100$ datapoints are simulated from the distribution $f = 0.50\,\mathrm{Expo}(r_1) + 0.50\,\mathrm{Expo}(r_2)$, with rates $r_1 = 2.00$ and $r_2 = 4.00$. The empirical $F_n(t)$ is the natural nonparametric estimator of the underlying (and typically unkwown) F.

(c) Since we know so much about the binomial, we quickly learn a few basic properties of the F_n. Show that $F_n(t)$ is unbiased for $F(t)$, and that its variance is $F(t)\{1 - F(t)\}/n$.

(d) Consider the process $Z_n(t) = \sqrt{n}\{F_n(t) - F(t)\}$. Show that it has mean zero, and that $Z_n(t) \to_d Z(t)$, say, where $Z(t)$ is a zero-mean normal with variance $F(t)\{1-F(t)\}$. Show also that

$$\mathrm{cov}\{Z_n(t), Z_n(t')\} = F(t)\{1 - F(t')\} \quad \text{for } t \le t'.$$

Compute and display the Z_n plot, using the same data values as for the previous figure; in other words, construct a version of Figure 3.1, right panel. [xx nils emil: we might contemplate putting comments such as the following in a 'comments' format, at the end

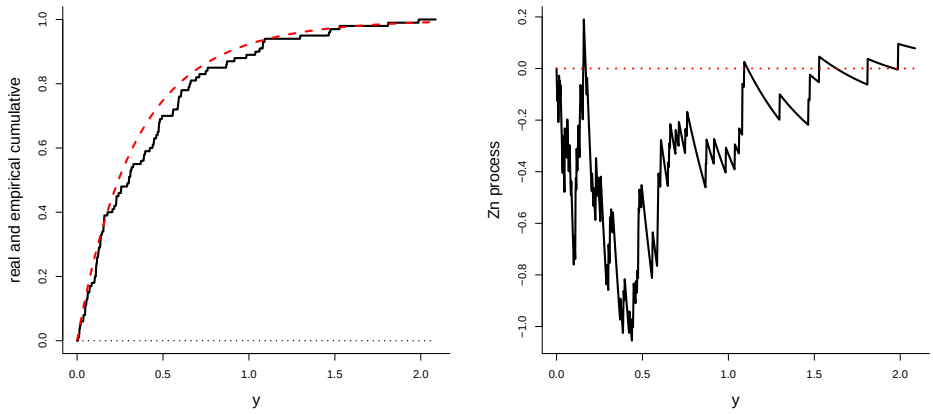


Figure 3.1: *Left panel: the real underlying data-generating $F(t)$ (dashed, red), with the empirical distribution function $F_n(t)$ (full line, black), computed from a sample of $n = 100$ data points from F. Right panel: the process $Z_n(t)$, computed for the data used for the same data. In 95 percent of such cases, the maximum absolute value of the Z_n process will be below 1.358.*

of certain exercises, with pointsrs to things to come, connections, etc. xx] Such plots may e.g. be used to check model adequacy – if the data come from a distribution not close to the F used to construct the plot, then the Z_n plot will deviate aignificantly from the zero line. To understand what might qualify as 'significantly different from the zero line' means we need theory for the behaviour of the full Z_n process, not merely the pointwise result that $Z_n(t) \to_d \mathrm{N}(0, F(t)\{1 - F(t)\})$.

(e) (xx some pointers: to Ch. 9. the 1.358 limit. kolmogorov-smirnov. and to Glivenko—Cantelli theorem, in Ex. 2.55. the F_n is used in CoW Story. there is full process convergence $Z_n \to_d Z$, a Gaussian zero-mean process with covariance function $F(y)(1-F(y'))$, see Ch. 9. kolmogorov-smirnov things. xx)

Ex. 3.10 *Estimating the normal density.* Most often the statistical interest lies in estimating some parameter related to, or expressed through, the normal distribution, like the mean, spread, or quantile, as illustrated above. In some situations one wishes to estimate the density itself. Consider once again a sample $Y_1, \ldots, Y_n$ from the normal $\mathrm{N}(\mu, \sigma^2)$.

(a) For the parameter σ, we shall again use $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_m$, with $m = n-1$, as in several previous exercises. With the traditional default estimator $\widehat{\sigma}^2 = Z/(n-1)$ of (1.5), find formulae for the mean of $1/\widehat{\sigma}^2$ and $\log \widehat{\sigma}$.

(b) Construct unbiased estimators for $1/\sigma^2$ and for $\log \sigma$, and then for the log-density function $\log f(y, \xi, \sigma) = -\log \sigma - \frac{1}{2}(y-\mu)^2/\sigma^2 - \frac{1}{2}\log(2\pi)$. In (xx pointer to exercise Ch8) we also construct an unbiased estimator on the direct scale $f(y, \xi, \sigma)$.

(c) For independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ from normal populations $\mathrm{N}(\xi_1, \sigma_1^2)$ and $\mathrm{N}(\xi_2, \sigma_2^2)$, construct an unbiased estimator for $d(y) = \log\{f_2(y)/f_1(y)\}$.

Confidence via normal approximations

Ex. 3.11 *Confidence intervals via normal approximations.* (xx make connection to Wald tests to come in Ex. 4.5. xx) As we've already seen in various situations of Ch. 2, there are often estimators for interest parameters for which there is approximate normality. Then various recipes under strict normality can still be used, but as approximations.

(a) Suppose ϕ is such a parameter of interest, for which there is an estimator $\widehat{\phi}$, being approximately normal, in the mathematical sense of $\sqrt{n}(\widehat{\phi} - \phi) \to_d \mathrm{N}(0, \tau^2)$, for some appropriate limiting variance τ^2. Suppose also that there is a consistent estimator $\widehat{\tau}$ of τ, with $\widehat{\tau}/\tau \to_{\mathrm{pr}} 1$. Show that $Z_n = \sqrt{n}(\widehat{\phi} - \phi)/\widehat{\tau} \to_d \mathrm{N}(0,1)$; you may check with Ex. 2.23.

(b) Show under these mild and very frequently met assumptions that

$$\Pr(\phi \in \widehat{\phi} \pm 1.96\,\widehat{\tau}/\sqrt{n}) \to 0.95.$$

In other words, the $[\widehat{\phi} - 1.96\,\widehat{\tau}/\sqrt{n}, \widehat{\phi} + 1.96\,\widehat{\tau}/\sqrt{n}]$ is an *approximate* or *asymptotic* 95 percent interval for ϕ. Note the grand generality here; this simple construction works in a large variety of situations, also in nonparametric setups, cases with dependent data, etc.

(c) The simplest interesting application of this standard recipe is for the unknown mean ξ of a population. Verify via the CLT of Ch. 2 that $\sqrt{n}(\bar{Y} - \xi)/\widehat{\sigma} \to_d \mathrm{N}(0,1)$. Hence the t-based interval $\bar{Y} \pm 1.96\,\widehat{\sigma}/\sqrt{n}$, for which we have very precise probability computations under normality, is large-sample correct even if the data are not at all normal.

(d) Suppose (X, Y, Z) is trinomial (n, p, q, r), with $p + q + r = 1$. Construct an approximate 90 percent confidence interval for $d = q - p$. – Check that you see how similar and not complicated tasks can be tended to in the examples of Ex. 2.43.

Ex. 3.12 *Confidence intervals for the standard deviation, outside normality.* Consider i.i.d. data $Y_1, \ldots, Y_n$, from which we compute the classical

$$\widehat{\xi} = \bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i \quad \text{and} \quad \widehat{\sigma} = \Big\{ \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \Big\}^{1/2}.$$

Here we illustrate the general large-sample methods by building confidence intervals for σ, with no assumptions on the distribution of the data, like normality. The only mild assumption we make is a finite fourth moment, in order for $\widehat{\sigma}$ to have a clear limit distribution. See Figure 3.2 for 100 simulated confidence intervals, all attempting to capture the true value, here $\sigma = 1$, for two different sample sizes.

(a) Make sure you understand and can prove that $\widehat{\xi}$ and $\widehat{\sigma}$ are consistent for ξ and σ, from the LLN theorems of Ch. 2.

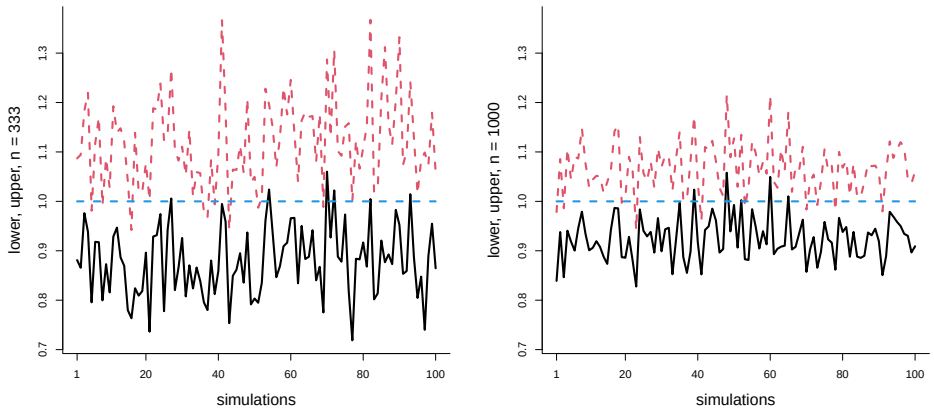


Figure 3.2: *Simulations, with datasets from the unit exponential, displaying lower and upper confidence points for 90 percent intervals; the intervals attempt to cover the true value* $\sigma = 1$*, and will succeed about 90 percent of the time. Left panel: with* $n = 333$*; right panel: with* $n = 1000$*.*

(b) For $S_n^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$, use Ex. 2.41 to establish that $\sqrt{n}(S_n^2 - \sigma^2)$ tends to $\mathrm{N}(0, \sigma^4(2+\gamma_4))$, in terms of the kurtosis parameter $\gamma_4 = \mathrm{E}\,\{(Y_i - \xi)/\sigma\}^4 - 3$. Then transform this, from variance to its square root, getting back to the real scale of the measurements: using the delta methods of Ch. 2, show that $\sqrt{n}(\widehat{\sigma} - \sigma) \to_d \mathrm{N}(0, (\frac{1}{2} + \frac{1}{4}\gamma_4)\sigma^2)$.

(c) Show that $\widehat{\gamma}_4 = (1/n)\sum_{i=1}^n \{(Y_i - \bar{Y})/\widehat{\sigma}\}^4 - 3$ is consistent for γ_4, and use this to construct an approximate 90 percent confidence interval for σ. Note that this is a nonparametric procedure, totally free of other distributional assumptions, like normality; *if* one assumes normality, as an extra condition, one may do more, of course.

(d) For an illustration, consider the unit exponential distribution; show that the standard deviation is 1 and that the kurtosis is $\gamma_4 = 6$. Simulate a suitably high number of datasets of size $n = 333$ from this distribution. For each simulated dataset, compute $\widehat{\gamma}_4$, to check how close it is to γ_4, along with the approximate 90 percent confidence interval for σ. Construct a version of Figure 3.2 (left panel for $n = 333$ and right panel for $n = 1000$). Examine in particular the coverage of your intervals: how often do they contain the correct σ? Use simulations to check How big n must be, in order for $\widehat{\gamma}_4$ to be inside $[5.8, 6.2]$ with probability at least 95 percent.

(e) Coming back to the general situation, show that

$$\begin{pmatrix} \sqrt{n}(\bar{Y} - \xi) \\ \sqrt{n}(S_n^2 - \sigma^2) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2, & \sigma^3\gamma_3 \\ \sigma^3\gamma_3, & \sigma^4(2+\gamma_4) \end{pmatrix}\Big),$$

and also that

$$\begin{pmatrix} \sqrt{n}(\widehat{\xi}-\xi) \\ \sqrt{n}(\widehat{\sigma}-\sigma) \end{pmatrix} \to_d \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1, & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3, & \frac{1}{2}+\frac{1}{4}\gamma_4 \end{pmatrix}),$$

in terms of kurtosis γ_4 and skewness $\gamma_3 = \mathrm{E}\,\{(Y-\xi)/\sigma\}^3$.

(f) Generate a dataset of size $n = 333$ from the unit exponential, and construct an approximate 90 percent confidence ellipsoid on your screen for (ξ, σ). Check if it contains the true values.

Ex. 3.13 *Inference for the normal density, via large-sample methods.* Consider again i.i.d. data $Y_1, \ldots, Y_n$ from a normal density $f(y, \xi, \sigma)$. In Ex. 3.10 we constructed an estimator $\widehat{f}(y)$ with the property that $\log \widehat{f}(y)$ is exactly unbiased for $\log f(y, \xi, \sigma)$, for any given n. Here we instead work out the large-sample approximations for the direct estimator $f^*(y) = f(y, \widehat{\xi}, \widehat{\sigma})$, having plugged in the usual empirical mean and empirical standard deviation.

(a) We have seen in Ex. 2.45 that $(\sqrt{n}(\widehat{\xi}-\xi), \sqrt{n}(\widehat{\sigma}-\sigma))$ tends to (A, B), independent zero-mean normals with variances σ^2 and $\frac{1}{2}\sigma^2$. For fixed y, wite $x = (y-\xi)/\sigma$. Use the delta method to explain that (i) $\sqrt{n}(\log\widehat{\sigma} - \log\sigma) \to_d (1/\sigma)B$; (ii) $\sqrt{n}((y-\widehat{\xi})/\widehat{\sigma} - (y-\xi)/\sigma) \to_d (1/\sigma)(-A - xB)$; (iii) $\sqrt{n}\{(y-\widehat{\xi})^2/\widehat{\sigma}^2 - x^2\} \to_d (1/\sigma)(-2xA - x^2B)$. Then combine these to show that

$$\sqrt{n}\{\log f^*(y) - f(y, \xi, \sigma)\} \to_d (1/\sigma)(-B + xA + \tfrac{1}{2}x^2B),$$

a zero-mean normal with variance $\tau^2 = x^2 + (\frac{1}{2}x^2 - 1)^2 = \frac{1}{2}(1 + x^4)$.

(b) Show that the log-unbiased estimator $\log \widehat{f}(y)$ is close enough to this $\log f^*(y)$ to make the scaled difference $\sqrt{n}\{\log f^*(y) - \log \widehat{f}(y)\}$ go to zero in probability; explain that the two estimators therefore have the same limit distribution.

(c) Explain how a pointwise 90 percent confidence band can be constructed for the log-density, of the type $\log f^*(y) \pm 1.645\,\widehat{\tau}(y)$, which then may be transformed back to the density scale.

Ex. 3.14 *Variance of the variance estimator.* Let $Y_1, \ldots, Y_n$ be i.i.d., with mean ξ, variance σ^2, and finite kurtosis $\gamma_4 = \mathrm{E}\,(Y_i - \xi)^4/\sigma^4 - 3$.

(a) With $A = \sum_{i=1}^n (Y_i - \bar{Y})^2$, show that $\mathrm{E}\,A = (n-1)\sigma^2$. This says that $\widehat{\sigma}^2 = A/(n-1)$ is unbiased for the variance, regardless of the underlying distribution.

(b) If the Y_i are actually normal, then $A \sim \sigma^2\chi^2_m$, with $m = n-1$. Show that $\mathrm{Var}\,\widehat{\sigma}^2 = 2\sigma^4/(n-1)$.

(c) Outside normality, work out an expression for $\mathrm{Var}\,A$, and show that

$$\mathrm{Var}\,\widehat{\sigma}^2 = \Big(3 + \gamma_4 - \frac{n-3}{n-1}\Big)\frac{\sigma^4}{n} = \frac{2\sigma^4}{n-1} + \frac{\gamma_4\sigma^4}{n}.$$

(xx check carefully. use Ex. 2.47. with normality, $\gamma_4 = 0$; show that it reduces to chi-squared based formula above. may perhaps check with O'Neill (2014). simulate a high number of samples of size $n = 12$ from the t distribution t_m, with say $m = 6$, and 'verify' the formula. xx)

(d) (xx tie this to large-sample results, with $(2 + \gamma_4)\sigma^4$ variances for limits, etc. xx)

Ex. 3.15 *The binomial, the normal approximation, and confidence intervals.* Consider Y_n, a binomial (n, p). To invoke the CLT it is practical to use the $Y_n = X_1 + \cdots + X_n$ representation, in terms of X_i being i.i.d. Bernoullis.

(a) Use the CLT of Ch. 2 to deduce that the normalised variable

$$W_n = \frac{\sum_{i=1}^n (X_i - p)}{\{\mathrm{Var} \sum_{i=1}^n (X_i - p)\}^{1/2}} = \frac{Y_n - np}{\{np(1-p)\}^{1/2}} = \frac{\widehat{p} - p}{\{p(1-p)/n\}^{1/2}}$$

converges to the standard normal $\mathrm{N}(0, 1)$ with increasing n. Discuss briefly the skewness result from Ex. 1.3 above in light of the limiting normality.

(b) With $\widehat{p}_B = (Y + 1)/(n + 2)$, show that the difference between W_n and $W_{n,B} = \sqrt{n}(\widehat{p}_B - p)$ is so small, for large n, that $W_{n,B}$ must have the same normal limit. The confidence intervals we construct below, based on $\widehat{p}$, can therefore alternatively be based on $\widehat{p}_B$.

(c) Verify from the above that $\Pr(-1.96 \le W_n \le 1.96) \to 0.95$ as sample size increases, and use this to construct an interval, based on having observed $Y_n = y$ in a given experiment with known n, which covers the true p with probability approximately 95 percent. binomial confidence

(d) There are actually several constructions of such confidence intervals, with this property. Here we shall point to one more such, since the method is famous and easy to use, and since carefully considering these matters for the simple binomial model paves and points the way to various partly related, partly similar findings and constructions in more complicated situations, covered later in this chapter. Considering the basic estimator $\widehat{p} = Y/n$ again, write $\sigma_n^2 = p(1-p)/n$ for its variance, and $\widehat{\sigma}_n^2 = \widehat{p}(1-\widehat{p})/n$ for its estimated variance. Verify that both asymptotic equivalence

$$W_n = (\widehat{p} - p)/\sigma_n \quad \text{and} \quad W_n' = (\widehat{p} - p)/\widehat{\sigma}_n$$

tend to the standard normal in distribution. Now show that the arguments above, used for W_n' in lieu of W_n, lead to the confidence interval $\widehat{p} \pm 1.96\,\widehat{\sigma}_n$ instead. Exemplify, with $n = 100$, for the three cases $y = 22$, $y = 55$, $y = 77$, where you compute both versions of the 95 percent confidence interval for p.

(e) Suppose certain details related to your applied research project require that you compute the probability p that $L \le 1.33\,R$, where

$$L = \{(G_1/G)(G_2/G)(G_3/G)(G_4/G)\}^{1/4}, \quad R = \{(G_5/G)(G_6/G)(G_7/G)\}^{1/3},$$

in terms of $G_1, \ldots, G_8$ being i.i.d. from the χ^2_{12} distribution (the chi-squared with degrees of freedom equal to 12), and $G = \sum_{i=1}^8 G_i$. Since it's hard to find an exact formula, or an exact answer in other ways, you *simulate* a high number sim of such vectors $(G_1, \ldots, G_8)$, and check for each simulation whether the event just described takes place or not. How large should sim be, in order for your simulation based estimate of p to be correct to three decimal places? Carry out such simulations and thus find p. Display also a histogram of simulated L/R.

Quantiles and sample quantiles

Ex. 3.16 *The sample median.* Let $Y_1, \ldots, Y_n$ be i.i.d. from a positive density f with true median $\theta = F^{-1}(\frac{1}{2})$.

(a) Suppose for simplicity that n is odd, say $n = 2m + 1$. Show that M_n has density of the form

$$g_n(y) = \frac{(2m+1)!}{m!\, m!} F(y)^m \{1 - F(y)\}^m f(y).$$

(b) Show then that the density of $Z_n = \sqrt{n}(M_n - \theta)$ can be written in the form $h_n(z) = g_n(\theta + z/\sqrt{n})/\sqrt{n}$. Prove that

$$h_n(z) \to (2\pi)^{-1/2} 2 f(\theta) \exp\{-\tfrac{1}{2} 4 f(\theta)^2 z^2\},$$

where the Stirling approximation formula of Ex. 2.39 may be of use. The limit is the density $h(z)$ of the normal $\mathrm{N}(0, \tau^2)$, with $\tau = \frac{1}{2}/f(\theta)$. We have hence proved $Z_n \to_d \mathrm{N}(0, \tau^2)$, by Scheffé's lemma; see Ex. 2.6.

(c) So when is the sample mean best, and when might the sample median be the better estimator, when it comes to estimating the centre point θ of a symmetric density? This is a matter of the ratio

$$\rho = \frac{\sigma}{\frac{1}{2}/f(\theta)} = 2\sigma f(\theta),$$

where σ is the standard deviation for f. Explain that if $\rho < 1$, then the sample mean is best, and that if $\rho > 1$, then the sample median is the best.

(d) Compare the limiting distributions for the sample mean and the sample median for the normal density, the double exponential density $\frac{1}{2}\exp(-|y|)$, and the Cauchy density $(1/\pi)/(1 + y^2)$.

(e) Consider t distribution, with degrees of freedom ν, see Ex. 1.46. find an expression for the ratio $\rho = \rho(\nu)$, plot $(\nu, \rho(\nu))$ in a diagram, and comment. Show that $\rho(\nu)$ approaches $(2/\pi)^{1/2} = 0.7979$ for large ν. Show that for $\nu < 4.678$, there is roughness at the top, and the median is best; whereas for $\nu > 4.678$, there is a smoother density at the top, and the mean is best. (xx See also Ex. 4.12. xx)

(f) Carry out a similar analysis for the binormal symmetric mixture model $f = \frac{1}{2}\mathrm{N}(-a, 1) + \frac{1}{2}\mathrm{N}(a, 1)$. For which values of a is the sample median a better estimator of the centre

point then the sample mean? [xx later on, another chapter: the estimator which says $\widehat{\theta}$ is sample median if A_n and sample mean if A_n^c, where A_n is the event that $\frac{1}{2}/\widehat{f}(\widehat{\theta}_0) < \widehat{\sigma}$. xx]

Ex. 3.17 *Uniform ordering.* Consider $U_1, \ldots, U_n$ i.i.d. from the uniform distribution. Order these, to $U_{(1)} < \cdots < U_{(n)}$.

(a) Show that $U_{(i)}$ has density

connection to Beta distributions

$$g_i(u) = \frac{n!}{(i-1)!\,1!\,(n-i)!} u^{i-1}(1-u)^{n-i} \quad \text{for } u \in (0,1).$$

Explain that $U_{(i)} \sim \text{Beta}(i, n-i+1)$, see Ex. 1.21, show that $\text{E}\,U_{(i)} = p_i = i/(n+1)$, and that $\text{Var}\,U_{(i)} = p_i(1-p_i)/(n+2)$.

(b) With $i < j$, show that $(U_{(i)}, U_{(j)})$ has density

$$g_{i,j}(u,v) = \frac{n!}{(i-1)!\,1!\,(j-i-1)!\,1!\,(n-j)!} u^{i-1}(v-u)^{j-i-1}(1-v)^{n-j}$$

for $u < v$. The idea behind the reasoning, and the ensuing notation, is that in order to see $U_{(i)} \in [u, u+\text{d}u]$ and $U_{(j)} \in [v, v+\text{d}v]$, there is a multinomial situation, with five boxes $[0,u], [u,u+\text{d}u], [u+\text{d}u, v], [v, v+\text{d}v], [v+\text{d}v, 1]$, inside which we need to find $i-1, 1, j-i-1, 1, n-j$ datapoints.

(c) For an i.i.d. uniform sample $U_1, \ldots, U_n$ on $[0,1]$, consider the uniform range $R_n = U_{(n)} - U_{(1)}$, where we know that $U_{(1)} \sim \text{Be}(1,n)$. Show that given $U_{(1)} = u$, $U_{(n)}$ can be represented as $u + Z$, where Z is the maximum of another uniform sample, of size $n-1$, on $[0, 1-u]$. Use this to show that the c.d.f. of R_n can be expressed as $H_n(r) = nr^{n-1}(1-r) + r^n = nr^{n-1} - (n-1)r^n$, and show that this is the $\text{Be}(n-1, 2)$ distribution. (xx pointer to exercises in Ch6, Ch7, or perhaps just to Story ii.5, depending on how Abel story is written out. xx)

(d) In general, if $Y_1, \ldots, Y_n$ are i.i.d. from a density f, show that the joint density for the full order statistic vector $(Y_{(1)}, \ldots, Y_{(n)})$ is $n!\, f(y_{(1)}) \cdots f(y_{(n)})$, on the set where $y_{(1)} < \cdots < y_{(n)}$. In particular, for order statistics from the uniform distribution, show that the joint density of $(U_{(1)}, \ldots, U_{(n)})$ is flat and equal to $n!$ on the set $u_{(1)} < \cdots < u_{(n)}$.

(e) Use this, in conjunction with Ex. 1.19, to demonstrate that

connection to Dirichlet

$$\begin{aligned}(U_{(1)}, U_{(2)}, \ldots, U_{(n)}) &=_d (D_1, D_1+D_2, \ldots, D_1+\cdots+D_n) \\ &=_d (V_1/S, (V_1+V_2)/S, \ldots, (V_1+\cdots+V_n)/S),\end{aligned}$$

with $V_1, \ldots, V_n, V_{n+1}$ being i.i.d. from the unit exponential, with sum $S = V_1 + \cdots + V_{n+1}$, and $(D_1, \ldots, D_n, D_{n+1})$ is a flat Dirichlet $(1, \ldots, 1, 1)$. The differences $D_i = U_{(i)} - U_{(i-1)}$ are called the *spacings*. We use '$=_d$' to signal equality in distributions. Show that this leads to the representation of the order statistics process as

$$U_{([nq])} = \sum_{i=1}^{[nq]} V_i \Big/ \sum_{i=1}^{n+1} V_i \quad \text{for } 0 \le q \le 1. \tag{3.5}$$

Here $[nq]$ is the largest integer less than or equal to nq. (xx check things and where they appear. Use the law of large numbers to show from this that $U_{([nq])} \to_{\rm pr} q$. point to things in Ch9 with full process convergence $\sqrt{n}(U_{[nq]} - q) \to_d W^0(q)$, the Brownian bridge. xx)

sample quantiles

Ex. 3.18 *Sample quantiles.* Suppose $Y_1, \ldots, Y_n$ are independent observations coming from the same distribution, with positive density f and c.d.f. F. The sample median estimates the population median $F^{-1}(0.50)$, and similarly the sample quantile $Q_n(q)$, at any prescribed level $q \in (0,1)$, estimates the population quantiles $F^{-1}(q)$. Built-in functions like `quantile(data,0.33)` in `R` find such sample quantiles directly, so users do not need the cumbersome linear interpolation fiddling between the two ordered observations coming closest to nq, or to care too much about ties in the data due to rounding-off errors. This exercise finds limit distributions for $\sqrt{n}\{Q_n(q) - F^{-1}(q)\}$, where the previous exercise corresponds to $q = 0.50$.

(a) Suppose $U \sim$ unif and let $Y = F^{-1}(U)$. Show that Y has distribution F, and hence density f.

(b) Explain that the full order statistic vector $Y_{(1)} < \cdots < Y_{(n)}$ may be represented via a correspondingly ordered sample of the uniform, as $F^{-1}(U_{(1)}) < \cdots < F^{-1}(U_{(n)})$, with $U_{(i)}$ being the ith ordered observations in an i.i.d. sample $U_1, \ldots, U_n$ from the uniform, studied in Ex. 3.17. In particular, $Y_{(i)}$ has the same distribution as $F^{-1}(U_{(i)})$.

(c) This also means that if we work out basic approximation results for the order statistics from the uniform, we are a modest delta method step away from similar results for the general case of a density f. In particular, suppose we manage to show $\sqrt{n}(U_{([nq])} - q) \to Z_q$, for some Z_q. Show that we then will have $\sqrt{n}\{Q_n(q) - F^{-1}(q)\} \to_d (F^{-1})'(q) Z_q$.

(d) To illustrate this point in a simple case first, show from what we already know in Ex. 3.16 that $\sqrt{n}(U_{([0.50\,n])} - 0.50) \to_d {\rm N}(0, 0.50^2)$, for the uniform median. Then show for a general density f that $\sqrt{n}\{Q_n(0.50) - \mu\} \to_d {\rm N}(0, 0.50^2/f(\mu)^2)$, with $\mu = F^{-1}(0.50)$ the population median. Here we used an exact expression for the density of the median. There are actually several other ways of proving this median of uniforms result. Such an alternative approach is to use the representation (3.5). Explain that $U_{([0.50\,n])} = A_n/(A_n + B_n)$, with A_n and B_n the averages of the first and second half of i.i.d. variables $V_1, \ldots, V_{n+1}$ from the unit exponential. Use the CLT for the joint limit distributions of $\sqrt{n}(A_n - \frac{1}{2})$ and $\sqrt{n}(B_n - \frac{1}{2})$, and then use the delta method to land the ${\rm N}(0, 0.50^2)$ limit.

(e) Then generalise to the case of any given quantile level q. Show first that $\sqrt{n}(U_{[nq]} - q) \to_d {\rm N}(0, q(1-q))$, and then that the limit distribution is ${\rm N}(0, q(1-q)/f(\mu_q)^2)$ for $\sqrt{n}\{Q_n(q) - \mu_q\}$, with $\mu_q = F^{-1}(q)$.

(f) For the case of two quantiles jointly, like the lower and upper sample quartiles, show for the uniform case that with $q_1 < q_2$,

$$\begin{pmatrix} \sqrt{n}\{Q_n(q_1) - q_1\} \\ \sqrt{n}\{Q_n(q_2) - q_2\} \end{pmatrix} \to_d {\rm N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} q_1(1-q_1), & q_1(1-q_2) \\ q_1(1-q_2), & q_2(1-q_2) \end{pmatrix} \Big).$$

Prove this via the explicit density for $(U_{(i)}, U_{(j)})$, given in Ex. 3.17.

(g) It is also instructive to use representation (3.5) via i.i.d. unit exponentials. Do this. Then generalise to the case of r quantiles, for levels $q_1 < \cdots < q_r$. Show for the uniform case that there is a joint multivariate normal limit, with variances $q_j(1-q_j)$ and covariances $-q_jq\ell$ for $j < l$. Then carry out the transformation arguments needed to prove that for the case of an underlying positive density f, there is limiting joint normality for $\sqrt{n}(Q_{n,j} - \mu_j\}$, where the limit has variances $q_j(1-q_j)/f(\mu_j)^2$ and covariances $q_j(1-q_\ell)/\{f(\mu_j)f(\mu_\ell)\}$ for $j < \ell$, where $\mu_j = F^{-1}(q_j)$. See Ex. 9.30 for convergence of a full quantile process.

(h) (xx something here, or in new separate not long exercise: checking $c(q)\{q(1-q)\}^{1/2} = \{q(1-q)\}^{1/2}/f(F^{-1}(q))$ for a few densities, which tells us the sizes of confidence intervals for quantiles, and more. link to q-q plots briefly discussed in Ch9. xx)

Ex. 3.19 *Min and max of two uniforms.* Suppose Y_1, Y_2 are i.i.d. from a density $f(y)$, and order them, to $V_1 < V_2$. (xx ask per august and martin why this particular probability calculation was of value. xx)

(a) Show that (V_1, V_2) has joint density $2f(v_1)f(v_2)$, on the set where $v_1 < v_2$.

(b) Then consider the special case of two datapoints from the uniform distribution on the unit interval, ordered to $V_1 < V_2$. Show that $R = V_1/V_2$ is another uniform on the unit interval, and that $W = Y_2 - Y_1$ is a Beta$(1, 2)$. Show that $\Pr(Y_2 - Y_1 \le c) = \Pr(Y_2 - Y_1 > c) = \frac{1}{2}$, for $c = 1 - 1/\sqrt{2} = 0.2929$.

(c) Find also the joint distribution for (R, W) here.

Ex. 3.20 *Good and bad estimators.* Suppose $X_1, \ldots, X_n$ are i.i.d. from the density $f(x, \theta) = \exp\{-(x-\theta)\}$ for $y \ge \theta$, i.e. a unit exponential starting at parameter θ.

(a) Explain that we have $X_i = \theta + Y_i$, with the Y_i being i.i.d. from the unit exponential, and hene that the order statistics can be represented as $X_{(i)} = \theta + Y_{(i)}$, cf. Ex. 1.13.

(b) For the smallest and largest observations, show that $\widehat{\theta}_A = X_{(1)} - 1/n$ and $\widehat{\theta}_B = X_{(n)} - s_n$ are unbiased estimators of θ, with $s_n = 1 + 1/2 + \cdots + 1/n$ the partial sum of the harmonic series. Find their variances.

(c) (xx a bit more. spell out that $\widehat{\theta}_B$ is not consistent. a bit on $X_{(i)} - c_i$ too, where $c_i = 1/n + \cdots + 1/(n-i+1) = s_n - s_{n-i}$. median is ok. xx)

Ex. 3.21 *Ratios of ordered uniforms.* (xx again, need checing and calibration, regarding what is told where. xx) Let $U_1, \ldots, U_n$ be an i.i.d. sample from the uniform distribution on the unit interval, and order these into $U_{(1)} < \cdots < U_{(n)}$. From these form the ratios

$$V_1 = U_{(1)}/U_{(2)},\ V_2 = U_{(2)}/U_{(3)},\ \ldots, V_{n-1} = U_{(n-1)}/U_{(n)},\ V_n = U_{(n)}/1.$$

(a) Show that the inverse transformation leads to the representation

$$U_{(n)} = V_n,\ U_{(n-1)} = V_nV_{n-1},\ \ldots, U_{(2)} = V_nV_{n-1}\cdots V_2,\ U_{(1)} = V_nV_{n-1}\cdots V_2V_1.$$

(b) Find the joint probability density for $(V_1, \ldots, V_n)$, and show in fact that these are independent, with

$$V_1 \sim \text{Beta}(1,1),\ V_2 \sim \text{Beta}(2,1),\ \ldots, V_{n-1} \sim \text{Beta}(n-1,1),\ V_n \sim \text{Beta}(n,1).$$

(c) Independently of the details above, find the density of $U_{(i)}$, and show that it is a $\text{Beta}(i, n-i+1)$. In particular, we have

$$\text{E}\, U_{(i)} = \frac{i}{n+1} \quad \text{and} \quad \text{Var}\, U_{(i)} = \frac{1}{n+2}\frac{i}{n+1}\Big(1 - \frac{i}{n+1}\Big).$$

The previous point then tells us that this $\text{Beta}(i, n-i+1)$ can be represented as a product of different independent Beta variables.

(d) It is of course a somewhat cumbersome simulation recipe for generating a uniform sample, but it is a useful exercise, opening doors & minds to fruitful generalisations: For $n = 10$, say, generate ordered uniform samples of size n in your computer via the representation above, in terms of products of Beta variables. Carry out some checks to see that each single $U_{(i)}$ then has the right distribution, i.e. as described in (c).

(e) Work with the following generalisation of the construction above: Let $X_1, \ldots, X_n$ be an i.i.d. sample from the distribution with density $f(x) = ax^{a-1}$, i.e. a $\text{Beta}(a,1)$. Again form the ratios $V_i = X_{(i)}/X_{(i+1)}$ as above, leading to $X_{(i)} = V_i V_{i+1} \cdots V_n$. Show that the V_i are again independent, now with $V_i \sim \text{Beta}(ai, 1)$.

(f) (xx just a bit more. indicate how this may be used to build more general models, possibly in BNP. xx)

Ex. 3.22 *Exercises with sample quantiles.* [xx various things, using the general results above. interquartile range $R_n = Q_n(0.75) - Q_n(0.25)$, e.g. for the normal. for the Cauchy, cool factoid: $\sqrt{n}(R_n - 2) \to_d \text{N}(0, \pi^2)$. Limit distribution of sample median, given the two 0.25 and 0.75 quartiles. a little link to the nonparametric quantile processes of Hjort and Petrone (2007) and the more general quantile pyramids of Hjort and Walker (2009). also pointer to fuller process result in Ch. 9. the limit is $(F^{-1})'(q)W^0(q)$. xx]

(a) (xx perhaps pushed to BNP chapter. xx) Consider building a model for $V_1 < V_2 < V_3$ as follows: (i) $V_2 \sim \text{unif}(0,1)$; (ii) given $V_2 = v_2$, let independently $V_1 \sim \text{unif}(0, v_2)$ and $V_3 \sim \text{unif}(v_2, 1)$. Show that (V_1, V_2, V_3) has mean $(1/4, 2/4, 3/4)$.

(b) Show that the densities of V_1 and V_3 become $g_1(v_1) = -\log v_1$ and $g_3(v_3) = -\log(1 - v_3)$ on the unit interval.

(c) Generalise to the case of $V_1 < \cdots < V_7$, thought of as random versions of the seven octiles: (i) $V_4 \sim \text{unif}(0,1)$; (ii) given V_4, let $V_2 \sim \text{unif}(0, V_4)$ and $V_6 \sim (V_4, 1)$, independently; (iii) given V_2, V_4, V_6, let V_1, V_3, V_5, V_7 be independent and uniform on the intervals $(0, V_2), (V_2, V_4), (V_4, V_6), (V_6, 1)$, respectively. Show that V_j has mean $j/8$, and find the densities for each individual V_j.

Ex. 3.23 *Which order statistics interval contains the true median?* (xx nilsrant, as of 12-August-2024, to be properly cleaned and with motivation. xx) Let $Y_1, \ldots, Y_n$ be i.i.d. from a positive and smooth density f, with cumulative F. With $Y_{(1)} < \cdots < Y_{(n)}$ the order statistics, which of the subintervals $(Y_{(i)}, Y_{(i+1)})$ will contain the true median, $\mu = F^{-1}(\frac{1}{2})$?

(a) Show that $p_i = \Pr\{\mu \in (Y_{(i)}, Y_{(i+1)})\} = \Pr\{\frac{1}{2} \in (U_{(i)}, U_{(i+1)})\}$, in terms of the order statistics from a uniform sample. We allow $i = 0, 1, \ldots, n$, here, for the $n+1$ possibilities for which interval shall contain μ, writing $u_{(0)} = 0$ and $u_{(n+1)} = 1$.

(b) Given $U_{(i)} = u$, show that the distribution of $U_{(i+1)}$ is the same as that of $u+(1-u)W$, where W is the smallest of $n-i$ observations from the uniforom in the unit interval. Use this to show that

$$\begin{aligned} p_i &= \Pr\{U_{(i)} < \tfrac{1}{2} < U_{(i+1)}\} = \int_0^{1/2} \Pr\{U_{(i+1)} > \tfrac{1}{2} \,|\, U_{(i)} = u\} g_i(u)\,\mathrm{d}u \\ &= \int_0^{1/2} \Big(\frac{\frac{1}{2}}{1-u}\Big)^{n-i} \mathrm{be}(u, i, n-i+1)\,\mathrm{d}u, \end{aligned}$$

involving a Beta density, as per Ex. 3.18. Show that this indeed leads to the explicit probability

$$p_i = (\tfrac{1}{2})^{n-i} \frac{n!}{(i-1)!\,(n-i)!} \int_0^{1/2} u^{i-1}\,\mathrm{d}u = \binom{n}{i} (\tfrac{1}{2})^n.$$

Hence we've reached the binomial probabilities, for a binom$(n, \frac{1}{2})$, via direct probability calculations. Try also to give a direct argument.

(c) (xx generalise to general quantile $\mu_p = F^{-1}(p)$. xx)

Moment matching methods

Ex. 3.24 *Moment matching estimators.* Suppose $Y_1, \ldots, Y_n$ are i.i.d. from some model $f(y, \theta)$, where $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{t}}$ is of dimension p. The *method of moments* consists in fitting the first p empirical moments to the theoretical ones. In detail, one computes

method of moments

$$M_1 = \bar{Y} = n^{-1} \sum_{i=1}^n Y_i,\ M_2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2,\ \ldots,\ M_p = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^p,$$

and solves the p equations $M_1 = g_1(\theta), \ldots, M_p = g_p(\theta)$, where $g_1(\theta) = \mathrm{E}_\theta\, Y$, $g_2(\theta) = \mathrm{E}_\theta\, \{Y - g_1(\theta)\}^2$, up to $g_p(\theta) = \mathrm{E}_\theta\, \{Y - g_1(\theta)\}^p$.

(a) For one-parameter models, explain that this amounts to fitting the empirical and theoretical mean. If $Y_1, \ldots, Y_n$ are i.i.d. geom(p), see Ex. 1.24, use $\mathrm{E}\, Y_i = 1/p$ to find the method of moments estimator for p. For another application, assume $Y_1, \ldots, Y_n$ follow the distribution with c.d.f. y^θ on $[0, 1]$. Find the method of moments estimator for θ.

(b) For two-parameter models, explain that the method of moments means fitting the empirical mean and variance to the theoretical ones. If $Y_1, \ldots, Y_n$ are i.i.d. Beta(a, b), see Ex. 1.18, find expressions for the method of moments estimators for a, b.

(c) Define now $h_j(\theta) = \mathrm{E}_\theta\, Y^j$ and $N_j = n^{-1}\sum_{i=1}^n Y_i^j$, for $j = 1, \ldots, p$. The 'method of direct moments' is to solve the equations $h_j(\theta) = N_j$ for $j = 1, \ldots, p$. For $p = 3$, set ut the two systems of three equations with three unknowns, i.e. $g_j(\theta) = M_j$ and $h_j(\theta) = N_j$ for $j = 1, 2, 3$, and show that the solutions $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3)$ are the same. Show in fact that the two methods, fitting centralised or direct methods, are identical, in the general case, for $p \geq 2$.

(d) For a given application one may choose one's method based on convenience and practicality. Sometimes formulae for the direct moments are more easily found that those for the centralised moments. The case for using centralised moments is partly numerical safety; the M_j numbers may be much smaller than the N_j. Exemplify this by seeting up the two equations, with two unknowns, both for the centralised moments and for the direct moments, for the case of the Beta(a, b) distribution.

(e) Generate $n = 100$ data points via the equation $y_i = [\exp\{a(\xi + \sigma N_i)\} - 1]/a$, for say $(a, \xi, \sigma) = (0.33, 0.55, 0.77)$, with the N_i being standard normal. This is a skewed extension of the usual normal model, which corresponds to $a \to 0$ here. Find formulae for the first three moments for this distribution. From your data, use the method of moments to estimate the three parameters.

Ex. 3.25 *Moment fitting estimators for the Gamma distribution.* We now apply the moment matching principle of Ex. 3.24 to the Gamma model with parameters (a, b), with density proportional to $y^{a-1}\exp(-by)$ for y positive; see Ex. 1.9, where we also give the mean, variance, skewness, kurtosis.

(a) With $\bar{Y}$ and V_n the usual sample mean and sample variance, find explicit formulae for the moment estimators $\widehat{a}, \widehat{b}$; show that $\widehat{a} = \bar{Y}^2/V_n$ and $\widehat{b} = \widehat{Y}/V_n$.

(b) Use skewness and kurtosis formulae, in combination with Ex. **??**, to show that

$$\begin{pmatrix} \sqrt{n}(\bar{Y} - a/b) \\ \sqrt{n}(V_n - a/b^2) \end{pmatrix} \to_d \begin{pmatrix} U \\ W \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} a/b^2, & 2a/b^3 \\ 2a/b^3, & (2 + 6/a)a^2/b^4 \end{pmatrix}).$$

Transform this, via the delta method, using Ex. **??**(d), to find the limit distribution for the moment estimators. (xx nils drafting for solution, all algebra needs checking: first $g_1(x, y) = x^2/y$, derivatives $2x/y, -x^2/y^2$ computed at the means become $(2a/b)/(a/b^2) = 2b$ and $-(a/b)^2/(a/b^2)^2 = -b^2$. so $\sqrt{n}(\widehat{a} - a) \to_d 2bU - b^2W$. then $g_2(x, y) = x/y$, derivatives $1/y, -x/y^2$ computed at the means become b^2/a and $-(a/b)/(a/b^2)^2 = -b^3/a$. so $\sqrt{n}(\widehat{b} - b) \to_d (b^2/a)U - (b^3/a)W$. xx)

(c) (xx application in Story ii.1. delta method for $g(\widehat{a}, \widehat{b})$, like the median. xx)

(d) xx

Ex. 3.26 *Moment method estimators for the exponential family.* Consider an exponential family type model, studied in Ex. 1.50, with density of the form $f(y, \theta) = \exp\{\theta^{\mathrm{t}} T(y) - k(\theta)\}h(y)$. Here $T(y) = (T_1(y), \ldots, T_p(y))^{\mathrm{t}}$ is a collection of data functions. As we saw in the exercise pointed to, many classical models are special cases. Now suppose $Y_1, \ldots, Y_n$ are i.i.d. from such a model.

(a) Spell out the basic moment matching method, fitting empirical and model based moments for $Y_i, Y_i^2, \ldots, Y_i^p$.

(b) The moment matching idea is however flexible enough to allow us to choose other data functions than those associated with $Y_i, \ldots, Y_i^p$. Show that matching the moments of $T_1(Y_i), \ldots, T_p(Y_i)$ lead to solving the equations

$$\bar{T}_j = \xi_j(\theta) = \partial k(\theta)/\partial\theta_j \quad \text{for } j = 1, \ldots, p,$$

with $\bar{T}_j = n^{-1}\sum_{i=1}^n T_j(Y_i)$.

(c) Consider the setup of Ex. 3.25, with the Gam(a, b) distribution. There we studied moment matching estimators based on sample mean $\bar{Y}$ and sample variance V_n. Show that the principle above, using the exponential family structure, leads to fitting $\bar{Y} = a/b$ and $n^{-1}\sum_{i=1}^n \log Y_i = \psi(a) - \log b$.

Ex. 3.27 *Mean and variance for background distribution for random sums.* Suppose there is a hidden background machine drawing first the number of dice N and then reporting the sum $Z = X_1 + \cdots + X_N$ of outcomes from having thrown those N dice (without reporting N).

(a) For the individual random X_i, show that the mean and variance are $\xi = 3.50$ and $\sigma^2 = 35/12 = 2.9167$. Use results from Ex. 1.37 to put of formulae for the mean and variance of Z.

(b) If $Z_1, \ldots, Z_m$ are observed, from such a two-layer-random machine, with mean 33.33 and standard deviation 12.12, estimate the mean and standard deviation for N.

Quantile matching methods

Ex. 3.28 *Quantile fitting estimators.* [xx will be used for GoT story. and for CoW story. fitting parameters by solving quantile matching equations. xx] An alternative to the method of moments, described in Ex. 3.24, we may fit empirical and theoretical quantiles (actually in several ways). If $Y_1, \ldots, Y_n$ are i.i.d. from a density $f(y, \theta)$, for a parameter vector of length p, with quantiles $Q(r, \theta) = F^{-1}(r, \theta)$, we choose quantile levels $r_1 < \cdots < r_p$, and solve the p equations $Q_n(r_j) = Q(r_j, \theta)$ with respect to the p unknown parameters, where $Q_n(r) = F_n^{-1}(r)$ is the empirical quantile.

method of quantiles

(a) Suppose the distribution to be fitted has c.d.f. $F(y) = y^\theta$ on the unit interval. Find the estimator corresponding to fitting the empirical to the theoretical median. Starting with the limit distribution for the median, see Ex. 3.16, find the limit distribution for $\sqrt{n}(\widehat{\theta} - \theta)$. More generally, find the estimator $\widehat{\theta}_r$ corresponding to fitting the r level quantile, and then the limit distribution for $\sqrt{n}(\widehat{\theta}_r - \theta)$.

(b) Consider $Y_1, \ldots, Y_n$ from the location Cauchy density $f_0(y - \theta)$, with $f_0(x) = (1/\pi)/(1+x^2)$ the standard Cauchy. Its c.d.f. is $F_0(x) = \frac{1}{2} + (1/\pi)\arctan x$, see Ex. 1.16. Show that the r level quantile is $\mu + F_0^{-1}(r)$, and that this leads to the estimator $\widehat{\mu}_r = Q_n(r) - F_0^{-1}(r)$. Find the limit distribution for $\sqrt{n}(\widehat{\mu}_r - \mu)$. What quantile level r leads to the sharpest estimator?

(c) (xx the normal, with median and interquartile range. more generally $Q_n(1-r) - Q_n(r)$, and find the best r. xx)

(d) (xx the Weibull, with two equations. xx)

(e) (xx point to more general versions, minimising $A_n(\theta) = \sum w_n(r_j)\{Q_n(r_j) - F^{-1}(r_j,\theta)\}^2$. xx)

Ex. 3.29 *Moment fitting and quantile fitting for the Weibull.* As a general illustration of moment and quantile fitting estimation methods, consider the Weibull distribution with c.d.f. $F(t) = 1 - \exp\{-(t/a)^b\}$ for $t \geq 0$, see Ex. 1.54.

(a) Take e.g. $(a,b) = (3.33, 1.44)$, and simulate $n = 100$ realisations. (i) Compute average and standard deviation for these, and compute estimates $(\widehat{a}_m, \widehat{b}_m)$. (ii) From 0.25 and 0.75 quantiles, fit the two relevant equations to compute $(\widehat{a}_q, \widehat{b}_q)$. Take the trouble to display three curves, the correct underlying cumulative hazard function $A(t) = (t/a)^b$ along with the two estimated versions.

(b) Repeat the experiment many times, to see how close the two $(\widehat{a}, \widehat{b})$ is to (a,b). Also, as an instance of a focused question, how close the two median estimates $\widehat{m} = \widehat{a}(\log 2)^{1/\widehat{b}}$ is closest to the real median? Which of the two estimating schemes is best? We should point here to the likelihood methodologies of Ch. 5; the maximum likelihood method will be the winning strategy, beating both moment and quantile fitting, under model conditions.

Minimum sum of squares and linear regression

Ex. 3.30 *Linear regression and least squares estimation.* Consider observed pairs (x_i, Y_i) for $i = 1, \ldots, n$, where Y_i conditionally on covariate x_i is modelled to have mean $a_0 + bx_i$ and common variance σ^2. This is classical linear regression, widely used in theoretical and applied statistics, most often analysed and used with the additional assumption that the Y_i are normally distributed. Importantly, the model is extended to the case of multiple covariates below, see Ex. 3.31, 3.32, with more to come regarding statistical testing (xx in Ex. 4.36 and one more xx). To see some of these classical methods in action, check out Stories iii.1 and iv.1.

(a) It is helpful to reparametrise the regression line from $a_0 + bx_i$ to $a + b(x_i - \bar{x})$. Show that minimising the sum of squares $Q(a,b) = \sum_{i=1}^n \{y_i - a - b(x_i - \bar{x})\}^2$ leads to

$$\widehat{a} = (1/n)\sum_{i=1}^n y_i = \bar{y}, \quad \widehat{b} = \sum_{i=1}^n (x_i - \bar{x})y_i/M_n, \quad \text{with } M_n = \sum_{i=1}^n (x_i - \bar{x})^2.$$

(b) Show that $\widehat{a}$ and $\widehat{b}$ are unbiased, with zero covariance, and variances σ^2/n and σ^2/M_n.

(c) Let $Q_0 = \min_{a,b} Q(a,b) = \sum_{i=1}^n \{y_i - \widehat{a} - \widehat{b}(x_i - \bar{x})\}^2$ be the minimum sum of squares. Show that

$$Q(a,b) = \sum_{i=1}^n \{y_i - a - b(x_i - x)\}^2 = Q_0 + n(\widehat{a} - a)^2 + M_n(\widehat{b} - b)^2.$$

Use this to show that $\widehat{\sigma}^2 = Q_0/(n-2)$ is an unbiased estimator of σ^2.

(d) We have found natural and unbiased estimators for a, b, σ^2, without yet making assumptions of the underlying distributions for Y_i, beyond means and variance. Assume now, however, that the distributions are normal, so that $Y_i \sim \mathrm{N}(a + b(x_i - \bar{x}), \sigma^2)$. Show that (i) $\widehat{a} \sim \mathrm{N}(a, \sigma^2/n)$; (ii) $\widehat{b} \sim \mathrm{N}(b, \sigma^2/M_n)$; (iii) $\widehat{\sigma}^2 \sim \sigma^2 \chi^2_m/m$, with $m = n - 2$; and (iv) that these three statistics are mutually independent. For this last point, it may be helpful to follow the line of proof for independence of sample mean and sample variance for normal data, used in Ex. 1.45, now with an orthonormal matrix A with first row $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$ and second row $((x_1 - \bar{x})/M_n^{1/2}, \ldots, (x_n - \bar{x})/M_n^{1/2})$.

(e) Construct confidence intervals for b, for σ, and for $\mathrm{E}\,(Y \,|\, x_0) = a + b(x_0 - \bar{x})$, the mean value at a given position x_0.

Ex. 3.31 *Linear multiple regression and least squares.* The celebrated linear multiple regression model remains a cornerstone success story of theoretical and applied statistics. It is a bag of tools for investigating the extent to which covariates x influence the outcomes of certain interest variables Y. The standard formulation of the model is as follows. The data collected can be organised into (x_i, Y_i), for individuals or objects $i = 1, \ldots, n$, where $x_i = (x_{i,1}, \ldots, x_{i,p})^{\mathrm{t}}$ is of dimension p and y_i of dimension one. The model then postulates that

$$Y_i = x_i^{\mathrm{t}}\beta + \varepsilon_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_i are i.i.d. from the normal $\mathrm{N}(0, \sigma^2)$. Thus there are $p+1$ parameters at work here, the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{t}}$ and the error distributiuon standard deviation σ. – Note that the very classical case of $y_i = a + bx_i + \varepsilon_i$, associated with a scatterplot of (x_i, Y_i), is a special case; see Ex. 3.30.

(a) With Y the vector of Y_i, ε the vector of ε_i, and X the $n \times p$ matrix having $(x_{i,1}, \ldots, x_{i,p})$ as its row i, show that

$$Y = X\beta + \varepsilon \sim \mathrm{N}_n(X\beta, \sigma^2 I),$$

with I the $n \times n$ identity matrix. This is a practical and compact linear algebra version of the model formulation. We do assume that X is of full rank p, so that the symmetric matrix $X^{\mathrm{t}}X$ is invertible. This amounts to there being at least p linearly independent covariate vectors in the X matrix; in particular, we must have $n \geq p$ to identify the β_j coefficients directly from data. [xx but quick pointers to later chapters with Bayes and to regularisation and to lasso and ridge here. xx]

(b) The least squares estimator $\widehat{\beta}$ is the minimiser of $Q(\beta) = \|Y - X\beta\|^2 = \sum_{i=1}^n (Y_i - x_i^{\mathrm{t}}\beta)^2$. Show that $\sum_{i=1}^n (Y_i - x_i^{\mathrm{t}}\widehat{\beta})x_i = 0$. With Σ_n the $p \times p$ matrix $n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}} = n^{-1}X^{\mathrm{t}}X$, show that

$$\widehat{\beta} = (X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}Y = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i Y_i.$$

Prove also that $Q(\beta) = Q_0 + n(\widehat{\beta} - \beta)^{\mathrm{t}}\Sigma_n(\widehat{\beta} - \beta)$, with $Q_0 = \min_{\text{all } \beta} Q(\beta) = \sum_{i=1}^n \widehat{\varepsilon}_i^2$, writing $\widehat{\varepsilon}_i = Y_i - x_i^{\mathrm{t}}\widehat{\beta}$ for the estimated residuals.

(c) Show that $\widehat{\beta}$ is unbiased and that its variance matrix can be written $\sigma^2(X^{\rm t}X)^{-1} = (\sigma^2/n)\Sigma_n^{-1}$.

(d) Show also that $\widehat{\beta}$ has a multinormal distribution, so that in fact $\widehat{\beta} \sim {\rm N}_p(\beta, (\sigma^2/n)\Sigma_n^{-1})$. This is the key result about the least squares estimators. We also need precise information for estimating σ; see Ex. 3.32.

Ex. 3.32 *The residuals and their variances.* The setup is an in the previous Ex. 3.31, the $Y \sim {\rm N}_n(X\beta, \sigma^2 I)$ linear regression model. Above we focused on the least squared method and the ensuing properties for the estimators of the regression coefficients, and found $\widehat{\beta} \sim {\rm N}_p(\beta, (\sigma^2/n)\Sigma_n^{-1})$. We also need to deal carefully with estimators of σ, the residual standard deviation, also since we encounter statistics of the type $(\widehat{\beta}_j - \beta_j)/\widehat{\sigma}$.

(a) From the basic $Y = X\beta + \varepsilon$ we may define the estimated residuals as

$$\widehat{\varepsilon} = Y - X\widehat{\beta} = (I - H)Y, \quad \text{where} \quad H = X(X^{\rm t}X)^{-1}X^{\rm t},$$

the so-called hat matrix, of size $n \times n$. Show that H is symmetric and idempotent, which means that $H^{\rm t} = H$ and $H^2 = H$. This also implies $(I - H)H = 0$.

(b) Now consider the random minimum achieved by the $Q(\beta)$ which was used in the least squares operation,

$$Q_0 = \min\{Q(\beta)\colon \text{all } \beta\} = Q(\widehat{\beta}) = \|Y - X\widehat{\beta}\|^2 = \sum_{i=1}^n (Y_i - x_i^{\rm t}\widehat{\beta})^2.$$

The main result, arrived at below, is that $Q_0/\sigma^2 \sim \chi^2_m$, with degrees of freedom $m = n-p$, and that Q_0 is independent of $\widehat{\beta}$. Show first that

$$\begin{pmatrix} X\widehat{\beta} \\ \widehat{\varepsilon} \end{pmatrix} = \begin{pmatrix} HY \\ (I-H)Y \end{pmatrix} \sim {\rm N}_{2n}(0, \sigma^2 \begin{pmatrix} H, & 0 \\ 0, & I-H \end{pmatrix}).$$

In particular, these two random vectors are independent; also, $Q_0 = \|\widehat{\varepsilon}\|^2 = Y^{\rm t}(I-H)Y$ is consequently independent of $X\widehat{\beta}$.

(c) Show that $(I-H)X = 0$, which implies $\widehat{\varepsilon} = (I-H)Y = (I-H)(Y - X\beta) = (I-H)\varepsilon$ and hence $Q_0 = \varepsilon^{\rm t}(I-H)\varepsilon$. We also reach the simple idenity

$$\|\varepsilon\|^2 = \varepsilon^{\rm t}H\varepsilon + \varepsilon^{\rm t}(I-H)\varepsilon,$$

where the left-hand side is a $\sigma^2\chi^2_n$ and the two terms on the right-hand side being independent. Show that the first term on the right-hand side is a $\sigma^2\chi^2_p$. Via independence and a moment-generating function argument show then that $Q_0 \sim \sigma^2\chi^2_{n-p}$. [xx pointer to Ex. A.33. might rearrange the sequence of exercises to have mgf before this. xx]

(d) (xx a few things regarding estimating σ. standard version is $\widehat{\sigma}^2 = Q_0/(n-p) \sim \sigma^2\chi^2_m/m$. make clear that things we've learned for the simple i.i.d. normal setup can be used here too, without further ado. xx)

(e) (xx can put in estimation of $\gamma = c^{\rm t}\beta$ things here, or in separate exercise. t distributions, intervals, tests. and how to predict y_0 for a new x_0. xx)

Ex. 3.33 *Confidence intervals for key parameters in linear regression models.* Consider the general linear regression setup of Ex. 3.31–3.32, where $Y_i = x_i^{\rm t}\beta + \varepsilon_i$ for $i = 1, \ldots, n$, and the ε_i being i.i.d. $\mathrm{N}(0, \sigma^2)$. The compact version of this is $Y \sim \mathrm{N}_n(X\beta, \sigma^2 I)$.

(a) Construct a 90 percent confidence interval for σ.

(b) From the general results obtained above, show that $\widehat{\beta}_j \sim \mathrm{N}(\beta_j, \sigma^2 r_j/n)$, with r_j the diagonal (j,j) element of Σ_n^{-1}. Show from this that $t_j = \sqrt{n}(\widehat{\beta}_j - \beta_j)/(r_j^{1/2}\widehat{\sigma})$ has a t_m distribution, with degrees of freedom $m = n - p$, and construct a confidence interval for β_j from this. If a 99 percent confidence interval for β_j is outside zero, how can this be interpreted and used?

(c) More generally, consider $\gamma = c^{\rm t}\beta = c_1\beta_1 + \cdots + c_p\beta_p$, a linear combnination of the regression coefficients. Find the distribution of $\widehat{\gamma} = c^{\rm t}\widehat{\beta}$, and construct a confidence interval. This may be used to assess the size of $\beta_1 - \beta_2$, of $\beta_1 - \frac{1}{2}(\beta_2 + \beta_3)$, and similar constrasts. (xx could go on to max over all contrasts, Scheffé things. xx)

(d) Show that $n(\widehat{\beta} - \beta)^{\rm t}\Sigma_n(\widehat{\beta} - \beta) \sim \sigma^2\chi^2_p$. Use this to construct a confidence region, actually a confidence ellipsoid, for the full β vector.

Ex. 3.34 *Predicting the next y.* Suppose linear regression analysis has been carried out, for a given dataset $(x_1, y_1), \ldots, (x_n, y_n)$. How can we predict what happens with Y_0, associated with another covariate vector x_0? This could be for an individual outside the dataset, or in a time context, speculating about the next datapoint in a sequence. An illustration of methods given below is in Story iv.1.

(a) Assume the regression data are of the form $y_i \sim \mathrm{N}(x_i^{\rm t}\beta, \sigma^2)$ for $i = 1, \ldots, n$, as with Ex. 3.31, and consider a new x_0, for which the not yet unobserved Y_0 is independent of the other data, and with distribution $\mathrm{N}(x_0^{\rm t}\beta, \sigma^2)$. Show that $\widehat{y}_0 = x_0^{\rm t}\widehat{\beta} \sim \mathrm{N}(x_0^{\rm t}\beta, x_0^{\rm t}\Sigma_n^{-1}x_0\,\sigma^2/n)$, and use this to form a confidence interval for the mean of Y_0, as opposed to for Y_0 itself.

(b) Then show that

$$Y_0 - \widehat{y}_0 \sim \mathrm{N}(0, \sigma^2(1 + n^{-1}x_0^{\rm t}\Sigma_n^{-1}x_0/n))$$

Construct a prediction confidence interval for Y_0 based on this. Comment on the situation with a large n, and on the difference between confidence for Y_0 and its mean.

(c) Then consider the classic linear regression case with only one variable studied in Ex. 3.30, with $Y_i \sim \mathrm{N}(a + b(x_i - \bar{x}), \sigma^2)$. For predicting a not yet observed $Y_0 \sim \mathrm{N}(a + b(x_0 - \bar{x}), \sigma^2)$, demonstrate that

$$Y_0 - \{\widehat{a} + \widehat{b}(x_0 - \bar{x})\} \sim \mathrm{N}(0, \sigma^2\{1 + 1/n + (x_0 - \bar{x})^2/M_n\}),$$

with $M_n = \sum_{i=1}^n (x_i - \bar{x})^2$. Show that this leads to the 95 percent confidence band

$$Y(x_0) \in \widehat{a} + \widehat{b}(x_0 - \bar{x}) \pm c\widehat{\sigma}\{1 + 1/n + (x_0 - \bar{x})^2/M_n\}^{1/2},$$

for an appropriate range of x_0 values, where $c = t_{0.975,n-2}$ the t upper quantile. Illustrate this with a simulated dataset, to see both where the band is tight and where it becomes very broad. (xx two more sentences. easiest to predict for individuals not far from the centre of the covariate distributions. xx)

Ex. 3.35 *Linear regression outside normality.* The aim here is to show and appreciate that the classical coefficient estimators in linear regression setups are still approximately normal, even when the error terms distribution is not normal. That this is so is essentially thanks to basic large-sample theory, as partly summarised in Ex. 3.11, and specifically to the Lindeberg type theorems of Ch. 2; see in particular Ex. 2.37. (xx pointer to other regression large-sample results in Ch. 5. xx)

(a) We first deal with a simple setup with a single regression coefficient. Suppose $y_i = x_i\beta + \varepsilon_i$ for $i = 1, \ldots, n$, with covariates x_i and error terms ε_i being i.i.d. from a zero-mean distribution with finite variance σ^2. Show that the estimator minimising $Q_n(\beta) = \sum_{i=1}^n (y_i - x_i\beta)^2$ is $\widehat{\beta} = \sum_{i=1}^n x_i y_i / M_n$, where $M_n = \sum_{i=1}^n x_i^2$. Show further that $\widehat{\beta}$ is unbiased with variance σ^2/M_n.

(b) Then consider $Z_n = M_n^{1/2}(\widehat{\beta} - \beta)$. Show that it has zero mean and variance σ^2, and that it can be written $\sum_{i=1}^n (x_i/M_n^{1/2})\varepsilon_i$.

(c) Deduce that $\widehat{\beta}$ is approximately normal, even if the ε_i are not normal, provided merely that $D_n = \max_{i\le n} |x_i|/M_n^{1/2} \to 0$. If in particular $(1/n)\sum_{i=1}^n x_i^2$ stays bounded, then the natural condition is $(1/\sqrt{n})\max_{i\le n}|x_i| \to 0$.

(d) Then consider the general linear regression model $Y_i = x_i^{\mathrm{t}}\beta + \varepsilon_i$ of Ex. 3.31, with the x_i being p-dimensional covariate vectors and β a p-dimensional vector of regression coefficients. We take the ε_i to be i.i.d. with mean zero and finite variance σ^2, but do not stipulate normality. The least squares estimator is $\widehat{\beta} = \Sigma_n^{-1}(1/n)\sum_{i=1}^n x_i Y_i$ with $\Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}}$. It is unbiased with variance matrix $(\sigma^2/n)\Sigma_n^{-1}$, assumed to have full rank. Assume $\Sigma_n \to \Sigma$, a full rank matrix. Show that $Z_n = \sqrt{n}(\widehat{\beta} - \beta) \to_d \mathrm{N}_p(0, \Sigma)$, under the condition $R_n = (1/\sqrt{n})\max_{i\le n}\|x_i\| \to 0$.

(e) Assume now that the x_i are drawn i.i.d. from a distribution over the covariate space, with finite variance matrix. Show that $R_n \to_{\mathrm{pr}} 0$.

(f) Argue that the fine-tuned finite-sample confidence methods, developed in Ex. 3.31–3.32 under exact normality, continue to hold in the large-sample sense, even if the distributions are not normal. Give conditions for such results to hold.

(g) Suppose the same type of phenomenon is studied in both Denmark and Sweden, with the same meaning for Y and covariates x_1 and x_2. This leads to regression estimators $\widehat{\beta}_D$ and $\widehat{\beta}_S$ for the two analyses. Construct confidence intervals for $d_j = \beta_{D,j} - \beta_{S,j}$, for $j = 1, 2$.

Ex. 3.36 *Least squares estimation in other setups.* (xx to come. point is that minimising $Q(\theta) = \sum_{i=1}^n \{y_i - \xi_i(\theta)\}^2$ is a general principle, also outside linear regression. xx)

(a) (xx some easy cases. binomial. normal mean. poisson. xx)

(b) Assume $y_1, \dots, y_n$ are i.i.d. and gamma distributed (a, b), with known a. Find the least squares estimator $\widehat{b}$. Then find the mean and variance of this estimator.

(c) (xx regression, with mean $a + bx_i$, then mean $a \exp(bx_i)$. xx)

Notes and pointers

(xx to come. we point to various matters, genesis of crucial concepts, and also point to chapters ahead. explain that yes, we've touched and used CLT and delta method and a bit more here, but with details and more material to come in Ch 4. xx)

[xx CLT for binomials: associated with the famous names de Moivre (who showed a version of this in 1733) and Laplace (who had a clearer and more general proof in 1812). xx]

Briefly genesis of неравенцво Маркова, the Markov, and the неравенцво Чебышёва, the Chebyshov (often anglisised to Chebyshev, but his name was really Чебышёв). mention Kahneman et al. (2020).

I.4

Testing, sufficiency, power

In previous chapters we have learned about classes of distributions, their parameters, ways of estimating these from data, along with assessment of precision and confidence intervals. The present chapter goes on to the fundamental statistical reporting tool of hypothesis testing. Statistical testing of hypotheses are data-based rules for when to reject (and hence, when not to reject) a hypothesis about the parameters of a model. Theory is developed to construct such tests in quite general setups. A test is constructed to have a certain significance level, like 0.05, the intended low probability of rejecting the hypothesis if it is in fact true. It also has has a power function, the probability of rejection as a function of how far the model parameters might be from the hypothesis. We learn the basics of the Neyman–Pearson theory for optimal testing, and see it panned out for many situations, including the general setting of exponential families. The fruitful concept of sufficiency, the notion that a lower-dimensional vector of summaries contains all statistical information about a model, is also developed in this chapter, with implications for both estimation and testing. (xx more to come in Chapters 5, 7, 8. xx)

Key words: ancillarity, completeness, conditional tests, exponential family, factorisation theorem, Neyman–Pearson optimality, power, p-value, sufficiency, testing, Wald ratios

In the broad context of analysing data from models, the previous chapter dealt generally speaking with estimators of relevant parameters, their precision, and comparisons, leading in particular to confidence intervals. The present chapter develops the methods and applications further, in the direction of *statistical testing of null hypotheses*. Studying such tests involves their interpretation, construction, properties, performance, along with connections to confidence and yet other themes.

Consider in general terms data y stemming from a model with a parameter vector $\theta = (\theta_1, \ldots, \theta_p)$, and suppose one wishes to test the null hypothesis H_0 that θ is inside a well-defined subset Θ_0 of the full parameter region Θ. Precisely what constitutes a null hypothesis is a matter of scientific and statistical context, often reflecting the intentions and the overall aims of the data collection and its analysis. The null hypothesis is typically a statement, concerning the nature of the mechanisms studied, the incorrect rejection

of which one attempts to avoid. A simple illustration is testing whether a particular regression coefficient is equal to zero (so H_0 could be '$\beta_3 = 0$'), or testing whether two parameters, perhaps for two groups, are equal (so H_0 would be '$\theta_A = \theta_B$').

A *test* for such a hypothesis is a rule saying 'H_0 is to be rejected if data y fall in the set R', along with the complementary statement 'H_0 is not rejected if data y fall in the set R^c'. We talk here of *the rejection region* R and *the acceptance region* R^c. A fundamental aspect of such a test to look for is its *significance level*, typically meant to be a relatively small probability, like 0.05 or 0.01 or even smaller. We say that the test has level α provided

the level of a test

$$\Pr_\theta(\text{reject } H_0) \le \alpha \quad \text{for all } \theta \in \Theta_0. \tag{4.1}$$

With level 0.01 one is guaranteed such a low chance of falsely claiming that H_0 is wrong, if it is indeed correct. So for the illustrations briefly alluded to above, when statisticians after careful analyses reject '$\beta_3 = 0$' or '$\theta_A = \theta_B$', therefore going on to claim that their alternatives are valid, with '$\beta_3 \neq 0$' and '$\theta_A \neq \theta_B$', these claims are seen as trustworthy (and may make into publications), since the probability of these claims being false is so low.

Often tests are carried out via appropriate *test statistics*, as when constructing a $T = T(y)$, a function of the data y, with the property that T ought to be inside some normal and well-understood range under H_0 conditions (we shall meet many cases of determining the *null distribution* of such test statistics), but bigger, if H_0 is wrong. In such cases, the rejection region takes the form $R = \{y\colon T(y) \ge t_0\}$, with t_0 the rejection threshold, chosen to have $\Pr_\theta(T(Y) \ge t_0) \le \alpha$ for all θ congruent with H_0.

test statistic, rejection threshold

Conceptually and operationally, there is a certain direction implied when setting up an hypothesis H_0 and its alternative. Rejecting H_0, with a test with low level testing level α, leads to a positive claim about the alternative being true; the observed data have landed in a region which if H_0 were true has low probability. On the other hand, *not rejecting* the null hypothesis should not be seen as 'verifying' H_0; one needs to be content with the not so bold statement 'the observed data do not provide sufficient evidence for claiming that H_0 does not hold'.

We are also keenly interested in *the power* of a test, which is the detection chance $\Pr_\theta(\text{reject } H_0)$ as a function of θ, in the alternative parameter domain $\Theta - \Theta_0$. Thus some tests are *stronger* than other tests with the same level, and we learn recipes for constructing such in the exercises below. The power of a test clearly depends on the quality of data, and typically of the sample size, as we shall see in exercises. Thus detecing that $\theta_A \neq \theta_B$, for parameters of two groups, in a setup where there really is a difference, becomes more likely with more data.

Below we also define, discuss, and use *p-values*, which are commonly quoted in most branches of applied statistics work, typically to indicate how clear a potential finding is. The idea is to quantify how unlikely it is, to observe what is actually observed, if some relevant null hypothesis H_0 is actually true. If the test is set up to reject the null if an appropriate test statistic T is sufficiently large, we're after $p = \Pr_{H_0}(T \ge t_{\text{obs}})$, with t_{obs} the observed T for the given dataset. Some care is needed since that probability might depend on parameters under H_0. The more careful version is

the p-value

$$p = \max\{\Pr_\theta(T \geq t_{\rm obs})\colon \theta \in \Theta_0\}. \tag{4.2}$$

A small p-value, like $p \leq 0.01$, casts serious doubt on H_0, since the observed $t_{\rm obs}$ is so unlikely. A rephrasing of the testing scheme, with significance level say 0.01, is to reject H_0 if $p \leq 0.01$.

A classic result for testing theory is the Neyman–Pearson Lemma, which in an idealised setup with just two possible densities identifies the most powerful method for testing one density against the other; see Ex. 4.7–4.9. This sharpens further questions for more general setups, and we manage to find optimal tests in a variety of setups, including for the broad exponential family class. This also necessitates exploring and developing certain themes of independent interest and use, those of sufficiency, ancillarity, completeness, conditional testing; see the string of exercises starting with Ex. 4.16. A core idea in that terrain is that of data compression; for various models, a low-dimensional vector of data summaries contains all relevant statistical information.

(xx then one paragraph with pointers to other chapters and perhaps to a few of the stories. xx)

Testing, testing

Ex. 4.1 *Testing a null hypothesis.* Here we introduce the notion of null hypotheses and their testing in a few simple setups.

(a) The probability $p = \Pr(A)$ of a certain event is meant to be $p_0 = 0.33$, if the machinery around it works as it should. To test this one carries out the relevant experiment $n = 100$ times, and the event takes place $y = 44$ times. Should you reject the 0.33 hypothesis? Show that with $Y \sim \text{binom}(n, p)$, the statistic $T = (Y - np_0)^2/\{np_0(1-p_0)\}$ is approximately a χ^2_1, under the null assumption that $p = p_0$. Show also that W tends to be bigger than a χ^2_1, if $p \neq p_0$. Show that when using T as test statistic, the p-value of (4.2) becomes $p = \Pr_{p_0}(T \geq t_{\rm obs})$, with $t_{\rm obs}$ the value of T seen with $y_{\rm obs} = 44$. Compute this p-value, and decide whether the $p = 0.33$ hypothesis should be rejected at the 0.05 level.

(b) Suppose (X, Y, Z) is trinomial with sum n and probabilities (p, q, r); see Ex. 1.5. For concreteness, suppose a theory holds that $(p, q, r) = (0.20, 0.30, 0.50)$, and that $(X, Y, Z) = (42, 47, 111)$ is observed. Is this enough to reject the theory in question? Show that

$$W = (X - np)^2/(np) + (Y - nq)^2/(nq) + (Z - nr)^2/(nr)$$

has mean equal to 2, and that it is approximately a χ^2_2, using the multidimensional CLT for (X, Y, Z). Explain how this may be used to test the theory mentioned, and carry out the test. This is actually the classic Pearson chi-squared test for multinomials; see Story vii.1 for details, generalisations, and discussion.

(c) You're rolling your die, but it takes you as many as $y_{\rm obs} = 15$ rolls to get your first '6'. Does this make you suspect that the probability p of a '6' is not 1/6, but lower? With testing level $\alpha = 0.05$, what is the set of suspicious outcomes of Y, the number of throws to get the first '6'?

Ex. 4.2 *Connections from confidence intervals to testing.* Though confidence intervals and testing are two different reporting tools, when summarising inference, there are clear connections. Suppose ϕ is a parameter of inference, perhaps a function of model parameters, for which we can build both confidence intervals and tests.

(a) Suppose one needs to test the one-point null hypothesis that $\phi = \phi_0$, a given value, and that $[L, U]$ is a 99 percent confidence interval. Show that the test consisting in rejecting, if ϕ_0 is outside this interval, has level 0.01.

(b) Suppose on the other hand that there is a well-defined 0.01 level test procedure for testing $\phi = \phi_0$, against $\phi \neq \phi_0$, for each candidate value ϕ_0. Gather together in a set A all the ϕ_0 values which are not rejected by the corresponding 0.01 level test. Show that $\Pr_\phi(\phi \in A) = 0.99$, making A a 99 percent confidence region.

(c) Go through the relevant details, from confidence interval to test and vice versa, for the simple prototype case of the observation being $Y \sim \mathrm{N}(\theta, 1)$.

Ex. 4.3 *Confidence intervals for quantiles.* Let $Y_1, \ldots, Y_n$ be i.i.d. from a continuous density, positive on its sample space. How can we construct confidence intervals for the median $\mu = F^{-1}(\frac{1}{2})$, and more generally for quantiles $\mu_q = F^{-1}(q)$? There are several approaches here, but here we give the basic method via the empirical c.d.f.

(a) Let F_n be the empirical c.d.f. for the data, see Ex. 3.9. If μ_0 is the true median, show that $F_n(\mu_0)$ is a simple B_n/n, with $B_n \sim \mathrm{binom}(n, \frac{1}{2})$, and that this implies $W_n(\mu_0) = \sqrt{n}\{F_n(\mu_0) - \frac{1}{2}\}/\frac{1}{2}$ being approximately a standard normal. Argue that a natural 0.05 level test for $\mu = \mu_0$ is to accept the hypothesis provided $|W_n(\mu_0)| \leq 1.96$.

(b) Following the general testing-to-confidence connection of Ex. 4.2, show that the associated 95 percent confidence interval becomes $\mathrm{ci}_n = \{\mu\colon |F_n(\mu) - \frac{1}{2}| \leq 1.96/(2\sqrt{n})\}$. In other words, we may read off the interval from a plot of F_n, without knowing or taking on board the details of the exact or approximate distribution of sample quantiles, as with Ex. 3.18.

(c) Generalise to the case of any quantile $\mu_q = F^{-1}(q)$. Show that the recipe above leads to the confidence interval $\mathrm{ci}_n = \{\mu_q\colon |F_n(\mu_q) - q| \leq z_\alpha\{q(1-q)\}^{1/2}/\sqrt{n}\}$, where $\Pr(|\mathrm{N}(0,1)| \leq z_\alpha) = \alpha$, the confidence leve. For an illustration of these methods, check Story i.6, where we plot the empirical c.d.f. and read off confidence intervals for quantiles $F^{-1}(q)$ at levels 0.10, 0.50, 0.90, for the weight of mothers pre pregnancy.

(d) Discuss ways in which more accurate confidence intervals can be constructed, using the exact binomial distribution; for the median, for example, $nF_n(\mu) \sim \mathrm{binom}(n, \frac{1}{2})$. This leads to slight modification of the horizontal bands when reading off intervals from the empirical c.d.f.

Ex. 4.4 *t testing, one and two samples.* Testing the mean based on a sample of normal observations is a recurring problem, in several guises, and with the famous t test being the canonical procedure; details are given below. We also go through the basics for testing the difference of means for two normals samples. Due to the connections to

confidence intervals discussed in Ex. 4.2 we also find accurate confidence intervals for the key parameters. Beyond their concrete relevance and repeated use in these standard setups, the t testing procedures are important since similar constructions can be worked with in large classes of more complicated setups, but then typically with approximations to key distributions, rather than the exact solutions found under these classic strict modelling assumptions.

(a) Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathrm{N}(\xi, \sigma^2)$, so far assuming σ to be known, and that one wishes to test $H_0 \colon \xi = \xi_0$ against the alternative that $\xi \neq \xi_0$, where ξ_0 is some appropriate given value, like zero. Using the exact distribution of $\bar{X}$, cf. Ex. 1.2, show that $Z = (\bar{X} - \xi_0)/(\sigma/\sqrt{n}) = \sqrt{n}(\bar{X} - \xi_0)/\sigma$ is standard normal under H_0, and that $|Z|$ will tend to be bigger than a normal if H_0 is not true. Explain that the test which rejects H_0 when $|Z| > 1.96$ has level 0.05.

(b) For the more realistic case of σ not being known, the natural construction is the t statistic $t = \sqrt{n}(\bar{X} - \xi_0)/\widehat{\sigma}$, with $\widehat{\sigma}$ the empirical standard deviation. Show from Ex. 1.46 that $t \sim t_{n-1}$ under H_0, and write down a precise 0.05 level test.

(c) Suppose now that the mean ξ for the population of X_i is to be compared with the mean η for another population, where we have i.i.d. data $Y_1, \ldots, Y_m \sim \mathrm{N}(\eta, \sigma^2)$. So the task is to test H_0 that $\xi = \eta$. Show first that $\bar{Y} - \bar{X} \sim \mathrm{N}(\eta - \xi, \sigma^2(1/m + 1/n))$. To build a t statistic we need an estimator for the denominator. Writing $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$ for the empirical deviances for samples 1 and 2, show that with

$$\widehat{\sigma}^2 = c_1\widehat{\sigma}_1^2 + c_2\widehat{\sigma}_2^2, \quad \text{using } c_1 = (n-1)/(n+m-2),\ c_2 = (m-1)/(n+m-2),$$

we have $\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n+m-2}/(n+m-2)$, independent of $\bar{X} - \bar{Y}$. Conclude that $t = (\bar{X} - \bar{Y})/R \sim t_{n+m-2}$ under H_0, where $R = \widehat{\sigma}(1/m + 1/n)^{1/2}$.

(d) Use these building blocks to also construct a 90 percent confidence interval for $d = \xi - \eta$.

Ex. 4.5 *Wald tests.* Here we present the basics of so-called Wald tests, a general and practical way of forming tests via approximate normality. Such tests are used very routinely when looking for and reporting findings about regression coeffiecients in all standard regression models; see Story i.6 for an illustration of this. The Wald tests can be constructed almost immediately, from the confidence interval construction of Ex. 3.11, via the interval-to-test connection of Ex. 4.2, but we tend to a few details to allow for potential simplifications of assumptions. – Assume $Y_1, \ldots, Y_n$ comprise the data (not necessarily i.i.d.), from a suitable model with vector parameter θ. Suppose further that $\phi = \phi(\theta)$ is a focus parameter, for which we need to test $\phi = \phi_0$, for a given null value ϕ_0.

(a) Suppose there is an estimator $\widehat{\phi}$ with the property that $\sqrt{n}(\widehat{\phi} - \phi) \to_d \mathrm{N}(0, \tau^2)$, and that there is a consistent estimator $\widehat{\tau}$ for this limit spread τ. Show as with Ex. 3.11 that $W_n = \sqrt{n}(\widehat{\phi} - \phi)/\widehat{\tau} \to_d \mathrm{N}(0,1)$. Show that the arguments go through, with a limiting standard normal, for $W_{n,0} = \sqrt{n}(\widehat{\phi} - \phi_0)/\widehat{\tau}_0$, at the null hypothesis, as long as $\widehat{\tau}_0$ is

consistent for τ at this null position; also, technically speaking, we only need to establish $\sqrt{n}(\widehat{\phi}-\phi_0) \to_d \mathrm{N}(0,\tau^2)$ at the null hypothesis. Conclude that the test which rejects $\phi=\phi_0$ when $|W_{n,0}| \geq 1.96$ has level approaching 0.05 (and of course similarly with other chosen testing levels; level 0.01 corresponds to threshold 2.576, etc.).

(b) Explain that $p = \Pr(|N(0,1)| \geq W_{n,0,\mathrm{obs}})$ is an approximation to the exact p-value.

(c) Suppose $X \sim \mathrm{binom}(100,p)$, with a need for testing $p=0.33$. Set up a Wald test, and compute the p-value, if $x_{\mathrm{obs}}=44$. Suppose then that there is an additional $Y \sim \mathrm{binom}(100,q)$, and that one wishes to test $p=q$. Set up a Wald test, and compute the p-value, if (X,Y) are observed to be $(44,55)$.

Ex. 4.6 *When confidence intervals for two parameters overlap.* Suppose confidence intervals for parameters a and b overlap. It might sound plausible that the null hypothesis $a=b$ will then not be rejected. To check aspects of this and related questions, consider a prototype setup where $\widehat{a} \sim \mathrm{N}(a,1)$ and $\widehat{b} \sim \mathrm{N}(b,1)$ are independent.

(a) With confidence level $1-\alpha$, show that the canonical intervals are $\widehat{a} \pm z_0$ and $\widehat{b} \pm z_0$, with $z_0 = \Phi^{-1}(1-\alpha/2)$, e.g. the standard 1.96 for level 0.95. Show that overlapping intervals corresponds to $|\widehat{d}| \leq 2z_0$, where $\widehat{d} = \widehat{b}-\widehat{a} \sim \mathrm{N}(b-a,2)$

(b) Consider the test of $a=b$ which rejects when the intervals for a and b are disjoint. Show that this test has considerably lower significance level than α; for 95 percent intervals this level is 0.0056.

(c) Yes, it is possible to have overlapping 95 percent intervals and yet reject $a=b$ at level 5 percent. Identify the range of $\widehat{d}$ where this happens.

Neyman–Pearson optimal testing

Ex. 4.7 *The Neyman–Pearson Lemma: the basics.* (xx finetune the intro prose here. xx) Suppose data y come from a density f, where there are just two possibilities: either $f=f_0$, which is the null hypothesis to be tested, or $f=f_1$, the alternative. Here there's an optimal strategy, made clear by the so-called Neyman–Pearson Lemma, part of the 49-page landmark paper Neyman and Pearson (1933). For simplicity of presentation we consider the continuous case, where f_0 and f_1 are densities on the relevant sample space $\mathcal{Y}$ (which can be multidimensional). – A *test function* is a $T\colon \mathcal{Y} \to [0,1]$, with $T(y)$ the probability of rejecting f_0 if the the data take on value y. This setup even allows the possibility of an element of randomisation, as in 'if y turns out be 3.33 I throw some coins and I reject H_0 with probability 0.77'. Once in a blue while this might be of relevance, with discrete data, but in practice such a test function $T(y)$ takes on only values 1, for a rejection set R, and 0, for the complementary acceptance set R^c.

(a) Show that the probability of rejecting the null, if the null is true, can be written $\Pr_{f_0}(\mathrm{reject}) = \int f_0 T \,\mathrm{d}y$.

(b) For a given testing level α, like 0.01, let T^* be the test which rejects when $f_1(y)/f_0(y) \geq c$, with c tuned such that

$$\Pr_{f_0}(T^* \text{ rejects}) = \int_{y:\ f_1(y)/f_0(y)\geq c} f_0(y)\,\mathrm{d}y = \alpha.$$

Let T be any other test function with the same level α. Show that the power difference at f_1 can be written

$$\pi_{T^*}(f_1) - \pi_T(f_1) = \int f_1(T^* - T)\,\mathrm{d}y = \int (f_1 - cf_0)(T^* - T)\,\mathrm{d}y.$$

the Neyman–Pearson Lemma

(c) Show that among all possible tests, with level α, the T^* has the strongest detection power at position f_1.

(d) (xx a bit more, to cover discrete case; and we may also allow competitors with $\int f_0 T < \alpha$. xx)

Ex. 4.8 *The Neyman–Pearson Lemma: more details.* Here we tend to some further details, related to the Neyman–Pearson Lemma and its proof, in Ex. 4.7.

(a) For two positive densities f_0 and f_1 defined on the same sample space, as with the Neyman–Pearson Lemma, consider the event $A_c = \{f_1(Y)/f_0(Y) \geq c\}$. Show that the function $p(c) = \Pr_{f_0}(A_c) = \int_{f_1(y)\geq cf_0(y)} f_0(y)\,\mathrm{d}y$ is a continuous and monotone function, starting at 1 and ending at 0, when c travels through $[0, \infty)$. Hence deduce that there for given α really is a unique c in the Neyman–Pearson recipe.

(b) Illustrate the $p(c) = \Pr_{f_0}(A_c)$ in a few concrete situations, including (i) $f_0 \sim \mathrm{N}(0, 1)$ and $f_1 \sim \mathrm{N}(1, 1)$; (ii) $f_0 \sim \mathrm{Gam}(2.2, 3.3)$ and $f_1 \sim \mathrm{Gam}(3.3, 2.2)$.

(c) (xx something about power at f_1, when testing f_0, is different from power at f_0, when testing f_1. link to other exercise. xx)

Ex. 4.9 *The Neyman–Pearson Lemma: applications.* For the simple two-possibilities setup we learn from the Neyman–Pearson lemma that there is a clear recipe for setting up the optimal test for $f = f_0$ against $f = f_1$. Here are some examples.

(a) Suppose $Y \sim \mathrm{N}(\theta, 1)$. Show that the optimal test of level $\alpha = 0.01$, for testing $\theta = 0$ vs. $\theta = 1.234$, is to reject if $Y \geq z_{0.99} = 2.326$, the upper 0.01 point of the standard normal.

(b) In this situation, verify that one finds the very same optimal 0.01 level test, for any alternative point $\theta_1 > 0$. Hence the $Y \geq z_{0.99}$ test is *uniformly most powerful*, against all positive alternative.

(c) Generalise this to the case of data $Y_1, \ldots, Y_n$ being i.i.d. from the normal $\mathrm{N}(\theta, \sigma^2)$, with known σ. Show that the test which rejects $\theta = 0$ against $\theta > 0$ when $Z_n = \sqrt{n}\bar{Y}/\sigma > z_{1-\alpha}$ is uniformly most powerful, among all tets with level α; here $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. Find its power function, and draw it in a diagram, for $\theta_0 = 1.234$, $\sigma = 1$, and for $n = 10, 20, 30$.

(d) Let $Y_1, \ldots, Y_n$ be i.i.d. from the $\mathrm{N}(\theta_0, \sigma^2)$ distribution, with θ_0 known. Consider the problem of testing $\sigma = \sigma_0$ against $\sigma > \sigma_0$, at level say 0.01, where σ_0 is a prescribed null value. Show that the test which rejects when $V_n = \sum_{i=1}^n (Y_i - \theta_0)^2 \geq \gamma_{n,0.99}$, the 0.99 quantile of the χ^2_n, is uniformly optimal. Find its power function.

(e) Consider f_0, the standard normal, and $f_1(y) = \frac{1}{2}\sqrt{2}\exp(-\sqrt{2}|y|)$; they have both zero mean and unit variance. Find the optimal test for f_0 against f_1, with level 0.05, and find its detection power at f_1. Then do the opposite, constructing the best test at level 0.05 for f_1 against f_0, and find the power at f_0.

(f) (xx with $n = 10$ data points, not merely one. put up the tests, find their powers. comment. make separate exercise to see optimal tests for f_0 against f_1, with n data points, KL approximations. xx)

(g) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the exponential $\theta\exp(-\theta y)$. Find the strongest 0.10 level test for $\theta = \theta_0$ against an alternative $\theta_1 > \theta_0$. Your test will not depend on the θ_1, as long as it is to the right of θ_0; hence this test is uniformly most powerful against these alternatives.

Ex. 4.10 *Density ratios and optimal testing: the normal and the Cauchy.* The Neyman–Pearson recipe is to reject when the density ratio $f_1(y)/f_0(y)$ is sufficiently big. This pans out differently in different situations, as illustrated here.

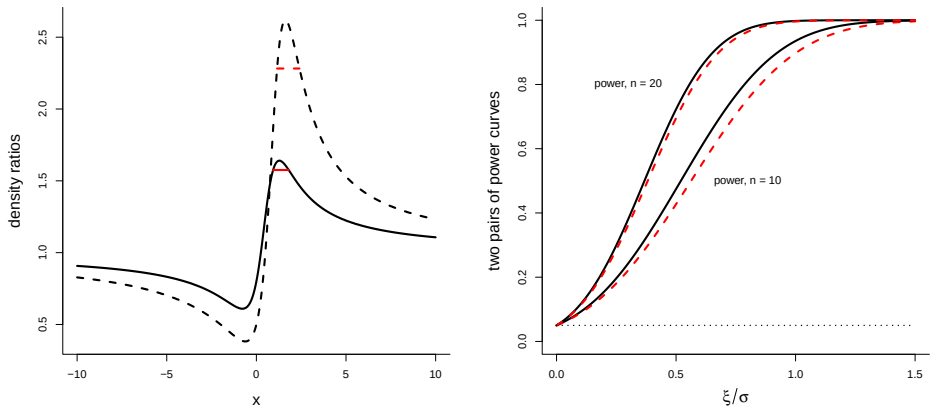


Figure 4.1: *Left panel: for the Cauchy density model, see Ex. 4.10, ratios $f(y,\theta_1)/f(y,0)$ (full line) and $f(y,\theta_2)/f(y,0)$ (slanted) are shown, for alternative values $\theta_1 = 0.50$ and $\theta_2 = 1.00$ to the null. Also indicated are the rejection intervals $[a_1, b_1]$ and $[a_2, b_2]$, for the optimal tests against θ_1 and θ_2. Right panel: two pairs of power. Testing $\xi = 0$ against $\xi > 0$, see Ex. 4.12, for a normal sample of size 10 (lower pair) and 20 (upper pair), at test level 0.05, as a function of ξ/σ. The lower power is for the t test (slanted, red); the upper power if for the normal based test (full, black), which requires that σ is known.*

(a) For a single observation Y, consider testing $f_0 = \mathrm{N}(0,1)$ against $f_1 = \mathrm{N}(\theta_1, 1)$, with θ_1 positive. Show that

$$\frac{f_1(y)}{f_0(y)} = \frac{\exp\{-\frac{1}{2}(y-\theta_1)^2\}}{\exp(-\frac{1}{2}y^2)} = \exp(\theta_1 y - \tfrac{1}{2}\theta_1^2).$$

Verify that this is a monotone function in y, regardless of the value of $\theta_1 > 0$. Argue that 'reject f_0 provided Y is big enough' becomes the uniformly optimal test. Exhibit the rejection threshold in $Y \geq c$, if the significance level is to be 0.10.

(b) The situation is rather different for the case of the Cauchy density $f(y,\theta) = (1/\pi)\{1+(y-\theta)^2\}^{-1}$. Suppose we wish to test $\theta = 0$ versus a positive θ_1. Show that

$$\frac{f_1(y)}{f_0(y)} = \frac{f(y,\theta_1)}{f(y,0)} = \frac{1+y^2}{1+(y-\theta_1)^2},$$

and draw this function in a diagram, for a few values of θ_1.

(c) For a concrete illustration, work through the alternative cases $\theta_1 = 0.50$ and $\theta_2 = 1.00$, for each case finding the rejection interval, say $[a_1, b_1]$ for the first and $[a_2, b_2]$ for the second, to give the optimal test, of level 0.10. (xx answers: $[0.933, 1.804]$ for θ_1; $[1.161, 2.388]$ for θ_2. construct a version of Figure 4.1, left panel. xx) – The point is that these rejection regions are different; the optimal test depends on the specific alternative, and there can be no uniformly optimal test.

(d) Again for the sake of concreteness, compute the optimal possible power, for any 0.10 level test, at $\theta_1 = 0.50$ and at $\theta_2 = 1.00$. Compare these powers with that of the simple test $Y \geq 3.078$.

(e) (xx there are two drastic differences here, between the simple normal and the non-simple Cauchy. the first is that the log-density-ratio $R(y,\theta_0,\theta_1) = \log f(y,\theta_1) - \log f(y,\theta_0)$ is monotone, for the normal, and not at all monotone for the Cauchy. the second is that of there being a simple one-dimensional sufficient statistic, in the case of $Y_1, \ldots, Y_n$ from the normal, whereas no such statistic exists for the Cauchy. where is sufficiency in kiosk? xx)

(f) (xx something about more regularity with n data points; above we just did $n = 1$. xx)

Ex. 4.11 *Optimal average power.* Supppose Y is observed, perhaps a full vector, from a density f, where one wishes to test the null hypothesis $f = f_0$, a given density. As we saw in Ex. 4.10, there are cases where there is no uniformly optimal test, against all or a subset of alternatives; the optimal test at f_1 might be different from the one at f_2. In one-parameter models, this is caused by the log-density-ratio not being monotone.

(a) Consider alternative densities $f_1, \ldots, f_m$, given nonnegative weights of importance $w_1, \ldots, w_m$. These may be taken to have sum 1. The *weighted average power* of a test T, at these points and with these weights, is

$$\bar\pi_T = \sum_{j=1}^m w_j \pi_T(f_j) = \sum_{j=1}^m w_j \int f_j(y) T(y)\,\mathrm{d}y,$$

with $\pi_T(f_j)$ the power at f_j. Let $\bar f(y) = \sum_{j=1}^m w_j f_j$. Show that this average power is maximised, among all test functions $T(y)$ with level α at the null, by the $T^*(y)$ which rejects H_0 when $\bar f(y)/f_0(y) \geq c$, with c tuned to give rejection probability α at the null.

(b) For a one-parameter model $f(y,\theta)$, consider testing of $\theta=\theta_0$ against $\theta>\theta_0$. For any test function $T(y)$ with rejection level α at the null, consider in general terms the weighted average power

$$\bar\pi_T=\int_{\theta>\theta_0}\pi_T(\theta)\,\mathrm{d}w(\theta),$$

with $\pi_T(\theta)=\int f(y,\theta)T(y)\,\mathrm{d}y$ the power of the test at position θ. Show that $\bar\pi_T=\int_{\theta>\theta_0}\bar f(y)T(y)\,\mathrm{d}y$, featuring the density $\bar f(y)=\int_{\theta>\theta_0}f(y,\theta)\,\mathrm{d}w(\theta)$. This is the model density averaged over all alternatives to the null, as weighted by the $\mathrm{d}w(\theta)$ measure.

(c) Show that the test maximising this weighted average power is rejecting the null if $\bar f(y)/f(y,\theta_0)\geq c$, with c tuned to have null level α.

(d) (xx a little more. the marginal density, or predictive density, with a link to Bayes, but specifically with a 'prior' over the alternative space. can take Cauchy with $a\exp(-a\theta)$ over the halfline. xx)

(e) (xx can look at $\mathrm{N}(\xi,\sigma^2)$ model, testing the null that $f=\mathrm{N}(0,1)$, against the alternative that $\xi>0$, or $\sigma>1$, or both. show first that the NP test against alternatives $(1.1,2.2)$ and $(2.2,3.3)$ are indeed different, so there is no uniformly most powerful test. then maximise average power, using a weight density we give our readers, with

$$\bar f(y_1,\ldots,y_n)=\int_{\xi>0,\sigma>1}\Big\{\prod_{i=1}^n f(y_i,\xi,\sigma)\Big\}\,\mathrm{d}w(\xi,\sigma).$$

perhaps a data example. xx)

(f) (xx make sure we have a good version on board of a lemma which says the Wilks test $D_n=2\{\ell_{n,\max}-\ell_n(\theta_0)\}$ is an approximation to this optimal weighted power test. can be in Ch 4, but then pointed to already here. xx)

Ex. 4.12 *The t test and its power.* Suppose $Y_1,\ldots,Y_n$ are i.i.d. from the $\mathrm{N}(\xi,\sigma^2)$, with testing of $\xi=0$ required against $\xi>0$. This is simple and standard, in the case of σ being known, but needs the t test in the case of σ being unknown and estimated from the data, as we have seen in Ex. 4.4, Setting up the test uses the relevant t distribution, from $t=\sqrt{n}\bar Y/\widehat\sigma\sim t_{n-1}$, but for studying the power function also the noncentral t distribution is required.

(a) Consider first the case of σ being known. Show that $z=\sqrt{n}\bar Y/\sigma$ is standard normal under the null, and that the 0.05 level test becomes that of rejecting when $z\geq z_0=\Phi^{-1}(0.95)=1.645$.

(b) Show that at a given $\xi>0$, we have $z\sim\mathrm{N}(\sqrt{n}\xi/\sigma,1)$, and that this leads to the power function $\pi_{n,N}(\xi/\sigma)=\Phi(\sqrt{n}\xi/\sigma-z_0)$. Compute and display this power function for the case of $n=10$ and $n=20$, as with the black full curves of Figure 4.1 (right panel).

(c) Then consider the more complex situation where σ is not known, needing the empirical standard deviation $\widehat{\sigma}$ of (1.5). We have seen in Ex. 1.46 that

$$t = \frac{\sqrt{n}\bar{Y}}{\widehat{\sigma}} = \frac{\sqrt{n}\bar{Y}/\sigma}{\widehat{\sigma}/\sigma} \sim \frac{\mathrm{N}(0,1)}{(\chi^2_{\mathrm{df}}/\mathrm{df})^{1/2}},$$

with nominator and denominator being independent, and where the degrees of freedom is $\mathrm{df} = n-1$. The probability density $g_{\mathrm{df}}(x)$ and cumulative distribution $G_{\mathrm{df}}(x)$ of this t_{df} distribution are moderately complicated, see the exercise mentioned, but that does not concern us much, as long as we can consult a table or run an algorithm to find associated quantiles and probabilities. Show hence that the t test, with level 0.05, must consist of rejecting when $t \geq t_0 = G^{-1}_{\mathrm{df}}(0.95)$. Using `qt(0.95,df)` in `R`, we find 1.833 and 1.729 for $n = 10$ and $n = 20$. Check that `qt(0.95,df)` becomes close to 1.645 as df increases, and explain why.

(d) For the power of the t test, show that $\pi_{n,t}(\xi/\sigma) = \Pr_{\xi/\sigma}(t \geq t_0)$, where

$$t = \frac{\sqrt{n}\bar{Y}}{\widehat{\sigma}} \sim \frac{\mathrm{N}(\sqrt{n}\xi/\sigma, 1)}{(\chi^2_{\mathrm{df}}/\mathrm{df})^{1/2}},$$

again with nominator and denominator independent. Explain that the power function therefore can be written

$$\pi_{n,t}(\xi/\sigma) = \Pr_{\xi/\sigma}(t \geq t_{0.m}) = 1 - G_m(t_{0,m}, \sqrt{n}\xi/\sigma),$$

with $G_m(x,\lambda)$ denoting the cumulative distribution for this noncentral t with degrees of freedom m and noncentrality parameter λ. This function is complicated, but can easily be found numerically via e.g. `pt(x,m,lambda)` in `R`. Construct a version of Figure 4.1, right panel, perhaps with other sample sizes than 10, 20. Comment on your findings.

(e) Describe how the two-sided tests and power functions pan out here, whem $\xi = 0$ is to be tested against $\xi \neq 0$. Make a corresponding version of Figure 4.1, right panel, with the relevant two-sided power functions.

Ex. 4.13 *Establishing that a parameter is close to zero.* Suppose δ is some effect parameter, where the scientific context relates to establishing that it is close to zero. In such situations it might be natural to test $H_0 \colon |\delta| \geq \varepsilon$, for some known small threshold ε, against the alternative that $\delta \in (-\varepsilon, \varepsilon)$. In some applications the δ might be the difference between two effect parameters, aiming to infer that these are close. This is turning the tables, somehow, compared to the more traditional setups, where the hypothesis to be tested is that a parameter is close to zero, against the alternative that it is some distance away.

(a) To study the main features of such a situation, consider a prototype setup with i.i.d. observations $Y_1, \ldots, Y_n$ being $\mathrm{N}(\delta, 1)$. Show that $\sqrt{n}\bar{Y} \sim \mathrm{N}(\sqrt{n}\delta, 1)$, which implies $Z_n = n\bar{Y}^2 \sim \chi^2_1(n\delta^2)$.

(b) Explain that it is natural to reject the null, and hence claim that $|\theta|$ is small, if $|\sqrt{n}\bar{Y}| \leq c$, or equivalently $Z_n \leq c^2$, for c calibrated to reach testing level e.g. $\alpha = 0.05$. This is equivalent to $Z_n \leq c^2$. Show that $\Pr_\delta(|\sqrt{n}\bar{Y}| \leq c) = \Gamma_1(c^2, n\delta^2)$, in terms of the c.d.f. for the noncentral χ^2_1.

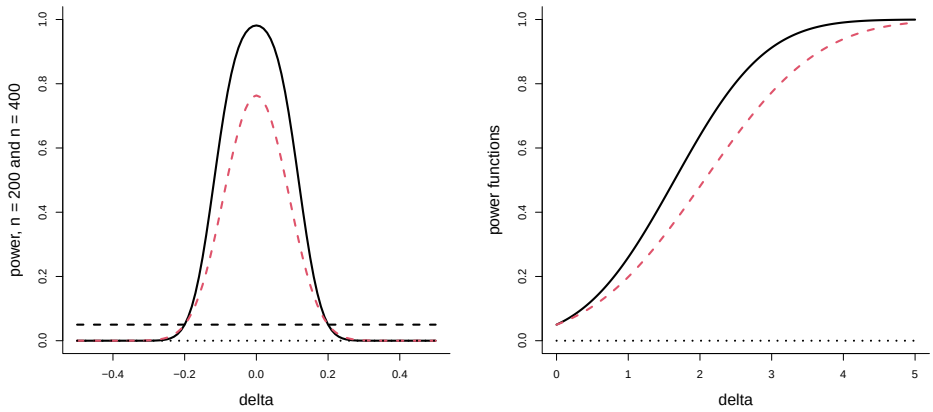


Figure 4.2: *Left panel: power functions for the bioequivalence tests of Ex. 4.13, with* $n = 200$ *(slanted curve) and* $n = 400$ *(full curve), and threshold* $\varepsilon = 0.20$*. Right panel: limiting local power functions for the two tests of Ex. 4.14, for* $\theta \leq \theta_0$ *against* $\theta > \theta_0$*, in terms of the local alternatives parameter* δ*, with* $\theta_n = \theta_0 + \delta/\sqrt{n}$*. The more powerful test is based on* $\bar{Y}$*, the other on the median* M_n*.*

(c) Explain that to have significance level 0.05, we need $\Gamma_1(c^2, n\varepsilon^2) = 0.05$. Find the full power curve $\pi_n(\delta) = \Pr_\delta(Z_n \leq c^2)$, and comment on the level of its maximum. Construct a version of Figure 4.2, left panel, for threshold $\varepsilon = 0.20$, for sample sizes $n = 200$ and $n = 400$.

Ex. 4.14 *Power and local power: a particular case.* This exercise studies a prototype situation in some detail; the type of calculations and results will be seen to be rather similar in a long range of different situations. Let $Y_1, \ldots, Y_n$ be i.i.d. data from $\mathrm{N}(\theta, \sigma^2)$. One wishes to test $H_0 \colon \theta = \theta_0$ vs. the alternative that $\theta > \theta_0$, where θ_0 is a known value (e.g. 3.14). Two tests will be considered, based on respectively $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ and the median M_n.

(a) For a given value of θ, prove that $\sqrt{n}(\bar{Y}_n - \theta) \to_d \mathrm{N}(0, \sigma^2)$ and $\sqrt{n}(M_n - \theta) \to_d \mathrm{N}(0, (\pi/2)\sigma^2)$. Note that the first result is immediate and actually holds with exactness for each n; the second result requires more care and follows from Ex. 3.16.

(b) Working under the null hypothesis $\theta = \theta_0$, show that

$$Z_n = \frac{\sqrt{n}(\bar{Y}_n - \theta_0)}{\widehat{\sigma}} \to_d \mathrm{N}(0,1), \quad Z_n^* = \frac{\sqrt{n}(M_n - \theta_0)}{(\pi/2)^{1/2}\widehat{\sigma}} \to_d \mathrm{N}(0,1),$$

where $\widehat{\sigma}$ is any consistent estimator of σ.

(c) With $z_{0.95} = \Phi^{-1}(0.95) = 1.645$, conclude from the above that the two tests that reject H_0 provided respectively

$$\bar{X}_n > \theta_0 + z_{0.95}\widehat{\sigma}/\sqrt{n} \quad \text{and} \quad M_n > \theta_0 + z_{0.95}(\pi/2)^{1/2}\widehat{\sigma}/\sqrt{n},$$

have the required asymptotic significance level 0.05; $\alpha_n = \Pr\{\text{reject } H_0 \,|\, \theta = \theta_0\} \to 0.05$. (There is one such α_n for the first test, and one for the other; both converge however to 0.05.)

(d) Our object is then to study the *local power*, the chance of rejecting the null hypothesis under alternatives of the type $\theta_n = \theta_0 + \delta/\sqrt{n}$, i.e. close to the null value. In generalisation of (b), show that

$$Z_n = \frac{\sqrt{n}(\bar{Y}_n - \theta_0)}{\widehat{\sigma}} \to_d \mathrm{N}(\delta/\sigma, 1), \quad Z_n^* = \frac{\sqrt{n}(M_n - \theta_0)}{(\pi/2)^{1/2}\widehat{\sigma}} \to_d \mathrm{N}((\pi/2)^{1/2}\delta/\sigma, 1),$$

where the convergence in question takes place under the indicated $\theta_0 + \delta/\sqrt{n}$ parameter values. Here we need to generalise the result of Ex. 3.16, to the case where θ_n moves with n. Writing F_n for the distribution of Y_i, you may do this by first showing that M_n has the same distribution as $F_n^{-1}(M_n^0) = \theta_0 + \delta/\sqrt{n} + \sigma\Phi^{-1}(M_n^0)$, where M_n^0 is the median for an i.i.d. sample from the uniform, and then using the delta method.

(e) Use these results to show that

$$\begin{aligned}\pi_n(\delta) &= \Pr\{\text{reject with } \bar{Y}_n \text{ test} \,|\, \theta_0 + \delta/\sqrt{n}\} \to \Phi(\delta/\sigma - z_{0.95}),\\ \pi_n^*(\delta) &= \Pr\{\text{reject with } M_n \text{ test} \,|\, \theta_0 + \delta/\sqrt{n}\} \to \Phi((2/\pi)^{1/2}\delta/\sigma - z_{0.95}),\end{aligned}$$

for the two power functions. Draw these in a diagram, and compare, as with Figure 4.2, right panel.

(f) Assume one wishes n to be large enough to secure that the power function is at least at level β for a certain alternative point θ_1. Using the local power approximation, show that the required sample sizes are respectively

$$n_A \doteq \frac{\sigma^2}{(\theta_1 - \theta_0)^2}(z_{1-\alpha} + z_\beta)^2 \quad \text{and} \quad n_B \doteq \frac{\sigma^2/c^2}{(\theta_1 - \theta_0)^2}(z_{1-\alpha} + z_\beta)^2$$

for tests A (based on the mean) and B (based on the median), with $c = \sqrt{2/\pi}$. With level $\alpha = 0.05$, compute these sample sizes for the case of $\beta = 0.95$ and $\theta_1 = \theta_0 + \frac{1}{2}\sigma$.

(g) One sometimes defines the ARE, the asymptotic relative efficiency of test B with respect to test A, as

$$\mathrm{ARE} = \lim n_A(\theta_1, \beta)/n_B(\theta_1, \beta),$$

the limit in question in the sense of alternatives θ_1 coming closer to the null hypothesis at speed $1/\sqrt{n}$. Show that the ARE in this particular situation becomes $c^2 = 2/\pi = 0.6366$; test A needs only ca. 64 percent as many data points to reach the same detection power as B needs.

Ex. 4.15 *Power and local power: general results.* In Ex. 4.14 we examined two particular estimators and tests, for the mean parameter in the normal distribution, and found precise limit distributions and local power functions, for alternatives of order $1/\sqrt{n}$ from the null hypothesis. Here we go through the appropriate generalities. The setup is that of

data, of sample size n, informing us about a key parameter θ, where the null hypothesis $H_0\colon \theta = \theta_0$ is to be tested against the alternative $\theta > \theta_0$, with θ_0 some specified value. We assume the data leads to one or more estimators $\widehat{\theta}$, with the basic premise that at the null parameter θ_0, we have $A_n = \sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}(0, \kappa^2)$, for the appropriate κ.

(a) With $\widehat{\kappa}$ an estimator for κ, under H_0 conditions, show that $Z_n = \sqrt{n}(\widehat{\theta} - \theta_0)/\widehat{\kappa}_0 \to_d \mathrm{N}(0,1)$ under H_0. Explain that the test which rejects when $Z_n \geq z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ has asymptotic testing level α, i.e. $\alpha_n = \Pr_0(Z_n \geq z_{1-\alpha}) \to \alpha$.

(b) We now make two more assumptions, concerning natural smoothness in a $O(1/\sqrt{n})$ size local neighbourhood around θ_0: (i) if the true parameter behind the data is $\theta_n = \theta_0 + \delta/\sqrt{n}$, then the same limit $\mathrm{N}(0, \kappa^2)$ obtains for $A'_n = \sqrt{n}\{\widehat{\theta} - (\theta_0 + \delta/\sqrt{n})\}$; (ii) the estimator $\widehat{\kappa}$ still converges to the same κ in probability. Show that $Z_n \to_d \mathrm{N}(\delta/\kappa, 1)$, and that this gives a limiting local power function,

$$\pi_n = \Pr(\text{reject} \,|\, \theta_0 + \delta/\sqrt{n}) \to \pi(\delta) = \Phi(\delta/\kappa - z_{1-\alpha}).$$

Show that this limit power function has derivative $\phi(z_{1-\alpha})/\kappa$ at zero.

(c) The above result is useful in its own right, providing an easy approximation to the power function for alternatives not far from the null hypothesis value. It may also be used for comparing different tests, built as above, for different estimators, say $\widehat{\theta}_j$. Assume the conditions above hold for these, with $\sqrt{n}(\widehat{\theta}_j - \theta_0) \to_d \mathrm{N}(0, \kappa_j^2)$, leading to estimators $\widehat{\kappa}_j$ for κ_j and to test statistics $Z_{n,j} = \sqrt{n}(\widehat{\theta}_j - \theta_0)/\widehat{\kappa}_j$. Explain that the best estimators lead to the best tests, with derivatives $\phi(z_{1-\alpha})/\kappa_j$ at zero.

(d) (xx briefly about ARE, as above. xx)

Sufficiency, factorisation theorem, completeness

Ex. 4.16 *Sufficiency.* (xx below we use the Bayes theorem; where should we point. xx) Flip a coin ten times and record the number of heads. It is intuitively clear that the number of heads in the ten tosses is as informative about the unknown $\theta = \Pr(\text{heads})$ of the coin, as the exact ordering in which the heads and tails occurred. In fact, the ordering in which the heads and tails occurred appear irrelevant for making inference about θ. This leads us to the notion of a sufficient statistic. If X is your data, stemming from a member of the family of distributions $\mathcal{P} = \{P_\theta\colon \theta \in \Theta\}$, and $T = T(X)$ is a statistic (i.e., any function of the data, real of vector valued, not depending on the parameter), then T is sufficient for $\mathcal{P}$, we often just say for θ, if the distribution of X given T is the same for all values of the parameter θ. We look more into this definition in Ex. 4.18.

sufficient statistic

(a) Here are a few examples. (i) Let $X_1, \ldots, X_n$ be independent Bernoulli(θ) random variables, and set $T = \sum_{i=1}^n X_i$. Use the Bayes theorem to show that $\Pr_\theta(X_1 = x_1, \ldots, X_n = x_n \,|\, T = t) = 1/\binom{n}{t}$ on $\{(x_1, \ldots, x_n)\colon x_i = 0, 1, \sum_{i=1}^n x_i = t\}$, from which we conclude that T is sufficient for θ. (ii) Let $Y_1, \ldots, Y_n$ be i.i.d. Pois(θ) and set $S = \sum_{i=1}^n Y_i$, and, again, use the Bayes theorem to show that $\Pr_\theta(Y_1 = y_1, \ldots, Y_n = y_n \,|\, S = s) = (\prod_{i=1}^n 1/y_i!)/(n^s/s!)$, on $\{(y_1, \ldots, y_n)\colon y_i = 0, 1, 2, \ldots, \sum_{i=1}^n y_i = s\}$,

which shows that S is sufficient for θ. (iii) Let $Z_1, \ldots, Z_n$ be i.i.d. unif$(0, \theta)$ and let $T = \max_{i \le n} Z_i$. Provide an intuitive argument for T being sufficient for θ. (iv) Let $W \sim \mathrm{N}(0, \sigma^2)$ and consider $T = |W|$. Again, provide an intuitive argument for why T is sufficient for σ^2.

(b) In view of Ex. (a) we may deduce the following result: Let $X_1, \ldots, X_n$ be discrete random variables and $T = T(X_1, \ldots, X_n)$ a statistic. Show that T is sufficient if and only if

$$\theta \mapsto \frac{\Pr_\theta\{X_1 = x_1, \ldots, X_n = x_n\}}{\Pr_\theta\{T(x_1, \ldots, x_n) = t\}} \tag{4.3}$$

is constant for every $x_1, \ldots, x_n$. In Ex. 4.18(f) we will see that this result holds more generally, that is, if $X = (X_1, \ldots, X_n)$ has density f_θ and $T = T(X)$ has density $g_\theta(t)$, then T is sufficient if and only if $\theta \mapsto f_\theta(x)/g_\theta(T(x))$ is constant for every x.

(c) The problem with the approach in (b) is that one has to make a guess at a sufficient statistic, find its distribution, and then compute the ratio in (4.3). The Fisher–Neyman factorisation theorem provides us with an automatic way for finding sufficient statistics. Suppose that $X_1, \ldots, X_n$ are random variables with joint density (e.g., joint p.m.f. or joint p.d.f.) $f_\theta(x_1, \ldots, x_n)$, and let $T = T(X_1, \ldots, X_n)$ be a statistic. The factorisation theorem says that T is sufficient if and only if there exists nonnegative functions h and g_θ so that for all θ and x

Fisher–Neyman factorisation theorem

$$f_\theta(x_1, \ldots, x_n) = g_\theta(T(x_1, \ldots, x_n))h(x_1, \ldots, x_n).$$

Prove the discrete version of this theorem, that is, the version where $f_\theta(x_1, \ldots, x_n) = \Pr_\theta(X_1 = x_1, \ldots, X_n = x_n)$. For a general proof of this theorem, i.e., one in which f_θ is any density, see Ex. 4.18.

(d) Use the factorisation theorem verify that the statistics from (a) are indeed sufficient. Find also a sufficient statistic based on an independent sample from the unif$(\theta, \theta + 1)$ distribution. Compared to the four other sufficient statistics of this exercise, what is particular about this latter?

(e) A sufficient statistic is not unique, and different sufficient statistics may provide varying degress of data compression. At one extreme are sufficient statistics not providing any compression of the data: (i) If $X_1, \ldots, X_n$ stem from a distribution with density f_θ, then then full sample is sufficient. Prove it. (ii) Let $X_1, \ldots, X_n$ be i.i.d. from an unknown continuous distribution F, and let $T = (X_{(1)}, \ldots, X_{(n)})$ be the order statistics. Show that the conditional distribution of $X_1, \ldots, X_n$ given T does not depend on F.

(f) For the lack of uniqueness, you can use the factorisation theorem to prove that any one-to-one transformation of a sufficient statistic is sufficient. And, for an example of increasing data compression, let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(0, \sigma^2)$ and consider the statistics $T_1 = (X_1, \ldots, X_n)$, $T_2 = (X_1^2, \ldots, X_n^2)$, $T_3 = (X_1^2 + \cdots + X_k^2, X_{k+1}^2 + \cdots + X_n^2)$, and $T_4 = X_1^2 + \cdots + X_n^2$. Clearly, T_4 is a function of T_3, T_3 is a function of T_2, and T_2 is a function of T_1, so the data compression is increasing in the indices. Use the factorisation theorem to prove that they are all sufficient.

Ex. 4.17 *Simulating data based on sufficient statistics.* A sufficient statistic T contains all the information provided by the original sample $X = (X_1, \ldots, X_n)$ about some parameter θ. Thus, given the sufficient statistic T, one may throw away the original data, and create an equally good data set $X' = (X'_1, \ldots, X'_n)$. What makes this possible is, of course, that the conditional distribution of X given T does not depend on θ. That X' is as good as X means that X' has the same distribution as X, so, for example, an estimator based on X' will be as good (i.e., same risk, see Ch. 8) as the same estimator computed from X. Let us look at a few examples.

(a) Let X and Y be independent Expo(θ). Show that $T = X + Y$ is sufficient for θ. Consider the random variables $X' = UT$ and $Y' = (1 - U)T$, where U is unif$(0, 1)$ and independent of X and Y. Think of U as a random variable you simulate on your computer knowing T. Show that $(X', Y') \sim (X, Y)$.

(b) Let X and Y be independent unif$(0, \theta)$ for some $\theta > 0$. Show that $T = \max(X, Y)$ is sufficient for θ. Consider the random variables $X' = \eta UT + (1 - \eta)T$ and $Y' = (1 - \eta)UT + \eta T$, where $U \sim$ unif$(0, 1)$ and $\eta \sim$ Bernoulli$(\frac{1}{2})$ are independent and independent of X and Y. Show that $(X', Y') \sim (X, Y)$. Find the conditional distribution of (X, Y) given $T = t$.

(c) Prove the general version of the above results, as discussed in the classical article Halmos and Savage (1949). That is, let $X \in \mathbb{R}^n$ be a random variable with distribution $\Pr_\theta$, and assume that T is sufficient for θ. Suppose we use a random number generator to simulate $X' \in \mathbb{R}^n$ from the conditional distribution $Q_t(B) = \Pr_\theta(X \in B \,|\, T = t)$. Show that $X' \sim X$ for all θ.

Ex. 4.18 *The factorisation theorem.* Above, in Ex. 4.16(b) we proved the factorisation theorem for discrete random variables. In this exercise we prove the general version, valid for any distribution dominated by a σ-finite measure (see Ex. A.2(a) for definition of σ-finiteness). First, we must be more formal in our definition of a sufficient statistic. Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability distributions on a measurable space $(\mathcal{X}, \mathcal{A})$. The statistic T is sufficient for $\{P_\theta : \theta \in \Theta\}$ if there is a function $p(A, x)$ of $A \in \mathcal{A}$ and $x \in \mathcal{X}$, not depending on θ, such that for all $A \in \mathcal{A}$ and $\theta \in \Theta$,

$$\int_G p(A, x)\, \mathrm{d}P_\theta(x) = \int_G I_A(x)\, \mathrm{d}P_\theta(x), \quad \text{for all } G \in \mathcal{G}.$$

Using the terminology introduced in Ex. **??** on conditional expectation, this means that $p(A, \cdot)$ is a *version* of the conditional probability $P_\theta(A \,|\, T)$ for all $A \in \mathcal{A}$ and $\theta \in \Theta$. Here, $P_\theta(A \,|\, T)$ is shorthand for the more cumbersome $P_\theta(A \,|\, \sigma(T))$, with $\sigma(T)$ the σ-algebra generated by T.

We now turn to the factorisation theorem. Suppose that the family $\{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ. For each θ, let f_θ be the density of P_θ with respect to μ. The statistic T is sufficient for $\{P_\theta : \theta \in \Theta\}$ if and only if there exist nonnegative functions h and g_θ such that

$$f_\theta(x) = g_\theta(T(x))h(x),$$

for all $\theta \in \Theta$. The proof of the factorisation theorem relies on the existence of a probability measure Q dominating $\{P_\theta : \theta \in \Theta\}$, i.e., $P_\theta \ll Q$ for all θ, with this dominating probability measure on the form $Q = \sum_{j=1}^{\infty} a_j P_{\theta_j}$, with $a_j > 0$ and each P_{θ_j} belonging to the family $\{P_\theta : \theta \in \Theta\}$. In the following string of exercises we first prove the factorisation theorem assuming the existence of such a probability measure Q, and defer the construction of Q to Ex. (d) [xx or perhaps the appendix? xx].

(a) Before we get to the proof of the factorisation theorem, let us work through some preliminaries. Let Q be as just described. First, show that Q is indeed a probability measure. Second, show that $P_\theta \ll \mu$ if and only if $Q \ll \mu$. Finally, show that when μ is σ-finite, $\mathrm{d}Q/\mathrm{d}\mu = \sum_{j=1}^{\infty} a_j \mathrm{d}P_{\theta_j}/\mathrm{d}\mu$.

(b) Suppose that $\{P_\theta : \theta \in \Theta\} \ll Q \ll \mu$, as described above. Assume that T is sufficient for $\{P_\theta : \theta \in \Theta\}$, i.e., there exists $p(A, \cdot)$ that is a version of $P_\theta(A \,|\, T)$ for every $\theta \in \Theta$. First, show that

$$\int_G Q(A \,|\, T)(x)\, \mathrm{d}Q(x) = \int_G p(A, x)\, \mathrm{d}Q(x),$$

for all $G \in \sigma(T)$. This shows that T is sufficient for the augmented family $\{P_\theta : \theta \in \Theta\} \cup \{Q\}$. Next, since $P_\theta \ll Q$, we can switch measure, $\mathrm{d}P_\theta = (\mathrm{d}P_\theta/\mathrm{d}Q)\, \mathrm{d}Q$ (see Ex. **??**). Use this measure switching in combination with the tower property of conditional expectation to show that

$$P_\theta(A) = \int g_\theta(T(x)) h(x)\, \mathrm{d}\mu(x),$$

for all $A \in \mathcal{A}$, where $h(x) = \mathrm{d}Q/\mathrm{d}\mu(x)$ and $g_\theta(T(x)) = \mathrm{E}_\theta\{\mathrm{d}P_\theta/\mathrm{d}Q \,|\, \sigma(T)\}(x)$, which proves (why?) one way of the factorisation theorem.

(c) To prove a converse of (b), still under the $\{P_\theta : \theta \in \Theta\} \ll Q \ll \mu$ assumption, show first that if, for all $\theta \in \Theta$, the density of P_θ with respect to Q only depends on x through $T(x)$, and is hence $\sigma(T)$-measurable, then $Q(A \,|\, \sigma(T))$ is a version of $P_\theta(A \,|\, \sigma(T))$ for all $\theta \in \Theta$. Next, assume that $f_\theta(x) = g_\theta(T(x))h(x)$ as described in the theorem. Appeal to (a) and Ex. **??** in the appendix, to show that

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}Q}(x) = \frac{g_\theta(T(x))}{\sum_{j=1}^{\infty} a_j g_{\theta_j}(T(x))},$$

and conclude that T is sufficient.

(d) [xx construction of Q here or in appendix xx]

(e) Suppose that $\{P_\theta : \theta \in \Theta\}$ satisfies the conditions of the factorisation theorem, and let T be a sufficient statistic, taking values in the measurable space $(\mathcal{T}, \mathcal{C})$. Thus, for every $\theta \in \Theta$, the density of P_θ with respect to μ is $f_\theta(x) = g_\theta(T(x))h(x)$. For every θ, we let $P_\theta^T(B) = P_\theta(\{x \in \mathcal{X} : T(x) \in B\})$ for $B \in \mathcal{C}$, be the distributions induced by T on $(\mathcal{T}, \mathcal{C})$. Let $Q = \sum_{j=1}^{\infty} a_j P_{\theta_j}$ be as described above, let $(QT^{-1})(B) = Q(\{x \in \mathcal{X} : (T(x) \in B\})$ be the measure induced on $(\mathcal{T}, \mathcal{C})$ via Q, and define a measure ν on

$(\mathcal{T},\mathcal{C})$ by $\nu(B) = \int_B \sum_{j=1}^\infty a_j g_{\theta_j}(t)\,\mathrm{d}(QT^{-1})(t)$, for $B \in \mathcal{C}$. Use what you found in (c) and the change of variable formula (see Ex. A.15(c)), to show that

$$P_\theta^T(B) = \int_B g_\theta(t)\,\mathrm{d}\nu(t),$$

for every $B \in \mathcal{C}$. This shows that P_θ^T has density $g_\theta(t)$ with respect to ν.

(f) Use (e) and the factorisation theorem to prove the general version of (4.3) in Ex. 4.16.

(g) Let us look at the result in (e) for a concrete example. Suppose $X_1,\ldots,X_n$ are i.i.d. Expo(θ), and let $T = \sum_{i=1}^n X_i$. Show that the joint density of $X_1,\ldots,X_n$ can be written $f_\theta(x_1,\ldots,x_n) = g_\theta(T(x_1,\ldots,x_n))h(x_1,\ldots,x_n)$, and conclude that T is sufficient. To find the marginal distribution of T, show that the m.g.f. of T is $\mathrm{E}_\theta\{\exp(aT)\} = (1-a/\theta)^{-n}$, $a<\theta$, from which we get that $T \sim \text{Gamma}(n,\theta)$. Find a measure ν on the range of T, with respect to which $P_\theta^T(B) = P_\theta(T\in B)$ has density $g_\theta(t)$. Convince yourself that ν is σ-finite.

Ex. 4.19 *The exponential family class, II.* (xx some technical but important things for the expo family. calibrate how and when we do sufficiency. xx)

(a) Suppose $Y_1,\ldots,Y_n$ are i.i.d. from an exponential family model $f(y,\theta) = \exp\{\theta^{\mathrm{t}}T(y) - k(\theta)\}h(y)$. From the factorisation theorem (Ex. 4.16(c)), the statistic $\bar T = (\bar T_1,\ldots,\bar T_p)^{\mathrm{t}}$, with $\bar T_j = (1/n)\sum_{i=1}^n T_j(Y_i)$ for $j=1,\ldots,p$, is clearly sufficient for the natural parameter $\theta = (\theta_1,\ldots,\theta_p)$. Use Ex. 4.18(e) to show that $\bar T$ has a distribution following the same exponential form, i.e., a distribution with density, say

$$g_n(t,\theta) = \exp\{\theta_1\bar t_1 + \cdots + \theta_p \bar t_p - k_n(\theta)\}h_n(y_1,\ldots,y_n),$$

for suitable k_n and h_n.

(b) (xx about conditioning. Point to Ex. 4.32. Rewrite this exercise xx). Let $\{P_{a,b}\colon (a,b)\in\Theta=\Theta_a\times\Theta_b\}$ be a family probability measures having densities of the exponential family form $f_{a,b}(y) = \exp\{a^{\mathrm{t}}U(y) + b^{\mathrm{t}}V(y) - k(a,b)\}h(y)$ with respect to μ. By the form of $f_{a,b}(y)$, you might, in view of the factorisation theorem, conjecture that for each a, V is sufficient for b. To see this, fix a, write the density $f_{a,b}$ as

$$f_{a,b}(y) = \exp\{b^{\mathrm{t}}V(y) - k_a(b)\}h_a(y),$$

where $k_a(b) = k(a,b)$ and $h_a(y) = \exp\{a^{\mathrm{t}}U(y)\}h(y)$, and appeal to the factorisation theorem. That V is sufficient means that there exists a version of the conditional probability $P_{a,b}(A\,|\,V)$ not depending on b, say $P_a^{U\,|\,V}(A)$. The distribution of U conditional on V is then $P_a^{U\,|\,V}(U\in B) = P_a^{U\,|\,V}U^{-1}(B)$, for B a measurable set in the range of U. It remains to construct a measure λ_v dominating P_aU^{-1}, and an expression for the conditional density $(\mathrm{d}P_a^{U|V}U^{-1}/\mathrm{d}\lambda_v)(u)$ belonging to the exponential class.

Ex. 4.20 *The exponential family class, III.* (xx more on the exponential family, more general parametrisations, examples. the moderate jump from (1.6) to

$$f(y,\theta) = \exp\{Q_1(\theta)T_1(y) + \cdots + Q_p(\theta)T_p(y) - k(\theta_1,\ldots,\theta_p)\}h(y),$$

with $Q_1(\theta),\ldots,Q_p(\theta)$. and, crucially, lift to regression models. xx)

(a)

(b)

Ex. 4.21 *Minimal sufficiency.* In Ex. 4.16(e) we saw that for any model there are many different sufficient statistics, often with some providing more compression of the data than others. Since the purpose of sufficient statistics is to compress the data, this naturally leads to a search for a sufficient statistic providing the maximum amount of data compression, while still retaining all the information about the unknown parameter of interest. Such a statistic is called a minimal sufficient statistic.

minimal sufficient

The formal definition is as follows: Let T be sufficient for $\{P_\theta \colon \theta \in \Theta\}$. Then T is minimal sufficient if for any other sufficient statistic S, there is a measurable function g so that $T = g(S)$ almost surely, for all values of θ. Another way of saying this is that if T is such that the implication 'if $S(x) = S(y)$ then $T(x) = T(y)$' holds for any sufficient statistic S, then T is minimal sufficient.

(a) (xx emil, what is U here. xx) Let $X \sim \mathrm{N}(0, \sigma^2)$. Show that both X and $|X|$ are sufficient for σ^2. Let $X' = U|X| + (1-U)|X|$, and show that $X \sim X'$. We see that $|X|$ provides more data compression than X, but is it minimal? We will soon have the tools to find out.

(b) Suppose that T is minimal sufficient, and let S be some sufficient statistic. Show that the σ-algebra generated by T must be contained in the σ-algebra generated by S. Show that any one-to-one function of a minimal sufficient statistic is minimal sufficient.

(c) The following theorem say the mapping from data to likelihood function, that is, $x \mapsto \{f_\theta(x) \colon \theta \in \Theta\}$, is minimal sufficient. The proof is based on the observation that from the factorisation $f_\theta(x) = f^{X \mid S}(x \mid s) f_\theta^S(s)$, the likelihood $\theta \mapsto f_\theta(x)$ is proportional to $\theta \mapsto f_\theta^S(s)$, for any sufficient statistic S. In other words, the likelihood function $f_\theta(x)$ is a function of the likelihood function $f_\theta^S(s)$ of any sufficient statistic S, and therefore the $f_\theta(x)$ is minimal sufficient.

Here is the theorem: Let $f_\theta(x)$ be the density of X. Suppose there is a function $T(x)$ is such that $T(x) = T(y)$ if and only if for some $h(x, y) > 0$

$$f_\theta(x) = f_\theta(x) h(x, y) \quad \text{for all } \theta.$$

Then $T(X)$ is minimal sufficient. To prove this, first, use the factorisation theorem to show that T is sufficient. Second, introduce another sufficient statistic S, and again use the factorisation theorem to show that T must be a function of S.

(d) (i) With $X \sim \mathrm{N}(0, \sigma^2)$, show that the absolute value $|X|$ is minimal sufficient. (ii) Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$, and show that $(\bar{X}_n, S_n)$ with $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $S_n = \sum_{i=1}^n (X_i - \bar{X})^2$ is minimal sufficient. (iii) Let $Y_1, \ldots, Y_n$ be i.i.d. from a distribution with density $f_\theta(y) = \exp(-(y-\theta))$ for $x > \theta$ and $\theta \in \mathbb{R}$. Find a minimal sufficient statistic for θ.

(e) Let $g(x)$ be a positive and integrable function on $(-\infty, \infty)$. Set $c(a, b)^{-1} = \int_a^b g(x)\,\mathrm{d}x$, and define $f_{a,b}(x) = c(a, b) g(x) I_{(a,b)}(x)$. Let $X_1, \ldots, X_n$ be i.i.d., from the distribution with density $f_{a,b}(x)$. Find a minimal sufficient statistic for (a, b).

(f) Let $X_1, \ldots, X_n$ be i.i.d. from a distribution with density $f_\theta(x) = 1/2 \exp(-|x - \theta|)$, $x, \theta \in \mathbb{R}$. Show that the order statistics are minimal sufficient.

(g) Let $Y_1, \ldots, Y_n$ be i.i.d. from a distribution with a density of the exponential class $f_\theta(y) = \exp\{\sum_{j=1}^p Q_j(\theta)T_j(y) - k(\theta_1, \ldots, \theta_p)\}h(y)$ of full rank (see Ex. 4.20). Show that $\bar{T} = (\bar{T}_1, \ldots, \bar{T}_p)$, where $\bar{T}_j = n^{-1} \sum_{i=1}^n T_j(Y_i)$ is minimal sufficient for $(\theta_1, \ldots, \theta_p)$. (xx coordinate this with Ex. 4.23 completeness below. xx) In fact, a stronger result holds, namely that $\bar{T}$ is complete (see, e.g., Schervish (1995, Theorem 2.74, p. 108) for a proof of this fact, and Ex. **??** for a proper treatment of completeness). That $\bar{T}$ being complete and sufficient (the latter follows from the factorisation theorem) is stronger than minimal sufficiency, is proven in Ex. **??**(g).

(h)

Ex. 4.22 *Ancillary statistics.* The opposite of sufficiency, in a sense, is ancillarity. If $X \sim P_\theta$, a statistic $U = U(X)$ is ancillary if its distribution is the same for all θ. In other words, U by itself does not provide any information about θ. This does not mean that U should be disregarded when making inference on θ. It just means that if you only learn $U = u$, you have not learned anything about θ.

ancillary statistic

(a)

(b) [xx ancillary stat in location families, and in scale families xx]

Ex. 4.23 *Completeness.* Often, models are so harmoniously constructed that there are clear one-to-one connections between estimators (perhaps based on a set of summary statistics) and estimands, in the sense that there for each estimand is only one unbiased estimator. Clarifying such regularity leads to the concept of *completeness*, which turns out to be useful also when coming to conditional testing and optimal power in exercises below. Technically, suppose some vector $T = (T_1, \ldots, T_p)^{\mathrm{t}}$ has a distribution $f(t, \theta)$, with the property that $\mathrm{E}_\theta\, h(T) = 0$ for all $\theta \in \Theta$ implies $\Pr_\theta\{h(T) = 0\} = 1$ for all θ, i.e. $h(t) = 0$ almost everywhere. We then say that T, or more formally its distribution, over the relevant parameter region, is complete.

complete

(a) Let $X \sim \mathrm{binom}(n, \theta)$, with $\theta \in (0, 1)$. Show that X is complete; zero is the only unbiased estimator of zero. (You may appeal to properties of power series.) Show that X is complete as long as the parameter region contains an open interval. Show similarly that if $X \sim \mathrm{geom}(p)$, see Ex. 1.24, then X is complete, again requiring only that the parameter range for p contains an open interval.

(b) With $Y_1, \ldots, Y_n$ i.i.d. from the uniform on $[0, \theta]$, consider $M = \max_{i \le n} Y_i$. Show that if the parameter region is the full halfline $\theta > 0$, then M is sufficient and complete, but that M is not complete if it is a priori known that $\theta \ge 1.234$, i.e. with a restricted parameter range.

(c) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the uniform distribution over $[\theta - 1, \theta + 1]$. Show that $(Y_{(1)}, Y_{(n)})$, the smallest and largest, is sufficient, but not complete.

(d) Consider Y being uniform on $\{1,\ldots,\theta\}$, where $\theta \in \{1,2,\ldots\}$ is an unknown parameter. Show that Y is complete. When the parameter region is e.g. $\{4,5,\ldots\}$, however, show that Y is not complete.

(e) If T is complete, show that any one-to-one transformation variable $T' = a(T)$ is also complete.

(f) Consider $Y_1,\ldots,Y_n$ an i.i.d. sample from the double exponential with density $f(y,\theta) = \frac{1}{2}\exp(-|y-\theta|)$. Show that the full set of order statistics $(Y_{(1)},\ldots,Y_{(n)})$ is sufficient, but not complete; do this, by exhibiting two different unbiased estimators of θ.

Ex. 4.24 *Completeness for the exponential family.* For the large class of exponential family models, see Ex. 1.50 (xx and follow-up exercises above xx), there is a completeness lemma, as follows. Suppose $Y_1,\ldots,Y_n$ are i.i.d. from the model $f(y,\theta) = \exp\{\theta^{\rm t}T(y) - k(\theta)\}h(y)$, with $T = (T_1,\ldots,T_p)^{\rm t}$ and $\theta = (\theta_1,\ldots,\theta_p)^{\rm t}$ varying in an open set, then the vector of sample averages $(\bar T_1,\ldots,\bar T_p)^{\rm t}$ is not merely sufficient, as seen in Ex. 4.21, but also complete. We shall freely use this lemma. (xx but point to proof, check BickelDoksum or Johansen or Brown or Schervish. perhaps requiring analyting continuation arguments. point is that $a(\theta) = {\rm E}_\theta\, h(\bar T)$ is a super smooth functions with all derivatives smooth. xx)

completeness lemma for exponential family

(a) Let $Y_1,\ldots,Y_n$ be i.i.d. from the ${\rm N}(\xi,1)$. Show that the full set $(Y_1,\ldots,Y_n)$ is not complete, but that the sample mean $\bar Y$ is.

(b) Consider $Y_1,\ldots,Y_n$ i.i.d. from the ${\rm Gam}(a,b)$ model. Show that $(\sum_{i=1}^n Y_i, \sum_{i=1}^n \log Y_i)$ is sufficient and complete. Identify similarly a sufficient and complete pair of statistics for a sample from the ${\rm Beta}(a,b)$.

(c) Consider an i.i.d. sample $Y_1,\ldots,Y_n$ from the ${\rm N}(\xi,\sigma^2)$. Show that $(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)$ is sufficient and complete, and also that $(\bar Y, \widehat\sigma)$ is sufficient and complete. Suppose then that the variance is postulated to be equal to the squared mean, so that the sample is from ${\rm N}(\theta,\theta^2)$. Construct two different unbiased estimators of θ, and show that this means that $(\bar Y,\widehat\sigma)$ is not complete. You may similarly construct two different unbiased estimators of θ^2.

(d) Consider the linear regression model with $Y_i \sim {\rm N}(a+bx_i,\sigma^2)$, studied in Ex. 3.30. With $\widehat a$, $\widehat b$ the least squares estimators, and $Q_0 = \sum_{i=1}^n \{Y_i - \widehat a - \widehat b(x_i-\bar x)\}^2$, show that the log-likelihood can be written $-n\log\sigma - \frac{1}{2}\{Q_0 + n(\widehat a - a)^2 + M_n(\widehat b - b)^2\}^2/\sigma^2$, with $M_n = \sum_{i=1}^n (x_i-\bar x)^2$. Write this in the exponential family fashion, with natural parameters $1/\sigma^2, a/\sigma^2, b/\sigma^2$. Argue via the general exponential family results that $(Q_0,\widehat a,\widehat b)$ is both sufficient and complete. Extend this, arguments and results, to the general multiple linear regression model of Ex. 3.31.

Ex. 4.25 *Basu's Lemma.* Consider a setting with data Y from some parametric family, indexed by $\theta \in \Theta$, where a suitable T is sufficient and complete.

(a) Assume that a suitable statistic $Z = Z(Y)$ has a distribution not depending on the θ. Show that Z is independent of T. This is called Basu's Lemma. You may follow this path: start writing $\Pr_\theta(Z \in A) = p$, by assumption not depending on θ. Show that $p = {\rm E}_\theta\, h(T)$, for all θ, where $h(t) = \Pr_\theta(Z \in A \,|\, T = t)$, another function not depending on θ. Explain that this implies $\Pr_\theta(Z \in A \,|\, T = t) = p$ for all θ.

Basu's lemma

(b) Consider the familiar setup with $Y_1, \ldots, Y_n$ being i.i.d. from $\mathrm{N}(\mu, \sigma^2)$. For σ fixed, show that the mean $\bar{Y}$ is sufficient and complete for μ; explain that all statistics Z with distribution not depending on μ must be independent of $\bar{Y}$. Explain hence in particular that $\sum_{i=1}^n (Y_i - \bar{Y})^2$ is independent of $\bar{Y}$. This has been demonstrated in different ways in Ex. 1.44 and 1.45.

(c) In this normal sample setting, argue that the classic t ratio $t = \sqrt{n}(\bar{Y} - \mu)/\widehat{\sigma}$, with $\widehat{\sigma}$ the sample standard deviation, is independent of $(\bar{Y}, \widehat{\sigma})$.

(d) We have seen in Ex. 1.51 that the five-parameter binormal distribution is inside the exponential family class. With $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. from the binormal, show first that $(\bar{X}, \bar{Y}, \widehat{\sigma}_1, \widehat{\sigma}_2, R_n)$ is complete and sufficient, with R_n the empirical correlation coefficient. Show that R_n is independent of $\bar{X}, \bar{Y}, \widehat{\sigma}_1, \widehat{\sigma}_2$ (xx check this with care xx).

Optimal conditional testing

Ex. 4.26 *Conditional tests.* Suppose in general terms that data y are observed from a model with parameter θ, where the null hypothesis $H_0\colon \theta \in \Theta_0$ is to be tested, against the alternative that $\theta \notin \Theta_0$. Assume one computes $U = U(y)$ and $V = V(y)$. A *conditional test*, with respect to V, with level α, is then to find a rejection region $R(v)$, using the distribution of U given $V(y) = v$, with

conditional tests

$$\Pr_\theta\{U(Y) \in R(v) \,|\, V(Y) = v\} \le \alpha \quad \text{for all } \theta \in \Theta_0.$$

Such tests are natural and important in multiparameter setups, as we shall see, and various constructions succeed in 'reducing the dimensionality' down to the analysis of a one-parameter family, where e.g. Neyman–Pearson more readily applies.

(a) Even when such a test has been constructed in a conditional modus, 'what is unlikely null behaviour of U given that $V = v$', it may of course be translated or paraphrased without the conditioning: one rejects if $U(Y) \in R(V)$. Show that the test also has unconditional level α.

(b) From unconditional to conditional: Conditional tests as above have the form $T(U, V) = I(U \in R(V))$, built to have $\mathrm{E}_\theta\{I(U \in R(v)) \,|\, v\} = \alpha$. Assume now that V is complete at the boundary $\partial\Theta_0$ of the null hypothesis parameter region; see Ex. 4.23. Show that if *any* test $T(U, V)$ has constant level α at this boundary, then it is a conditional test, with this level; $\mathrm{E}_\theta\{T(U, V) \,|\, V = v\} = \alpha$ for all $\theta \in \bar{\Theta}_0$. (It might be useful to check the function $h(v) = \mathrm{E}_\theta\{T(U, V) \,|\, V = v\} - \alpha$.)

Ex. 4.27 *Conditional tests: pairs of exponentials.* Suppose $X \sim \mathrm{Expo}(a)$ and $Y \sim \mathrm{Expo}(a + \delta)$, and that one wishes to test $\delta = 0$, i.e. equal distributions, against $\delta > 0$.

(a) Show that the joint density may be written $a(a + \delta)\exp(-az - \delta y)$, with $z = x + y$. Find the distribution of $Z = X + Y$, and show that the distribution of Y given $Z = z$ has the density

$$g_\delta(y \,|\, z) = \frac{\delta \exp(-\delta y)}{\int_0^z \delta \exp(-\delta y')\,\mathrm{d}y'} = \frac{\delta \exp(-\delta y)}{1 - \exp(-\delta z)} \quad \text{for } 0 \le y \le z.$$

In particular, it does not depend on the θ. For the null hypothesis case of $\delta = 0$, show that $Y \,|\, z$ is uniform on $[0, z]$.

(b) The natural conditional 0.05 level test is then to first compute z, and then to reject if $y \leq 0.05\, z$. Show that it indeed has level 0.05, and that is the power optimal test among all conditional tests, using Y given z. Verify that this conditional test is the same as the unconditional test of rejecting when $R = Y/(X+Y) \leq 0.05$. Compute the power function of the $T^* = I(Y \leq 0.05\, Z)$ test (in the testing function parlance of Ex. 4.7), conditional on z, and unconditionally.

(c) At the boundary of the null, where $\delta = 0$, show that Z is complete. Show hence that *any* test with level 0.05 also must be a conditional on Z 0.05 level test, via Ex. 4.27.

(d) We know that $T^*(y, z) = I(y \leq 0.05\, z)$ is the most powerful conditional test with level 0.05; we now wish to extend this statement to T^* actually being the most powerful among *all* tests with level 0.05. For any competing test $T(Y, Z)$ with level 0.05, show, since it must be a z-conditional 0.05 level test, where it cannot beat T^*, that

$$\mathrm{E}_{a,\delta}\,\{T^*(Y,Z) \,|\, z\} \geq \mathrm{E}_{a,\delta}\,\{T(Y,Z) \,|\, z\} \quad \text{for all } \delta > 0, z > 0.$$

There is equality, to 0.05, at $\delta = 0$. Show from this that T^* is more powerful than such T, unconditionally; in suitable power function symbols, $\pi_{T^*}(a, \delta) \geq \pi_T(a, \delta)$ for all $\delta > 0$.

(e) Suppose now that there are m independent pairs, $X_i \sim \mathrm{Expo}(a_i)$ and $Y_i \sim \mathrm{Expo}(a_i + \delta)$, with sums $Z_i = X_i + Y_i$; there are hence $m+1$ parameters with $2m$ data points. Show that the optimal test is to reject when $U_m = Y_1 + \cdots + Y_m$ is small, given $z_1, \ldots, z_m$. Explain how the null distribution of U_m can be evaluated via simulations. For an illustration, suppose three pairs (x_i, y_i) are observed: $(0.927, 0.819)$, $(1.479, 0.408)$, $(3.780, 1.311)$. Carry out the test of δ, and compute the p-value.

Ex. 4.28 *Conditional tests: normal.* (xx various situations with distribution of $U \,|\, (V = v)$, followed by natural conditional test. xx) Consider a pair of normals, where interest lies in assessing their difference in means. This may of course be parametrised in different ways, but one natural way is $x \sim \mathrm{N}(\theta, 1)$ and $y \sim \mathrm{N}(\theta + \delta, 1)$. One wishes to test $\delta = 0$ vs. $\delta > 0$, equivalent, of course, to testing equality of the means vs. $\mathrm{E}\, y > \mathrm{E}\, x$.

(a) Show that the joint likelihood can be written

$$\begin{aligned} f(x, y, \theta, \delta) &= (2\pi)^{-1} \exp[-\tfrac{1}{2}\{(x-\theta)^2 + (y-\theta-\delta)^2\}] \\ &= (2\pi)^{-1} \exp\{\theta z + \delta y - \tfrac{1}{2}x^2 - \tfrac{1}{2}y^2 - \tfrac{1}{2}\theta^2 - \tfrac{1}{2}(\theta+\delta)^2\}, \end{aligned}$$

where $z = x + y$, and with the main interaction between parameters and data being in the $\theta z + \delta y$ part.

(b) Show that (y, z) is a binormal, and set up its mean vector and variance matrix. Then use Ex. 1.41, or other algebraic methods, to show that $y \,|\, z \sim \mathrm{N}(\frac{1}{2}(z+\delta), \frac{1}{2})$; in particular, its conditional distribution does not depend on θ.

(c) Through the conditioning on z the testing problem has been reduced from a two-parameter to a one-parameter situation. For $y \,|\, z \sim \mathrm{N}(\frac{1}{2}(z+\delta), \frac{1}{2})$, show that the optimal test is to reject when $y - \frac{1}{2}z > (1/\sqrt{2})c$, with $c = \Phi^{-1}(1-\alpha)$ the standard normal quantile.

(d) Show that the above test, constructed to be optimal in the model for $y \,|\, z$, is equivalent to that of rejecting whem $D = y - x > \sqrt{2}c$. (xx so the conditional test is an ordinary unconditional test in disguise, or vice versa, in this particular situation. the point is the general principle. xx)

(e) (xx Consider m pairs of normal data, of the form $x_i \sim \mathrm{N}(\theta_i, 1)$ and $y_i \sim \mathrm{N}(\theta_i + \delta, 1)$. do the math, with the steps above. log joint density $\sum_{i=1}^m (\theta_i z_i + \delta y_i)$, with $z_i = x_i + y_i$. conditional test, $\sum_{i=1}^k y_i$ big given $z_1, \ldots, z_k$. xx)

(f) (xx something re power. xx)

Ex. 4.29 *Conditional tests: Poisson.* (xx various situations with distribution of $U \,|\, (V = v)$, followed by natural conditional test. xx)

(a) We start with a single pair of Poissons, x with mean θ, y with mean $\theta\gamma$. Show that the joint distribution becomes $\exp(-\theta - \theta\gamma)\theta^{x+y}\gamma^y/(x!\,y!)$. This inspires inspecting the distribution of y given $z = x + y$. Show that $y \,|\, z \sim \mathrm{binom}(z, \gamma/(\gamma+1))$.

(b) To test $\gamma = 1$ against $\gamma > 1$, describe in details the natural conditional test which rejects when y is big, given $z = x + y$.

(c) Next consider independent Poisson pairs x_i, y_i for $i = 1, \ldots, m$, where x_i has mean θ_i and y_i mean $\theta_i\gamma$. The model hence has $m + 1$ parameters for the $2m$ observations, with γ the common multiplicative factor. Show that the joint distribution may be written

$$f = \exp\Big\{-\sum_{i=1}^m (\theta_i + \theta_i\gamma) + \sum_{i=1}^m \{(x_i + y_i)\log\theta_i + y_i \log\gamma\}\Big\} \frac{1}{x_1!\,y_1! \cdots x_m!\,y_m!}.$$

With $z_i = x_i + y_i$, find the distribution of $y_i \,|\, z_i$, and also the distribution of $S = \sum_{i=1}^m y_i$ given $z_1, \ldots, z_m$.

(d) Find the power optimal test for $\gamma = 1$ against $\gamma > 1$, among all those based on S given $z_1, \ldots, z_m$.

(e) (xx more, rounding off. something with limit. point to Ch7 optimal CD. also do $Y \sim \mathrm{Pois}(m_0\theta_0)$ and $Y_1 \sim \mathrm{Pois}(m_1\theta_1)$, with m_0 and m_1 exposure time. With interest being in the ratio parameter $\gamma = \theta_1/\theta_0$, show that $Y_1 \,|\, (Z = z)$ is binomial $(z, m_1\gamma/(m_0 + m_1\gamma))$. xx)

Ex. 4.30 *Conditional tests:* 2×2 *tables.* (xx various situations with distribution of $U \,|\, (V = v)$, followed by natural conditional test. xx)

(a) Consider two binomials $y_0 \sim \mathrm{binom}(m_0, p_0)$ and $y_1 \sim \mathrm{binom}(m_1, p_1)$. The outcomes in such situations are often presented as a two-by-two table,

$$\begin{array}{ll} y_0, & m_0 - y_0 \\ y_1, & m_1 - y_1 \end{array}$$

Consider the so-called logistic parametrisation

$$p_0 = H(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} \quad \text{and} \quad p_1 = H(\theta+\gamma) = \frac{\exp(\theta+\gamma)}{1+\exp(\theta+\gamma)}.$$

Show that $\theta = \log\{p_0/(1-p_0)\}$ and $\theta+\gamma = \log\{p_1/(1-p_1)\}$ in terms of the so-called log-odds. Show that the joint distribution can be written

$$f = \binom{m_0}{y_0}\binom{m_1}{y_1}\frac{\exp(\theta(y_0+y_1))}{\{1+\exp(\theta)\}^{m_0}}\frac{\exp(\gamma y_1)}{\{1+\exp(\theta+\gamma)\}^{m_1}}.$$

(b) (xx in view of ... check, calibrate. xx) This inspires reaching inference for γ via the conditional distribution of y_1 given $z = y_0+y_1$. Show that this distribution becomes

$$g_\gamma(y_1 \mid z) = \binom{m_0}{z-y_1}\binom{m_1}{y_1}\exp(\gamma y_1)\Big/\sum_{y_1'\le\min(m_1,z)}\binom{m_0}{z-y_1'}\binom{m_1}{y_1'}\exp(\gamma y_1')$$

for $y_1 = 0,1,\ldots,\min(m_1,z)$. In particular, this so-called excentric hypergeometric distribution depends on γ but not θ. We recognise the ordinary hypergeometric for $\gamma = 0$; see Ex. 1.62.

(c) Show that the optimal 0.05 level conditional test for the null hypothesis of equality, $p_0 = p_1$, is to reject when $y_1 > c(z)$, with $c(z)$ the highest number with $\sum_{0\le y_1\le c(z)} g_0(y_1 \mid z) \le 0.95$.

(d) (xx the power. xx)

(e) (xx to k two-by-two tables, $p_{i,0} = H(\theta_i)$ and $p_{i,1} = H(\theta_i+\gamma)$, $k+1$ parameters. optimal conditional test for $\sum_{i=1}^k y_{1,i}$ given $z_1,\ldots,z_k$, with $z_i = y_{i,0}+y_{i,1}$. point to Story i.10. xx)

Ex. 4.31 *The t test as an optimal conditional test.* Let $Y_1,\ldots,Y_n$ be i.i.d. from the normal (ξ,σ^2), where we wish to test $\xi = 0$ against $\xi > 0$. The canonical classical test, of level say 0.05, is based on $t = \sqrt{n}\bar{Y}/\sigma$, rejecting if $t \ge t_{n-1,0.95}$, the upper 0.05 point of the t_{n-1} distribution; see also Ex. 3.7 and (3.4). We cannot use the Neyman–Pearson lemma directly to demonstrate optimality of the t test, however. One of several optimality properties may be derived via conditioning.

(a) Write $U = \sqrt{n}\bar{Y}$ and $V = \sum_{i=1}^n Y_i^2$, so that in particular $W = \sum_{i=1}^n (Y_i-\bar{Y})^2 = V-U^2$. Show that the joint density of the data can be written

$$\begin{aligned} f &= \frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp\Big\{-\tfrac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^n (y_i-\xi)^2\Big\} \\ &= \frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp\Big\{\frac{\xi}{\sigma^2}\sqrt{n}U - \tfrac{1}{2}\frac{1}{\sigma^2}V - \tfrac{1}{2}\frac{\xi^2}{\sigma^2}\Big\}. \end{aligned}$$

Note that the testing problem is equivalent to testing $\lambda = 0$ against $\lambda > 0$, with $\lambda = \xi/\sigma^2$, the mathematics indicating that this is a parameter easier to work with than ξ.

(b) Find the distribution of $U \,|\, (V = v)$, and show in particular that it depends on the parameters only via $\lambda = \xi/\sigma^2$. It is convenient here to work with

$$T = \frac{U}{\{W/(n-1)\}^{1/2}} = (n-1)^{1/2} \frac{U}{(V - U^2)^{1/2}}.$$

(c) Show that the power optimal test, among all tests based on $U \,|\, (V = v)$, is to reject when U is big, say $U \geq c(v)$, with $\Pr_0\{U \geq c(v) \,|\, V = v\} = 0.05$. Then show that this is actually the same as the t test.

(d) Show via arguments used in Ex. 4.27 and in other places above that the t test is not merely optimal among all conditional tests, for given level α, i.e. given V, but among *all* tests with level α.

Ex. 4.32 *Conditional tests: multiparameter exponential models.* In the rather simple situation of Ex. 4.27, with the exponential pair $X \sim \text{Expo}(a)$ and $Y \sim \text{Expo}(a + \delta)$, with sum $Z = X + Y$, we learned that (i) there is a clear level α conditional test for $\delta = 0$ vs. $\delta > 0$, in terms of $Y \,|\, Z$; (ii) that test is uniformly most powerful against all $\delta > 0$, among all conditional tests; and (iii) all other level α competitors are in fact also Z-conditional. Hence the winning test, reject if $Y \leq \alpha Z$, is the uniformly most powerful level α test. – We shall see now that the same arguments essentially go through for the wide class of all exponential families. Consider data Y from a density of the form $f(y, a, b) = \exp\{aU(y) + b^{\text{t}}V(y) - k(a, b)\}h(y)$, as in Ex. 4.19, with one-dimensional U and p-dimensional V. Suppose we need to test $a = a_0$ against $a > a_0$, for some given null hypothesis value a_0.

(a) We have seen in the exercise pointed to that $U \,|\, (V = v)$ has a density depending on a but not b, and that it has an exponential form. Assume for simplicity that the distribution of U is continuous; mild formalistic additional arguments are required if the distribution is discrete. Deduce that there is a most powerful conditional level α test, say $T^*(y) = I\{U > c(V)\}$, with $c(v)$ determined from $\Pr_{a_0}(U > c(v) \,|\, V = v) = \alpha$, and with consequent power function $\pi(a, b) = \Pr_{a,b}\{U > c(V)\}$.

(b) Then consider *any* competing test $T(U, V)$ with level α. From $\text{E}_{a_0,b}\, T(U(Y), V(Y)) = \alpha$, for all b, use completeness in b of $V(Y)$ for fixed a_0 to prove that

$$\text{E}_{a_0,b}\{T(U(Y), V(Y)) \,|\, V(y) = v)\} = \alpha$$

for all v (except perhaps in a region of probability zero), and for all b. Thus the T competitor is also a level α V-conditional test, and we have proved that the conditional test is uniformly most powerful among *all* level α tests.

(c) The theory extends fruitfully to the case of testing $H_0 \colon a \leq a_0$ against $a > a_0$. Show that the test $T^* = I\{U > c(V)\}$ above, with $c(v)$ determined from $\Pr_{a_0}(U > c(v) \,|\, V = v) = \alpha$ at the boundary, is still of level α. Then show that this test is uniformly most powerful against all competing tests with constant level α at the boundary $a = a_0$; one says that such tests are *unbiased*. This latter very mild limitation is in order for the completeness argument to go through.

(d) (xx briefly about two-sided tests. still based on $U \,|\, (V = v)$. xx)

(e) When $U \,|\, (V = v)$ has a discrete distribution, the arguments still go through, but one cannot expect to find $c(v)$ with e.g. $\Pr_{a_0}\{U > c(v) \,|\, V = v\} = 0.05$. There are two ways out of this mild quandary. The first is to be satisfied with level 0.042, say, if that is how close one comes to 0.050, by appropriate choice of $c(v)$. The other, if one pedantically insists on 0.05, is to finetune $c(v)$ such that $\Pr_{a_0}\{U > c(v) \,|\, V = v\}$ is just below 0.05, and then identify the probability r such that

$$\Pr_{a_0}\{U > c(v) \,|\, V = v\} + r \Pr_{a_0}\{U = c(v) \,|\, V = v\} = 0.05,$$

So one rejects if $U > c(V)$, or, but then with probability r, if $U = v(C)$.

(f) (xx go through the previous exercises about conditional tests, once more, make sure that the tests found there are really uniformly most powerful among all unbiased tests. xx)

Ex. 4.33 *Optimal t testing in linear regression.* Consider the classic linear regression setup with $Y_i \sim \mathrm{N}(a + bx_i, \sigma^2)$ for $i = 1, \ldots, n$, see Ex. 3.30, and for simplicity of presentation here we assume the x_i have been centred, so $\bar{x} = 0$. The three estimators are then $\widehat{a} = \bar{Y}$, $\widehat{b} = \sum_{i=1}^n x_i Y_i / M_n$, with $M_n = \sum_{i=1}^n x_i^2$, and $\widehat{\sigma}^2 = Q_0/(n-2)$, with $Q_0 = \sum_{i=1}^n (Y_i - \widehat{a} - \widehat{b}x_i)^2$.

(a) With $V = \sum_{i=1}^n Y_i^2$, show that $Q_0 = V - n\widehat{a}^2 - M_n\widehat{b}^2$, and that the likelihood may be written

$$\begin{aligned} L &= \frac{1}{\sigma^n} \exp\Big\{-\tfrac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - a - bx_i)^2\Big\} \\ &= \frac{1}{\sigma^n} \exp\Big\{-\tfrac{1}{2}\frac{V}{\sigma^2} + \frac{a}{\sigma^2} n\widehat{a} + \frac{b}{\sigma^2} M_n \widehat{b} - \tfrac{1}{2}\frac{na^2 + M_n b^2}{\sigma^2}\Big\}, \end{aligned}$$

which is in the exponential family form, with natural parameters $1/\sigma^2$, a/σ^2, b/σ^2.

(b) Explain from the general exponential family testing theory that the uniformly most powerful test for $b = 0$ against $b > 0$, among all level α tests is to reject when $\widehat{b}$ is large enough, in its conditional distribution given $(V, \widehat{a})$. We need to find $c(V, \widehat{a})$ such that

$$\Pr_{a,\sigma}(\widehat{b} > c(V, \widehat{a}) \,|\, V, \widehat{a}) = \alpha$$

under $b = 0$. But this may be transformed to

$$W = \frac{\widehat{b}}{(V - n\widehat{a}^2)^{1/2}} = \frac{\widehat{b}}{(Q_0 + M_n\widehat{b}^2)^{1/2}} = \frac{\widehat{b}/Q_0^{1/2}}{(1 + M_n\widehat{b}^2/Q_0)^{1/2}} > \frac{c(V, \widehat{a})}{(V - n\widehat{a}^2)^{1/2}}.$$

But W is a smooth increasing function of $t = \widehat{b}/(\widehat{\sigma}/M_n)$, the classic t ratio, which has the t_{n-2} distribution under $b = 0$. Argue that W is therefore independent of $(V, \widehat{a})$, and that the optimal test is to reject when $t > t_0$, the upper α in the t_{n-2} distribtion. You have now shown that the traditional t test is optimal.

(c) Generalise the above to general linear multiple regression models, giving the result that the traditional t tests, for each of the β_j coefficients, are optimal.

Ex. 4.34 *A generalised Poisson distribution.* For a count variable Y, consider the model with point probabilities

$$f(y,\lambda,\gamma) = k(\lambda,\gamma)^{-1}\lambda^y/(y!)^\gamma \quad \text{for } y = 0,1,2,\ldots,$$

where $k(\lambda,\gamma)$ is the normalisation constant $\sum_{y=0}^\infty \lambda^y/(y!)^\gamma$. For $\gamma = 1$ we're back to ordinary Pois(λ), with $k(\lambda,1) = \exp(\lambda)$. This two-parameter generalised Poisson model is from Schweder and Hjort (2016, Examples 4.18, 8.16). A regression version model of this type is used in Story iv.6, to assess potential overdispersion in Poisson counts.

(a) Pick some λ, and compute and display curves of the mean $\xi(\lambda,\gamma)$ and the variance-to-mean ratio $\rho(\lambda,\gamma)$, for an interval of γ around 1. Show that this ratio is decreasing in γ; hence $\gamma < 1$ indicates overdispersion and $\gamma > 1$ underdispersion, relative to the Poisson. Also show that the mean of $\log(Y!)$ is decreasing in γ.

(b) Show that the distribution is of the exponential family form, and that the sufficient statistics, after having observed a sample $Y_1,\ldots,Y_n$, is $T = \sum_{i=1}^n Y_i$ and $U = \sum_{i=1}^n \log(Y_i!)$. Show also that the joint distribution of these two must take the form

$$g_n(t,u) = \exp\{t\log\lambda - u\gamma - r_n(\lambda,\gamma)\}h_n(y_1,\ldots,y_n),$$

for appropriate functions r_n and h_n.

(c) For an observed sample $Y_1,\ldots,Y_n$, to test the Poisson assumption, against overdispersion, show that the optimal test is to reject when U is sufficiently small, given $T = t$. In other words, with level the classic 0.05, for example, we reject when $U \le u_0(t)$, where $u_0(t)$ is the 0.95 quantile of the distribution of U given $T = t$, computed at $\gamma = 1$, i.e. under Poisson conditions. (xx this needs more care; distribution of $U \,|\, (T = t)$ needs a formula or two, so we see that U significantly small indicates $\gamma < 1$. xx)

(d) There is no table or simple formula for the distribution of $U \,|\, (T = t)$, but show that it depends on γ, but not λ. Show that under $\gamma = 1$, $(Y_1,\ldots,Y_n) \,|\, (T = t)$ is a multinomial with count t and probabilities $(1/n,\ldots,1/n)$. Explain then how the distribution of $U \,|\, (T = t)$ may be simulated under Poisson conditions.

(e) (xx give them a dataset. nils checks the football matches dataset. decide later if this is a Story or an exercise. xx)

Ex. 4.35 *Testing for correlation.* Consider binormal i.i.d. pairs (X_i, Y_i), with parameters $\xi_1,\xi_2,\sigma_1,\sigma_2,\rho$, as in Ex. 1.51. Suppose we need to test $\rho = 0$ vs. $\rho > 0$.

(a) Argue as in the exercise pointed to that the binormal distribution is inside the exponential family class, with natural parameters

$$\frac{\xi_1}{(1-\rho^2)\sigma_1^2}, \frac{1}{(1-\rho^2)\sigma_1^2}, \frac{\xi_2}{(1-\rho^2)\sigma_2^2}, \frac{1}{(1-\rho^2)\sigma_2^2}, \frac{\rho}{(1-\rho^2)\sigma_1\sigma_2},$$

associated with data functions $X_i, X_i^2, Y_i, Y_i^2, X_iY_i$. Explain that the testing problem is equivalent to testing $\lambda = 0$ vs. $\lambda > 0$, where $\lambda = \rho/\{(1-\rho^2)\sigma_1\sigma_2\}$.

(b) Then argue from the general testing theory for the exponential class that there for given testing level α is a uniformly most powerful test, consisting in rejecting the null provided $A_n = n^{-1}\sum_{i=1}^n X_iY_i$ is big enough, in its conditional null distribution given $(\bar X, n^{-1}\sum_{i=1}^n X_i^2, \bar Y, n^{-1}\sum_{i=1}^n Y_i^2)$. Formally, the rejection threshold c, which depends on these variables, is determined by

$$\Pr_0\{A_n > c \,|\, (\bar X, n^{-1}\sum_{i=1}^n X_i^2, \bar Y, n^{-1}\sum_{i=1}^n Y_i^2)\} = \alpha,$$

with the footscript 0 indicating probability under $\rho = 0$.

(c) Write $\widehat\sigma_1^2 = n^{-1}\sum_{i=1}^n X_i^2 - \bar X^2$ and $\widehat\sigma_2^2 = n^{-1}\sum_{i=1}^n Y_i^2 - \bar Y^2$ for the empirical variances. Explain that in the conditional situation, given values for $\bar X, \bar Y, \widehat\sigma_1, \widehat\sigma_2$, the requirement above can be transformed to $\Pr_0\{R_n > d \,|\, (\bar X, \bar Y, \widehat\sigma_1, \widehat\sigma_2)\} = \alpha$, with R_n the empirical correlation coefficient, see (2.11).

(d) Then explain that R_n is actually independent of $\bar X, \bar Y, \widehat\sigma_1, \widehat\sigma_2$, under the null, so that the optimal test becomes the simpler one, of rejecting if $R_n > d$, where $\Pr_0(R_n > d) = \alpha$.

(e) For this optimal test regime to be specified fully, it remains to find the null distribution of R_n. Explain that this null distribution does not depend on the actual values of $(\xi_1, \xi_2, \sigma_1, \sigma_2)$, which we therefore may take to be $(0, 0, 1, 1)$. For practical purposes one may simulate a high number of R_n and then read off the d quantile; explain also that $\sqrt{n}R_n \to_d \mathrm{N}(0,1)$, from Ex. 2.48, so that $z_{1-\alpha}/\sqrt{n}$ is an approximation to the upper α point. We show however in the following point that $T_n = m^{1/2}R_n/(1-R_n^2)^{1/2}$ has a t_m distribution, under the null, with $m = n-2$ degrees of freedom. With e.g. $n = 25$, find the 0.95 quantile of the R_n distribution.

(f) So let us find the null distribution for R_n. Write $s_x^2 = \sum_{i=1}^n (X_i - \bar X)^2$, $s_y^2 = \sum_{i=1}^n (Y_i - \bar Y)^2$, $s_{xy} = \sum_{i=1}^n (X_i - \bar X)(Y_i - \bar Y)$, so that $R_n = s_{xy}/(s_x s_y)$. We may now use results from linear regression of the Y_i with respect to the X_i. Study in particular the least squares estimators $\widehat a, \widehat b$, minimising $Q(a,b) = \sum_{i=1}^n \{Y_i - a - b(X_i - \bar X)\}^2$, as with Ex. 3.30. Using results from that exercise, show (i) that $\widehat a = \bar Y$ and $\widehat b = s_{xy}/s_x^2$; (ii) that $Q_0 = Q(\widehat a, \widehat b) = s_y^2 - s_x^2\widehat b^2$; and (iii) that $T = \widehat b/\{(Q_0/m)^{1/2}/s_x\} = m^{1/2}s_x\widehat b/Q_0^{1/2}$ has the t_m distribution, with $m = n-2$ degrees of freedom. Show that this implies

$$R_n = \widehat b\frac{s_x}{s_y} = \frac{s_x\widehat b}{(Q_0 + s_x^2\widehat b^2)^{1/2}} = \frac{T/m^{1/2}}{(1 + T^2/m)^{1/2}}.$$

Solve for T to get the $T = m^{1/2}R_n/(1-R_n^2)^{1/2}$, pointed to and used in the previous point. Go on to work out the density of R_n, via the density g_m for the t_m; show that it becomes

$$h_n(r) = g_m\Big(\frac{m^{1/2}r}{(1-r^2)^{1/2}}\Big)\frac{m^{1/2}}{(1-r^2)^{3/2}} = \frac{\Gamma(\frac12(m+1))}{\Gamma(\frac12 m)\sqrt{\pi}}(1-r^2)^{(m-2)/2}$$

for $r \in (-1,1)$. As a curiosum, check that it is U-shaped for $n = 3$, uniform for $n = 4$, and then bell-shaped for $n \geq 5$.

Ex. 4.36 *Inference for linear multiple regression.* [xx to be done. point back to Ex. 3.31, 3.32, 3.33, and point to one or two stories. xx] in this exercise we show typical and not so typical inference methods for the linear multiple regression model, using the key results reached in the previous exercise. confidence intervals, tests, also for σ, for a p quantile $F^{-1}(q\,|\,x_0) = x_0^{\rm t}\beta + z_q\sigma$, and delta method for things like $\Pr(Y \le y_0\,|\,x_0)$. and prediction. xx]

(a) (xx typical things first. show that $\widehat{\beta}_j \sim {\rm N}(\beta_j, k_j^2\sigma^2/n)$, where $k_j^2 = \sigma_n^{j,j}$ the diagonal elements of Σ_n^{-1}. from this show $t_j = (\widehat{\beta}_j - \beta_j)/(k_j\widehat{\sigma}/\sqrt{n})$ is a t_{n-p}. then ci for each β_j. and test of $\beta_j = 0$. also ok for $\beta_j - \beta_k$ etc. xx)

(b) (xx inference for σ. xx)

(c) For a given individual, with covariate vector x_0, the outcome Y_0 has the distribution ${\rm N}(x_0^{\rm t}\beta, \sigma^2)$. Consider the inference task for a quantile in this distribution. Show that the q-quantile becomes $\gamma_q = x_0^{\rm t}\beta + z_q\sigma$, with $\Phi(z_q) = q$. With estimator $\widehat{\gamma}_q = x_0^{\rm t}\widehat{\beta} + z_q\widehat{\sigma}$, show that

$$W_q = \frac{\widehat{\gamma}_q - \gamma_q}{\widehat{\sigma}} = \frac{x_0^{\rm t}(\widehat{\beta}-\beta) + z_q(\widehat{\sigma}-\sigma)}{\widehat{\sigma}} = \frac{{\rm N}(0, x_0^{\rm t}\Sigma_n^{-1}x_0/n) + z_q\{(\chi^2_m/m)^{1/2} - 1\}}{(\chi^2_m/m)^{1/2}}.$$

(xx round off. the point is that W_q can be simulated. also approximated with a normal. give data example. xx)

(d) (xx prediction, what will Y_0 be, at position x_0. also $\Pr(Y \le y_0\,|\,x_0)$. xx)

Ex. 4.37 *How much of the variance is explained?* (xx the below to be polished and illustrated, and with a clearer link to R^2. exact cc(ρ) to come in Ch. 7. point to illustration in Story i.6. xx) In the linear regression model, the extent to which the covariates influence the outcomes may be assessed in several ways, one of which is to decompose the variance of the outcomes into a covariate part and a 'remaining variability' part. Such assessments relate also to 'signal plus noise' viewpoints; how strong is the signal?

(a) We start out writing the regression model as

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i = \beta_0 + x_i^{\rm t}\beta + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

again with the ε_i seen as i.i.d. ${\rm N}(0, \sigma^2)$, and further assume that the covariates have been centred, having their means subtracted, so that $\sum_{i=1}^n x_{i,j} = 0$ for $j = 1, \ldots, p$. This gives β_0 the interpretation as the overall mean of the Y_i. With $\Sigma_n = (1/n)\sum_{i=1}^n x_i x_i^{\rm t}$ the empirical $p \times p$ variance matrix for the x_i, show that the least squares estimators become

$$\widehat{\beta}_0 = \bar{Y} \sim {\rm N}(\beta_0, \sigma^2/n), \quad \widehat{\beta} = \Sigma_n^{-1}(1/n)\sum_{i=1}^n x_i Y_i \sim {\rm N}_p(\beta, (\sigma^2/n)\Sigma_n^{-1}),$$

and that these two are independent.

(b) Write $\widehat{\mu}_i = \widehat{\beta}_0 + x_i^{\rm t}\widehat{\beta}$ for the model based estimate of the outcome at x_i. For the sum of squared residuals, show that $Q_0 = \sum_{i=1}^n (Y_i - \widehat{\mu}_i)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 - n\widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta}$. In other words,

$$V_n = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \widehat{\mu}_i)^2 + n\widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta},$$

a neat decomposition of the full variability of outcomes as a sum of squared residuals and the covariate part $n\widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta}$.

(c) (xx place a little caveat below, regarding interpretation; we need to think about the population being sampled from. xx) Standard themes for the linear regression model are developed with analyses carried out conditional on the covariates. Allow now a change in this narrative, where the x_i are seen as having their own covariate distribution, with mean zero and variance matrix Σ_n. Show that a randomly selected outcome Y_i then has variance $\beta^{\rm t}\Sigma_n\beta + \sigma^2$. Show that the covariate part of the full variability becomes

$$\rho = \frac{\beta^{\rm t}\Sigma_n\beta}{\beta^{\rm t}\Sigma_n\beta + \sigma^2} = \frac{\lambda}{\lambda+1}, \quad \text{with } \lambda = \beta^{\rm t}\Sigma_n\beta/\sigma^2.$$

With $\widetilde{\sigma}^2 = Q_0/n$ (rather than the unbiased $\widehat{\sigma}^2 = Q_0/(n-p-1)$), show that this leads to

$$\widetilde{\rho} = \frac{\widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta}}{\widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta} + \widetilde{\sigma}^2} = \frac{V_n - \sum_{i=1}^n (Y_i - \widehat{\mu}_i)^2}{V_n} = 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{\mu}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

This is often called the coefficient of determination, or R^2.

(d) To carry out precise inference for ρ, show first that $n\widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta}/\sigma^2 \sim \chi^2_p(n\lambda)$; check with Ex. 1.48. Then show that with $\widehat{\lambda} = \widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta}/\widehat{\sigma}^2$, we have

$$F = n\widehat{\lambda}/p = n\widehat{\beta}^{\rm t}\Sigma_n\widehat{\beta}/(p\widehat{\sigma}^2) \sim F(p, m, n\lambda),$$

the noncentral F, see Ex. **??**, with $m = n - (p+1)$ the degrees of freedom for $\widehat{\sigma}^2$.

(e) Explain how this may be used to set confidence intervals for λ and hence for ρ. You may check with Ex. 7.15 how this may be used to construct full confidence distributions for the fraction ρ of variation explained by the covariates; for an illustration, see Story i.6.

Ex. 4.38 *Inference for ratios of standard deviations.* Suppose two independent samples, of sizes n_1 and n_2, come from two populations, with standard deviations σ_1 and σ_2. From the empirical standard deviations $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$, form the ratio $R = \widehat{\sigma}_1/\widehat{\sigma}_2$, to be used for inference about the underlying ratio $\rho = \sigma_1/\sigma_2$.

(a) Suppose first that the two distributions are normal. We then saw in Ex. 1.49 that $R^2 = \rho^2 F$, where $F \sim F_{m_1,m_2}$, an F distribution with degrees of freedom $(m_1, m_2) = (n_1 - 1, n_2 - 1)$. Construct a 95 percent confidence interval for ρ based on this. Also give a 0.05 level test for the equality hypothesis $\sigma_1 = \sigma_2$. this gives c.i. for ρ, and tests for $\rho = 1$. (xx answer: with $\Pr(a \le F \le b) = 0.95$, with (a, b) found from quantiles of the F, we find $[R/b^{1/2}, R/a^{1/2}]$. test: accept if $a < R^2 < b$. xx)

(b) Then consider inference for the ratio ρ outside the assumption of normally distributed data. From Ex. 3.12, find the representation

$$R = \frac{\widehat{\sigma}_1}{\widehat{\sigma}_2} \doteq \frac{\sigma_1}{\sigma_2} \frac{1 + (1/n_1)^{1/2}(\frac{1}{2} + \frac{1}{4}\gamma_{4,1})^{1/2} N_{n_1}}{1 + (1/n_2)^{1/2}(\frac{1}{2} + \frac{1}{4}\gamma_{4,2})^{1/2} N_{n_2}},$$

in terms of the kurtoses $\gamma_{4,1}$ and $\gamma_{4,2}$, where N_{n_1} and N_{n_2} are independent variables tending to standard normals as sample sizes increase. Use the delta method to deduce that $R/\rho \approx_d \mathrm{N}(1, \tau^2)$, with $\tau^2 = (1/n_1)(\frac{1}{2} + \frac{1}{4}\gamma_{4,1}) + (1/n_2)(\frac{1}{2} + \frac{1}{4}\gamma_{4,2})$. For normal data, show that this matches the distributional approximation $F^{1/2} \approx_d \mathrm{N}(1, 1/(2n_1) + 1/(2n_2))$.

(c) Construct an approximate 95 percent confidence interval for σ_1/σ_2, valid also outside normal data. For an application, see the Bach and Reger Story ii.12.

(d) (xx one more point. xx)

Ex. 4.39 *Testing equality across groups.* (xx mention below that $k = 2$ is simple, with pairwise comparisons, t type things. xx) Suppose a parameter has the same interpretation across groups, say θ_j for groups $j = 1, \ldots, k$. associated with estimators, perhaps of different precision. How can we test the null hypothesis of no difference, i.e. $H_0 \colon \theta_1 = \cdots = \theta_k$?

(a) Suppose first that $Y_j \sim \mathrm{N}(\theta_j, \sigma^2)$, with independence, and the same known precision, i.e. σ. Explain from classic results of Ex. 1.44–1.45, that $Z = \sum_{j=1}^k (Y_j - \bar{Y})^2/\sigma^2 \sim \chi^2_{k-1}$, with $\bar{Y} = (1/k)\sum_{j=1}^k Y_j$ the natural estimator for the common parameter under H_0. Explain that in the case of σ not being known, we still have $Z^* = \sum_{j=1}^k (Y_j - \bar{Y})^2/\widehat{\sigma}^2 \to_d \chi^2_{k-1}$, as long as $\widehat{\sigma}$ is consistent, with growing information for each group.

(b) A useful generalisation of the basic results for i.i.d. normal data is as follows. Suppose $Y_j \sim \mathrm{N}(0, 1/a_j)$, with the a_j positive and $a = \sum_{j=1}^k a_j$. Define $Y^* = \sum_{j=1}^k (a_j/a) Y_j$ and then $Z = \sum_{j=1}^k a_j (Y_j - Y^*)^2$. Show that $Z = \sum_{j=1}^k a_j Y_j^2 - a(Y^*)^2$, that it is independent of Y^*, and that its distribution is a χ^2_{k-1}. The classical case, leading to the distribution of the empirical variance and then to the t test, corresponds to all $a_j = 1$.

(c) A natural variation of the previous point is as follows. Assume in general terms that $Y_j \sim \mathrm{N}(\theta, \sigma_j^2)$, with independence, for $j = 1, \ldots, k$. Show that the minimiser of $Q(\theta) = \sum_{j=1}^k (Y_j - \theta)^2/\sigma_j^2$ is $\widehat{\theta} = (\sum_{j=1}^k Y_j/\sigma_j^2)/(\sum_{j=1}^k 1/\sigma_j^2)$, and that $Q_{\min} = Q(\widehat{\theta})$ is χ^2_{k-1}, independent of $\widehat{\theta}$. Demonstrate also that of all unbiased estimators $\theta^* = \sum_{j=1}^k c_j Y_j$ of θ, $\widehat{\theta}$, with weights proportionao to inverse variances, has the smallest variance. Again, if the σ_j are consistently estimated, as opposed to known, explain that with $\theta^* = (\sum_{j=1}^k Y_j/\widehat{\sigma}_j^2)/(\sum_{j=1}^k 1/\widehat{\sigma}_j^2)$, we still have $Q^* = \sum_{j=1}^k (Y_j - \theta^*)^2/\widehat{\sigma}_j^2$ tending to χ^2_{k-1}.

(d) Consider correlations $\rho_1, \ldots, \rho_k$ for k groups of binormal data, with sample sizes $n_1, \ldots, n_k$, summing to the total n. Via Ex. 2.48, explain that $n_j^{1/2}(\widehat{\zeta}_j - \zeta_j) \to_d \mathrm{N}(0, 1)$, where $\widehat{\zeta}_j = \frac{1}{2}\log\{(1 + \widehat{\rho}_j)/(1 - \widehat{\rho}_j)\}$ is Fisher's zeta transform. Deduce that $Q = \sum_{j=1}^k n_j(\widehat{\zeta}_j - \zeta^*)^2 \to_d \chi^2_{k-1}$, with $\zeta^* = \sum_{j=1}^k (n_j/n)\widehat{\zeta}_j$. This is accordingly a natural test statistic for $\rho_1 = \cdots = \rho_k$.

(e) Situations similar to that of the previous point abound in applied statistics, when needing to compare more groups than two. Suppose $\widehat{p}_j = Y_j/n_j$ is estimating a binomial p_j, for independent experiments, for $j = 1, \ldots, k$. Then form $\widehat{p} = \sum_{j=1}^k (n_j/n)\widehat{p}_j$, and show that under $p_1 = \cdots = p_k$ conditions, $Q = \sum_{j=1}^k n_j(\widehat{p}_j - \widehat{p})^2 \to_d p(1-p)\chi^2_{k-1}$, using p to denote the common value of the p_j. Construct a test for equality of the p_j.

(f) (xx one more. could test equality of medians in k groups. xx)

Ex. 4.40 *Testing equality across groups, vector parameter case.* In Ex. 4.39 we developed a recipe for testing equality across groups for any one-dimensional parameter. Here we lift those methods to the case where the $\theta_1, \ldots, \theta_k$ in question have dimension $p \geq 2$. We wish again to test equality of these. (xx to be used in Story iii.7. xx)

(a) Assume $Y_j \sim \mathrm{N}_p(\theta_j, \Sigma_j)$ for $j = 1, \ldots, k$, with known positive definite variance matrices. Under H_0 of $\theta_1 = \cdots = \theta_k$, show that $Q(\theta) = \sum_{j=1}^k (Y_j - \theta)^{\mathrm{t}} \Sigma_j^{-1} (Y_j - \theta)$ is a χ^2_{kp}, and that it is minimised by $\widehat{\theta} = A^{-1} \sum_{j=1}^k \Sigma_j^{-1} Y_j$, with $A = \sum_{j=1}^k \Sigma_j^{-1}$. Show further (i) that $\widehat{\theta} \sim \mathrm{N}_p(\theta, A^{-1})$; (ii) that the full ensemble of $Y_j - \theta$ is independent of $\widehat{\theta}$; and (iii) that the natural test statistic $Q_0 = Q(\widehat{\theta}) = \sum_{j=1}^k (Y_j - \widehat{\theta})^{\mathrm{t}} \Sigma_j^{-1} (Y_j - \widehat{\theta})$ is a $\chi^2_{(k-1)p}$.

(b) This Q_0 may be used as a test statistic for $\theta_1 = \cdots = \theta_k$, with an exact chi-square null distribution, provided the Σ_j matrices are known. In various applications these are estimated from data, with $\widehat{\Sigma}_j$ using data for group j. This leads to $\widehat{\theta} = \widehat{A}^{-1} \sum_{j=1}^k \widehat{\Sigma}_j^{-1} Y_j$, with $\widehat{A} = \sum_{j=1}^k \widehat{\Sigma}_j^{-1}$. Show that if these $\widehat{\Sigma}_j$ are consistent, then $Q^* = Q^*(\widehat{\theta}) = \sum_{j=1}^k (Y_j - \widehat{\theta})^{\mathrm{t}} \widehat{\Sigma}_j^{-1} (Y_j - \widehat{\theta})$ tends to the $\chi^2_{(k-1)p}$, under the null.

(c) Suppose we have normal datasets with sample sizes n_j, yielding the usual parameter estimates $(\widehat{\xi}_j, \widehat{\sigma}_j)$, for groups $j = 1, \ldots, k$, From these compute $\widehat{\xi} = \sum_{j=1}^k (n_j/n)\widehat{\xi}_j$, the overall mean, and $\widehat{\sigma} = \prod_{j=1}^k \widehat{\sigma}_j^{n_j/n}$. Show that

$$Q = \sum_{j=1}^k \{n_j(\widehat{\xi}_j - \widehat{\xi})^2 + 2n_j(\log \widehat{\sigma}_j - \log \widehat{\sigma})^2\}$$

tends to $\chi^2_{2(k-1)}$ under the hypothesis of common (ξ, σ). Write also down the Wilks test, based on attainde log-likelihood maxima. (xx do both, for the mothers and babies story, for the three ethnic groups. xx)

(d) Consider a probability distribution $b = (b_1, \ldots, b_s)$ for suitable outcomes $1, \ldots, s$, and suppose we have multinomial estimators $\widehat{b}_j = (\widehat{b}_{j,1}, \ldots, \widehat{b}_{j,s})$ for each of groups $j = 1, \ldots, r$, resulting from sample sizes $n_1, \ldots, n_k$, with total sum n. In other words, $\widehat{b}_j = (N_{j,1}, \ldots, N_{j,s})/n_j$, for the multinomial experiment for group j. The hypothesis to be tested is that of equal probability distribution across groups. Explain that $\widehat{b}_j \approx_d \mathrm{N}_s(b, \Sigma/n_j)$, where Σ_j has $b_j(1 - b_j)$ on the diagonal and $-b_j b_\ell$ outside. As in Story vii.1, write Σ_0 for the $(s-1) \times (s-1)$ submatrix of Σ, needed since Σ does not have full rank, with $\widehat{b}_0$ and $\widehat{b}_{j,0}$ similarly being the shorter $(s-1)$-length vectors, so that

the general method above can be applied. Then show that the test statistic Q_0 from the general recipe above can be written

$$\sum_{j=1}^k n_j(\widehat{b}_{j,0}-\widehat{b}_0)^{\mathrm{t}}\widehat{\Sigma}_0^{-1}(\widehat{b}_{j,0}-\widehat{b}_0)=\sum_{j=1}^k n_j\sum_{r=1}^s\frac{(\widehat{b}_{j,r}-\widehat{b}_r)^2}{\widehat{b}_r}=\sum_{j=1}^k\sum_{r=1}^s\frac{(N_{j,r}-E_{j,r})^2}{E_{j,r}},$$

which then tends to $\chi^2_{(k-1)(s-1)}$ under the null. Show that this is actually precisely identical to the Pearson statistic from Story iii.7, for testing independence between the two factors 'group' and 'bin'.

Ex. 4.41 *One-way layout.* (xx a point about $k=2$ at the end, where F i suppose becomes simply t^2, with usual t test. xx) The most prominent special case of the setup from Ex. 4.39 is that with an i.i.d. normal sample for each of the groups in question, with the same variance for all data. Consider therefore $Y_{j,1},\ldots,Y_{j,n_j}$ i.i.d. from $\mathrm{N}(\xi_j,\sigma^2)$, for group j, and with overall sample size $n=\sum_{j=1}^k n_j$. This is the so-called normal one-way layout, with the chief hypothesis to be tested is $H_0\colon \xi_1=\cdots=\xi_k$.

(a) Let $\bar{Y}_j=(1/n_j)\sum_{r=1}^{n_j}Y_{j,r}$ be the group averages. Explain that we for the overall average have $\bar{Y}=(1/n)\sum_{j=1}^k\sum_{r=1}^{n_j}Y_{j,r}=\sum_{j=1}^k(n_j/n)\bar{Y}_j$. Show then that $Q=\sum_{j=1}^k n_j(\bar{Y}_j-\bar{Y})^2$ has $\mathrm{E}\,Q=(k-1)\sigma^2+\sum_{j=1}^k n_j(\xi_j-\bar{\xi})^2$, with $\bar{\xi}=\sum_{j=1}^k(n_j/n)\xi_j$. Explain that $Q\sim\sigma^2\chi^2_{k-1}$, under equality of means, so that the test statistic $W=Q/\widehat{\sigma}^2$ approximately has the χ^2_{k-1} distribution under H_0, as long as $\widehat{\sigma}$ is consistent. This is already a satisfactory answer to the one-way layout testing problem, without further fine-tuning.

(b) In this setup further fine-tuning is available, however. Show that $Q_0=\sum_{j=1}^k\sum_{r=1}^{n_j}(Y_{j,r}-\bar{Y}_j)^2\sim\sigma^2\chi^2_{n-k}$, making $\widehat{\sigma}^2=Q_0/(n-k)$ the natural unbiased estimator of σ^2. Explain also that Q_0 is independent of Q. Under the null of equal means, then, deduce that

$$F=\frac{Q/(k-1)}{Q_0/(n-k)}=\frac{Q/(k-1)}{\widehat{\sigma}^2}=\frac{\sum_{j=1}^k n_j(\bar{Y}_j-\bar{Y})^2/(k-1)}{\sum_{j=1}^k\sum_{r=1}^{n_j}(Y_{j,r}-\bar{Y}_j)^2/(n-k)}$$

has the F distribution $F_{k-1,n-k}$. Checking whether an observed F is too big, compared to this null distribution, is then a way to assess the hypothesis $\xi_1=\cdots=\xi_k$. For an illustration, see Story i.6.

(c) (xx a short thing estimating contrasts. xx)

Ex. 4.42 *Testing equality of multinormal means.* (xx a brief thing, used in Story ii.7, can point to ML theory too. establish the following, then apply in two-three settings. xx) Suppose $A\sim\mathrm{N}_p(a,\Sigma_1)$ and $B\sim\mathrm{N}_p(b,\Sigma_2)$ are independent multinormal data vectors, with known variance matrices. How can we test $a=b$? Show that $W=(B-A)^{\mathrm{t}}(\Sigma_1+\Sigma_2)^{-1}(A-B)\sim\chi^2_p$ under the null.

Notes and pointers

(xx confidence intervals. testing. connections. power. Neyman–Pearson. point to Lehmann. and to later chapters, Ch. 7 for CDs. point to interplay between modelling, probability calculus, thinking, a bit of philosophy, and practice. xx)

(xx one-way layout: primarily to test equality of means, so 'anova' is arguably a slight misnomer. xx)

Basu's lemma: from Basu (1955), see also Ghosh (2002).

ToDo notes, as of 12-August-2024:

Lots, though the chapter is shaping up. There are lots of 'test $\theta = \theta_0$ against $\theta > \theta_0$ prose in exercises, since it's easiest and cleanest, with NP etc. But we need to say a couple of times that all of this generalises to $\theta \neq \theta_0$ etc.

the Lindqvist and Taraldsen things, for simulating from $U(x) \,|\, \{V(x) = v\}$. do the handball model from CLP.

Do the bioequivalence: test the H_0 that θ is outside $[-\varepsilon, \varepsilon]$ against the alternative that it is inside. Diffent type of tests, and different looking power function.

Do sample size things, and efficiency; $\pi_n(\theta) \doteq \Phi(\sqrt{n}(\theta - \theta_0)/\sigma - z_0)$, with informative derivative $\phi(0)\sqrt{n}/\sigma$ at θ_0. Efficiency things.

Make an example or two, perhaps with Cauchy again, to see that the confidence region might not be an interval.

Point to Cox (1958) and also interviews with him regarding conditional stuff.

I.5

Minimum divergence and maximum likelihood

Consider the joint density of a dataset Y from a parametric model, say $f_{\text{full}}(y, \theta)$. The likelihood function, a fundamental concept in parametric inference, is just this density, but seen as a function of θ, with y fixed at the observed dataset y_{obs}. In this chapter we go through the fundamental likelihood inference methods, in particular with results for the maximum likelihood estimator, the maximiser of the likelihood, and for the attained maximum value itself. It is useful to see this general likelihood theory as a special case of the more flexible machinery of minimum divergence methods. Here statistical divergences, often seen as a distance from one fixed model to a collection of approximation models, lead to empirical divergence functions, then to be minimised to find the best approximation. In fact we develop general theory for minimum divergences first, below, after which likelihod theory is a relatively easy consequence, in that maximum likelihood corresponds to one particular divergence, namely the Kullback–Leibler. The divergence and likelihood methods are practical and versatile, as is demonstrated in several exercises and stories, also for say non-standard regression models. With models outside the most familiar ones, inference analysis essentially flows from being able to programme the log-likelihood function or divergence function. Further material connected to likelihood theory are Cramér–Rao information inequalities, Wilks theorems, influence functions, and certain flexible robustification methods. Crucially, the theory developed does not in general presuppose that the parametric model worked with is correct, as results are established both under and outside the precise model conditions. (xx perhaps there is room for empirical likelihood. xx)

Key words: BHHJ, Cramér–Rao lower bounds, Fisher information matrix, influence function, Kullback–Leibler, least false parameters, log-likelihood, maximum likelihood, minimum divergence, regression models, score function, Wilks tests

In earlier chapters we have met and worked with several methods for estimation parameters in different settings. Sometimes one estimates population parameters directly from data, like the mean, the median, the standard deviation, the skewness, the correlation, the median difference between groups, etc., without necessarily using parametric models. Very frequently, however, the most fruitful data analyses involve fitting some parametric model to the data, as in regression models, where regression coefficients are estimated and assessed to learn how covariates influence the main outcomes. In Chs. 3-4 we have

already worked with methods associated with moment fitting, quantile fitting, and least squares for regression models, but the present chapter has a wider aim and applicability, specifically developing methodology for *minimum divergence function estimators*, and its primary special case, *maximum likelihood estimators*.

In general terms, suppose data y have been modelled via a parametric model, leading to a joint probability density $f_{\rm full}(y,\theta)$, with θ in some relevant parameter region Θ. Then the *likelihood function* is $L(\theta) = f_{\rm full}(y_{\rm obs},\theta)$, studied as a function of the parameters, with y held fixed at the observed $y_{\rm obs}$. The maximum likelihood (ML) estimator is the value $\widehat\theta$ maximising the likelihood, or equivalently the *log-likelihood function* $\ell(\theta) = \log f_{\rm full}(y_{\rm obs},\theta)$. The simplest setup is that of i.i.d. observations $Y_1,\ldots,Y_n$ to be fitted to some parametric $f(y,\theta)$, with θ a parameter vector inside its relevant parameter region. The log-likelihood function is then

the ML estimator

log-likelihood function

$$\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i,\theta). \tag{5.1}$$

The apparatus of minimum divergence function and likelihood estimation carries over to regression models too, where the density of Y_i given a covariate vector x_i is modelled via some $f(y_i \,|\, x_i,\theta)$.

Generally speaking there are two valid viewpoints when developing the required theory. The first takes *the model to be correct*, so there is a true parameter θ_0 to be estimated, with associated inference. The second takes the parametric model to be a sensible approximation to the real and unknown data-generating mechanism. With i.i.d. data, the model density $f(y,\theta)$ is consequently seen as an approximation to the real underlying density $g(y)$. Estimation and inference then involves the *least false parameter* θ_0, in a suitable sense making $f(y,\theta_0)$ coming as close to the real $g(y)$ as possible.

A generic and powerful statistical idea is to estimate parameters by minimising a relevant distance or divergence from the true mechanism to the data. Our chapter starts with general ways in which to construct relevant *distance functions*, often tied to divergence ideas about the distance from one density to another, which in various settings define the best parameter θ_0 to be minimiser of such functions. This is then followed up by constructing the parameter estimator $\widehat\theta$ as the minimiser of an empirical version of the same distance function. The arguably most important estimation method, the *maximum likelihood method*, is the special case associated with the *Kullback–Leibler divergence* $\mathrm{KL}(g,f_\theta)$. Apart from clarifying the concepts, defining parameters and estimators, in such ways, developing the theory amounts to establishing clear results for the behaviour and performance of these estimators and related data-driven tests, confidence methods, etc. This again involves limiting normality, assessing and estimating variability, and so on. All of this leads to a fruitful and indeed practical general theory, which also in new situations, with models constructed for new purposes, allows the statistician to estimate, to reach relevant confidence statements, test and compare, predict, etc.

In brief, this chapter has these connected parts, with associated groups of exercises.

(i) Motivating and building an apparatus for defining parameters and their estimators via minimum divergence functions, initially for i.i.d. setups. The most important method, the maximum likelihood (ML) method, briefly described above, is a special case,

associated with the KL divergence.

(ii) Analysing ML methods involves studying score functions, information functions, and the Fisher information matrix, for general parametric models. This again relates to 'information inequalities', specifically of the Cramér–Rao type, establishing lower bounds for how small variances can be, with a given model and a given sample size n. The ML methods achieve these lower bounds, under model conditions, for growing n.

(iii) Methods are developed for deriving the basic properties, concerning limiting normality of estimators, chi-squared type results for profiled versions for focus parameters, and quadratic forms for classes of test criteria, via the mechanics of minimisers of random functions.

(iv) Importantly, the i.i.d. setup is then lifted to classes of regression models. Intrigruingly, with attention to certain crucial details, these extensions turn out to be not too strenuous, partly with Lindeberg theorem arguments replacing CLT details for the simpler case. The general likelihood theory in particular then makes it relatively straightforward to define and work with not only multiple linear regression, but logistic regression, Poisson regression, gamma type regressions, and more.

Carrying out data analyses using tools from this chapter is often even surprisingly straightforward, even in new situations with new models. This is partly due the general well-working theory, but also to the modern conveniences of software packages for numerical optimisation, easy calculation of derivatives and second derivatives, and so on, after having programmed the basic empirical distance function to be used. For maximum likelihood methodology, in particular, this is showcased in Story i.6, with a logistic full-data model for how covariates influences birthweights, Story vii.4, concerning models for time-to-failure of machine components, and in Story iv.6, with regression models for the number of bird species on islands outside Ecuador. These applications are also meant to inspire the invention of new parametric models in new situations, perhaps along with context relevant distance functions.

(xx can briefly point to extensive use of likelihood theory in Ch. 7, 10, 11, 12. to be clearer, also using Notes: M-estimators, Z-estimators. the general wondrous recipes around $\widehat{\theta} \approx_d \mathrm{N}_p(\theta_0, \widehat{J}_{\rm obs}^{-1})$. make robustness aspects clearer. xx)

Divergences, Kullback–Leibler, likelihoods

Ex. 5.1 *Maximising the log-likelihood.* (xx check that we don't repeat too similar things later on. xx) Here we work through some examples of setting up the log-likelihood function and finding the ML estimator.

(a) Suppose $Y \sim \text{binom}(100, \theta)$ and that you observe $y_{\rm obs} = 22$. Set up the log-likelihood function $\ell(\theta)$ and plot it. Show that its maximiser is $\widehat{\theta} = 0.22$.

(b) Simulate ten values of Y from the binomial $(100, 0.25)$, and plot the resulting ten log-likelihood functions. Note how they vary, giving different ML estimates of the underlying true $\theta_0 = 0.25$. With your code, experiment a bit with different sizes of the binomial n. For bigger n, the peak is sharper; show this mathematically.

(c) With i.i.d. observations $Y_1, \ldots, Y_n$ from some parametric $f(y, \theta)$, show that the log-likelihood function becomes $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$. With such data points from the

two-parameter normal $\mathrm{N}(\xi, \sigma^2)$ model, give a clear formula for the log-likelihood function $\ell_n(\xi, \sigma)$. Show that the ML estimators are $\widehat{\xi} = \bar{Y}$ and $\widehat{\sigma}^2 = Q_0/n$, with $Q_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Compared to the classical empirical variance $Q_0/(n-1)$, see Ex. 1.45, we learn that the ML estimator has a small negative bias, but that the difference is small for moderate to large n.

(d) With data $y_1, \ldots, y_{100}$ from the exponential density $\theta \exp(-\theta y)$, set up the log-likelihood function and find a formula for its maximiser $\widehat{\theta}$.

(e) Simulate 50 points from the uniform $[0, \theta]$, with true value $\theta_0 = 1$. Find and plot the log-likelihood function, and read off the ML estimate $\widehat{\theta}$.

(f) Simulate 100 values y_i from the $\mathrm{Gam}(a_0, 1)$ distribution (see Ex. 1.9), with e.g. $a_0 = 3.33$. Set up and plot the log-likelihood function. Set up a little experiment where you keep track of the ML estimators $\widehat{a}$ in repeated experiments from the samme $\mathrm{Gam}(a_0, 1)$.

Ex. 5.2 *The likelihood and log-likelihood functions.* Consider the one-parameter model with density $f(y, \theta) = \exp(-\theta y^{1/2})\theta/(2y^{1/2})$ for $y > 0$, and assume $n = 12$ data points have been observed:

```
0.233 0.334 0.067 0.148 0.007 0.639 0.017 0.298 0.030 0.120 0.061 0.063
```

(a) Show that $f(y, \theta)$ indeed is a density, write down the log-likelihood function $\ell_n(\theta)$, and show that it is maximised at $\widehat{\theta} = 1/W_n$, with $W_n = n^{-1} \sum_{i=1}^n y_i^{1/2}$. Find also a formula for the Hessian at the maximum position, $\widehat{J}_{\mathrm{obs}} = -\ell_n''(\widehat{\theta})$.

(b) Find the limit distribution of $\sqrt{n}(W_n - 1/\theta)$, using the CLT of Ch. 2, and use the delta method to find that the limit distribution of $\sqrt{n}(\widehat{\theta} - \theta)$ is $\mathrm{N}(0, \theta^2)$. You may verify already that this is what comes out of the general ML theory developed below, see e.g. Ex. 5.17.

(c) Anticipating general ML theory to come, check again with Ex. 5.17, it will be seen that $\widehat{\theta} \approx_d \mathrm{N}(\theta, 1/\widehat{J})$, under model conditions. Explain that this leads to approximate confidence intervals of the type $\widehat{\theta} \pm z_0/\widehat{J}^{1/2} = \widehat{\theta}(1 \pm z_0/\sqrt{n})$, with z_0 the appropriate normal quantile. We've actually simulated these few data points above from the model, with true parameter $\theta_0 = 3.33$. Construct a version of Figure 5.1, left panel.

(d) In general the ML estimator might have a complicated distribution (though it is approximately normal, as we have seen here). In this particular model its precise distribution may be worked out, however; show that $\widehat{\theta} \sim \theta\,(2n)/\chi^2_{2n}$. Use this to find a precise 90 confidence interval for θ, and compare to the approximation given above.

(e) To simulate data from this model, show that Y_i is equal in distribution to $(V_i/\theta_0)^2$, with the V_i i.i.d. from the unit exponential. Make a computer programme to simulate n points from the $f(y, \theta_0)$ model, and which then finds the log-likelihood function, the ML estimator, and the approximate 90 percent confidence interval, as above. Run such a programme for say 88 more points, forming a bigger dataset with $n = 100$ datapoints, and comment on what you find. Produce a version of Figure 5.1, right panel. Comment on

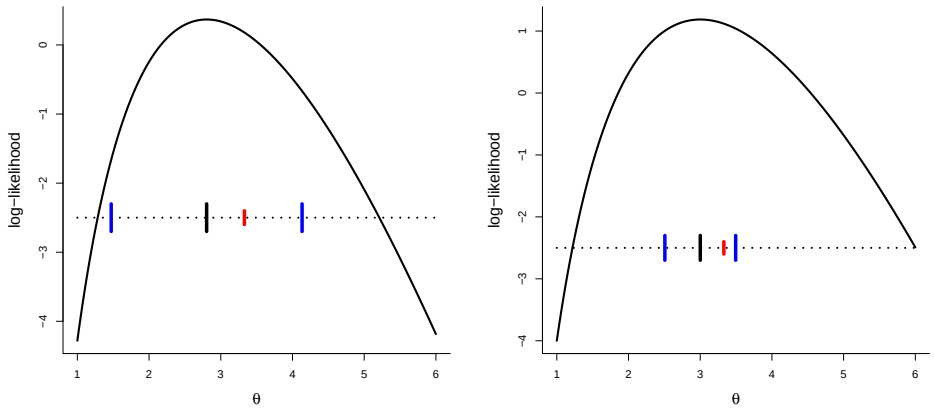


Figure 5.1: *Left panel: The log-likelihood function* $\ell_n(\theta)$ *for the simple* $n = 12$ *dataset of Ex. 5.2. The maximum is attained at* $\widehat{\theta} = 2.803$. *The blue bars indicate the 90 percent confidence interval coming from standard ML estimation theory, see Ex. 5.17, and is* $[1.472, 4.134]$. *The true value behind the data,* $\theta_0 = 3.33$, *is indicated as the red bar. Right panel: as for the left, but now with the bigger data set of* $n = 100$ *datapoints, from the same distribution, where the first 12 are as above. The minus second derivative* $\widehat{J}$ *at the ML position has increased from 1.527 to 11.103, causing the ML based confidence interval to become considerably tighter, and is* $[2.507, 3.494]$.

the main features here, including that $\widehat{J}$ becomes bigger with more data, yielding sharper confidence intervals. We would usually plot the log-likelihood and related aspects, as the confidence curves of Ch. 7 for a shorter range of parameter values than in this right panel, but we here choose to plot using the same range for both $n = 12$ and $n = 100$.

Ex. 5.3 *Minimising* L_2 *distance.* Suppose i.i.d. data $Y_1, \ldots, Y_n$ come from some data generating density g which we wish to approximate with some parametrically modelled f_θ. The L_2 distance between g and f_θ is

$$D(g, f_\theta) = \int (g - f_\theta)^2 \, \mathrm{d}y = \int f_\theta^2 \, \mathrm{d}y - 2 \int g f_\theta \, \mathrm{d}y + a(g),$$

where $a(g)$ does not depend on θ.

(a) Use this to motivate what we may call the minimum L_2 estimator $\widehat{\theta}$, the minimiser of $D_n(\theta) = \int f_\theta^2 \, \mathrm{d}y - 2n^{-1} \sum_{i=1}^n f(Y_i, \theta)$. Operationally, this is easiest when there is a closed-form formula for $\int f_\theta^2 \, \mathrm{d}y$, but numerical minimisation might be carried out even without this. Simulate 100 detapoints from a $\mathrm{Gam}(a, b)$, where you choose (a, b) as you wish, and estimate these using this method.

(b) Carry out a similar simple experiment with 100 datapoints drawn from a normal, i.e. estimate the mean and standard deviation using the minimum L_2 method. Then do the following variation: push one of the 100 datapoints far away from the others, which

will cause the traditional estimates of μ and σ to be off; show however that the minimum L_2 estimates are far less affected. – The minimum L_2 method is a special case of the BHHJ method, see Ex. 5.9.

Ex. 5.4 *Estimating means via minimum divergence.* Suppose again i.i.d. data $Y_1, \ldots, Y_n$ stem from some data generating density g. The point to convey here is that key parameters can be seen as minimisers of natural distance functions, or divergences, leading to recipes for estimating them.

(a) For any distribution G, let $\xi(G)$ be the minimiser of $H(\xi) = \mathrm{E}_G\,(Y-\xi)^2$, with estimator the minimiser $\widehat{\xi}$ of the empirical version $H_n(\xi) = n^{-1}\sum_{i=1}^n (Y_i-\xi)^2$. Show, for this simple illustration of minimum divergence function methods, that $\xi(G)$ is the mean $\mathrm{E}_G\, Y$ and that $\widehat{\xi} = \bar{Y}$, the sample mean.

(b) For characterising and then estimating both mean and standard deviation, consider $h(y,\xi,\sigma) = (y-\xi)^2 + \{y^2 - (\xi^2+\sigma^2)\}^2$, and let (ξ,σ) be the minimiser of the distance function $H(\xi,\sigma) = \mathrm{E}_G\, h(Y,\xi,\sigma)$. Show that $\xi(G)$, $\sigma(G)$ are the mean and standard deviation, and work out expressions for $\widehat{\xi}$, $\widehat{\sigma}$, the minimisers of the empirical distance function $H_n(\xi,\sigma) = n^{-1}\sum_{i=1}^n h(Y_i,\xi,\sigma)$.

(c) For positive Y_i data, to be fitted to the Gam(a,b), consider the function $h(y,a,b) = (y-a/b)^2 + (y^2 - (a/b^2 + a^2/b^2))^2$, constructed in view of a/b and a/b^2 being the formulae for the mean and variance for a Gamma. Explain that the empirical version of the distance function $H(a,b) = \mathrm{E}_G\, h(Y,a,b)$ becomes $H_n(a,b) = n^{-1}\sum_{i=1}^n \{(Y_i-a/b)^2 + (Y_i^2 - (a/b^2 + a^2/b^2)\}^2$. Show that the minimum divergence function estimators $(\widehat{a},\widehat{b})$ are equivalent to the moment estimators, worked with in Ex. 3.25, i.e. the solution to the two equations $\bar{Y} = a/b$ and $\widehat{\sigma}^2 = a/b^2$.

Ex. 5.5 *Minimum divergence function estimators: general setup.* We learn from Ex. 5.3–5.4 that classes of estimators may be formed via minimisation of suitable empirical distance functions; these estimate the corresponding minimisers of distance functions operating on the underlying distributions. We call these *minimum divergence function estimators* or *minimum distance estimators*. In general terms, for observations $Y_1, \ldots, Y_n$ from some distribution G, consider a parameter $\theta_0 = \theta(G)$ defined as the minimiser of the function $H(\theta) = \mathrm{E}_G\, h(Y,\theta)$, for a suitable $h(y,\theta)$. It is assumed that θ_0 thus defined, which may also be multidimensional, is the unique minimiser. The empirical version of $H(\theta)$ is $H_n(\theta) = n^{-1}\sum_{i=1}^n h(Y_i,\theta)$, so a natural estimator for θ_0 is $\widehat{\theta} = \mathrm{argmin}(H_n)$. In fact many important estimators are of this or related types, perhaps minimising somewhat more complicated random functions, as we shall see in this chapter. In exercises below we shall develop clear results for how the minimum divergence function estimators behave, under sets of natural assumptions, but the present exercise is meant to illustrate the basic construction via different types of examples.

minimum divergence function estimators

(a) Explain that $H_n(\theta)$ can be written $\int h(y,\theta)\,\mathrm{d}G_n(y)$, with G_n the empirical distribution, having mass $1/n$ at each datapoint; see Ex. 3.9. Explain why $H_n(\theta) \to_{\mathrm{pr}} H(\theta)$ for each θ, and find the limit distribution of $\sqrt{n}\{H_n(\theta) - H(\theta)\}$. What we need, tended to in several exercises to follow, are conditions under which $\widehat{\theta} = \mathrm{argmin}(H_n)$ tends to $\theta_0 = \mathrm{argmin}(H)$, along with a limit distribution.

(b) For one-dimensional Y_i, work through the details of $h(y,\theta) = (y-\theta)^2$. Then consider $h(y,\theta) = [\exp\{c(y-\theta)\}-1-c(y-\theta)]/c^2$, with c a balance parameter. Draw 100 datapoints from a normal $\mathrm{N}(\theta,1)$, with θ of your choice, and estimate θ in this minimum H_n fashion, for a few values of the balance parameter c. Show that c close to zero corresponds to the mean.

(c) Let generally $h(y,\theta) = \{r(y)-\theta\}^{\mathrm{t}} V \{r(y)-\theta\}$, for some $r(y) = (r_1(y),\ldots,r_p(y))$ and a symmetric positive definite matrix V. Show that $\theta_0 = \mathrm{E}_G\, r(Y)$ and that $\widehat{\theta} = n^{-1}\sum_{i=1}^n r(Y_i)$.

(d) Consider $h(y,\xi,\tau) = p(\tau) + \frac{1}{2}(y-\xi)^2/\tau^2$, where $p(\tau)$ is a smooth increasing function of $\tau > 0$. Find a recipe for computing the estimates $(\widehat{\xi},\widehat{\tau})$ associated with the distance function $n^{-1}\sum_{i=1}^n h(Y_i,\xi,\tau)$. Check in particular the case of $p(\tau) = \log\tau$.

(e) Consider $h_0(x) = x\arctan x - \frac{1}{2}\log(1+x^2)$, and define θ_0 as the minimiser of $\mathrm{E}_G\, h_0(Y-\theta)$. Show that $\widehat{\theta}$, the minimiser of $H_n(\theta) = n^{-1}\sum_{i=1}^n h_0(Y_i-\theta)$, is also the unique solution to $\sum_{i=1}^n \arctan(Y_i-\theta) = 0$. (xx so connection from minimum divergence estimator to M estimator. round off. xx)

(f) There are connections here to the moment matching estimation method, as worked with in Ex. 3.24. For a one-parameter model first, with $\mathrm{E}\,Y = m(\theta)$, consider $h(y,\theta) = \{y-m(\theta)\}^2$. Show that the implied best parameter is for $m(\theta_0) = \mathrm{E}\,Y$, or $\theta_0 = m^{-1}(\mathrm{E}\,Y)$, and that the minimum divergence function method leads to solving $m(\theta) = \bar{Y}$. Generalise to the case of there being two parameters in the model, and then to the general vector parameter case.

Ex. 5.6 *The Kullback–Leibler divergence and the maximum likelihood method.* With i.i.d. data $Y_1,\ldots,Y_n$ from a density g, to be approximated with a parametric f_θ, the particularly important *maximum likelihood estimation method* is worked with here, seen as the natural cousin to the Kullback–Leibler divergence as a measure of distance from the true density to the parametric approximation. Development in exercises below give a precise description of how the method actually behaves.

the Kullback–Leibler distance

(a) For two densities g and f, defined on a common support, the Kullback–Leibler distance, interpreted to be 'from the first density to the second', is

$$\mathrm{KL}(g,f) = \int g \log\frac{g}{f}\,\mathrm{d}y. \tag{5.2}$$

It is an important concept and tool for communication and information theory, as for probability theory and statistics. The $\log(g/f)$ term will be both positive and negative, in different parts of the domain. Show nevertheless that indeed $\mathrm{KL}(g,f) \geq 0$, perhaps via the Jensen inequality, and that $\mathrm{KL}(g,f) = 0$ only when the two densities are equal a.e. In Ex. 5.7 we learn more details about the KL distance, and look into illustrations, but here the main point is to see its close connection to ML estimation.

(b) We now apply the general minimum divergence machinery of Ex. 5.5, with basic function $h(y,\theta) = -\log f(y,\theta)$. Show that this defines $\theta_0 = \theta(G)$ as the minimiser of

$\mathrm{KL}(g, f_\theta)$, and that minimising the implied distance function $H_n(\theta)$ is the same as maximising the function $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$, as in (5.1). This function is sufficiently famous and pervasive to have earned its own name, the *log-likelihood function* (here for i.i.d. observations). As explained above, the ML estimation method, seen here to be equivalent to minimum divergence with the underlying divergence being the Kullback–Leibler, from true density to parametric approximation.

maximum likelihood

(c) As a simple illustration (to be returned to in Ex. 5.9), generate say $n = 100$ points from the uniform distribution, and use the parametric density $f(y, \theta) = \theta y^{\theta-1}$ on the unit interval. Write down and plot the log-likelihood function, and find the ML estimate. Also show that the implied best parameter value, if the data stem from some g, rather than from the model, is $\theta_0 = 1/\mathrm{E}_g \log(1/Y)$.

Ex. 5.7 *The Kullback–Leibler distance: details and illustrations.* (xx repair here. xx) In Ex. 5.6 and other exercises above we have seen that the machinery of maximum likelihood is intimately related to the KL distance $\mathrm{KL}(g, f) = \int g \log(g/f)\,\mathrm{d}y$. Here we work on illustrations to learn more.

(a) For two normal densities, $\mathrm{N}(a, 1)$ and $\mathrm{N}(b, 1)$, show that the KL distance becomes $\frac{1}{2}(b-a)^2$. Prove also the somewhat more general result, that with $g \sim \mathrm{N}(\xi_1, \sigma^2)$ and $f \sim \mathrm{N}(\xi_2, \sigma^2)$, the KL distance is $\frac{1}{2}(\xi_2 - \xi_1)^2/\sigma^2$.

(b) Find the KL distance from one Poisson to another.

(c) The KL distance is also perfectly well-defined and meaningful in higher dimension. Show that the KL distance from $\mathrm{N}_p(\xi_1, \Sigma)$ to $\mathrm{N}_p(\xi_2, \Sigma)$ can be expressed as $\frac{1}{2}\delta^2$, where $\delta = \{(\xi_2 - \xi_1)^{\mathrm{t}}\Sigma^{-1}(\xi_2 - \xi_1)\}^{1/2}$ is the so-called Mahalanobis distance between the two populations.

the Mahalanobis distance

(d) For several of these examples we find KL distances being symmetric, between the two densities in question, but this is not true in general. Compute the KL distance from $\mathrm{N}(\xi, \sigma_1^2)$ to $\mathrm{N}(\xi, \sigma_2^2)$, and compare to the reciprocal case.

(e) (xx may consider a little reshuffle of exercises. xx) Consider a parametric density $f(y, \theta)$, with score function $u(y, \theta) = \partial \log f(y, \theta)$ and information matrix $J(\theta) = \mathrm{Var}_\theta\, u(Y, \theta)$; see Ex. 5.14 for more on these. Show here that

$$\mathrm{KL}(f(\cdot, \theta), f(\cdot, \theta + \varepsilon)) = \tfrac{1}{2}\varepsilon^{\mathrm{t}} J(\theta)\varepsilon + O(\varepsilon^3).$$

(f) Start from $d(g, f) = -\int g \log\{1 + (f/g - 1)\}\,\mathrm{d}y$, for densities which are not far from each other, and use Taylor expansion to find

$$\mathrm{KL}(g, f) \approx \tfrac{1}{2}\int g(f/g - 1)^2\,\mathrm{d}y = \tfrac{1}{2}\Big(\int f^2/g\,\mathrm{d}y - 1\Big).$$

(xx some words indicating that the root-KL might have an easier interpretation. xx)

(g) (xx a bit of text, more than a question. xx) As noted the KL distance is not symmetric, so 'distance' has a direction. In various statistical setups it makes sens to

interpret $d(g, f)$ as the the distance from 'home density g' to 'approximation candidate f'. As also becoming clear from examples above, it's somehow quadratic in nature, so when numbers are involved, measuring the KL distances, it would typically make more sense to give their square roots, as with $\{d(g, f_\theta)\}^{1/2}$, the degree of closeness of the parametric approximant f_θ to the ground truth g.

Ex. 5.8 *KL approximation.* (xx to be edited. xx) For the following cases the point is to set up a data generating density g, and then check how well a certain parametric family $f(y, \theta)$ does the approximation job. For each case, this tells us how well the ML can do its job, with enough data. For the various cases, find the minimiser, i.e. the best approximation; find the minimum square-root distance $d(g, f(\cdot, \theta_0))^{1/2}$ (since this gives a better picture than on the KL scale itself); and plot the true g alongside the parametric approximant.

(a) Let $g = 0.33\,\mathrm{N}(-1, 1) + 0.67\,\mathrm{N}(1, 1)$. Find the best normal approximation.

(b) Let g be a Gamma with parameters $(2.22, 3.33)$. Find the best Weibull approximant, and also the best log-normal approximant. Similarly, start with a Weibull distribution, with parameters say $(3.33, 2.22)$, and find the best Gamma distribution approximation.

(c) Let $g = 0.95\,\mathrm{Expo}(1) + 0.05\,\mathrm{Expo}(0.01)$, which roughly means that about five percent of the data come from a distribution which much higher mean than the mainstream exponential data. Find the best exponential model approximation, and also the best Gamma and Weibull approximations. Display the true g and these three best parametric approximations in the same diagram.

(d) Suppose data really come from $\mathrm{N}(0.333, \sigma_1^2)$, with $\sigma_1 = 1.111$, where a statistician fits the simpler $\mathrm{N}(0, \sigma^2)$ model. First, find out what happens to the ML estimator. Secondly, illustrate 'what goes on' by drawing e.g. ten samples of size $n = 50$ from the true density, and then display the ten versions of $n^{-1}\ell_n(\sigma)$, along with its limit $C(\sigma)$. Comment on your findings.

BHHJ, the minimum divergence process, and its limit

Ex. 5.9 *The BHHJ density power divergence method.* Here we set up the basics for the so-called *density power divergence method*. It involves raising the density function $f(y, \theta)$ to some power a, as one of its ingredients. In the literature it is sometimes called the *BHHJ divergence method*, from its inventors Basu, Harris, Hjort, Jones (Basu et al. (1998), Jones et al. (2001)). We shall see in later exercises that it amounts to a robust modification of the ML method.

(a) For a density g, in what follows to be seen as the true underlying data-generating model, consider measuring the distance to an approximate $f_\theta(y) = f(y, \theta)$ density as

$$d_a(g, f_\theta) = \int \Big\{ f_\theta^{1+a} - \Big(1 + \frac{1}{a}\Big) g f_\theta^a + \frac{1}{a} g^{1+a} \Big\}\, \mathrm{d}y, \tag{5.3}$$

with a a positive tuning parameter. Show that $d_a(g, f_\theta) \geq 0$, and that the distance is zero only when $g = f_\theta$ a.e. Note that this for $a = 1$ is the same as the L_2 distance of Ex. 5.3.

(b) Use Taylor expansion for f_θ^a and g^a for small a, to demonstrate that the integrand in (5.3) may be written

$$f_\theta - g + g\log(g/f_\theta) - a(g-f_\theta)\log f_\theta + \tfrac{1}{2}ag\{(\log g)^2 - (\log f_\theta)^2\} + O(a^2).$$

Hence show that for a small, we have $d_a(g, f_\theta) = \mathrm{KL}(g, f_\theta) + O(a)$, with the Kullback–Leibler distance $\int g\log(g/f_\theta)\,\mathrm{d}y$, assuming that the functions $g\log f_\theta$, $f_\theta\log f_\theta$, $g(\log f_\theta)^2$, $g(\log g)^2$ have finite integrals.

(c) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from some unknown g, and that we wish to estimate θ by making the distance $d_a(g, f_\theta)$ small. The point is now that the third term of d_a does not depend on θ, and that we may accurately estimate the two first, using

$$H_{n,a}(\theta) = \int f(y,\theta)^{1+a}\,\mathrm{d}y - (1+1/a)\,n^{-1}\sum_{i=1}^n f(y_i,\theta)^a. \tag{5.4}$$

Show that this is also coming out of the general minimum divergence function apparatus of Ex. 5.5, with $h(y,\theta) = \int f_\theta^{1+a}\,\mathrm{d}y - (1+1/a)f(y,\theta)^a$. Show indeed that $H_n(\theta)$ is an unbiased estimator of the two first terms of $d_a(g, f_\theta)$, and give an expression for its variance. We call the minimiser $\widehat{\theta}$ of $H_n(\theta)$ the BHHJ estimator.

the BHHJ estimator

(d) Return to the ML estimator illustration of Ex. 5.6, with $n = 100$ points drawn from the uniform, fitted to the $\theta y^{\theta-1}$ density. For some values of a, compute and plot the $H_{n,a}$ function, finding also the BHHJ estimate $\widehat{\theta}_a$. After having computed this for a grid of a, plot the resulting $\widehat{\theta}_a$ as a function of a, and comment.

Ex. 5.10 *Integrals for BHHJ estimation.* Using the BHHJ method of Ex. 5.9 for estimating the parameters of models $f(y,\theta)$, we are very much helped, algorithmically and numerically, by having formulae for the term $A(a) = \int f_\theta^{1+a}\,\mathrm{d}y$.

(a) For the normal $\mathrm{N}(\xi,\sigma^2)$, show that

$$A(a) = \int f(y,\xi,\sigma)^{1+a}\,\mathrm{d}y = (2\pi)^{-a/2}(1+a)^{-1/2}\sigma^{-a}.$$

Generalise to the multinormal case of $\mathrm{N}_p(\xi,\Sigma)$, with

$$A(a) = \int f(y,\xi,\Sigma)^{1+a}\,\mathrm{d}y = (2\pi)^{-ap/2}(1+a)^{-p/2}|\Sigma|^{-a/2}.$$

(b) For the gamma model, with $Y \sim \mathrm{Gam}(a,b)$, show that

$$\int\Big\{\frac{b^a}{\Gamma(a)}y^{a-1}\exp(-by)\Big\}^{1+\alpha}\,\mathrm{d}y = \frac{\Gamma(a+\alpha a-\alpha)}{\Gamma(a)^{1+\alpha}}\frac{b^\alpha}{(1+\alpha)^{a+\alpha a-\alpha}}.$$

(c) For the log-normal model, where Y is such that $\log Y \sim \mathrm{N}(\xi,\sigma^2)$: Show that

$$\int_0^\infty f(y,\xi,\sigma)^{1+a}\,\mathrm{d}y = (2\pi)^{-a/2}\sigma^{-a}(1+a)^{-1/2}\exp\{-a\xi + \tfrac{1}{2}a^2\sigma^2/(1+a)\}.$$

(d) For the Weibull, with c.d.f. $1-\exp\{-(y/a)^b\}$ for $y \geq 0$, show that

$$\int_0^\infty f(y,a,b)^{1+\alpha}\,\mathrm{d}y = \Big(\frac{b}{a}\Big)^\alpha \frac{\Gamma(1+\alpha-\alpha/b)}{(1+\alpha)^{1+\alpha-\alpha/b}}.$$

(e) For robust estimation of the three-parameter t distribution, consider the density $f(y,\xi,\sigma,\nu) = g_\nu((y-\xi)/\sigma)(1/\sigma)$ for $Y = \xi + \sigma t_\nu$, with g_ν the t density with ν degrees of freedom. Find

$$\begin{aligned}\int f^{1+a}\,\mathrm{d}y &= \frac{1}{\sigma^a}\int g_\nu^{1+a}\,\mathrm{d}x \\ &= \frac{1}{\sigma^a}\frac{\Gamma(((1+a)\nu+a)/2)}{\Gamma((1+a)(\nu+1)/2)}\Big\{\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\Big\}^{1+a}\frac{1}{(\nu\,\pi)^{a/2}}.\end{aligned}$$

Ex. 5.11 *Maximum weighted likelihood estimation.* Another fruitful generalisation of the ML method of Ex. 5.6, in addition to the BHHJ method of Ex. 5.9, starts from adding a weight function to the Kullback–Leibler $\mathrm{KL}(g,f)$ divergence.

(a) Show that $\mathrm{KL}(g,f)$ may be written $\int\{g\log(g/f)-(g-f)\}\,\mathrm{d}y$, that the integrand is nonnegative, and is zero only if $g=f$ a.e. With any nonnegative weight function $w(y)$ on the sample space, deduce that

$$\mathrm{KL}_w(g,f) = \int w(y)\Big[g(y)\{\log\frac{g(y)}{f(y)} - \{g(y)-f(y)\}\Big]\,\mathrm{d}y$$

is a divergence, i.e. nonnegative, and is zero only if $g=f$ on the support set of w. In particular, for g fixed and parametric $f(\cdot,\theta)$, show that $\mathrm{KL}_w(g,f_\theta)$ can be expressed as $\int w(-g\log f_\theta + f_\theta)$ plus other terms not depending on θ.

(b) For data $Y_1,\ldots,Y_n$ from a generating density g, to be approximated with a parametric $f(y,\theta)$, explain that the general minimum divergence function method leads to maximisation of

$$\ell_{n,w}(\theta) = \sum_{i=1}^n w(y_i)\log f(Y_i,\theta) - nB(\theta),$$

with $B(\theta) = \int w f_\theta\,\mathrm{d}y$. We call the maximiser $\widehat{\theta}_w$ the maximum weighted likelihood estimator, associated with weight function $w(y)$. The classic ML estimator then corresponds to constant weight function $w(y)=1$. Properties of $\widehat{\theta}_w$ are derived in Ex. 5.20.

(c) For an illustration, access the birthweight dataset for Oslo children, reported on in Story i.5. For the boys and the girls, fit normal distributions, using maximum weighted likelihood, with weight function $w(y)$ being one on $[2.5,4.5]$ kg and zero outside. Plot the estimated densities in a diagram, This requires programming a function for optimisation. Compare with curves estimated via full ML.

Ex. 5.12 *Minimum divergence function estimators: a limit process and main heuristics.* After having motivated and worked through particular instances of the minimum divergence function estimators, we now return to the general case, aiming to demonstrate

both limiting normality and other associated results, finding recipes for large-sample approximations in the process. The aim of the present exercise is to go through the basic ideas, involving also two main heuristics. These then apply to large classes of estimators and associatead minimum divergence function minima. To make these heuristics precise we need more conditions and details, to which we return in further exercises below. In particular, results derived here and in a few of the following exercises will apply immediately to ML estimators, BHHJ estimators, MWL estimators, via their definitions in Ex. 5.6, 5.9, 5.11.

So $\theta_0 = \theta_0(G)$ is the minimiser of $H(\theta) = \mathrm{E}_G\, h(y,\theta)$, and $\widehat{\theta}$ is its estimator, the minimiser of $H_n(\theta) = \int h(y,\theta)\,\mathrm{d}G_n(y)$, with G_n the empirical distribution for an observed i.i.d. sample $Y_1,\ldots,Y_n$ from G. To present the basic ideas and main heuristics we shall start with these regularity conditions: (i) The true $\theta_0 = \theta_0(G)$ is an inner point in its parameter space inside $\mathbb{R}^p$. (ii) The $h(y,\theta)$ is smooth in θ, in a neighbourhood around θ_0, with at least two derivatives, say $h'(y,\theta)$ (a vector, with components h'_a), $h''(y,\theta)$ (a matrix, with components $h''_{a,b}$). (iii) The matrix $J = \mathrm{E}_G\, h''(Y,\theta_0)$ is finite and positive definite. (iv) The first derivative $h'(Y,\theta_0)$ has finite variance matrix K.

(a) A key idea is to work with the following random function. Write

$$\begin{aligned} A_n(s) &= n\{H_n(\theta_0 + s/\sqrt{n}) - H_n(\theta_0)\} \\ &= \sum_{i=1}^n \{h(Y_i,\theta_0 + s/\sqrt{n}) - h(Y_i,\theta_0)\} = U_n^{\mathrm{t}} s + \tfrac{1}{2} s^{\mathrm{t}} J_n s + r_n(s), \end{aligned}$$

with $U_n = (1/\sqrt{n})\sum_{i=1}^n h'(Y_i,\theta_0)$, $J_n = n^{-1}\sum_{i=1}^n h''(Y_i,\theta_0)$. Show that $A_n(s)$ is properly defined for all large enough n, that U_n has mean zero and tends to $U \sim \mathrm{N}_p(0,K)$, and that $J_n \to_{\mathrm{pr}} J$.

(b) It is also clear from the Taylor expansion argument that with moderate further regularity, the remainder term $r_n(s)$ will go to zero in probability, for each s. There is then convergence $A_n(s) \to_d A(s) = U^{\mathrm{t}} s + \frac{1}{2} s^{\mathrm{t}} J s$, for each s. Heuristic One is then to go from $A_n \to_d A$ to $\mathrm{argmin}(A_n) \to_d \mathrm{argmin}(A)$. Explain that this then leads to $\sqrt{n}(\widehat{\theta} - \theta_0) \to_d -J^{-1}U$, which is a $\mathrm{N}_p(0, J^{-1}KJ^{-1})$. We call this limit distribution variance matrix $J^{-1}KJ^{-1}$ the sandwich matrix.

the sandwich matrix

(c) Similarly, Heuristic Two is to go from $A_n \to_d A$ to $\min(A_n) \to_d \min(A)$. Argue that this entails $W_n = 2n\{H_n(\theta_0) - H_n(\widehat{\theta})\} \to_d W = U^{\mathrm{t}} J^{-1} U$. Show for this quadratic form that its mean and variance are $\mathrm{Tr}(J^{-1}K)$ and $2\,\mathrm{Tr}(J^{-1}KJ^{-1}K)$.

(d) (xx give an easy example, to identify J and K. xx)

Ex. 5.13 *Minimum divergence function estimators: from heuristics to proofs.* (xx nils and emil check details carefully. repair and edit. xx) The general setup is as with the previous Ex. 5.12, with minimum divergence function estimator $\widehat{\theta}$ minimising $H_n(\theta) = n^{-1}\sum_{i=1}^n h(Y_i,\theta)$, with observations coming from G. In addition to regulatity conditions (i)–(iv) given there, we postulate (v), that also third derivatives of $h(y,\theta)$ exist, say $h'''_{a,b,c}(y,\theta)$, and that these have finite means in a neighbourhood around θ_0. We use the process $A_n(s) = n\{H_n(\theta_0 + s/\sqrt{n}) - H_n(\theta_0)\}$ from the previous exercise.

(a) Explain first that the minimiser of A_n is $\alpha_n = \sqrt{n}(\widehat{\theta} - \theta_0)$, where we shall also study the overall minimum $A_{n,\min} = A_n(\alpha_n)$ below. Let $B_n(s) = U_n^{\mathrm{t}} s + \frac{1}{2} s^{\mathrm{t}} J_n s$ be the quadratic approximation to A_n, with minimiser β_n and overall minimum $B_{n,\min} = \min\{B_n(s) \colon \text{all } s\}$. Show that

$$\begin{aligned} \beta_n &= -J_n^{-1} U_n \to_d -J^{-1} U \sim \mathrm{N}_p(0, J^{-1} K J^{-1}), \\ B_{n,\min} &= -\tfrac{1}{2} U_n^{\mathrm{t}} J_n^{-1} U_n \to_d -\tfrac{1}{2} U^{\mathrm{t}} J^{-1} U. \end{aligned}$$

So things are simple and clear for the quadratic approximation B_n; we need to show that the same results obtain for the real thing, the A_n.

(b) Supposing $|r_n(s)| \le \delta$ for all s in a subset S, show from $A_n = B_n + r_n$ that

$$|\min_{s \in S} A_n(s) - \min_{s \in S} B_n(s)| \le \delta.$$

We next establish that α_n cannot be far away. Show that when $\|s\| \ge c n^{1/8}$, $B_n(s) \ge D_n n^{1/4}$, with D_n positive and bounded in probability; and that when $\|s\| \le c n^{1/8}$, then $|r_n(s)| \le E_n / n^{1/8}$, with E_n also bounded in probability. Show from this that $\alpha_n = O_{\mathrm{pr}}(n^{1/8})$, and that $A_{n,\min} - B_{n,\min} \to_{\mathrm{pr}} 0$.

(c) Then consider B_n a certain distance away from the minimum. For given small ε, show that for v with $\|v\| \ge \varepsilon$, we have

$$B_n(\beta_n + v) = B_{n,\min} + \tfrac{1}{2} v^{\mathrm{t}} J_n v \ge B_{n,\min} + \tfrac{1}{2} j_n \varepsilon^2,$$

where j_n is the smallest eigenvalue of J_n. Show then that the event Ω_n, where $A_n(\beta_n + v) \ge B_{n,\min} + \frac{1}{4} j_n \varepsilon^2$ for all v with $\|v\| \ge \varepsilon$, must have $\Pr(\Omega_n) \to 1$. Prove from these established statements that $\alpha_n - \beta_n$ must tend to zero in probability. Conclude that

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &\to_d -J^{-1} U \sim \mathrm{N}_p(0, J^{-1} K J^{-1}), \\ 2n\{H_n(\theta_0) - H_n(\widehat{\theta})\} &\to_d W = U^{\mathrm{t}} J^{-1} U. \end{aligned} \tag{5.5}$$

Score function, Fisher information matrix, Cramér–Rao lower bounds

Ex. 5.14 *Score functions and the Fisher information matrix.* Consider a parametric model with density $f(y, \theta)$ with respect to some measure μ, where $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{t}}$, the parameter of the model is contained in some open parameter space Θ. Introduce

score function

$$u(y, \theta) = \partial \log f(y, \theta) / \partial \theta \quad \text{and} \quad i(y, \theta) = \partial^2 \log f(y, \theta) / \partial \theta \, \partial \theta^{\mathrm{t}},$$

called the *score function*, with p components, and the *information function*, a $p \times p$ matrix. These partial derivatives are assumed to exist and, for the maximum likelihood theory below, they must be continuous; [xx check this with Nils xx] note that this concerns smoothness in the parameter θ, not necessarily smoothness in y. We also assume that the *support* for the distribution, the smallest closed set for which the density is positive, does not depend on θ. Cases falling outside such assumptions are, e.g., the uniform on an unknown interval $[0, \theta]$. Finally, we assume that $\int f(y, \theta) \, \mathrm{d}\mu(y)$ can be differentiated under the integral sign with respect to each coordinate of θ.

Fisher information regularity conditions

(a) The score function has mean zero: show that $\mathrm{E}_\theta\, u(Y,\theta) = \int f(y,\theta)u(y,\theta)\,\mathrm{d}\mu(y) = 0$. Let next

the Fisher information matrix

$$K(\theta) = \mathrm{Var}_\theta\, u(Y,\theta) \quad \text{and} \quad J(\theta) = -\mathrm{E}_\theta\, i(Y,\theta),$$

and show that indeed $J(\theta) = K(\theta)$, the so-called Bartlett identity. This matrix is often called *the Fisher information matrix* for the model. It provides a measure of how much information about a parameter a dataset provides. Note that the calculation of both $J(\theta)$ and $K(\theta)$ is taking place under the assumption that the model is actually correct.

the Bartlett identity

(b) For the exponential model, with density $\theta\exp(-\theta y)$, find the score function, and compute the Fisher information function in two ways. The second (derivative) way of computing the Fisher information here was quite simple.

(c) Consider the general exponential family, with its natural parametrisation $f(y,\theta) = \exp\{\theta^{\mathrm{t}}T(y) - k(\theta)\}h(y)$, see Ex. 1.50. Explain that the score function becomes $T(y) - k'(\theta)$, with Fisher matrix $J(\theta) = k''(\theta)$, that of the second order derivatives of $k(\theta)$.

(d) For the normal $\mathrm{N}(\xi,\sigma^2)$ model, show that the score function can be expressed as

$$u(y,\xi,\sigma) = \begin{pmatrix} \frac{1}{\sigma}(y-\xi)/\sigma \\ \frac{1}{\sigma}\{(y-\xi)^2/\sigma^2 - 1\} \end{pmatrix} = \frac{1}{\sigma}\begin{pmatrix} z \\ z^2-1 \end{pmatrix},$$

writing $z = (y-\xi)/\sigma$, which is a standard normal when y comes from the model. Demonstrate that the Fisher information matrix becomes

$$J(\xi,\sigma) = \mathrm{Var}_{\xi,\sigma}\, u(Y,\xi,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

(e) (xx Check with a few more of your favourite parametric models, where you find the score function and the information function, and where then formulae for both $J(\theta)$ and the variance matrix $K(\theta)$ of the score function, verifying that they are the same. ask for poisson, for binomial with parametrisation $p = \exp(\theta)/\{1+\exp(\theta)\}$, geometric. xx)

(f) (xx more care here, since we do CR a bit later. xx) If Y has the uniform distribution on $[0,\theta]$, which of the regularity conditions listed above fail? In this situation, one might try to define the Fisher information to be $1/\theta^2$. Assuming that this is indeed the Fisher information, use the Cramér–Rao lower bound to derive a contradiction.

Ex. 5.15 *Cramér–Rao lower bounds for estimators.* A certain basic and classic inequality provides a lower bound for the variance of any unbiased estimator of a given parameter. There are various versions and generalisations, some of which we go through here. Inequalities of the type encountered here are sometimes called 'information inequalities', as they may be used to define and analyse how much information there can be in a finite set of data. We also discover clear links to log-likelihoods and ML estimation; the behaviour of ML estimators for growing sample size, described in Ex. 5.17, exactly matches the Cramér–Rao lower bound for variances.

(a) To begin simply, suppose Y is an observation from the density $f(y,\theta)$, assumed smooth in its one-dimensional parameter. Let $u(y,\theta) = \partial \log f(y,\theta)/\partial\theta$ be the score function, with finite variance, by definition equal to the Fisher information, $J(\theta) = \int f(y,\theta)u(y,\theta)^2\,\mathrm{d}y$. Let $T = T(Y)$ be any estimator unbiased for θ, and assume that $f(y,\theta)$ satisfies the conditions of Ex. A.12(g). From $\mathrm{E}_\theta\, T = \int T(y)f(y,\theta)\,\mathrm{d}y = \theta$, use the conditions on $f(y,\theta)$ to deduce that $(\mathrm{d}/\mathrm{d}\theta)\int T(y)f(y,\theta)u(y,\theta)\,\mathrm{d}y = 1$. Show that $\mathrm{cov}_\theta\,(T, u(Y,\theta)) = 1$, and, consequently $1 \le \mathrm{Var}_\theta\, T\,\mathrm{Var}\, u(Y,\theta)$, from which the we get one-dimensional classic Cramér–Rao inequality

$$\mathrm{Var}_\theta\, T \ge 1/J(\theta).$$

(b) It is easy to generalise the above to the more interesting case of having more observation than one. Suppose $Y_1,\ldots,Y_n$ are i.i.d. from the parametric model $f(y,\theta)$. Show that the arguments above still hold, essentially since $Y = (Y_1,\ldots,Y_n)$ can be considered a single datum from the model with joint density $f(y_1,\theta)\cdots f(y_n,\theta)$. Show that the score function now becomes $u(y_1,\ldots,y_n,\theta) = \sum_{i=1}^n u_0(y_i,\theta)$, writing for emphasis $u_0(y,\theta) = \partial \log f(y,\theta)/\partial\theta$ for the score function for a single observation. Deduce that the combined Fisher information for the full sample is $J_n = \mathrm{Var}_\theta\, u(Y_1,\ldots,Y_n) = nJ_0$, with $J_0 = \mathrm{Var}_\theta\, u_0(Y_i,\theta)$ the information in a single observation.

Cramér–Rao lower bound

(c) Show from this that if $T = T(Y_1,\ldots,Y_n)$ is an unbiased estimator for θ, then

$$\mathrm{Var}_\theta\, T \ge \frac{1}{nJ_0(\theta)} = \frac{1/J_0(\theta)}{n}.$$

This says that there is a clear limit to how well one might estimate a parameter in a model, with n observations. If you're not entirely satisfied with $\mathrm{Var}_\theta\, T = 0.10$, say, and wish for variance 0.05 instead, then shell out more money to get twice as many observations.

(d) Show more generally that if $T = T(Y_1,\ldots,Y_n)$ is an estimator for θ, with mean $\mathrm{E}_\theta\, T = \theta + b(\theta)$, i.e. with a certain bias $b(\theta)$, then

$$\mathrm{Var}_\theta\, T \ge \frac{1}{n}\frac{\{1+b'(\theta)\}^2}{J_0(\theta)}.$$

In particular, show that there's a lower bound on the mean squared error for *any* estimator (i.e. not merely the unbiased ones):

$$\mathrm{mse}(\theta) = \mathrm{E}\,\{T(Y_1,\ldots,Y_n) - \theta\}^2 \ge n^{-1}\{1+b'(\theta)\}^2/J_0(\theta) + b(\theta)^2.$$

(e) Go through the following examples, in each case finding the score function, the information $J_0(\theta)$, and the lower bound for any unbiased estimator of the model parameter. (i) y is binomial (n,θ). (ii) y is Poisson θ. (iii) y is normal (θ,σ^2), with σ known. (iv) y is normal (θ,σ^2), with θ known, and σ to be estimated. Comment on the implications of your findings.

Ex. 5.16 *Cramér–Rao bounds for the multidimensional case.* In generalisation of the above situation to the case of multiparameter models, assume first that y is a single

observation from the model $f(y,\theta)$, with $\theta=(\theta_1,\ldots,\theta_p)$ of dimension p. Let $u_0(y,\theta)=\partial\log f(y,\theta)/\partial\theta$ be the score function, for such a single y, with the $p\times p$ Fisher information matrix $J_0(\theta)=\mathrm{Var}_\theta\, u_0(Y,\theta)$ assumed positive definite.

(a) For a symmetric $p\times p$ matrix A we write $A\geq 0$ provided it is nonnegative definite, i.e. that $c^{\mathrm{t}}Ac\geq 0$ for all c. Show that the covariance matrix of a random vector is necessarily nonnegative, and that $A\geq 0$ is is equivalent to its eigenvalues being nonnegative. Explain that $a_{ii}\geq 0$, for the diagonal elements, but that we may still have $a_{i,j}<0$ for some off-diagonal elements. With two symmetric matrices, we write $A\geq B$ if $A-B\geq 0$, which is the ordering of variance matrices we use below.

(b) Assume that $T=T(Y)$ is an unbiased estimator of θ, which also means that $\mathrm{E}_\theta\, T_j(Y)=\theta_j$ for each component j. With I_p the identity matrix of size $p\times p$, deduce from $\mathrm{E}_\theta\, T=\int T(y)f(y,\theta)\,\mathrm{d}y$ that

$$(\partial/\partial\theta)\,\mathrm{E}_\theta\, T=\int T(y)f(y,\theta)u_0(y,\theta)^{\mathrm{t}}\,\mathrm{d}y=I_p.$$

(c) Then work out that

$$\mathrm{Var}_\theta\,\{T-J_0(\theta)^{-1}u_0(Y,\theta)\}=\mathrm{E}_\theta\,\{T-\theta-J_0(\theta)^{-1}u_0(Y,\theta)\}\{T-\theta-J_0(\theta)^{-1}u_0(Y,\theta)\}^{\mathrm{t}}$$

can be expressed as $\mathrm{Var}_\theta\, T-J_0(\theta)^{-1}$. We have then shown a multidimensional version of the Cramér–Rao inequality, that $\mathrm{Var}_\theta\, T\geq J_0(\theta)^{-1}$.

(d) Generalise the above to the case of n i.i.d. observations $Y_1,\ldots,Y_n$ from the model. Show that the information matrix for the full data set becomes

$$J_n(\theta)=\mathrm{Var}_\theta\,\frac{\partial\log\{f(Y_1,\theta)\cdots f(Y_n,\theta)\}}{\partial\theta}=nJ_0(\theta),$$

and that for *any* unbiased estimator $T=T(Y_1,\ldots,Y_n)$ of θ, we must have

Cramér–Rao lower bound, matrix case

$$\mathrm{Var}_\theta\, T\geq\{nJ_0(\theta)\}^{-1}=n^{-1}J_0(\theta)^{-1}.$$

More generally, if independent observations $Y_1,\ldots,Y_n$ come from densities $f_1(y,\theta),\ldots,f_n(y,\theta)$, with Fisher information matrices $J_1(\theta),\ldots,J_n(\theta)$, show that any unbiased estimator T of θ has $\mathrm{Var}_\theta\, T\geq\{J_1(\theta)+\cdots+J_n(\theta)\}^{-1}$.

(e) Suppose now that $\phi=\phi(\theta_1,\ldots,\theta_p)$ is a one-dimensional parameter in focus, and that $T=T(Y)$ is an unbiased estimator. With $c(\theta)=\partial\phi(\theta)/\partial\theta$ the $p\times 1$ gradient, show that

$$\mathrm{Var}_\theta\,\{T-c(\theta)^{\mathrm{t}}J(\theta)^{-1}u_0(Y)\}=\mathrm{Var}_\theta\, T-c(\theta)^{\mathrm{t}}J(\theta)^{-1}c(\theta),$$

and conclude the lower bound $\mathrm{Var}_\theta\, T\geq\lambda(\theta)=c(\theta)^{\mathrm{t}}J(\theta)^{-1}c(\theta)$. Generalise to the case of $Y_1,\ldots,Y_n$ being i.i.d. from $f(y,\theta)$: if $T_n=T(Y_1,\ldots,Y_n)$ is unbiased for ϕ, show that $\mathrm{Var}_\theta\, T\geq\lambda(\theta)/n$. We learn also a bit more from these arguments. Show that

$$\mathrm{Var}_\theta\,\Big\{T_n-n^{-1}\sum_{i=1}^n c(\theta)^{\mathrm{t}}J(\theta)^{-1}u_0(Y_i,\theta)\Big\}=\mathrm{Var}_\theta\, T_n-\lambda(\theta)/n,$$

so T_n having a variance coming close to the lower bound means $T_n-\phi$ being equal to or close to the random variable $n^{-1}\sum_{i=1}^n c(\theta)^{\mathrm{t}}J(\theta)^{-1}u_0(Y_i,\theta)$.

(f) (xx something here, tying these matters to ML. consider $T_n^* = \theta + n^{-1} J(\theta)^{-1} \sum_{i=1}^n u_0(Y_i, \theta)$. show that this T_n^* is unbiased and achieves the Cramér–Rao lower bound. this happens in exponential families, give pointer. the point is then that ML in the general case is close to this with growing n. xx)

(g) Also other estimators for θ deserve to be studied, even when they are not exactly unbiased. We start with a single observation Y from $f(y, \theta)$, with score function $u_0(y, \theta)$ as above, and then generalise to n observations afterwards. Assume therefore that $T = T(Y)$ is such that

$$\mathrm{E}_\theta\, T = \int T(y) f(y,\theta)\,\mathrm{d}y = \theta + b(\theta) = \begin{pmatrix} \theta_1 + b_1(\theta) \\ \vdots \\ \theta_p + b_p(\theta) \end{pmatrix},$$

for suitable bias functions $b_1(\theta), \ldots, b_p(\theta)$, perhaps not far from zero. Show that

$$(\partial/\partial\theta)\,\mathrm{E}_\theta\, T = \begin{pmatrix} 1 + \partial b_1(\theta)/\partial\theta \\ \vdots \\ 1 + \partial b_p(\theta)/\partial\theta \end{pmatrix} = \int T(y) f(y,\theta) u_0(y,\theta)^{\mathrm{t}}\,\mathrm{d}y.$$

Then work with $\mathrm{Var}_\theta\,[T - \{I_p + b'(\theta)\} J_0(\theta)^{-1} u_0(Y,\theta)]$ to demonstrate that

$$\mathrm{Var}_\theta\, T \geq \{I_p + b'(\theta)\} J_0(\theta)^{-1} \{I + b'(\theta)\}^{\mathrm{t}}.$$

(h) Generalise to the case of n i.i.d. observations, to reach

$$\mathrm{Var}_\theta\, T \geq n^{-1} \{I_p + b'(\theta)\} J_0(\theta)^{-1} \{I + b'(\theta)\}^{\mathrm{t}}.$$

(i) xx a bit more to round it off. an example or two. CR lower bound not always attained, in some models only for growing n, but that's ok. i make a separate point that the arguments also lead to bounds of type

$$\mathrm{Var}_\theta\, T \geq \{J_1(\theta) + \cdots + J_n(\theta)\}^{-1},$$

in cases with different situations or types of information sources, for the same θ. tie it all to the large-sample ML results. xx

Ex. 5.17 *Maximum likelihood estimators.* Thanks to the general efforts in exercises above, in particular 5.12–5.13, we may already learn the basic properties of the most important of all estimation methods, namely the ML method, introduced in Ex. 5.6. We do return to further details, extensions, results, illustrations, applications in exercises to come. The basic setup, to be generalised later, is that of observations $Y_1, \ldots, Y_n$ being i.i.d. from some $g(y)$, to be fitted to a parametric $f_\theta(y) = f(y,\theta)$, and with $\widehat{\theta}$ the maximiser of $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$. The log-density derivatives $u(y,\theta)$ and $i(y,\theta)$ are those studied in Ex. 5.14. There is a short list of regularity conditions to secure results reached below, inherited from those called (i)-(v) in Ex. 5.13.

(a) Arguably the first natural question, for any estimation method, is what it aims for. Show that $\widehat{\theta} \to_{\rm pr} \theta_0$, the minimiser of $\mathrm{KL}(g, f_\theta)$, assumed here to be unique and an inner point in the parameter space. So we've uncovered what goes on in the mindset of the ML operator; it aims for this *least false parameter*, the θ_0 minimising the KL distance from the truth to the model approximation.

the least false parameter value

(b) With p the dimension of θ, show that the two crucial $p \times p$ matrices J and K, from the general treatment of Ex. 5.13, become

$$J = -\mathrm{E}_G\, i(Y, \theta_0) \quad \text{and} \quad K = \mathrm{Var}_G\, u(Y, \theta_0).$$

We have seen in Ex. 5.14 that these are equal under model conditions, i.e. that $g(y) = f(y, \theta_0)$ for all y.

(c) Via efforts in previous exercises, show (i) that $U_n = n^{-1/2} \sum_{i=1}^n u(Y_i, \theta_0) \to_d U \sim \mathrm{N}_p(0, K)$; (ii) that $J_n = -n^{-1} \sum_{i=1}^n i(Y_i, \theta_0) \to_{\rm pr} J$; and the basic log-likelihood process convergence result

$$\begin{aligned} A_n(s) &= \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) \\ &= U_n^{\rm t} s - \tfrac{1}{2} s^{\rm t} J_n s + r_n(s) \to_d A(s) = U^{\rm t} s - \tfrac{1}{2} s^{\rm t} J s, \end{aligned} \tag{5.6}$$

for each s. This drives much of the largs-sample results for likelihood inference, including the important Wilks theorems we return to in Ex. 5.28, along with further generalisations for e.g. regression models. Show also that

$$B_n(s) = \ell_n(\widehat{\theta} + s/\sqrt{n}) - \ell_n(\widehat{\theta}) = -\tfrac{1}{2} s^{\rm t} \widehat{J}_n s + r'_n(s) \to_d B(s) = -\tfrac{1}{2} s^{\rm t} J s,$$

for each s, i.e. $r'_n(s) \to_{\rm pr} 0$, where $\widehat{J}_n = -n^{-1} \partial^2 \ell_n(\widehat{\theta}) / \partial\theta\, \partial\theta^{\rm t}$ is the *normalised observed Fisher information matrix*.

(d) Let $\ell_{n,\max} = \ell_n(\widehat{\theta})$ be the maximised log-likelihood. From (5.5), deduce that

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &\to_d J^{-1} U \sim \mathrm{N}_p(0, J^{-1} K J^{-1}), \\ 2\{\ell_{n,\max} - \ell_n(\theta_0)\} &\to_d W = U^{\rm t} J^{-1} U, \end{aligned}$$

in which $U \sim \mathrm{N}_p(0, K)$. Remarkably, we have been able to reach these general results for ML estimators without going into special cases, and without caring about there being explicit formulae for these or not.

(e) As simple corollaries, we are also reaching the important and very frequently used consequences for ML estimation in general smooth parametric models, under model conditions: with $J = J(\theta_0)$ the Fisher information matrix, at the underlying true parameter value, show that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}_p(0, J(\theta_0)^{-1}) \quad \text{and} \quad 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d \chi^2_p. \tag{5.7}$$

Ex. 5.18 *The J, K, and sandwich matrices for the BHHJ method.* (xx place inside prose here that we also need estimation of these matrices, to which we return in Ex. 5.27. xx) When fitting a parametric family $f(y, \theta)$ to observations, we have seen in Ex. 5.12-5.13

that the behaviour of minimum divergence function estimators is characteriserd by (i) the implied least false parameter $\theta_0 = \theta_0(G)$ and then (ii) the crucial matrices J and K, leading to the sandwich matrix $\Sigma = J^{-1}KJ^{-1}$. As above G, with density g, is the data generating mechanism. Here we consider the BHHJ estimation method Ex. 5.9, with the estimator $\widehat{\theta}$ minimising $H_{n,a}(\theta)$ of (5.4). Again, a is a positive finetuning parameter, and for small a the method is close to ML estimation. For the following points, we assume regularity conditions (i)-(v) are in place, from Ex. 5.12-5.13, and in particular that there is a unique minimiser $\theta_0 = \theta_{0,a} = \theta_{0,a}(G)$ of the $d_a(g, f_\theta)$ of (5.3).

(a) Explain that the theory developed in these earlier exercises implies that $\widehat{\theta} \to_{\rm pr} \theta_0$. In the following points, write for simplicity $f_0(y) = f(y, \theta_0)$, $u_0(y) = u(y, \theta_0)$, $i_0(y) = i(y, \theta_0)$. Show also that θ_0 is characterised as the solution to $\int f_\theta^{1+a} u_\theta \, {\rm d}y = \int g f_\theta^a u_\theta \, {\rm d}y$. For this vector, evaluated at θ_0, we write

$$\xi_a = \int g f_0^a u_0 \, {\rm d}y = \int f_0^{1+a} u_0 \, {\rm d}y.$$

(b) As seen in Ex. 5.9, the BHHJ method corresponds to using $h(y, \theta) = \int f_\theta^{1+a} \, {\rm d}y - (1 + 1/a) f(y, \theta)^a$ in the general minimum divergence function setup. Find the derivative $h'(y, \theta)$ and deduce that

$$K_a = (1+a)^2 \Big\{ \int g f_0^{2a} u_0 u_0^{\rm t} - \xi_a \xi_a^{\rm t} \Big\},$$

(xx find K_a. with simplified at model. xx)

(c) (xx check this carefully. also: when estimating J_a, we don't need such a formula, as we get it as the Hessian of the minimisation. xx) Find then an expression for the second derivative $h''(y, \theta)$, and derive a formula for the J matrix:

$$J_a = (1+a) \int \{ f_0^{1+a} u_0 u_0^{\rm t} + (f_0^{1+a} - g f_0^a)(i_0 + a u_0 u_0^{\rm t}) \} \, {\rm d}y.$$

Ex. 5.19 *BHHJ precision under model conditions.* The BHHJ method is designed to do well when the parametric model used is not perfect. We may however also study its precision when the model actually holds. For matrices studied in Ex. 5.18 there are simplified expressions under model conditions, i.e. when $g(y) = f(y, \theta_0)$ for all y.

(a) Under such model conditions, and with f_0, u_0 notation from Ex. 5.18, show that

$$J_a = (1+a) \int f_0^{1+a} u_0 u_0^{\rm t} \, {\rm d}y, \quad K_a = (1+a)^2 \Big(\int f_0^{1+2a} u_0 u_0^{\rm t} \, {\rm d}y - \xi_a \xi_a^{\rm t} \Big),$$

with a consequent simplified sandwich matrix $\Sigma_a = J_a^{-1} K_a J_a^{-1}$. Show that for $a \to 0$, $\Sigma_a \to J_0^{-1}$, with $J_0 = J(\theta_0)$ the Fisher information matrix for the model.

(b) We may now check the loss of efficiency of the BHHJ method, with tuning parameter a, compared to the ML method, under model conditions, by comparing $J_a^{-1} K_a J_a^{-1}$ to the inverse Fisher information matrix. Carry out the relevant computations for estimation θ in the exponential $\theta \exp(-\theta y)$ model. For the case of the $\mathrm{N}(\xi, \sigma^2)$ model, check with Story vii.5. For small a, there is significant gain in robustness at a low cost in efficiency.

(c) For the following, assume

$$A = \int f_0 \log f_0 \, u_0 u_0^{\mathrm{t}} \, \mathrm{d}y, \quad B = \int f_0 (\log f_0)^2 \, u_0 u_0^{\mathrm{t}} \, \mathrm{d}y, \quad c = \int f_0 \log f_0 \, u_0 \, \mathrm{d}y$$

are finite; A and B are $p \times p$ matrices and c a $p \times 1$ vector. Below we carry out Taylor expansions to order a^2, and use $\doteq$ to indicate the consequent approximations for small a. Show that to this order

$$J_a \doteq (1+a)(J_0 + aA + \tfrac{1}{2}a^2 B), \quad K_a \doteq (1+a)^2 \{J_0 + 2aA + a^2(2B - cc^{\mathrm{t}})\}.$$

As a useful interlude, consider $p \times p$ symmetric matrices M and ε, where M is positive definite and ε is smaller, in the technical sense that the eigenvalues of $M^{-1}\varepsilon$ are smaller than 1 in absolute value. Show that

$$(M+\varepsilon)^{-1} = M^{-1} - M^{-1}\varepsilon M^{-1} + M^{-1}\varepsilon M^{-1}\varepsilon M^{-1} + \cdots .$$

Use this to reach

$$(J_0 + aA + \tfrac{1}{2}a^2 B)^{-1} \doteq J_0^{-1} - aJ_0^{-1}AJ_0^{-1} + a^2(J_0^{-1}AJ_0^{-1}AJ_0^{-1} - \tfrac{1}{2}J_0^{-1}BJ_0^{-1}).$$

It is then a matter of bureaucratic algebra to reach an informative approximation to the sandwich matrix

$$\Sigma_a \doteq (J_0 + aA + \tfrac{1}{2}a^2 B)^{-1} \{J_0 + 2aA + a^2(2B - cc^{\mathrm{t}})\} (J_0 + aA + \tfrac{1}{2}a^2 B)^{-1}.$$

Show that this leads to $\Sigma_a \doteq J_0^{-1} + a^2 D$, with $D = J_0^{-1}(B - cc^{\mathrm{t}} - AJ_0^{-1}A)J_0^{-1}$. This indicates indeed that for small a, the efficiency loss is small.

Ex. 5.20 *The J, K, and sandwich matrices for the maximum weighted likelihood method.* (xx check all this. xx) Just as we for the BHHJ estimation method worked out the basics for J_a, K_a, $J_a^{-1}K_aJ_a^{-1}$ in Ex. 5.18, we here tend to their parallels for the maximum weighted likelihood method of Ex. 5.11. For a parametric model $f(y,\theta)$, and with a given weight function $w(y)$, it consists in maximising the weighted log-likelihood function $\ell_{n,w}(\theta) = \sum_{i=1}^n w(y_i) \log f(Y_i,\theta) - nB(\theta)$, with $B(\theta) = \int w(y) f(y,\theta) \, \mathrm{d}y$. We assume the regularity conditions of Ex. 5.13 are in force.

(a) When observations $Y_1, \ldots, Y_n$ stem from G, with density g, show that the MWL estimator $\widehat{\theta}$ tends to $\theta_0 = \theta_0(G)$, the minimiser of the weighted KL divergence $\mathrm{KL}_w(g, f_\theta)$. When g is not inside the parametric family, this least false parameter value depends also on the weight function. Show also that θ_0 is characterised as the solution to $\int w(y)\{g(y) - f(y,\theta_0)\} u(y,\theta_0) \, \mathrm{d}y = 0$. As for Ex. 5.18 we write f_0, u_0, i_0 for the functions $f(y,\theta_0), u(y,\theta_0), i(y,\theta_0)$, and $\xi_0 = \int w g u_0 \, \mathrm{d}y = \int w f_0 u_0 \, \mathrm{d}y$.

(b) Explain that maximising $\ell_{n,w}(\theta)$ is the same as minimising $n^{-1}\sum_{i=1}^n h(Y_i,\theta)$, with $h(y,\theta) = w(y) \log f(y,\theta) - B(\theta)$, or solving $n^{-1}\sum_{i=1}^n h'(Y_i,\theta) = 0$, with $h'(y,\theta) = w(y)u(y,\theta) - \xi(\theta)$, where $\xi(\theta) = \partial B(\theta)/\partial\theta = \int w(y) f(y,\theta) u(y,\theta) \, \mathrm{d}y$. Show that

$$K_w = \int w^2 g u_0 u_0^{\mathrm{t}} \, \mathrm{d}y - \xi_0 \xi_0^{\mathrm{t}}.$$

Under model conditions, we have $\xi_0 = 0$, and the above simplies to $K_w = \int w^2 f_0 u_0 u_0^{\mathrm{t}} \, \mathrm{d}y$.

(c) Then show that $J_w = -\int wgi_0\,\mathrm{d}y + \partial^2 B(\theta_0)/\partial\theta\,\partial\theta^{\mathrm{t}}$. Under model conditions, $J_w = -\int wf_0 i_0\,\mathrm{d}y$.

(d) (xx find sandwich, and compare with J^{-1} under model. xx)

Ex. 5.21 *Maximum weighted likelihood for multinomial models.* The weighted likelihood ideas of Ex. 5.11, 5.20 can also be used to estimate parameters in models for multinomial probabilities, allowing different weights of importance for different outcomes. (xx pointer to Story ii.8. xx)

(a) Show first that $p\log p/q-(p-q)$ is always nonnegative, for p, q in $(0,1)$. Now consider $(Y_1,\ldots,Y_k)$ being multinomial with count n and probability vector $p=(p_1,\ldots,p_k)$, as with Ex. 1.5. Then

$$d_w(p,p_\theta)=\sum_{j=1}^k w_j\{p_j\log(p_j/p_{j,\theta})-(p_j-p_{j,\theta})\},$$

is a proper nonnegative divergence, from p to some p_θ, as long as the weights $w_1,\ldots,w_k$ are nonnegative. With equal weights, show that this is the Kullback–Leibler divergence $\mathrm{KL}(p,p_\theta)$. If $p_\theta=(p_1(\theta),\ldots,p_k(\theta))$ is some postulated model, with θ of dimension r, say, show that the minimum divergence method amounts to maximising the weighted and modified log-likelihood

$$\ell_{n,w}(\theta)=n\sum_{j=1}^k w_j\{\widehat{p}_j\log p_j(\theta)-p_j(\theta)\}=\sum_{j=1}^k w_jY_j\log p_j(\theta)-n\sum_{j=1}^k w_jp_j(\theta),$$

with $\widehat{p}_j=Y_j/n$. This generalises the usual ML method, which corresponds to equal weights.

(b) Let $p=(p_1,\ldots,p_k)$ denote the true probability vector. Explain that the weighted likelihood estimator $\widehat{\theta}_w$, maximising $\ell_{n,w}(\theta)$, tends in probability to the least false parameter maximising $\sum_{j=1}^k w_j\{p_j\log p_j(\theta)-p_j(\theta)\}$, i.e. to the minimiser of the weighted KL distance $d_w(p,p_\theta)$. Writing $u_j(\theta)=\partial\log p_j(\theta)/\partial\theta$, show that the least false parameter is also the solution to $\sum_{j=1}^k w_j\{p_j-p_j(\theta)\}u_j(\theta)=0$, assumed here to be unique. Show furthermore that

$$U_n=n^{-1/2}\frac{\partial\ell_{n,w}(\theta_0)}{\partial\theta}=\sum_{j=1}^k w_j\sqrt{n}\{\widehat{p}_j-p_j(\theta_0)\}u_j(\theta_0)\to_d U=\sum_{j=1}^k w_jZ_ju_j(\theta_0),$$

using results and notation from the multinomial CLT worked with in Ex. 2.44. Deduce that $U\sim\mathrm{N}_r(0,K_w)$, with

$$K_w=\sum_{j=1}^k w_j^2p_ju_j(\theta_0)u_j(\theta_0)^{\mathrm{t}}-\xi_0\xi_0^{\mathrm{t}},\quad\text{with }\xi_0=\sum_{j=1}^k w_jp_ju_j(\theta_0).$$

(c) Next demonstrate that the normalised Hessian matrix $\widehat{J}_{n,w}=-n^{-1}\partial^2\ell_{n,w}(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ tends in probability to a well-defined J_w. In fact, writing $i_j(\theta)=\partial^2\log p_j(\theta)/\partial\theta\partial\theta^{\mathrm{t}}$,

show that $J_w = J_w(\theta_0)$, where

$$\begin{aligned} J_w(\theta) &= -\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{t}}}\sum_{j=1}^{k} w_j\{p_j \log p_j(\theta) - p_j(\theta)\} \\ &= \sum_{j=1}^{k} w_j\big[p_j(\theta)u_j(\theta)u_j(\theta)^{\mathrm{t}} - \{p_j - p_j(\theta)\}i_j(\theta)\big], \end{aligned}$$

with the expression simplifying under model conditions. Use this in conjunction with the general minimum divergence theory to establish that $\sqrt{n}(\widehat{\theta}_w - \theta_0) \to_d J_w^{-1}U \sim \mathrm{N}_r(0, \Sigma_w)$, with the sandwich matrix $\Sigma_w = J_w^{-1}K_w J_w^{-1}$.

(d) (xx a simple example here, with pointer to Story ii.8. xx)

Ex. 5.22 *Completing increasingly simpler tasks.* In a certain game of learning a player needs to complete tasks $1, 2, \ldots, n$, which become increasingly simpler with each passing of a new level. Assume that the time needed to complete these tasks are $V_1, \ldots, V_n$, with these being independent with $V_i \sim \mathrm{Expo}(i/\theta)$, where θ is an unknown parameter. – For the following questions, you may encounter the partial sums

$$a_n = 1 + 1/2 + 1/3 + \cdots + 1/n, \quad b_n = 1 + 1/2^2 + 1/3^2 + \cdots + 1/n^2.$$

Here the first is slowly divergent, with $a_n \doteq \log n + 0.5772$, and the second is convergent, with $b_n \to \pi^2/6$, as shown by Euler in 1734, bringing him instant world fame.

(a) Find expressions for the mean and variance of $T_n = V_1 + \cdots + V_n$, the time it takes the player to complete all tasks. In particular, show that T_n has mean $a_n\theta$. Put up the unbiased estimator based on T_n, say $\widehat{\theta}$. Find its variance, and show that the estimator is consistent.

(b) Then work out a formula for the log-likelihood function, based on having observed not merely the total time T_n, but the individual waiting times $V_1, \ldots, V_n$. Find the maximum likelihood estimator, say θ^*. Show that also this estimator θ^* is unbiased, and compare its variance to that of $\widehat{\theta}$. Find also the Cramér–Rao lower bound for variances of unbiased estimators for θ, and comment.

(c) Assume the game goes on, up to level $2n$, and consider the time a player needs to pass the last half of these levels, i.e. $T_n^* = T_{2n} - T_n$. Show that T_n^* tends in probability to a certain limit as n grows.

Ex. 5.23 *Profiling quadratic functions.* Consider some quadratic function $A(s) = \frac{1}{2}s^{\mathrm{t}}Js$, with J symmetric and positive definite of dimension $p \times p$. It is useful to sort out minima of A under different types of side constraints.

(a) We start out examining the minimum of $A(s)$ over all s with $c^{\mathrm{t}}s = x$, for some given vector c and level x. The Lagrange multiplier way of solving such a problem is to minimise the function $\frac{1}{2}s^{\mathrm{t}}Js - \lambda(c^{\mathrm{t}}s - x)$, with no constraint on s, and in the process find the λ agreeing with the constraint. Taking derivatives, show that minimum occurs for $s_0 = \lambda J^{-1}c$, leading to $c^{\mathrm{t}}s_0 = \lambda c^{\mathrm{t}}J^{-1}c$, which should then equal x. Explain that this leads to minimiser $s_0 = x/(c^{\mathrm{t}}J^{-1}c)$ and attained minimum $A_{\min} = \frac{1}{2}x^2/(c^{\mathrm{t}}J^{-1}c)$.

(b) Generalising the above, from a scalar to a vector, consider a $p_0 \times p$ matrix C and an x of dimension p_0, where it is assumed that $CJ^{-1}C^{\mathrm{t}}$ has full rank. The task is to work out the minimiser and minimum of $\frac{1}{2}s^{\mathrm{t}}Js$ over all s with $Cs = x$. Work out that the minimiser is

$$s_0 = J^{-1}C^{\mathrm{t}}\lambda = J^{-1}C^{\mathrm{t}}(CJ^{-1}C^{\mathrm{t}})^{-1}x \text{ with } A_{\min} = \tfrac{1}{2}x^{\mathrm{t}}(CJ^{-1}C^{\mathrm{t}})^{-1}x.$$

(xx we should find the following from this; round off. xx) with $\phi = k(\theta) = (\phi_1, \ldots, \phi_{p_0})$,

$$n(H_{n,\min,\mathrm{narr}} - H_{n,\min,\mathrm{wide}}) \to_d \tfrac{1}{2}U^{\mathrm{t}}J^{-1}C(C^{\mathrm{t}}J^{-1}C)^{-1}C^{\mathrm{t}}J^{-1}U.$$

Ex. 5.24 *Profiling a minimum divergence function, I.* For data $Y_1, \ldots, Y_n$ from some distribution G, we have in Ex. 5.5 considered estimating a parameter $\theta_0 = \mathrm{argmin}(H)$, for $H(\theta) = \mathrm{E}_G\, h(Y, \theta)$, by minimising the distance function $H_n(\theta) = n^{-1}\sum_{i=1}^n h(Y_i, \theta)$. For a focus parameter $\phi = k(\theta)$, a smooth function of $\theta = (\theta_1, \ldots, \theta_p)$, it is useful to work with the associated profile function

profiling a distance function

$$H_{n,\mathrm{prof}}(\phi) = \min\{H_n(\theta)\colon k(\theta) = \phi\}.$$

(a) As an introductory illustration, consider estimating the parameters (a, b) of a Gamma distribution via minimum L_2, as in Ex. 5.3. Simulate 100 datapoints from a gamma; compute $(\widehat{a}, \widehat{b})$; and compute and display also the profile function $H_{n,\mathrm{prof}}(\mu)$ for the mean parameter $\mu = a/b$.

(b) From the full model, with ensuing minimum divergence function estimator $\widehat{\theta}$, show that the consequent $\widehat{\phi} = k(\widehat{\theta})$ becomes normal. With setup and notation as in Ex. 5.13, prove indeed that $\sqrt{n}(\widehat{\phi} - \phi_0) \to_d c^{\mathrm{t}}J^{-1}U$, with $c = \partial k(\theta_0)/\partial\theta$, and that the limit is a zero-mean normal with variance $c^{\mathrm{t}}J^{-1}KJ^{-1}c$. Show also that this $\widehat{\phi}$ is identical to the minimiser of the profile function.

(c) In other words we already know the basic story for any focus parameter estimator $\widehat{\phi}$, thanks to the delta method. It is however fruiful to work with representations and approximations stemming from examining the associated profile function. With methods from Ex. 5.13, show that

$$n\{H_n(\widehat{\theta} + s/\sqrt{n}) - H_n(\widehat{\theta})\} = \tfrac{1}{2}s^{\mathrm{t}}J_n s + r_n(s), \quad \text{with } r_n(s) = O_{\mathrm{pr}}(\|s\|^3/\sqrt{n}).$$

For the profiling, therefore, we must minimise this expression over all s such that $k(\theta) = k(\widehat{\theta} + s/\sqrt{n}) = \phi$. With $k(\theta) = \widehat{\phi} + c_n^{\mathrm{t}}s/\sqrt{n} + O_{\mathrm{pr}}(\|s\|^2/n)$, here writing $c_n = \partial k(\widehat{\theta})/\partial\theta$, the essence is to minimise $\frac{1}{2}s^{\mathrm{t}}J_n s$ under $c_n^{\mathrm{t}}s = \sqrt{n}(\phi - \widehat{\phi}) = x_n$, say. Appealing to Ex. 5.23, show that this minimum becomes $\frac{1}{2}x_n^2/c_n^{\mathrm{t}}J_n^{-1}c_n = n(\widehat{\phi} - \phi)^2/c_n^{\mathrm{t}}J_n^{-1}c_n$. Fill in more details to prove that with $\phi_0 = k(\theta_0)$ the true parameter in question,

$$2n\{H_{n,\mathrm{prof}}(\phi_0) - H_{n,\mathrm{prof}}(\widehat{\phi})\} = \frac{n(\widehat{\phi} - \phi_0)^2}{c_n^{\mathrm{t}}J_n^{-1}c_n} + o_{\mathrm{pr}}(1) \to_d \frac{(c^{\mathrm{t}}J^{-1}U)^2}{c^{\mathrm{t}}J^{-1}c} \sim \kappa\chi_1^2,$$

with $\kappa = c^{\mathrm{t}}J^{-1}KJ^{-1}c/c^{\mathrm{t}}J^{-1}c$.

(d) (xx explain that this often leads to better approximations than the direct limiting normality thing. then an illustration of this. and pointer to Wilks. and pointer to regression versions of these methods and results; the Y_i need not at all be i.i.d. xx)

Ex. 5.25 *Profiling a distance function, II.* In Ex. 5.5, 5.13, 5.24 we have considered parameters defined as minimisers of functions $H(\alpha) = \mathrm{E}_G\, h(Y,\alpha)$, and developed the basic theory for the associated minimum divergence function estimators. We now consider situations where some of the components of the argmin(H) parameter are specified. Such occur when one tests for lower-dimensional structure, etc. This invites setting up the following framework, with a wide model having $\alpha = (\theta,\gamma)$ of length $p+q$, and the narrow model considered has (θ,γ_0), with θ unknown but $\gamma = \gamma_0$ fixed. Estimators $(\widehat\theta,\widehat\gamma)$ in the wide model minimise $H_n(\theta,\gamma) = \int h(y,\theta,\gamma)\,\mathrm{d}G_n(y)$ whereas $\widetilde\theta$ for the narrow model minimises $H_n(\theta,\gamma_0) = \int H(y,\theta,\gamma_0)\,\mathrm{d}G_n(y)$. The theory developed in the previous exercises mentioned holds for the wide and the narrow models, separately, and below we postulate that the regularity conditions (i)–(v) put up in Ex. 5.13 are in force. Efforts of linear and matrix algebra are required in order to handle these models jointly, however. Define therefore

$$J_{\rm wide} = \mathrm{E}_G \frac{\partial^2 H(Y,\theta_0,\gamma_0)}{\partial\alpha\partial\alpha^{\rm t}} = \begin{pmatrix} J_{00}, & J_{01} \\ J_{10}, & J_{11}\end{pmatrix} \quad \text{with inverse} \quad J_{\rm wide}^{-1} = \begin{pmatrix} J^{00}, & J^{01} \\ J^{10}, & J^{11}\end{pmatrix},$$

where $J_{00} = J_{\rm narr}$ is of size $p\times p$, etc. There is similarly a $(p+q)\times(p+q)$ matrix $K_{\rm wide}$, with submatrices $K_{00},K_{01},K_{10},K_{11}$, the variance matrix of $U = (U_0^{\rm t},U_1^{\rm t})^{\rm t}$, the first derivative $\partial H(Y,\theta_0,\gamma_0)/\partial\alpha$.

(a) The following developements are under the $\gamma=\gamma_0$ constraint, so $\alpha_0 = (\theta_0,\gamma_0)$ is the true parameter, determined by the distribution G. Argue that

$$\begin{pmatrix} \sqrt{n}(\widehat\theta-\theta_0) \\ \sqrt{n}(\widehat\gamma-\gamma_0)\end{pmatrix} \to_d -J_{\rm wide}^{-1}\begin{pmatrix} U_0 \\ U_1\end{pmatrix}, \quad \sqrt{n}(\widetilde\theta-\theta_0) \to_d -J_{00}^{-1}U_0.$$

Show also, again using results reached earlier, that

$$\begin{aligned} n\{H_{n,\rm wide} - H_n(\theta_0,\gamma_0)\} &\to_d -\tfrac12 U^{\rm t}J_{\rm wide}^{-1}U, \\ n\{H_{n,\rm narr} - H_n(\theta_0,\gamma_0)\} &\to_d -\tfrac12 U_0^{\rm t}J_{00}^{-1}U_0, \end{aligned}$$

with $H_{n,\rm wide} = H_n(\widehat\theta,\widehat\gamma)$ and $H_{n,\rm narr} = H_n(\widetilde\theta,\gamma_0)$. Deduce that

$$W_n = 2n(H_{n,\rm narr} - H_{n,\rm wide}) \to_d W = U^{\rm t}J_{\rm wide}^{-1}U - U_0^{\rm t}J_{00}^{-1}U_0. \tag{5.8}$$

Show that this limit variable has mean $\mathrm{Tr}(J_{\rm wide}^{-1}K_{\rm wide}) - \mathrm{Tr}(J_{00}^{-1}K_{00})$.

(b) (xx clean and simplify this. xx) We tend to a few matrix and submatrix identities here, as they come in handy for some of the technical arguments below. The $q\times q$ matrix J^{11} has an important role, here and on later occasions (as with the FIC in Ch. 11). Show that

$$Q = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}. \tag{5.9}$$

Similarly, we have $J^{00} = J_{00}^{-1} + J_{00}^{-1}J_{01}QJ_{10}J_{00}^{-1}$. Show also that $J^{00} - J_{00}^{-1} = J^{01}J_{10}J_{00}^{-1}$, $J^{10} = -QJ_{10}J_{00}^{-1}$.

(c) We now use the structure of K_{wide} to transform (U_0, U_1) to (U_0, V), with $V = U_1 - K_{10}K_{00}^{-1}U_0$, the point being that U_0 and V become independent. Work through the details of

$$\begin{pmatrix} U_0 \\ V \end{pmatrix} = \begin{pmatrix} U_0 \\ U_1 - K_{10}K_{00}^{-1}U_0 \end{pmatrix} \sim \mathrm{N}_{p+q}(0, \begin{pmatrix} K_{00} & 0 \\ 0 & K_{11} - K_{10}K_{00}^{-1}K_{01} \end{pmatrix}).$$

Show also that the variance of V is the same as $(K^{11})^{-1}$ (xx check with care xx).

(d) With this transformation, work out the following formula for W, in terms of the independent U_0 and V (xx check all this xx):

$$\begin{aligned} W &= U_0^{\mathrm{t}}(J^{00} - J_{00}^{-1})U_0 + (V + K_{10}K_{00}^{-1}U_0)^{\mathrm{t}}Q(V + K_{10}K_{00}^{-1}U_0) \\ &\quad + 2\,U_0^{\mathrm{t}}J^{01}(V + K_{10}K_{00}^{-1}U_0) \\ &= V^{\mathrm{t}}QV + U_0^{\mathrm{t}}(J^{00} - J_{00}^{-1} + K_{00}^{-1}K_{01}QK_{10}K_{00}^{-1} + 2\,J^{01}K_{10}K_{00}^{-1})U_0 \\ &\quad + U_0^{\mathrm{t}}(J^{01} + K_{00}^{-1}K_{01}Q)V + V^{\mathrm{t}}(J^{10} + QK_{10}K_{00}^{-1})U_0. \end{aligned}$$

(e) There are additional informative and insightful representations of the W above. Start by showing $\sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d -Z$, where

$$\begin{aligned} Z &= J^{10}U_0 + J^{11}U_1 \\ &= J^{10}U_0 + Q(V + K_{10}K_{00}^{-1}U_0) = QV + (J^{10} + QK_{10}K_{00}^{-1})U_0, \end{aligned}$$

Show that $Z \sim \mathrm{N}_q(0, \Sigma_{11})$, with $\Sigma = J^{-1}KJ^{-1}$ the sandwich matrix. The point is now to demonstrate that W above is identical to $W' = Z^{\mathrm{t}}Q^{-1}Z$. Verify first that its mean $\mathrm{Tr}(Q^{-1}\Sigma_{11})$ is identical to the formula found above for $\mathrm{E}\,W$. Work out that

$$\begin{aligned} W' &= [QV + (J^{10} + QK_{10}K_{00}^{-1})U_0]^{\mathrm{t}}Q^{-1}[QV + (J^{10} + QK_{10}K_{00}^{-1})U_0] \\ &= V^{\mathrm{t}}QV + U_0^{\mathrm{t}}(J^{01} + K_{00}^{-1}K_{01}Q)Q^{-1}(J^{10} + QK_{10}K_{00}^{-1})U_0 \\ &\quad + U_0^{\mathrm{t}}(J^{01} + K_{00}^{-1}K_{01}Q)V + V^{\mathrm{t}}(J^{10} + QK_{10}K_{00}^{-1})U_0. \end{aligned}$$

Prove $W = W'$ by by checking the separate terms. One needs to verify that $A = A'$, in

$$\begin{aligned} A &= J^{00} - J_{00}^{-1} + K_{00}^{-1}K_{01}QK_{10}K_{00}^{-1} + J^{01}K_{10}K_{00}^{-1} + K_{00}^{-1}K_{01}J^{10}, \\ A' &= (J^{01} + K_{00}^{-1}K_{01}Q)Q^{-1}(J^{10} + QK_{10}K_{00}^{-1}). \end{aligned}$$

(xx nils cleans and checks all of this. xx)

(f) For the special case $J = K$, which we meet for ML estimation under model conditions, show that $W \sim \chi_q^2$. For the case $K = cJ$, which turns up in certain overdispersion setups, show that $W \sim c\chi_q^2$.

(g) (xx give an example. can simulate from limit distribution. clarify connections to the case of narrow model $p - 1$, wide model p, i.e. profiling over a 1-dimensional $\phi = k(\theta)$. xx)

Ex. 5.26 *ML and minimum divergence function estimators in practice.* In previous exercises we have learned that $\widehat{\theta}$, the minimiser of $H_n(\theta) = n^{-1}\sum_{i=1}^n h(Y_i,\theta)$, is a natural estimator for θ_0, the minimiser of $\mathrm{E}_G\, h(Y,\theta)$, and that its distribution approaches normality. In order to use such results in practice, for testing, setting confidence intervals, etc., we need to estimate the two crucal matrices $J = \mathrm{E}_G\, h''(Y,\theta_0)$ and $K = \mathrm{Var}_G\, h'(Y,\theta_0)$. These comments apply in particular to ML estimation, with $J = -\mathrm{E}_G\, i(Y,\theta_0)$ and $K = \mathrm{Var}_G\, u(Y,\theta_0)$, see Ex. 5.17.

(a) Consider first, in general terms, some function $p(y,\theta)$ with finite mean in a neighbourhood of the true θ_0, with $\widehat{\theta}$ an estimator of θ_0. Explain first that $p_n = n^{-1}\sum_{i=1}^n p(Y_i,\theta_0)$ tends to $p_0 = \mathrm{E}_G\, p(Y,\theta_0)$. The best we may do for estimating p_0 in practice is $\widehat{p}_n = n^{-1}\sum_{i=1}^n p(Y_i,\widehat{\theta})$. Show that if $|p(y,\theta_0+\varepsilon) - p(y,\theta_0)| \le M(y)\|\varepsilon\|$, for all small $\|\varepsilon\|$, for some function $M(y)$ with finite mean, then indeed $\widehat{p}_n \to_{\mathrm{pr}} p_0$.

(b) Give conditions under which the natural estimators

$$\widehat{J} = n^{-1}\sum_{i=1}^n h''(Y_i,\widehat{\theta}) \quad \text{and} \quad \widehat{K} = n^{-1}\sum_{i=1}^n h'(Y_i,\widehat{\theta})h'(Y_i,\widehat{\theta})^{\mathrm{t}}$$

are consistent for J and K. Deduce that when such hold, the empirical sandwich matrix $\widehat{\Sigma} = \widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$ is consistent for $\Sigma = J^{-1}KJ^{-1}$. Note also that $\widehat{J} = H_n''(\widehat{\theta})$ is the Hessian matrix of the criterion function H_n, often computed directly when using numerical minimisation methods for finding $\widehat{\theta}$ in the first place. (xx nils emil small note: where do we say that $A_n \to_{\mathrm{pr}} A$ and $B_n \to_{\mathrm{pr}} B$ for matrices implies $A_nB_n \to_{\mathrm{pr}} AB$, etc.? xx)

(c) Explain how confidence intervals for the components of θ may be read off from this. More generally, for any focus parameter $\phi = k(\theta)$, with estimator $\widehat{\phi} = k(\widehat{\theta})$, show that $\widehat{\phi} \pm 1.96\,\widehat{\kappa}/\sqrt{n}$ is an approximate 95 percent interval for ϕ, where $\widehat{\kappa}^2 = \widehat{c}^{\,\mathrm{t}}\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}\widehat{c}$, and $\widehat{c} = \partial k(\widehat{\theta})/\partial\theta$.

(d) The case of ML estimation is a special case of the setup above, see Ex. 5.17. Show that the estimated matrices become

$$\widehat{J} = -n^{-1}\sum_{i=1}^n \frac{\partial^2 \log f(Y_i,\widehat{\theta})}{\partial\theta\partial\theta^{\mathrm{t}}} \quad \text{and} \quad \widehat{K} = n^{-1}\sum_{i=1}^n \widehat{u}_i\widehat{u}_i^{\mathrm{t}},$$

where $\widehat{u}_i = u(Y_i,\widehat{\theta})$. Note also that $\widehat{J} = -n^{-1}\ell_n''(\widehat{\theta})$, the normalised Hessian matrix of the log-likelihood function computed at the ML position.

(e) To illustrate how the above machinery works in practice, simulate 100 points from the standard normal, and then estimate the two normal parameters (ξ,σ) via minimum L_2, as in Ex. 5.5. Explain that this means minimising the empirical distance function

$$H_n(\xi,\sigma) = \int f(y,\xi,\sigma)^2\,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^n f(y_i,\xi,\sigma) = \frac{\frac12/\pi^{1/2}}{\sigma} - \frac{2}{n}\sum_{i=1}^n \frac{1}{\sigma}\phi\Big(\frac{y_i-\xi}{\sigma}\Big).$$

Carry out this minimisation using e.g. `nlm` in R, a non-linear minimisation algorithm, which finds both $(\widehat{\xi},\widehat{\sigma})$ and the Hessian $\widehat{J}$. Compute also $\widehat{K}$, and find confidence intervals for ξ, for σ, and for $p(y_0) = \Pr(Y \ge y_0)$, with say $y_0 = 1.00$.

(f) Change one or two of your simulated datapoints to somewhat far-off values, e.g. $y_{99} = d$ and $y_{100} = d$, with $d = 5.00$ (which indeed is really far off for the standard normal). Observe what then happens to the ordinary ML estimators, and compare with what happens with the minimum L_2 estimators. The point is the the minimum L_2 method is much more robust than the ML method.

(g) For the ML method, we know from the Bartlett identity that the two matrices J and K are equal under model conditions. For various models we would then have two different estimators, with the same aim, and perhaps different precision; this does not matter for the first-order large-sample theory, since consistency is what matter for $\widehat{J}$ and $\widehat{K}$, and specifically for the estimated sandwich matrix $\widehat{\Sigma} = \widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$. For illustration, consider Y_i from the Pois(θ) model. Show that $J = 1/\theta_0$ and $K = \tau_0^2/\theta_0^2$, in terms of the true mean and variance θ_0 and τ_0^2 of the underlying distribution. The recipes above lead to $\widehat{J} = 1/\bar{Y}$ and $\widehat{K} = V_n/\bar{Y}^2$, in terms of sample mean and sample variance. Under model conditions, both aim for $1/\theta_0$. Show also that the consequent estimated sandwich becomes $\widehat{\Sigma} = V_n$. The general recipes hence lead to two somewhat different confidence intervals for θ, namely $\widehat{\theta} \pm z_0\widehat{\kappa}/\sqrt{n}$, with either $\widehat{\kappa} = \bar{Y}^{1/2}$ of $\widehat{\kappa} = V_n^{1/2}$. Argue that both are valid, with both aiming for the same quantity under Poisson conditions, whereas the second option might be preferred if there might be overdispersion.

(h) (xx profiling too, to illustrate, for $p(y_0)$. again handled by the general theory. intervals need $\kappa = c^{\mathrm{t}}J^{-1}KJ^{-1}c/c^{\mathrm{t}}J^{-1}c$. xx)

Ex. 5.27 *BHHJ analysis in practice.* We have seen how the behaviour of the model robust BHHJ estimators, defined in Ex. 5.9, is described via the two crucial matrices J_a and K_a worked with in Ex. 5.18. Here we go into the details of their estimation. The setting is having i.i.d. data $Y_1, \ldots, Y_n$ from some density g, fitted to a parametric $f(y, \theta)$ by minimising $H_{n,a}(\theta)$ of (5.4). We let $\widehat{\theta}$ be this BHHJ estimator, computed for the given balance parameter a. Below, we write $\widehat{u}_i = u(Y_i, \widehat{\theta})$ and $\widehat{f}_i = f(Y_i, \widehat{\theta})$.

(a) Show that $\widehat{\xi}_a = n^{-1}\sum_{i=1}^n \widehat{f}_i^a \widehat{u}_i$ is equal to $\int f(y,\widehat{\theta})^{1+a} u(y,\widehat{\theta})\,\mathrm{d}y$, and that it is consistent for the ξ_a defined in Ex. 5.18.

(b) Show next that $\widehat{J}_a = H''_{n,a}(\widehat{\theta})$, the Hessian matrix associated with minimisation of the criterion function, is consistent for the J_a matrix.

(c) Show that

$$\widehat{K}_a = (1+a)^2\Big\{n^{-1}\sum_{i=1}^n \widehat{f}_i^{2a}\widehat{u}_i\widehat{u}_i^{\mathrm{t}} - \widehat{\xi}_a\widehat{\xi}_a^{\mathrm{t}}\Big\} = (1+a)^2\, n^{-1}\sum_{i=1}^n (\widehat{f}_i^a\widehat{u}_i - \widehat{\xi}_a)(\widehat{f}_i^a\widehat{u}_i - \widehat{\xi}_a)^{\mathrm{t}}$$

is consistent for K_a.

Ex. 5.28 *Wilks theorems.* (xx repair and round off. log-likelihood profiling, deviance functions, Wilks theorems. setup as in Ex. 5.17. we harvest from earlier profiling efforts. xx)

(a) Consider a one-dimensional focus parameter $\phi = k(\theta)$, and the consequent profile log-likelihood function

$$\ell_{n,\text{prof}}(\phi) = \max\{\ell_n(\theta)\colon k(\theta) = \phi\}.$$

Show first that its maximum is reached for $\widehat{\phi} = k(\widehat{\theta})$, so this $\widehat{\phi}$ is also maximising the profile function, and is rightly the ML estimator of ϕ. With $\phi_0 = k(\theta_0)$, the least false parameter value for ϕ, write also $c = \partial k(\theta_0)/\partial\theta$ for the gradient vector. Show from Ex. 5.24 that

the deviance function

$$D_n(\phi_0) = 2\{\ell_{n,\max} - \ell_{n,\text{prof}}(\phi_0)\} \to_d (c^{\rm t}J^{-1}U)^2/c^{\rm t}J^{-1}c \sim \kappa\chi^2_1,$$

with $\kappa = c^{\rm t}J^{-1}KJ^{-1}/c^{\rm t}J^{-1}c$. This $D_n(\phi)$, when computed for a range of ϕ values, is called the *deviance function* for that parameter. Under model conditions, $D_n(\phi_0) \to_d \chi^2_1$, which is one of several so-called Wilks theorems.

(b) Explain that this Wilks theorem may be used rather directly, without the need to estimate the J matrix, to construct confidence intervals for the focus parameter $\phi = \phi(\theta)$, assuming that the $f(y,\theta)$ model holds. With $C_n(\alpha) = \{\phi\colon \Gamma_1(D_n(\phi)) \le \alpha\}$, writing $\Gamma_1(\cdot)$ for the c.d.f. of the χ^2_1, show in fact that $\Pr_\theta\{\phi \in C_n(\alpha)\} \to \alpha$, for any desired confidence level α. This is a key method for constructing full confidence distributions, the core topic of Ch. 7.

(c) To formulate and prove a Wilks theorem for testing a submodel within a bigger model, via ML estimation, consider $Y_1,\ldots,Y_n$ i.i.d. from a parametric model of the form $f(y,\alpha)$, with $\alpha = (\theta,\gamma)$. We call this the wide model, of dimension $p+q$, with p and q the dimensions of θ and γ, and then study the narrow model, of dimension p, corresponding to $\gamma = \gamma_0$ for some fixed γ_0. Let $\ell_{n,\max,\text{wide}}$ and $\ell_{n,\max,\text{narr}}$ be the log-likelihood maxima for the wide and the narrow models. Show now via Ex. 5.25 that if the narrow model holds, with data arising from $f(y,\theta_0,\gamma_0)$ for some θ_0, that

$$W_n = 2(\ell_{n,\max,\text{wide}} - \ell_{n,\max,\text{narr}}) \to_d W = Z^{\rm t}Q^{-1}Z.$$

Here $Z \sim \text{N}_q(0,Q)$, with $Q = J^{11}$ the lower right $q\times q$ submatrix of J^{-1}, as in (5.9). Under the narrow model conditions, then, $W_n \to_d \chi^2_q$, another Wilks theorem.

(d) (xx choose a simple illustration, one covered in score function exercise above. may point to a couple of stories. xx)

Ex. 5.29 *log-likelihood and ML for the binomial, trinomial, multinomial.* Working through the likelihood mechanics of the binomial, trinomial, multinomial models provide good illustrations of the developed methodology. We already know the basic large-sample behaviour for the natural estimators in these models, via Ex. 2.44, but here we connect such results to the general likelihood theory. See also Story vii.1.

(a) For $X \sim \text{binom}(n,p)$, a sum of independent Bernoulli variables, show that its log-likelihood function is $\ell_n(p) = X\log p + (n-X)\log(1-p)$, with maximiser $\widehat{p} = X/n$. Show that $nJ = n/\{p(1-p)\}$, here with J defined relative to a single Bernoulli variable.

Explain that we may hence read off the limit distribution of $\sqrt{n}(\widehat{p}-p)$ being $\mathrm{N}(0,p(1-p))$, without necessarily even knowing the X/n formula, or about the CLT for binomials. One may even argue the other way, starting with the X/n formula: since we know the limit is $\mathrm{N}(0,p(1-p))$, via the CLT, we must have $J^{-1}=p(1-p)$, hence J must be $1/\{p(1-p)\}$.

(b) For further illustration of the binomial likelihood mechanics, draw the log-likelihood function $\ell_n(p)$, computing both its maximiser, its maximum, and its second derivative at the maximum, for cases (i) $n=20, y=4$, (ii) $n=40, y=8$, (iii) $n=200, y=40$. Note how $\widehat{J}_{\rm obs}=-\ell_n''(\widehat{p})$ becomes bigger and the implied estimated standard deviance $1/\widehat{J}_{\rm obs}^{1/2}$ becomes smaller.

(c) Let (X,Y) be trinomial (n,p,q), and use $r=1-p-q$ and $Z=n-X-Y$. Show that the log-likelihood function becomes $X\log p+Y\log q+Z\log(1-p-q)$, and for the Fisher information matrix that

$$J=\begin{pmatrix}1/p+1/r, & 1/r\\ 1/r, & 1/q+1/r\end{pmatrix} \quad\text{with}\quad J^{-1}=\begin{pmatrix}p(1-p), & -pq\\ -pq, & q(1-q)\end{pmatrix}.$$

Conclude, even before finding or perhaps without caring that there are clear formulae $\widehat{p}=X/n$ and $\widehat{q}=Y/n$ for the ML estimators, that $(\sqrt{n}(\widehat{p}-p),\sqrt{n}(\widehat{q}-q))$ tends to the binormal zero-mean distribution with the J^{-1} covariance matrix. This may of course also be shown, for $(X/n,Y/n)$, without knowing that these are ML estimators.

(d) Generalise the above to the full multinomial model, with data $(X_1,\ldots,X_k)$ with sum n and probabilities $p_1,\ldots,p_k$ summing to 1 over k boxes. The model has $k-1$ parameters, since p_k is known when $(p_1,\ldots,p_{k-1})$ is known. Find the J and J^{-1} matrices, of size $(k-1)\times(k-1)$.

Ex. 5.30 *The maximum likelihood estimator: examples.* Here we work through some examples, where the task is to set up the log-likelihood function and if feasible also explicit formulae for the ML estimators.

(a) With $Y_1,\ldots,Y_n$ from the normal model $\mathrm{N}(\mu,\sigma^2)$, write down the log-likelihood function. Find the ML estimator for σ when μ is a known value, and find also the ML estimators $(\widehat{\mu},\widehat{\sigma})$ in the case where both parameters are unknown.

(b) Suppose $Y_i\sim\mathrm{Pois}(w_i\theta)$, with known exposure times w_i, and that the observations are independent, for $i=1,\ldots,n$. Write down the log-likelihood function, find the ML estimator, and find its mean and variance.

(c) (xx one or two more, with explicit formulae for ML. xx)

(d) Let $Y_1,\ldots,Y_n$ be i.i.d. from the uniform model on $[0,\theta]$, with θ the unknown endpoint. Set up the likelihood function and find the ML estimator.

Ex. 5.31 *Maximum likelihood for the Beta and Gamma models.* Consider the Beta and Gamma two-parameter models, with densities

$$\mathrm{be}(y,a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1} \text{ and } g(y,a,b)=\frac{b^a}{\Gamma(a)}y^{a-1}\exp(-by),$$

for $y\in(0,1)$ and $y>0$, respectively. The task is in each case to estimate the parameters based on an i.i.d. sample $Y_1,\ldots,Y_n$.

(a) We start with the Beta distribution, see Ex. 1.18, where we in particular have found formulae for the mean and variance in terms of (a, b). With empirical mean and variance $\bar{y}$ and $\widehat{\sigma}^2$, show how (a, b) can be fitted by solving the two equations $\bar{y} = \mathrm{E}\, Y$ and $\widehat{\sigma}^2 = \mathrm{Var}\, Y$. With solutions $(\widehat{a}_m, \widehat{b}_m)$ for these moment estimators, and assuming the Beta model is correct, explain how you can find limit distributions of $\sqrt{n}(\widehat{a}_m - a, \widehat{b}_m - b)$.

(b) Write down the log-likelihood function, say $\ell_n(a, b)$. Show that the ML estimators $(\widehat{a}, \widehat{b})$ are the solutions to the two equations

$$n^{-1} \sum_{i=1}^{n} \log Y_i = \psi(a) - \psi(a+b), \quad n^{-1} \sum_{i=1}^{n} \log(1 - Y_i) = \psi(b) - \psi(a+b),$$

where $\psi(x) = \partial \log \Gamma(x)/\partial x$ is the so-called digamma function. There are no explicit formulae here, but the two equations may be easily solved numerically. Explain how the limit distribution for $\sqrt{n}(\widehat{a} - a, \widehat{b} - b)$ may be found, via the two-dimensional central limit theorem. – Here it turns out (i) that the ML estimators are more precise than the moment estimators, and (ii) that finding the limit distribution is rather easier via the general results about ML behaviour, already sorted out in Ex. 5.17.

(c) Then turn attention to the two-parameter Gamma model, where arguments and results will be similar. Show that the mean and variance are a/b and a/b^2, and find moment estimators $\widehat{a}_m, \widehat{b}_m$ based on this. Find the limit distribution of $\sqrt{n}(\widehat{a}_m - a, \widehat{b}_m - b)$ using Ex. **??**.

(d) Then write down the log-likelihood function, take derivatives, and show that the ML estimators $(\widehat{a}, \widehat{b})$ are the solutions to the two equations

$$\bar{Y} = a/b, \quad n^{-1} \sum_{i=1}^{n} \log Y_i = \psi(a) - \log b.$$

Explain how the limit distribution of $\sqrt{n}(\widehat{a} - a, \widehat{b} - b)$ may be obtained. Again, this is rather easier via the general recipes worked out in Ex. 5.17; also, we shall again find that ML estimators under model conditions are more precise than the moment estimators.

Ex. 5.32 *Log-linear mixing of densities.* How far have we moved, from A to B? Suppose i.i.d. data $Y_1, \ldots, Y_n$ come from a density which is percieved of as being 'betweeen' given densities f_A and f_B. There are various ways of creating probabilistic bridges from A to B, via which one then can estimate the position of the current density, e.g. to assess whether it is close to A or to B. One such model takes $f(y, \lambda) = f_A(y)^{1-\lambda} f_B(y)^{\lambda}/R(\lambda)$, with $R(\lambda)$ the normalisation constant $\int f_A^{1-\lambda} f_B^{\lambda}\, \mathrm{d}y$. This creates such a bridge, with $\lambda \in [0, 1]$, and with A and B corresponding to $\lambda = 0$ and $\lambda = 1$.

(a) Explain that $R(0) = R(1) = 1$, and that the derivative of $R(\lambda)$, at the endpoints 0 and 1, can be written

$$R'(0) = -\mathrm{KL}(f_A, f_B), \quad R'(1) = \mathrm{KL}(f_B, f_A).$$

in terms of the Kullback–Leibler distances discussed in Ex. 5.7 and later on. In particular, the $R(\lambda)$ has negative derivative at the start and positive derivative at the end.

(b) For a concrete illustration, suppose f_A is $\mathrm{N}(0,1)$ and f_B is $\mathrm{N}(c,1)$, with c positive. Show that $R(\lambda) = \exp\{-\frac{1}{2}c^2\lambda(1-\lambda)\}$. From this explain that the score function is $u(y,\lambda) = cy - \lambda c^2$ and that the Fisher information becomes the constant c^2. Show indeed that $f(y,\lambda)$ is $\mathrm{N}(\lambda c, 1)$, that the ML estimator becomes $\widehat{\lambda} = \bar{Y}/c$, truncated to $[0,1]$. From general ML theory, explain that if the real λ is inside $(0,1)$, then $\sqrt{n}(\widehat{\lambda} - \lambda) \to_d \mathrm{N}(0, 1/c^2)$. For c small the variance will be very big, signifying that it is a hard task to estimate the balance parameter when f_A and f_B are close.

(c) Returning to the general setup, write $S(y) = \log f_B(y) - \log f_A(y)$. Show that the log-likelihood function becomes

$$\ell_n(\lambda) = n\{\lambda \bar{S}_n - \log R(\lambda)\}, \quad \text{with } \bar{S}_n = n^{-1}\sum_{i=1}^n S(y_i).$$

Explain that the ML estimator $\widehat{\lambda}$ hence is the solution to $\bar{S}_n = R'(\lambda)/R(\lambda)$. (xx a bit more. and example. limiting normality. evaluate for $\lambda = 0$ and $\lambda = 1$. xx)

(d) Bayes posterior: $\pi(\lambda)\exp[n\{\lambda S_n - \log R(\lambda)\}]$.

(e) (xx there is perhaps hope for the semiparametric construction $\widehat{f}(y) = f_0(y)^{1-\lambda} f_n(y)^{\lambda}/R_n(\lambda)$, resembling the Hjort–Glad estimator, see Hjort and Glad (1995). xx)

Convex processes, log-concave likelihoods

Ex. 5.33 *Minimisers of convex processes, I.* We have seen useful constructions, methods, and results for minimum divergence function estimators in Ex. 5.5 and 5.24, in particular when applied to ML estimators, as with the general apparatus of Ex. 5.17. There are issues worth refining and generalising, however. The regularity conditions required for the Taylor expansion based arguments to go fully through are a bit cumbersome, and there are important constructions where the distance function $h(y,\theta)$ in $H(\theta) = \mathrm{E}_G\, h(Y,\theta)$ is not smooth. Here we give the basics for how matters simplify, with weaker conditions, if the distance function is convex.

(a) From pointwise to uniform: Suppose $A_n(s)$ is a sequence of convex random functions defined on an open convex set $\mathcal{S}$ of $\mathbb{R}^p$, which convergences in probability to some $A(s)$, for each $s \in \mathcal{S}$. Show that the convergence is automatically uniform; $\max_{s\in\mathcal{S}} |A_n(s) - A(s)| \to_{\mathrm{pr}} 0$.

(b) Nearness of argmins: Suppose $A_n(s)$ is convex and is approximated by $B_n(s)$. Let α_n and β_n be the argmins of A_n and B_n. Then there is a probabilistic bound on how far these minimisers can be from each other: show that

$$\Pr(\|\alpha_n - \beta_n\| \geq \delta) \leq \Pr\{\Delta_n(\delta) \geq \tfrac{1}{2}h_n(\delta)\},$$

in which

$$\Delta_n(\delta) = \sup_{\|s-\beta_n\|\leq\delta} |A_n(s) - B_n(s)| \quad \text{and} \quad h_n(\delta) = \inf_{\|s-\beta_n\|=\delta} B_n(s) - B_n(\beta_n).$$

(c) Basic corollary: Suppose $A_n(s)$ is convex and can be represented as $\frac12 s^{\rm t}Js + U_n^{\rm t}s + C_n + r_n(s)$, where J is symmetric and positive definite, U_n is stochastically bounded, C_n is arbitrary, and $r_n(s) \to_{\rm pr} 0$ for each s. For the approximation $B_n(s) = \frac12 s^{\rm t}Js + U_n^{\rm t}s + C_n$, show that $\beta_n = -J^{-1}U_n$ is its argmin. Then demonstrate that their minimisers as well as their minima are close. Specifically, show (i) that $\alpha_n - \beta_n \to_{\rm pr} 0$; and (ii) that $A_{n,\min} - B_{n,\min} \to_{\rm pr} 0$.

(d) Show that if in addition $U_n \to_d U$, then $\alpha_n \to_d -J^{-1}U$, and that $B_{n,\min} - C_n$ as well as $A_{n,\min} - C_n$ tend to $-\frac12 U^{\rm t}J^{-1}U$. These two statements are what we worked hard for in Ex. 5.13, 5.25 see (5.5)–(5.8), now obtained in a simpler fashion and with weaker smoothness assumptions, though bought with the extra convexity condition.

(e) Prove the following modest but useful generalisation of the above: the statements continue to hold if a random matrix J_n replaces V, provided $J_n \to_{\rm pr} J$.

Ex. 5.34 *Minimisers of convex processes, II.* The framework worked with now is as in Ex. 5.5 and 5.24, with $Y_1, \ldots, Y_n$ being i.i.d. from some G, a possibly multidimensional parameter θ_0 defined as the minimiser of $H(\theta) = \int h(y,\theta)\,{\rm d}G(y)$, with estimator $\widehat\theta$ the minimiser of the distance function $H_n(\theta) = \int h(y,\theta)\,{\rm d}G_n(y) = n^{-1}\sum_{i=1}^n h(Y_i,\theta)$. Here we put in one more condition, however, that $h(y,\theta)$ is convex in θ. The point is that this both simplifies various technical arguments, via methods of Ex. 5.33, and allows for nonsmooth distance functions.

(a) With $h(y,\mu) = |y-\mu|$, show that $\mu_0 = {\rm med}(G)$, the median, with $\widehat\mu = M_n$, the sample median. More generally, for some $q \in (0,1)$, consider

$$h_q(y,\mu) = q(y-\mu)_+ + (1-q)(\mu-y)_+ = \begin{cases} q(y-\mu) & \text{if } y \ge \mu, \\ (1-q)(\mu-y) & \text{if } y \le \mu. \end{cases}$$

Show that $\mu_0 = G^{-1}(q)$, the q quantile.

(b) For an $\alpha \ge 1$, consider the parameter θ_0 being the minimiser of ${\rm E}_G\,|Y-\theta|^\alpha$. Show that the distance function indeed is convex, and that the special cases $\alpha = 1$ and $\alpha = 2$ correspond to the median and the mean, respectively.

(c) We now work through regularity conditions ensuring control over the behaviour of such estimators. Part of the point is that we avoid needing smooth derivatives in θ. Suppose that

$$h(y,\theta_0+\varepsilon) - h(y,\theta_0) = D(y)^{\rm t}\varepsilon + R(y,\varepsilon),$$

for a $D(y)$ with mean zero under G, and that

$${\rm E}\,\{h(Y,\theta_0+\varepsilon) - h(Y,\theta_0)\} = {\rm E}\,R(Y,\varepsilon) = \tfrac12\varepsilon^{\rm t}J\varepsilon + o(\|\varepsilon\|^2) \quad \text{as } \varepsilon \to 0$$

for a positive definite J. Assume furthermore that the variance matrix $K = {\rm Var}_G\,D(Y)$ is finite and that ${\rm Var}\,R(Y,\varepsilon) = o(\|\varepsilon\|^2)$. Show that $\sqrt n(\widehat\theta-\theta_0) = -J^{-1}(1/\sqrt n)\sum_{i=1}^n D(Y_i) + o_{\rm pr}(1)$. In particular, it tends to ${\rm N}_p(0, J^{-1}KJ^{-1})$. Show also that

$$W_n(\theta_0) = 2n\{H_n(\theta_0) - H_n(\widehat\theta)\} \to_d U^{\rm t}J^{-1}U,$$

and explain how this may be used to find a confidence region for θ_0.

(d) The median: Suppose $Y_1, \ldots, Y_n$ are i.i.d. from a distribution G with a density g positive at the median μ. For the median distance function $|y-\mu|$, show that

$$|y-(\mu+\varepsilon)| - |y-\mu| = D(y)\varepsilon + R(y,\varepsilon),$$

with $D(y) = I(y \le \mu) - I(y > \mu)$ and

$$R(y,\varepsilon) = \begin{cases} 2\{\varepsilon - (y-\mu)\}\, I(\mu \le y \le \mu+\varepsilon) & \text{if } \varepsilon > 0, \\ 2\{(y-\mu) - \varepsilon\}\, I(\mu+\varepsilon \le y \le \mu) & \text{if } \varepsilon < 0, \end{cases}$$

with $R(y,0) = 0$. Verify from this that $\mathrm{E}\, R(Y,\varepsilon) = g(\mu)\varepsilon^2 + o(\varepsilon^2)$ and $\mathrm{E}\, R(Y,\varepsilon)^2 = (4/3)g(\mu)|\varepsilon|^3 + o(|\varepsilon|^3)$. Deduce that $\sqrt{n}(M_n - \mu) \to_d \mathrm{N}(0, 1/\{4g(\mu)^2\})$.

(e) Generalise to the case of $\mu_q = G^{-1}(q)$, with $Q_{n,q} = G_n^{-1}(q)$ the empirical q quantile. Show in fact that

$$Z_n(q) = \sqrt{n}(Q_{n,q} - \mu_q) = -g(\mu_q)^{-1}\sqrt{n}\{G_n(\mu_q) - q\} + \varepsilon_n(q),$$

where $\varepsilon_n(q) \to_{\mathrm{pr}} 0$ for each q. Derive from this that the limit is a $\mathrm{N}(0, q(1-q)/g(\mu_q)^2)$. We saw this in Ex. 3.18, but here the technique is distinctly different, and we find an interesting representation in terms of the empirical $G_n(\mu_q)$.

(f) Let ξ_α be the minimiser of $\mathrm{E}\,|Y_i - \xi|^\alpha$, with estimator $M_{n,\alpha}$ minimising $\sum_{i=1}^n |Y_i - \xi|^\alpha$. Show that

$$\sqrt{n}(M_{n,\alpha} - \xi_\alpha) \to_d \mathrm{N}(0, \tau_\alpha^2) \quad \text{with } \tau_\alpha^2 = \frac{\mathrm{E}\,|Y-\xi_\alpha|^{2(\alpha-1)}}{\{(\alpha-1)\,\mathrm{E}\,|Y-\xi_\alpha|^{\alpha-2}\}^2}.$$

Explain that if the distribution is symmetric, then the ξ_α is the same for each α. Compute and display τ_α, for $\alpha \in [1,2]$, i.e. the range from median to mean, for the normal and for the Laplace,

Ex. 5.35 *Maximum likelihood asymptotics with log-concave likelihood.* (xx edit with care. regularity condition on R. point to expofamily in later exercise. xx) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from a density g, modelled as $f(y,\theta)$, where $\log f(y,\theta)$ is concave in θ for each y; the $\theta = (\theta_1, \ldots, \theta_p)$ is allowed to be multidimensional. Let $\ell_n(\theta)$ be the log-likelihood, with ML estimator $\widehat{\theta}$. From efforts of Ex. 5.17 we know $\widehat{\theta}$ is consistent for the least false parameter θ_0, the KL minimiser, assumed here to be an inner point in the parameter space.

log-concave density

(a) Show first that ℓ_n is concave. To start with mild conditions, assume merely that

$$\log f(y,\theta_0+\varepsilon) - \log f(y,\theta_0) = D(y)^{\mathrm{t}}\varepsilon + R(y,\varepsilon), \tag{5.10}$$

for a $D(y)$ with mean zero, and that $\mathrm{E}\, R(Y,\varepsilon) = \frac{1}{2}\varepsilon^{\mathrm{t}} J\varepsilon + o(\|\varepsilon\|^2)$ as $\varepsilon \to 0$ for some positive definite J, along with $\mathrm{Var}\, R(Y,\varepsilon) = o(\|\varepsilon\|^2)$. Show via the convexity driven methods of Ex. 5.33 and 5.34 that $\sqrt{n}(\widehat{\theta}-\theta_0) = J^{-1}U_n + o_{\mathrm{pr}}(1)$, where $U_n = (1/\sqrt{n})\sum_{i=1}^n u(Y_i,\theta_0)$. Here $U_n \to_d U \sim \mathrm{N}_p(0,K)$, with $K = \mathrm{Var}_g\, u(Y,\theta_0)$ assumed finite. Deduce the two fundamental results

$$\begin{aligned} Z_n &= \sqrt{n}(\widehat{\theta}-\theta_0) \to_d \mathrm{N}_p(0, J^{-1}KJ^{-1}), \\ W_n &= 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d U^{\mathrm{t}} J^{-1} U, \end{aligned} \tag{5.11}$$

with simplifications under model conditions. Show also that consistency of the ML estimator for θ_0 is a simple consequence of the first statement. We have seen such results before, see Ex. 5.17 and in particular (5.7); the point is that they have now been derived more simply and under weaker conditions, as long as there is log-density concavity.

(b) Work through the details for the case of the Laplace distribution with $f(y,\theta) = \frac{1}{2}\exp(-|y-\theta|)$. The point is that with log-concavity, we reach the required results of (5.11) even without needing full smoothness in the parameter; here the score function is not defined for all values, etc. Usually, though, the $D(y)$ is the score function $u(y,\theta_0) = \partial \log f(y,\theta_0)/\partial\theta$ and $J = J(\theta_0)$ is the variance matrix of this score function, i.e. the Fisher information matrix, evaluated at the true position in the parameter space.

(c) (xx a couple of illustrations. for each, find the limit distribution of $\sqrt{n}(\widehat{\theta}-\theta_0)$. the poisson. gamma. beta. normal. also something like $Y_i \sim \text{Pois}(e_i\theta)$, with different exposures w_i, the point is non-i.i.d. nils does this after mild extra editing for Ex. 5.33 and 5.34, with a bit of Lindeberg too. xx)

(d) (xx to polish. xx) something nontrivial. can do

$$f(y,\theta) = (1/k)\exp\{(y-\theta)\arctan(y-\theta)\}/\{1+(y-\theta)^2\}^{1/2},$$

which has log-density $(y-\theta)\arctan(y-\theta) - \frac{1}{2}\log\{1+(y-\theta)^2\}$ and nice score function $\arctan(y-\theta)$. something nontrivial II. and $f = (1/k)\exp(-|y-\theta|^{1.5})$.

Ex. 5.36 *Differentiability in quadratic mean.* The key to proving asymptotic normality of the ML estimator for the log-concave densities in Ex. 5.35 was the assumption that $\log f(y,\theta_0+\varepsilon) - \log f(y,\theta_0) = D(y)^{\mathrm{t}}\varepsilon + R(y,\varepsilon)$, where $D(y)$ and $R(y,\varepsilon)$ satisfy the conditions specified in (a) of that exercise. This raises the question of what conditions are needed for such an expansion of the log-likelihood ratio to hold.

(a) Suppose that $\theta \mapsto \log f(\theta,y)$ is three times continuously differentiable for every y. Assume that $u(\theta,y)$ is square integrable, that $J(\theta_0)$ exists and is nonsingular, and that the third derivative of $\log f(\theta,y)$ is bounded by some integrable function $k(y)$ (not depending on θ). Provided that $\widehat{\theta}_n$ is consistent for θ_0 you may now Taylor expand $0 = U_n(\widehat{\theta}_n)$ around θ_0 to show that $\sqrt{n}(\widehat{\theta}_n-\theta_0) = J(\theta_0)^{-1}n^{-1/2}U_n(\theta_0) + o_{\mathrm{pr}}(1)$. Do it. Classical maximum likelihood conditions

(b) Retain the classical assumptions from (a), except the consistency assumption (as it plays no role here). Show that the expansion in (5.10) of Ex. 5.35(a) holds.

(c)

Ex. 5.37 *The exponential family class and ML.* The ML and log-likelihood machinery works particularly well for the exponential family class of models, see Ex. 1.50, 4.19, 4.20. This is due the log-linear structure for parameters and sufficient statistics, and also to the consequent log-concavity of densities. Consider i.i.d. data $Y_1,\ldots,Y_n$ from a density g, modelled with the generic parametric $f(y,\theta) = \exp\{\theta^{\mathrm{t}}T(y) - k(\theta)\}h(y)$, with $\theta^{\mathrm{t}}T(y) = \theta_1T_1(y) + \cdots + \theta_pT_p(y)$ involving basic data functions $T_1,\ldots,T_p$.

(a) Show that the score function is $u(y,\theta) = T(y) - \xi(\theta)$, with $\xi(\theta) = \partial k(\theta)/\partial\theta$. Also, the information function becomes $i(y,\theta) = -J(\theta)$, with $J(\theta) = \partial^2 k(\theta)/\partial\theta\,\partial\theta^{\mathrm{t}}$.

(b) Under model conditions, show that $\mathrm{E}_\theta\, T(Y) = \xi(\theta)$ and $\mathrm{Var}_\theta\, T(Y) = J(\theta)$. Outside model conditions, with g not belonging to the f_θ, assume merely that $T(Y)$ has true mean ξ_0 and true variance K. Show that the least false parameter θ_0 minimising $\mathrm{KL}(g, f(\cdot,\theta))$ is characterised by $\xi_0 = \xi(\theta_0)$. This may also be written $\theta_0 = M(\xi_0)$, with M the inverse map. Write also $J = J(\theta_0)$, the Fisher information matrix computed at the least false parameter position.

(c) Show that the log-likelihood function becomes $\ell_n(\theta) = n\{\theta^{\mathrm{t}}\bar{T} - k(\theta)\}$, in terms of sample averages $\bar{T} = (\bar{T}_1, \ldots, \bar{T}_p)^{\mathrm{t}}$, and that it is concave. Deduce that the ML estimator has $\xi(\widehat{\theta}) = \bar{T}$, of $\widehat{\theta} = M(\bar{T})$.

(d) We have $U_n = \sqrt{n}(\bar{T} - \xi_0) \to_d U \sim \mathrm{N}_p(0, K)$, by the CLT. For the basic log-likelihood function process, show that

$$\begin{aligned} A_n(s) &= \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = \sqrt{n}\bar{T}^{\mathrm{t}}s - n\{k(\theta_0 + s/\sqrt{n}) - k(\theta_0)\} \\ &= U_n^{\mathrm{t}}s - \tfrac{1}{2}s^{\mathrm{t}}J(\theta_0)s + o(\|s\|^3/\sqrt{n}) \to_d A(s) = U^{\mathrm{t}}s - \tfrac{1}{2}s^{\mathrm{t}}Js. \end{aligned}$$

This is as with (5.6), but now established rather simply and directly, with a minimum of regularity conditions. By appealing to Ex. 5.12, 5.13, deduce also that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}(0, J^{-1}KJ^{-1}), \quad 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d W = U^{\mathrm{t}}J^{-1}U.$$

Under model conditions, the sandwich matrix is $J(\theta_0)^{-1}$ and $W \sim \chi^2_p$. Also, the Wilks theorems of Ex. 5.28 hold, for focus parameters and for testing submodels.

(e) (xx a bit more. xx) From $\widehat{\theta}_j = M_j(\bar{T})$, for components $M_1, \ldots, M_p$ of the mapping $\theta_0 = M(\xi_0)$, write

$$\widehat{\theta}_j = M_j(\xi_0) + M_j'(\xi_0)^{\mathrm{t}}(\bar{T} - \xi_0) + \tfrac{1}{2}(\bar{T} - \xi_0)^{\mathrm{t}}M_j''(\xi_0)(\bar{T} - \xi_0) + O_{\mathrm{pr}}(1/n^{3/2}).$$

Show that this implies $\mathrm{E}\,\widehat{\theta}_j = \theta_{0,j} + \frac{1}{2}c_j/n + o(1/n)$, with $c_j = \mathrm{Tr}(M_j''(\xi_0)K)$.

(f) (xx a simple example here. xx)

Ex. 5.38 *ML asymptotics under model conditions: applications.* (xx cleaning required, and more of the directly useful up front; we estimate parameters and have confidence, almost automatically, via normality and delta method. xx) Results reached in Ex. 5.17 are central in applied statistics. The versatile ML machinery allows the statistician to construct good estimators for even complicated functions of parameters in new models, and to supplement these estimators with confidence intervals, tests, etc. We will also see that the general ML asymptotics results may be used to verify what we already knew, so to speak, regarding estimators in the more familiar models. (xx check all this to make sure that we don't become repetitive. xx)

(a) Let Y be binomial (n, p). Even before you find a formula for the ML estimator for p, show that $\sqrt{n}(\widehat{p} - p) \to \mathrm{N}(0, p(1-p))$. By all means, show also that $\widehat{p} = Y/n$.

(b) In a similar vein, study the classic case of $Y_1, \ldots, Y_n$ being i.i.d. from the normal (ξ, σ^2). Using the Fisher information matrix found in Ex. 5.14, show that the ML estimators $\widehat{\xi}$ and $\widehat{\sigma}$ must be independent, in the limit, with approximate distributions

$\mathrm{N}(\xi, \sigma^2/n)$ and $\mathrm{N}(\sigma, \frac{1}{2}\sigma^2)$. Remarkably, these results follow from the ML apparatus even without or before knowing any formulae for the estimators, and without or before knowing any finite-sample theory for these. As we know (xx crossref here xx) there is *exact* independence here, and the distribution for $\widehat{\xi}$ is exactly correct, for each n.

(c) (xx something with $\mathrm{Gam}(a, b)$, and approximate distribution for $\widehat{\mu}$, estimator for the median $\mu = \mu(a, b)$. illustrate also Wilks. which can be used without explicit formulae. xx)

(d) (xx the Weibull $F(t) = 1 - \exp\{-(t/a)^b\}$. perhaps an earlier exercise where we find $J(a, b)$. xx)

(e) Consider random i.i.d. pairs (X_i, Y_i) from the standardised binormal distribution with zero means, unit variances, and correlation ρ. Set up the log-likelihood function $\ell_n(\rho)$, show that the Fisher information becomes $J(\rho) = (1 + \rho^2)/(1 - \rho^2)^2$, and find the limiting distribution of the ML estimator. How much better is the ML estimator compared to the usual empirical correlation coefficient R_n? (xx kladd to be pushed to solutions follows. xx)

$$\log f(x, y, \rho) = -\tfrac{1}{2}\log(1 - \rho^2) - \tfrac{1}{2}(x^2 + y^2 - 2\rho xy)/(1 - \rho^2),$$

score function

$$u(x, y, \rho) = \frac{1}{(1 - \rho^2)^2}\{\rho - \rho^2 + (1 + \rho^2)xy - \rho(x^2 + y^2)\},$$

where we may check that the mean is zero. We find $\mathrm{Var}\, XY = 1 + \rho^2$, $\mathrm{Var}\,(X^2 + Y^2) = 4(1 + \rho^2)$, $\mathrm{cov}(XY, X^2 + Y^2) = 4\rho$, and this leads to Fisher information $J(\rho) = (1 + \rho^2)/(1 - \rho^2)^2$.

Ex. 5.39 *Examples of agnostic ML operations.* It is useful to go through a list of special cases, to see how the agnostic ML theory pans out in practice. Note that convergence to the normal $\mathrm{N}_p(0, J^{-1}KJ^{-1})$ takes place in general, model after model after model (including those you might invent next week), without any need for working with explicit formulae for the ML estimators etc.

(a) For the exponential model $\theta \exp(-\theta y)$, show that the score function is $u(y, \theta) = 1/\theta - y$, that its least false parameter value is $\theta_0 = 1/\xi_0$, in termas of the true mean $\xi_0 = \mathrm{E}\, Y$. Show that $\sqrt{n}(\widehat{\theta} - \theta_0)$ has limit distribution $\mathrm{N}(0, \sigma_0^2\theta_0^4)$, where σ_0^2 is the true variance. Show that this generalises the 'usual result' derived under model conditions.

(b) Then do the normal: assume data follow some density g, and the normal $\mathrm{N}(\xi, \sigma^2)$ model is used. We already know that the least false parameters are ξ_0 and σ_0, the true mean and standard deviation (i.e. even if g is far from the normal). Assume that the fourth moment is finite, so that skew $= \mathrm{E}\, Z^3$ and kurt $= \mathrm{E}\, Z^4 - 3$ are finite, with $Z = (Y - \mathrm{E}\, Y)/\mathrm{sd}(Y) = (Y - \xi_0)/\sigma_0$. Working with the score function, and the second order derivatives, show that

$$J = \frac{1}{\sigma_0^2}\begin{pmatrix} 1, & 0 \\ 0, & 2 \end{pmatrix} \quad \text{and} \quad K = \frac{1}{\sigma_0^2}\begin{pmatrix} 1, & \gamma_3 \\ \gamma_3, & 2 + \gamma_4 \end{pmatrix}.$$

(c) For the ML estimators $\widehat{\xi}$ and $\widehat{\sigma}$, show from this that

$$\begin{pmatrix} \sqrt{n}(\widehat{\xi}-\xi) \\ \sqrt{n}(\widehat{\sigma}-\sigma) \end{pmatrix} \to_d \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1, & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3, & \frac{1}{2}+\frac{1}{4}\gamma_4 \end{pmatrix}).$$

Note that this is a 'rediscovery' of what we found in Ex. **??** and 2.46, but here we managed to find the limit distribution fully without knowing (or caring) about the exact expressions for the ML estimators.

(d) (xx one more case to come here. xx)

Ex. 5.40 *An average power optimality property.* (xx we shall see how this pans out, and how it can be best told. make connection to BIC of Ch. 11. xx) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from a smooth parametric model $f(y, \theta)$, where we need to test $\theta = \theta_0$, a given value, against $\theta \neq \theta_0$. In general there is no uniformly most powerful test. We have seen in Ex. 4.11, however, that there is a well-defined test maximising the weighted average power $\bar{\pi}_n = \int \pi_n(\theta)\,\mathrm{d}w(\theta)$, with $\pi_n(\theta)$ the power at position θ and $\mathrm{d}w(\theta)$ a given probability measure on the alternative region, here $\theta \neq \theta_0$. This optimal strategy is to use the Neyman–Pearson Lemma for the marginal density $\bar{f}(y_1, \ldots, y_n) = \int \prod_{i=1}^n f(y_i, \theta)\,\mathrm{d}w(\theta)$.

(a) Let $\ell_n(\theta)$ be the log-likelihood function, with $\widehat{\theta}$ the ML estimator, and write also f_0 for the model at the null value θ_0. Show that the Neyman–Pearson ratio can be expressed as

$$\begin{aligned} R_n = \frac{\bar{f}(y_1, \ldots, y_n)}{f_0(y_1, \ldots, y_n)} &= \int \exp\{\ell_n(\theta) - \ell_n(\theta_0)\}\,\mathrm{d}w(\theta) \\ &= \exp\{\ell_n(\widehat{\theta}) - \ell_n(\theta_0)\} \int \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta})\}\,\mathrm{d}w(\theta). \end{aligned}$$

(b) For θ close to $\widehat{\theta}$, use Taylor expansion to get

$$\ell_n(\theta) - \ell_n(\widehat{\theta}) \doteq -\tfrac{1}{2} n(\theta - \widehat{\theta})^{\mathrm{t}} J_n (\theta - \widehat{\theta}),$$

with $J_n = -n^{-1}\partial^2 \ell_n(\widehat{\theta})/\partial\theta\,\partial\theta^{\mathrm{t}}$ the normalised Hessian matrix at the max point.

(c) The optimal test consists in rejecting when R_n is above its null distribution threshold. Show that the above leads to

$$\begin{aligned} R_n &\doteq \exp(\tfrac{1}{2} D_n)(2\pi)^{p/2} |n J_n|^{1/2} w(\widehat{\theta}), \\ 2\log R_n &\doteq D_n - p\log n + \log|J_n| + p\log(2\pi) + 2\log w(\widehat{\theta}). \end{aligned}$$

Here $D_n = 2\{\ell_{n,\max} - \ell_n(\theta_0)\}$ is the Wilks or log-likelihood-ratio test statistic, with its χ^2_p limiting null distribution.

(d) Conclude that the D_n test is an approximation to the maximum averaged power test, almost regardless of the weighting measure.

(e) (xx round this off. example. xx)

Divergences and likelihoods in regression models

Ex. 5.41 *Extending theory and methods to regression setups, I.* Above we have dealt with likelihood methods, involving ML estimation, limit distributions under and outside model conditions, the Wilks theorem for profiled log-likelihoods, broadly valid for all smooth parametric models, etc. – but after all under simple i.i.d. conditions. Crucially, most of these concepts, methods, and results extend to classes of general regression models. Here we go through the various steps to see how the scene broadens and to learn the appropriate extensions for concepts, techniques, and results.

Consider in general terms regression data of the form (x_i, Y_i), with x_i a covariate vector, of length say p, thought to influence the main outcome Y_i. We assume here that the Y_i are independent given the covariates. Let $f(y_i \,|\, x_i, \theta)$ be a suitable density for y_i given x_i, with score function $u(y_i \,|\, x_i, \theta) = \partial \log f(y_i \,|\, x_i, \theta)/\partial\theta$ and information function $i(y_i \,|\, x_i, \theta) = \partial^2 \log f(y_i \,|\, x_i, \theta)/\partial\theta\,\partial\theta^{\rm t}$. The θ could comprise both regression coefficients and parameters describing the shape of the distributions. In this exercise we assume that the model holds, with θ_0 denoting the true parameter, an inner point in the parameter spoace. We also postulate *the ergodic condition*, that averages over covariates stabilise, with increasing sample size; formally, for each bounded $h(x)$, there is a well-defined limit $h_0 = \int h(x)\,{\rm d}R(x)$ for $n^{-1}\sum_{i=1}^n h(x_i)$, for an appropriate distribution R on the covariate space. For theory and applications, we do not need to model this R, or take it explicitly into account, beyond postulating its existence.

ergodic conditions

(a) First of all, there is a log-likelihood function, also in these regression setups, $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i \,|\, x_i, \theta)$. The ML estimator $\widehat\theta$ is its maximiser, satisfying also $U_n(\widehat\theta) = 0$, with $U_n(\theta) = \sum_{i=1}^n u(y_i \,|\, x_i, \theta)$. Secondly, to extend theory and results for the i.i.d. case, see Ex. 5.17, we need to understand $U_n = n^{-1/2}U_n(\theta_0) = n^{-1/2}\sum_{i=1}^n u(Y_i \,|\, x_i, \theta_0)$. Show that it has mean zero and variance matrix $J_n = n^{-1}\sum_{i=1}^n J(x_i)$, where $J(x_i) = {\rm Var}_{\theta_0} u(Y_i \,|\, x_i, \theta_0)$. Under ergodic assumptions, there is convergence $J_n \to J$, say. Give Lindeberg type conditions under which $U_n \to_d U \sim {\rm N}_p(0, J)$.

(b) Extend techniques from Ex. 5.17 to deduce the natural parallel to (5.6), in this general regression setting, that

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = U_n^{\rm t}s - \tfrac12 J_n s + r_n(s) \to_d A(s) = U^{\rm t}s - \tfrac12 s^{\rm t}Js,$$

Set up clear but mild regularity conditions, as in Ex. 5.13, to secure the required $r_n(s) \to_{\rm pr} 0$ for each s. Explain as in Ex. 5.17 that this leads to the two fundamental results

$$\sqrt{n}(\widehat\theta - \theta_0) \to_d {\rm N}_p(0, J^{-1}) \quad \text{and} \quad 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d \chi^2_p,$$

under these Lindeberg type conditions. Show also that the observed Fisher information matrix

observed Fisher information matrix

$$\widehat J_{n,\rm full} = -\partial^2 \ell_n(\widehat\theta)/\partial\theta\,\partial\theta^{\rm t}, \tag{5.12}$$

i.e. the Hessian mastrix associated with the maximisation of the log-likelihood, satisfies $\widehat J_n = n^{-1}\widehat J_{\rm full} \to_{\rm pr} J$. Deduce from this that $\widehat\theta \approx_d {\rm N}_p(\theta_0, \widehat J_{\rm full}^{-1})$. (xx need to point to Ch2 thing with weak LLN for averages of non-i.i.d. xx)

(c) (xx rounding this off. also log-likelihood profiling and deviance and χ^2_1 limit and CDs. the point is that i.i.d. results all extend to the broad regression cases. point also to exercise below with outside the model. xx)

Ex. 5.42 *Linear regression revisited.* (xx edit and clean. xx) Consider the linear regression model of Ex. 3.31, with $Y_i \,|\, x_i \sim \mathrm{N}(x_i^{\mathrm{t}}\beta, \sigma^2)$, for which exact finite-sample theory has been well developed. We now take another look at this classical model, with the general likelihood tools.

(a) For the log-likelihood, show that $\ell_n(\beta,\sigma) = -n\log\sigma - \frac{1}{2}Q(\beta)/\sigma^2 - \frac{1}{2}n\log(2\pi)$, with $Q(\beta) = \sum_{i=1}^n (y_i - x_i^{\mathrm{t}}\beta)^2$ Show that the ML estimator for β is the least squares estimator $\widehat{\beta} = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i Y_i$, with $\Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}}$, see the exercise mentioned, and that $\widehat{\sigma} = (Q_0/n)^{1/2}$, with $Q_0 = Q(\widehat{\beta})$ the minimum of $Q(\beta)$.

(b) Show that the score function becomes

$$u(y_i \,|\, x_i, \beta, \sigma) = \begin{pmatrix} (1/\sigma^2)(y_i - x_i^{\mathrm{t}}\beta)x_i \\ -1/\sigma + (1/\sigma^3)(y_i - x_i^{\mathrm{t}}\beta)^2 \end{pmatrix} = \begin{pmatrix} (1/\sigma)\varepsilon_i x_i \\ (1/\sigma)(\varepsilon_i^2 - 1) \end{pmatrix}$$

in terms of $\varepsilon_i = (y_i - x_i^{\mathrm{t}}\beta)/\sigma$, which are independent standard normals under the model. With (β_0, σ_0) the true parameters, deduce that the $(p+1)\times(p+1)$ Fisher information matrix becomes

$$J_n = n^{-1}\sum_{i=1}^n \mathrm{Var}_{\beta_0,\sigma_0} u(Y_i \,|\, x_i, \beta_0, \sigma_0) = (1/\sigma_0^2)\,\mathrm{diag}(\Sigma_n, 2),$$

with $\Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}} = n^{-1}X^{\mathrm{t}}X$. Show also that the observed Fisher information matrix becomes $\widehat{J}_{n,\mathrm{full}} = (n/\widehat{\sigma}^2)\,\mathrm{diag}(\Sigma_n, 2)$.

(c) (xx then on to what likelihood theory implies for $\widehat{\beta}$ and $\widehat{\sigma}$. the J_n and $\widehat{J}_n$. a t_{n-p} vs. approximate normality. we reproduce the $\widehat{\beta}$ distribution, and come close for $\widehat{\sigma}^2$. xx)

Ex. 5.43 *Logistic regression.* Consider binary outcome data, where the values 0-1 for Y_i are influenced by a covariate vector x_i, of dimension say p. The logistic regression model takes the probabilities to be

logistic regression

$$p_i = \Pr(Y_i = 1 \,|\, x_i) = H(x_i^{\mathrm{t}}\beta) = \frac{\exp(x_i^{\mathrm{t}}\beta)}{1+\exp(x_i^{\mathrm{t}}\beta)} \quad \text{for } i = 1,\ldots,n, \tag{5.13}$$

with $H(u) = \exp(u)/\{1+\exp(u)\}$ the logistic transform, studied in Ex. 1.57. One may interpret the model via underlying or latent i.i.d. variables Z_i, having the logistic distribution, with the individuals having different thresholds; $p_i = \Pr(Z_i \le x_i^{\mathrm{t}}\beta)$ says that outcomes are '1' for those individuals whose Z_i fall to the left of the threshold.

(a) Show that $H(u) = p$ means $u = H^{-1}(p) = \log\{p/(1-p)\}$, so that the model can be represented as $\log\{p_i/(1-p_i)\} = x_i^{\mathrm{t}}\beta$.

(b) Show that $\Pr(Y_i = y \,|\, x_i) = p_i^y(1-p_i)^{1-y}$, for the two outcomes, and deduce that the log-likelihood function can be written

$$\ell_n(\beta) = \sum_{i=1}^n \{y_i \log p_i + (1-y_i)\log(1-p_i)\} = \sum_{i=1}^n \big[y_i x_i^{\mathrm{t}}\beta - \log\{1+\exp(x_i^{\mathrm{t}}\beta)\}\big].$$

Show from this that the estimation equation, giving rise to the ML estimator $\widehat{\beta}$, is $\sum_{i=1}^n (y_i - p_i)x_i = 0$, and that

$$J_{n,\text{full}}(\beta) = -\frac{\partial^2 \ell_n(\beta)}{\partial\beta\partial\beta^{\mathrm{t}}} = \sum_{i=1}^n p_i(1-p_i)x_i x_i^{\mathrm{t}} = \sum_{i=1}^n H(x_i^{\mathrm{t}}\beta)\{1-H(x_i^{\mathrm{t}}\beta)\}x_i x_i^{\mathrm{t}}.$$

This matrix is assumed here to be positive definite, which in particular requires $n \geq p$. Explain that the log-likelihood function is concave, with $\widehat{\beta}$ the unique maximiser.

(c) Show that, under model conditions, $\widehat{\beta} \approx_d \mathrm{N}_p(\beta, \widehat{J}_{n,\text{full}}^{-1})$, where $\widehat{J}_{n,\text{full}} = J_{n,\text{full}}(\widehat{\beta})$ is the observed Fisher information matrix (the Hessian matrix of minus the normalised log-likelihood function, at the ML position).

(d) Consider an individual, perhaps outside the dataset, with covariate vector x_0. Show that $x_0^{\mathrm{t}}\widehat{\beta}$ is approximately a normal $(x_0^{\mathrm{t}}\beta, x_0^{\mathrm{t}}\widehat{J}_{n,\text{full}}^{-1}x_0)$, and use this to construct a confidence interval for $p(x_0) = \Pr(Y_0 = 1 \,|\, x_0)$.

(e) (xx repair a bit here. xx) Consider the important special case of a single x_i recorded for Y_i, where we write the model equation as $p_i = H(a + bx_i)$, corresponding to 2-size vectors $(1, x_i)^{\mathrm{t}}$ in the more general notation used above. Show that $(\widehat{a}, \widehat{b})$ are the solutions to $\sum_{i=1}^n (y_i - p_i) = 0$ and $\sum_{i=1}^n (y_i - p_i)x_i = 0$, and that

$$J_n(a,b) = \sum_{i=1}^n \frac{\exp(a+bx_i)}{\{1+\exp(a+bx_i)\}^2}\begin{pmatrix} 1, & x_i \\ x_i, & x_i^2 \end{pmatrix}.$$

(f) The logistic regression model uses the H transform, which is symmetric around zero. A more flexible model takes $p_i = H(x_i^{\mathrm{t}}\beta)^\kappa$, allowing skewness in the underlying distribution for the latent variables. Write down the log-likelihood function, explain that standard theory applies for the ML estimator $(\widehat{\beta}, \widehat{\kappa})$. (xx point perhaps to illustration, Polish girls story, where κ is significantly larger than 1. xx)

Ex. 5.44 *Probit regression.* Consider again data (x_i, Y_i), with covariates x_i and 0-1 outcomes Y_i. For the logistic regression of Ex. 5.43 we modelled the 1-probabilities as $p_i = \Pr(Z_i \leq x_i^{\mathrm{t}}\beta) = H(x_i^{\mathrm{t}}\beta)$, with H the logistic c.d.f. Clearly there are alternatives, and the more famous one among these is *probit regression*, which takes the Z_i to be standard normal, i.e. $p_i = \Phi(x_i^{\mathrm{t}}\beta)$.

probit regression, from probability of unit

(a) Explain that the log-likelihood becomes $\ell_n(\beta) = \sum_{i=1}^n [y_i \log p_i(\beta) + (1-y_i)\log\{1 - p_i(\beta)\}]$. The ML estimator $\widehat{\beta}$ is its maximiser.

(b) Explain also that standard regression likelihood theory applies, with $\widehat{\beta} \approx_d \mathrm{N}_p(\beta, \widehat{J}_n^{-1})$, and give a recipe for computing the variance matrix.

(c) To compare analyses from logistic and probit regressions, we may scale the normal above to have the same variance $\tau^2 = \pi^2/3$ as the logistic. Explain that this leads to an equivalent model, though with scaled parameters, namely $p_i = \Pr(Z_i' \leq x_i^{\mathrm{t}}\beta) = \Phi(x_i^{\mathrm{t}}\beta/\tau)$.

(d) Simulate a simple dataset of size $n = 1000$ pairs, with x_i taken $\mathrm{N}(0, \kappa^2)$ with $\kappa = 2$ and true probabilities $p_i = H(a_0 + b_0 x_i)$ with $a_0 = -0.33$ and $b_0 = 0.77$. The litt point with $\kappa = 2$ is to secure a fair range for the x_i, in view of the probabilities. (i) Estimate (a, b) in the logistic regression model $p_i = H(a + bx_i)$; and then (ii) estimate (a, b) in the probit model $p_i = \Phi((a + bx_i)/\tau)$. Draw the two estimated regression curves, say $H(\widehat{a} + \widehat{b}x)$ and $\Phi((\widetilde{a} + \widetilde{b}x)/\tau)$. Comment on what you find.

Ex. 5.45 *Poisson regression.* Consider independent count data $y_1, \ldots, y_n$, influenced by covariate vectors $x_1, \ldots, x_n$. The Poisson regression model, in its standard form, takes $Y_i \sim \mathrm{Pois}(\mu_i)$, with $\mu_i = \exp(x_i^{\mathrm{t}}\beta)$.

(a) Show that the log-likelihood function becomes

$$\ell_n(\beta) = \sum_{i=1}^n \{-\mu_i + y_i \log(\mu_i)\} = \sum_{i=1}^n \{y_i x_i^{\mathrm{t}}\beta - \exp(x_i^{\mathrm{t}}\beta)\},$$

and that the equations $\sum_{i=1}^n \{y_i - \mu_i(\beta)\}x_i = 0$ define the ML estimators.

(b) Show that $-\partial^2 \ell_n(\beta)/\partial\beta\,\partial\beta^{\mathrm{t}} = \sum_{i=1}^n \mu_i x_i x_i^{\mathrm{t}}$, leading to the observed Fisher information matrix $\widehat{J}_{n,\mathrm{full}} = \sum_{i=1}^n \widehat{\mu}_i x_i x_i^{\mathrm{t}}$, where $\widehat{\mu}_i = \exp(x_i^{\mathrm{t}}\widehat{\beta})$.

(c) For a case with covariate vector x_0, estimate the associated expected count $\mu_0 = \exp(x_0^{\mathrm{t}}\beta)$, and construct a confidence interval.

(d) Sometimes there is overdispersion, compared to how the counts y_i should behave under Poisson conditions. Such overdispersion could e.g. reflect 'hidden covariates' not taken on board in the model. A method handling overdispersion is to take $Y_i \,|\, \mu \sim \mathrm{Pois}(\mu_i)$, but modelling potential extra randomness via $\mu_i \sim \mathrm{Gam}(\exp(x_i^{\mathrm{t}}\beta)/\tau, 1/\tau)$. This is an application of Poisson-gamma mixtures from Ex. 1.26. Show that μ_i has mean $\exp(x_i^{\mathrm{t}}\beta)$ and variance $\tau \exp(x_i^{\mathrm{t}}\beta)$; τ small means getting back to ordinary Poisson regression. Show altso that Y_i has mean $\exp(x_i^{\mathrm{t}}\beta)$ variance $(1+\tau)\exp(x_i^{\mathrm{t}}\beta)$. Furthermore, work out that an expression for the log-likelihood function $\ell_n(\beta, \tau)$ is

$$\sum_{i=1}^n \{(a_i/\tau)\log(1/\tau) - \log\Gamma(a_i/\tau) + \log\Gamma(a_i/\tau + y_i) - (a_i/\tau + y_i)\log(1/\tau + 1)\},$$

in which $a_i = \exp(x_i^{\mathrm{t}}\beta)$. In applications it is often fruitful to profile out the β, and studying $\ell_{n,\mathrm{prof}}(\tau) = \max_{\mathrm{all}\ \beta} \ell_n(\beta, \tau)$. (xx pointer to Story iv.6. xx)

Ex. 5.46 *GLM regression.* (xx to be polished. xx) basic expofamily model $f(y, \theta, \kappa) = \exp\{\theta T(y) + \kappa U(y) - k(\theta, \kappa)\}h(y)$, here with one-dimensional θ, κ. now regression with $\theta_i = x_i^{\mathrm{t}}\beta$.

$$\ell_n(\beta, \kappa) = \sum_{i=1}^n \{x_i^{\mathrm{t}}\beta T(y_i) + \kappa U(y_i) - k(x_i^{\mathrm{t}}\beta, \kappa)\}.$$

then ML $\widehat{\beta}, \widehat{\kappa}$, and more.

(a)

(b) (xx example. xx)

Ex. 5.47 *Wald tests.* (xx something here, re Wald tests, used e.g. in regression models. p-value. more on power for the two variations with two nevnere. point back to Ex. 4.5, and to a couple of stories where we use this tool. xx)

Ex. 5.48 *A heteroscedastic linear regression model.* (xx edit and clean. xx) In various linear regression type applications for (x_i, y_i) data the linear mean assumption can be reasonable, whereas the variance might not be taken constant across covariates. Consider therefore the model with independent $Y_i \,|\, (x_i, w_i) \sim \mathrm{N}(x_i^{\mathrm{t}}\beta, \sigma_i^2)$, for $i = 1, \ldots, n$, with covariate vectors x_i of length p and variance related covariates w_i of length q, influencing $\sigma_i = \sigma \exp(\gamma^{\mathrm{t}} w_i)$. These w_i could be a subset of the x_i or functions thereof. It is convenient to normalise these such that $\bar{w} = n^{-1}\sum_{i=1}^n w_i = 0$, which also means that σ is the standard deviation for an average individual, with w_i equal to $\bar{w}$.

(a) (xx log-likelihood. score function. J_n and $\widehat{J}_n$. approximations. pointers. xx) Show that the log-likelihood function can be written

$$\ell_n(\beta, \gamma) = -n \log \sigma - \tfrac{1}{2}(1/\sigma^2) Q(\beta, \gamma), \quad \text{with } Q(\beta, \gamma) = \sum_{i=1}^n \frac{(y_i - x_i^{\mathrm{t}}\beta)^2}{\exp(2w_i^{\mathrm{t}}\gamma)}.$$

Show that minimising $Q(\beta, \gamma)$ over β, for fixed γ, takes place for

$$\widehat{\beta}(\gamma) = \Big\{ \sum_{i=1}^n \frac{x_i x_i^{\mathrm{t}}}{\exp(2\gamma^{\mathrm{t}} w_i)} \Big\}^{-1} \sum_{i=1}^n \frac{x_i y_i}{\exp(2\gamma^{\mathrm{t}} w_i)}.$$

Demonstrate that this leads to the profiled log-likelihood $\ell_{n,\mathrm{prof}}(\gamma) = -n \log \widehat{\sigma}(\gamma) - \frac{1}{2}n$, where $\widehat{\sigma}(\gamma)^2 = Q_0(\gamma)/n$, with $Q_0(\gamma) = Q(\widehat{\beta}(\gamma), \gamma)$ the minimum sum of squares. Deduce from this that a recipe for finding the ML estimators consists in (i) minimising $Q_0(\gamma)$ over γ, yielding $\widehat{\gamma}$; (ii) reading off $\widehat{\beta} = \widehat{\beta}(\widehat{\gamma})$ and $\widehat{\sigma} = \widehat{\sigma}(\widehat{\gamma})$.

(b) (xx calibrate this with Wilks things. xx) Is it worthwhile, turning from classic linear regression, to include the extra layer of variance heterogeneity sophistication? Show that the log-likelihood-ratio test becomes that of comparing $D_n = 2n \log\{\widehat{\sigma}(0)/\widehat{\sigma}(\widehat{\gamma})\}$ to the χ^2_q, in which $\widehat{\sigma}(0)^2 = Q_0(0)/n$ is the standard estimator for σ^2 under variance constancy.

(c) For the $p + q + 1$-parameter model, with parameters β, γ, σ, show that the score function becomes

$$u(y_i \,|\, x_i) = \begin{pmatrix} (1/\sigma^2)(y_i - x_i^{\mathrm{t}}\beta) x_i / \exp(2\gamma^{\mathrm{t}} w_i) \\ -w_i + (1/\sigma^2)(y_i - x_i^{\mathrm{t}}\beta)^2 w_i / \exp(2\gamma^{\mathrm{t}} w_i) \\ -1/\sigma + (1/\sigma^3)(y_i - x_i^{\mathrm{t}}\beta)^2 / \exp(2\gamma^{\mathrm{t}} w_i) \end{pmatrix} = \begin{pmatrix} (1/\sigma)\varepsilon_i x_i / \exp(\gamma^{\mathrm{t}} w_i) \\ (1/\sigma)(\varepsilon_i^2 - 1) w_i \\ (1/\sigma)(\varepsilon_i^2 - 1) \end{pmatrix},$$

in terms of $\varepsilon_i = (y_i - x_i^{\mathrm{t}}\beta)/\{\sigma \exp(\gamma^{\mathrm{t}} w_i)\}$. Show from this that the normalised Fisher information matrix becomes $J_n = (1/\sigma_0^2)\mathrm{diag}(\Sigma_n(\gamma_0), M_n, 2)$, at the true parameters $(\beta_0, \gamma_0, \sigma_0)$, in terms of $\Sigma_n(\gamma) = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}} / \exp(2\gamma^{\mathrm{t}} w_i)$ and $M_n = n^{-1}\sum_{i=1}^n w_i w_i^{\mathrm{t}}$.

(d) (xx check if $\widehat{J}_{n,\mathrm{full}}$ has these off-diagonal zeroes, or if it only holds for the information calculus. spell out nice behaviour for ML estimators. $\widehat{\gamma} \approx_d \mathrm{N}_q(\gamma, (\sigma^2/n) M_n^{-1})$. xx)

(e) (xx round off. confidence for $\mu(x_0, w_0)$, $x_0^{\rm t}\widehat{\beta} \pm 1.96\,\widehat{\sigma}\exp(\widehat{\gamma}^{\rm t}w_0)$. pointer to Story iv.3. xx)

Ex. 5.49 *Nonlinear regression.* (xx calibrate this with both classic linear regression and what we've said with genereal regression models. xx) Consider in general terms the model with independent $y_i \sim {\rm N}(m_i(\beta), \sigma^2)$ for $i = 1, \ldots, n$, wherethe means $m_i(\beta)$ are perhaps nonlinear functions of an appropriate vector prarameter β, involving also covariates.

(a) Show that the log-likelihood function becomes $-n\log\sigma - \frac{1}{2}Q_n(\beta)/\sigma^2$, with $Q_n(\beta) = \sum_{i=1}^n \{y_i - m_i(\beta)\}^2$. Show that the ML estimator for β is the minimiser of Q_n, and that $\widehat{\sigma}^2 = Q_n(\widehat{\beta})/n$.

(b) Show that the normalised Fisher information matrix becomes

$$J_n(\beta,\sigma) = \frac{1}{\sigma^2}\begin{pmatrix}\Sigma_n & 0\\ 0 & 2\end{pmatrix}, \quad \text{with } \Sigma_n = n^{-1}\sum_{i=1}^n m_i^*(\beta)m_i^*(\beta)^{\rm t},$$

in which $m_i^*(\beta) = \partial m_i(\beta)/\partial\beta$.

(c) Deduce that $\sqrt{n}(\widehat{\beta} - \beta) \to_d {\rm N}(0, \sigma^2\Sigma^{-1})$, under Lindeberg type conditions, where Σ is the limit of Σ_n. (xx more here. also the case with different normalisation, for the time series cyclic thing. xx)

Ex. 5.50 *We can do things.* The spirit of this exercise is to see that the log-likelihood machinery is useful, versatile, flexible, and not too hard to use in new situations, outside ordinary textbook terrain.

(a) One records the number of a certain event, say per week, over a time period. Suppose most of these counts are Poisson-like, with some parameter θ, but that a fraction come from another Poisson with a higher parameter. A model for such data is that Y_i stems from the mixture distribution $(1-p)\,{\rm Pois}(\theta) + p\,{\rm Pois}(c\theta)$, with $c > 1$. (i) Find the mean and variance for this distribution. (ii) Generate such a dataset, say $y_1, \ldots, y_n$ with $n = 250$, $\theta = 10$, $c\theta = 20$. Taking first $c = 2$ known, estimate the parameters (p, θ), along with confidence intervals. (iii) Using the same data, now with c unknown, estimate all three parameters, with confidence intervals. Briefly investigate how much is earned in precision for estimating (p, θ) when c is known compared to it being unknown.

(b) Suppose data Y_i^0 are generated according to some normal (ξ, σ^2), but that we are only able to see those falling inside some observation window $[a, b]$. Write down the log-likelihood function, for the observed data Y_i. To see how it works, generate say $n_0 = 1000$ points from the standard normal, keep the n amongst these that fall inside $[1.0, 3.0]$, and estimate the two parameters from these. Find also approximate confidence intervals for the two parameters.

(c) For two distributions with densities f_1, f_2 and c.d.f.s F_1, F_2, consider the bivariate density

$$f(x, y, a) = f_1(x)f_2(y)[1 + a\{F_1(x) - \tfrac{1}{2}\}\{F_2(y) - \tfrac{1}{2}\}].$$

What is the range of the dependence parameter a? Show that the marginal distributions are f_1, f_2. Generate first binormal data with $\mathrm{N}(0,1)$ marginals, and estimate a, with confidence interval for a. Then generate another binormal dataset, and estimate the five parameters $(\xi_1, \xi_2, \sigma_1, \sigma_2, a)$, along with confidence intervals.

(d) Tired of normality? Then generalise it. With $\mathrm{Be}_{a,b}(x)$ the c.d.f. for the $\mathrm{Beta}(a,b)$ distribution, consider the four-parameter model with c.d.f. $F(y) = \mathrm{Be}_{a,b}(\Phi((y-\xi)/\sigma))$. Write up an expression for the log-likelihood function. Generate first $n = 500$ datapoints from the standard normal, and check if the ML metod succeeds in coming close to the true parameters $(0,1,1,1)$. Argue that you here have a test for normality, and spell out the ingredients. Then simulate another dataset, from $(a,b) = (1.23, 2.34)$, compute ML estimates and their estimated standard deviations.

Ex. 5.51 *Extending theory and methods to regression setups, II.* (xx smooth and polish. regression outside model conditions. careful with notion, point to previous for u, i. then $\theta_{0,n}$, J_n, K_n, sandwich. xx) For i.i.d. setups we have seen how the behaviour of ML estimators can be accurately described also when the data-generating density g is not inside the parametric model f_θ, via a clearly defined least false parameter θ_0 and the two matrices J and K, see Ex. 5.17. Under model conditions, the two are equal. Here we extend these notions and results to regression models. The setting is partly as in Ex. 5.41, with a parametric model $f(y_i \,|\, x_i, \theta)$ for $Y_i \,|\, x_i$, but notions and results need to be lifted to the agnostic state of affairs when the real density $g(y_i \,|\, x_i)$ is not necessarily inside the model.

(a) For given x, there is a KL distance

$$\mathrm{KL}_x(g(\cdot \,|\, x), f(\cdot \,|\, x, \theta)) = \int g(y \,|\, x) \log \frac{g(y \,|\, x)}{f(y \,|\, x, \theta)} \, \mathrm{d}y.$$

Consider the overall KL distance, weighted over the covariate vectors in the sample,

$$\mathrm{KL} = n^{-1} \sum_{i=1}^n \mathrm{KL}_{x_i} = \int \int g(y \,|\, x) \log \frac{g(y \,|\, x)}{f(y \,|\, x, \theta)} \, \mathrm{d}y \, \mathrm{d}R_n(x),$$

with R_n the empirical distribution over these covariate vectors. Show that this is a divergence, i.e. nonnegative and equal to zero only if $g(y \,|\, x) = f(y \,|\, x, \theta)$ for all y and all $x_1, \ldots, x_n$. Let $\theta_{0,n}$ be the least false parameter, minimising this KL, for the given covariate vectors, and show that this is the same as the maximiser of

$$M_n(\theta) = n^{-1} \sum_{i=1}^n \int g(y \,|\, x_i) \log f(y \,|\, x_i, \theta) \, \mathrm{d}y.$$

Explain also that $M_n(\theta) \to M(\theta)$, under ergodic conditions, with this limit involving the distribution R over covariate space, using $\int h \, \mathrm{d}R_n \to \int h \, \mathrm{d}R$. Letting θ_0 be the maximiser of this limit $M(\theta)$, show also that $\theta_{0,n} \to \theta_0$, under weak conditions.

(b) The following arguments support the notion that the ML estimator $\widehat{\theta}$ aims at this least false $\theta_{0,n}$. With $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i \,|\, x_i, \theta)$ the log-likelihood function, explain that $n^{-1}\ell_n(\theta)$ has mean $M_n(\theta)$. the mean of $n^{-1}\ell_n(\theta)$ is $M_n(\theta)$, and that its variance goes to zero. Now show that $\widehat{\theta} - \theta_{0,n} \to_{\mathrm{pr}} 0$.

(c) As in the earlier and simpler case of (5.6), it is fruitful to work with the log-likelihood function process. Show that

$$A_n(s) = \ell_n(\theta_{0,n} + s/\sqrt{n}) - \ell_n(\theta_{0,n}) = U_n^{\mathrm{t}} s - \tfrac{1}{2} s^{\mathrm{t}} J_n s + r_n(s),$$

now with $U_n = n^{-1/2} \sum_{i=1}^n u(Y_i \,|\, x_i, \theta_{0,n})$ and $J_n = -n^{-1} \sum_{i=1}^n i(Y_i \,|\, x_i, \theta_{0,n})$. Let $K(x_i) = \mathrm{Var}\, u(Y_i \,|\, x_i, \theta_{0,n})$. Under Lindeberg type conditions, show that $U_n \to_d U \sim \mathrm{N}_p(0, K)$, with K the limit of $n^{-1} \sum_{i=1}^n K(x_i)$. Explain that $A_n(s) \to_d A(s) = U^{\mathrm{t}} s - \frac{1}{2} s^{\mathrm{t}} J s$.

(d) (xx then read off limits. proof of pudding lies in eating, i.e. applications, below. estimates $\widehat{J}$ and $\widehat{K}$. xx) $\widehat{J} = -n^{-1} \partial^2 \ell_n(\widehat{\theta}) / \partial\theta\, \partial\theta^{\mathrm{t}}$, for estimating K, use the estimated scores $\widehat{u}_i = u(Y_i \,|\, x_i, \widehat{\theta})$, noting that these sum to zero. then use $\widehat{K} = n^{-1} \sum_{i=1}^n \widehat{u}_i \widehat{u}_i^{\mathrm{t}}$. sandwich $\widehat{\Sigma} = \widehat{J}^{-1} \widehat{K} \widehat{J}^{-1}$.

Ex. 5.52 *Linear regression: agnostic analysis.* (xx to be repaired and cleaned and calibrated with next. xx) When a regression model is used, without being fully correct, the general theory of Ex. 5.51 explains (i) how the ML estimators implicitly aim for the best parametric approximation, defined via Kullback–Leibler divergences weighted over the available covariate vectors, and (ii) how the approximate distributions are affected. Here we consider such consequences for the classic linear normal regression model. For a dataset of pairs (x_i, Y_i), assume merely that $\mathrm{E}\, Y_i = m(x_i)$, for some smooth $m(x)$ function, with i.i.d. errors ε_i with mean zero and variance σ_0^2. The linear model approximates this via the standard $Y_i \,|\, x_i \sim \mathrm{N}(x_i^{\mathrm{t}} \beta, \sigma^2)$, for which we have found the ML estimators $\widehat{\beta}$ and $\widehat{\sigma}$ in Ex. 5.42.

(a) Explain that minimising the KL distance, as per Ex. 5.51, is the same as maximising

$$n^{-1} \sum_{i=1}^n \{-\log\sigma - \tfrac{1}{2} \frac{1}{\sigma^2} \mathrm{E}\,(Y_i - x_i^{\mathrm{t}}\beta)^2\} = -\log\sigma - \tfrac{1}{2} \frac{1}{\sigma^2} n^{-1} \sum_{i=1}^n [\{m(x_i) - x_i^{\mathrm{t}}\beta\}^2 + \sigma_0^2].$$

Argue from this that the least false parameter $\beta_{0,n}$ is the one minimising $Q_n(\beta) = n^{-1} \sum_{i=1}^n \{m(x_i) - x_i^{\mathrm{t}}\beta\}^2$, and show that this means $\beta_{0,n} = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n m(x_i) x_i$, where $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^{\mathrm{t}}$. If the model is perfect, there is a true β_{true} for which $\mathrm{E}\,(Y_i \,|\, x_i) = x_i^{\mathrm{t}} \beta_{\mathrm{true}}$, in which case $\beta_{0,n} = \beta_{\mathrm{true}}$. Show next that the least false $\sigma_{0,n}$ is determined by $\sigma_{0,n}^2 = Q_{0,n} + \sigma_0^2$, where $Q_{0,n}$ is the minimum of $Q_n(\beta)$. If the linear mean is a good model, the $Q_{0,n}$ is small and the $\sigma_{0,n}$ aimed at by the ML estimator $\widehat{\sigma}$ is not much bigger than σ_0. If the linear mean is not a good approximation to the real $m(x)$, then the $\widehat{\sigma}$ implicitly picks up both data variability and the difference between $x_i^{\mathrm{t}}\beta$ and the real $m(x_i)$.

(b) For the two matrices J and K determining the behaviour of the ML estimators we have found in Ex. 5.42 that $\widehat{J} = (1/\widehat{\sigma}^2)\, \mathrm{diag}(\Sigma_n, 2)$, and that the score vectors, computed at the least false values, can be written $u(y_i \,|\, x_i, \theta_0) = (1/\sigma_{0,n})(x_i z_i, z_i^2 - 1)^{\mathrm{t}}$, in which $z_i = (y_i - x_i^{\mathrm{t}}\beta_{0,n})/\sigma_{0,n}$. Note that we for the estimated standardised residuals $\widehat{z}_i = (y_i - x_i^{\mathrm{t}}\widehat{\beta})/\widehat{\sigma}$ have $\sum_{i=1}^n \widehat{z}_i x_i = 0$ and $n^{-1} \sum_{i=1}^n \widehat{z}_i^2 = 1$. For the K matrix, and with

notation from Ex. 5.51, show that

$$n^{-1}\sum_{i=1}^n K(x_i) = n^{-1}\sum_{i=1}^n \frac{1}{\sigma_{0,n}^2}\mathrm{Var}\begin{pmatrix} z_i x_i \\ z_i^2-1\end{pmatrix}.$$

(xx round this off. two substories, depending on what is assumed for z_i. with $y_i = m(x_i)+\sigma_0\varepsilon_i$, the z_i are not i.i.d., but we may cope with that too, via the right $\widehat{K}$. simplification if we postulate $m(x_i)=x_i^{\mathrm{t}}\beta_0$, then involving only $\widehat{\gamma}_3$ and $\widehat{\gamma}_4$ for the $\widehat{z}_i$. xx) a somewhat messy model-robust approach would use

$$\widehat{K} = \frac{1}{\widehat{\sigma}^2}n^{-1}\sum_{i=1}^n \begin{pmatrix} \widehat{z}_i^2 x_i x_i^{\mathrm{t}}, & \widehat{z}_i^3 x_i \\ \widehat{z}_i^3 x_i^{\mathrm{t}}, & \widehat{z}_i^4-1\end{pmatrix}.$$

(c)

Ex. 5.53 *Logistic regression, Poisson regression: agnostic analysis.* (xx to be repaired and cleaned. xx) When the regression model used does not necessarily hold up, ML estimators aim for the relevant least false parameters, the general theory of Ex. 5.51 shows that $\widehat{\theta}\approx_d \mathrm{N}(\theta_{0,n},\widehat{\Sigma}/n)$, for the relevant least false parameter and with estimated sandwich matrix $\widehat{\Sigma}_n = \widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$. Under model conditions, matters simplify, the underlying J_n and K_n matrices are equal, and we have the standard result $\widehat{\theta}\approx_d \mathrm{N}(\theta_0,\widehat{J}^{-1}/n)$, used extensively in statistical software packages for a range of regression models. Here we check what the model agnostic setup leads to, for securing model robust inference, for standard logistic and Poission regression models; see also the agnostic analysis for the linear regression model in Ex. 5.52.

(a) Consider logistic regression (xx point back xx), with $p_i = H(x_i^{\mathrm{t}}\beta)$: show that $u(y_i\,|\,x_i,\beta) = (y_i-p_i)x_i$, and that

$$\widehat{J}_n = n^{-1}\sum_{i=1}^n \widehat{p}_i(1-\widehat{p}_i)x_i x_i^{\mathrm{t}}, \quad \widehat{K}_n = n^{-1}\sum_{i=1}^n (y_i-\widehat{p}_i)^2 x_i x_i^{\mathrm{t}}.$$

(b) Then look at Poisson regression (xx point back xx), with $\mu_i = \exp(x_i^{\mathrm{t}}\beta)$: show that $u(y_i\,|\,x_i,\beta) = (y_i-\mu_i)x_i$, with

$$\widehat{J}_n = n^{-1}\sum_{i=1}^n \widehat{\mu}_i x_i x_i^{\mathrm{t}}, \quad \widehat{K}_n = n^{-1}\sum_{i=1}^n (y_i-\widehat{\mu}_i)^2 x_i x_i^{\mathrm{t}}.$$

If there is overdispersion, with $(y_i-\widehat{\mu}_i)^2$ tending to be bigger than $\widehat{\mu}_i$, this is picked up here, with sandwich matrix bigger than $\widehat{J}^{-1}$. The potential error of applying straightforward Poisson regression, perhaps via standard packages, is that variability is underestimated, with confidence intervals becoming too narrow. For an illustration of this, see Story iv.6.

(c) With the Poisson-gamma overdispersion regression model studied in Ex. 5.45, where $Y_i\,|\,x_i$ has mean $\exp(x_i^{\mathrm{t}}\beta)$ and variance $(1+\tau)\exp(x_i^{\mathrm{t}}\beta)$, show that $\widehat{K}_n \to_{\mathrm{pr}} (1+\tau)J$, in terms of the limit J of $\widehat{J}_n$. Deduce that if $\widehat{\beta}$ is the ML estimator computed for the standard Poisson regression model, then $\sqrt{n}(\widehat{\beta}-\beta_{0,n})\to_d \mathrm{N}_p(0,(1+\tau)J^{-1})$.

Influence functions

Ex. 5.54 *Influence functions.* For a distribution function F, consider some associated parameter, say $\theta = T(F)$, with T the appropriate functional mapping the distribution to the parameter value in question. Examples include the mean, the standard deviation, the skewness, the interquartiale range, a threshold probability. The *influence function* for $\theta = T(F)$ is a very useful quantity, as we shall see. It is defined as

$$\mathrm{IF}(F, y) = \lim_{\varepsilon \to 0} (1/\varepsilon)\{T((1-\varepsilon)F + \varepsilon\delta(y)) - T(F)\}. \tag{5.14}$$

Here $\delta(y)$ is the measure putting full mass 1 at the point y, and $(1-\varepsilon)F + \varepsilon\delta(y)$ the consequent mixture distribution. A variable Y_ε drawn from this mixture is from F with probability $1-\varepsilon$ and is equal to y with probability ε.

(a) Consider $\theta(F) = \mathrm{E}_G\, h(Y) = \int h(y)\,\mathrm{d}F(y)$, the mean of $h(Y)$. Show that $\mathrm{IF}(F, y) = h(y) - \theta(F)$. In particular, the influence is bounded when h is, but unbounded e.g. in the case of the plain mean $h(y) = y$, which signifies a potential lack of robustness of this mean parameter functional $\theta = \mathrm{E}_G\, Y$.

(b) Then consider the class of smooth functions of means. For mean type parameters $\gamma_1 = \mathrm{E}_G\, h_1(Y), \ldots, \gamma_k = \mathrm{E}_G\, h_k(Y)$, let $\theta = T(F) = A(\gamma_1(F), \ldots, \gamma_k(F))$, where $A(u_1, \ldots, u_k)$ is smooth in a neighbourhood of $(\gamma_1(F), \ldots, \gamma_k(F))$. Show that this parameter has influence function

$$\begin{aligned} \mathrm{IF}(F, y) &= c_1(F)\,\mathrm{IF}_{\gamma_1}(F, y) + \cdots + c_k(F)\,\mathrm{IF}_{\gamma_k}(F, y) \\ &= c_1(F)\{h_1(y) - \gamma_1(F)\} + \cdots + c_k(F)\{h_k(y) - \gamma_k(F)\}, \end{aligned}$$

with $c_j(F)$ is the partial derivative $\partial A(u_1, \ldots, u_k)/\partial u_j$, evaluated at $(\gamma_1(F), \ldots, \gamma_k(F))$.

(c) Writing $\mu_F = \mathrm{E}_G\, Y$ for the mean, show for the variance parameter $\sigma_F^2 = \mathrm{E}_G\, Y^2 - \mu_F^2$ that its influence function becomes

$$\mathrm{IF}(F, y) = -2\,\mu_F(y - \mu_F) + y^2 - \mathrm{E}_G\, Y^2 = (y - \mu_F)^2 - \sigma_F^2.$$

Then for the standard deviation parameter $\sigma(F)$ itself, show that its influence function becomes

$$\mathrm{IF}_\sigma(F, y) = \tfrac{1}{2}(1/\sigma_F)\{(y - \mu_F)^2 - \sigma_F^2\}.$$

(d) For a given parametric family $f(y, \theta)$, consider the ML functional $T(F)$, mapping a given F with density f to the least parameter value $\theta_0 = \theta_0(F)$, the minimiser of the information distance $\mathrm{KL}(f, f(\cdot, \theta))$, or the maximiser of $\int \log f(y, \theta)\,\mathrm{d}F(y)$. With F_n the empirical distribution of the data, placing probability $1/n$ on each data point, cf. Ex. 3.9, Show that $T(F_n)$ is the ML estimator, and that its influence function becomes $\mathrm{IF}(F, y) = J^{-1}u(y, \theta_0)$.

Ex. 5.55 *Influence for quantiles.* Assume the c.d.f. F has a smooth density f.

(a) Consider $\mu = F^{-1}(\frac{1}{2})$, the median. Show that the influence function $\mathrm{IF}(F, y)$ becomes $-\frac{1}{2}/f(\mu)$ for $y \le \mu$ and $\frac{1}{2}/f(\mu)$ for $y > \mu$. Verify that the influence function has mean zero and variance $\frac{1}{4}/f(\mu)^2$.

(b) More generally, let $\mu_q = F^{-1}(q)$, for some quantile level $q \in (0,1)$. Show that $\mathrm{IF}(F, y)$ is $-(1-q)/f(\mu_q)$ for $y \le \mu_q$ and $q/f(\mu_q)$ for $y > \mu_q$. Comment on the fact that these are bounded.

(c) Then consider a smooth function of quantiles, say $\gamma(F) = A(Q_1(F), \ldots, Q_k(F))$, where $Q_j(F) = F^{-1}(q_j)$. Show that its influence function is $\mathrm{IF}(F, y) = c_1(F)\,\mathrm{IF}_1(F, y) + \cdots + c_k(F)\,\mathrm{IF}_k(F, y)$, in terms of $\mathrm{IF}_j(F, y)$ the influence function of $F^{-1}(q_j)$ calculated as above, and where $c_j(F) = \partial A(Q_1, \ldots, Q_k)/\partial Q_j$ computed at $(F^{-1}(q_1), \ldots, F^{-1}(q_k))$. As an illustration, find and graph the influence function for $\gamma(F) = F^{-1}(0.90) - F^{-1}(0.10)$.

Ex. 5.56 *An estimator represented via its influence function.* Consider an i.i.d. sequence $Y_1, Y_2, \ldots$ from F, with $\theta = T(F)$ a parameter of interest. It may be estimated nonparametrically using $\widehat{\theta} = T(F_n)$, with F_n the empirical distribution. Here we work towards a representation of $\widehat{\theta} - \theta = T(F_n) - T(F)$ in terms of the influence function.

(a) Consider the case of $\theta = A(\gamma(F))$, where $\gamma(F) = \mathrm{E}_G\, h(Y) = \int h\,\mathrm{d}F$. Show that $\widehat{\theta} = T(F_n)$ is equal to $A(\bar{h})$, with $\bar{h} = \int h\,\mathrm{d}F_n = n^{-1}\sum_{i=1}^n h(y_i)$. Assuming $A(u)$ smooth, with two derivatives, show that

$$\widehat{\theta} = A(\gamma_0) + A'(\gamma_0)(\bar{h} - \gamma_0) + \tfrac{1}{2}A''(\gamma_0)(\bar{h} - \gamma_0)^2 + o_{\mathrm{pr}}(1/n),$$

with $\gamma_0 = \gamma(F)$. Deduce that $\mathrm{E}\,\widehat{\theta} = \theta + \frac{1}{2}A''(\gamma_0)\tau^2/n + o(1/n)$, in terms of $\tau^2 = \mathrm{Var}\, h(Y_i)$, and that $\widehat{\theta} - \theta = n^{-1}\sum_{i=1}^n \mathrm{IF}(F, y_i) + b/n + o_{\mathrm{pr}}(1/\sqrt{n})$, where $b = \frac{1}{2}A''(\gamma_0)\tau^2$.

(b) Generalise to the case of $T(F) = A(\gamma_1(F), \ldots, \gamma_p(F))$ being a smooth function of several means, as studied also in Ex. 5.54, with $\gamma_j(F) = \int h_j\,\mathrm{d}F$. Show that $\mathrm{E}\,\widehat{\theta} = \theta + b/n + o(1/n)$, with $b = \frac{1}{2}\mathrm{Tr}(A''(\gamma_0)K)$, with K the variance matrix of $(h_1(Y_i), \ldots, h_p(Y_i))^{\mathrm{t}}$, and $A''(\gamma_0)$ the second order derivative matrix of A, computed at γ_0. Show that

$$\widehat{\theta} - \theta = T(F_n) - T(F) = n^{-1}\sum_{i=1}^n \mathrm{IF}(F, Y_i) + b/n + \varepsilon_n, \tag{5.15}$$

with $\varepsilon_n = o_{\mathrm{pr}}(1/\sqrt{n})$, i.e. small enough to have $\sqrt{n}\varepsilon_n \to_{\mathrm{pr}} 0$. Show

(c) The powerful representation (5.15) actually holds quite generally, as long as $T(F)$ is a moderately smooth functional (xx find refs, Shao (1991), Jullum and Hjort (2017) xx), though with no easy general formula for the bias component b/n. Deduce that $\sqrt{n}(\widehat{\theta} - \theta)$ has the limit distribution $\mathrm{N}(0, \kappa^2)$, with κ^2 the variance of $\mathrm{IF}(Y_i, \theta)$, and with the bias part b/n disappearing in this normal limit.

(d) Use the above to find the limit distribution of $\sqrt{n}(\widehat{\sigma} - \sigma)$. This gives a new and partly simpler proof of things proved in Ex. 2.43.

Ex. 5.57 *Influence functions for BHHJ and for weighted likelihood estimators.* For some smooth parametric model $f_\theta(y) = f(y, \theta)$, consider the BHHJ estimation method

of Ex. 5.9 and 5.18. The setup involves the data-generating density g, and for the given tuning parameter a, the defining parameter is $\theta_0 = \theta_{0,a}$, the minimiser of $\int f_\theta^{1+a}\,\mathrm{d}y - (1+1/a)\int g f_\theta^a\,\mathrm{d}y$, also the solution to $\int (f_\theta^{1+a} - g f_\theta^a) u_\theta\,\mathrm{d}y = 0$. Here $u_\theta(y) = u(y,\theta)$ is the score function, and we will also need the information function $i_\theta(y) = i(y,\theta)$ below. As in Ex. 5.18 we use f_0, u_0, i_0 for these functions at θ_0.

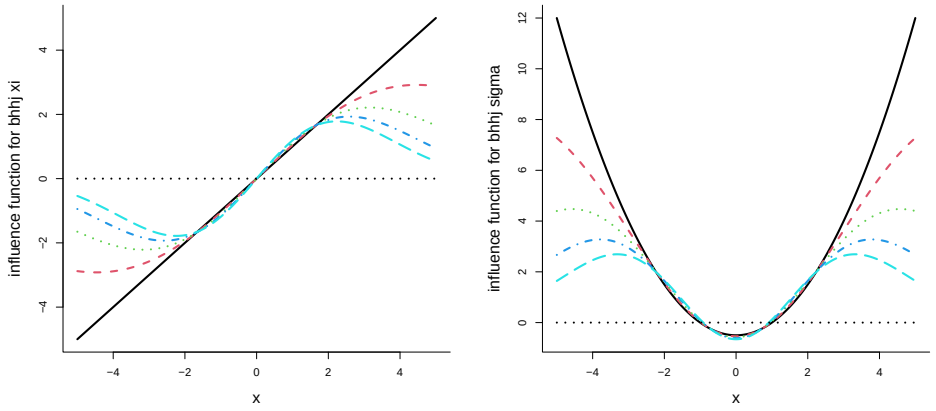


Figure 5.2: *Influence functions for the BHHJ estimator, for the normal* (ξ, σ^2) *model, for values* $a = 0, 0.05, 0.10, 0.15, 0.20$*; the full black curves are for the ML case* $a = 0$*. Left panel: for* $\widehat{\xi}_a$*; right panel: for* $\widehat{\sigma}_a$*.*

(a) Explain that to find the influence function $\mathrm{IF}_a(G, y)$, we need for a small ε to assess the solution $\theta_\varepsilon = \theta_0 + \delta$ to the equation

$$\int f_{\theta_0+\delta}^{1+a} u_{\theta_0+\delta}\,\mathrm{d}y = (1-\varepsilon)\int g f_{\theta_0+\delta}^a u_{\theta_0+\delta}\,\mathrm{d}y + \varepsilon f(y,\theta_0+\delta)^a u(y,\theta_0+\delta).$$

The programme is now for the given small ε to carry out first order Taylor expansion for δ small and then solve for δ. Some details are as follows: show first

$$f_{\theta_0+\delta}^a \doteq f_0^a(1 + a\delta^{\mathrm{t}} u_0), \quad f_{\theta_0+\delta}^{1+a} \doteq f_0^{1+a}\{1 + (1+a)\delta^{\mathrm{t}} u_0\}, \quad u_{\theta_0+\delta} \doteq u_0 + i_0\delta.$$

Again with $\xi_a = \int f_0^{1+a} u_0\,\mathrm{d}u = \int g f_0^a u_0\,\mathrm{d}y$, use this to write the left and right hand sides of the equation as

$$\begin{aligned} L &\doteq \xi_a + \Big\{(1+a)\int f_0^{1+a} u_0 u_0^{\mathrm{t}}\,\mathrm{d}y + \int f_0^{1+a} i_0\,\mathrm{d}y\Big\}\delta, \\ R &\doteq \xi_a + \Big\{\int g f_0^a (u_0 u_0^{\mathrm{t}} + i_0)\,\mathrm{d}y\Big\}\delta + \varepsilon\{f(y,\theta_0)^a u(y,\theta_0) - \xi_a\}. \end{aligned}$$

Conclude that $\mathrm{IF}_a(G, y) = M_a^{-1}\{f_0(y)^a u_0(y) - \xi_a\}$, with the matrix

$$M_a = (1+a)\int f_0^{1+a} u_0 u_0^{\mathrm{t}}\,\mathrm{d}y + \int f_0^{1+a} i_0\,\mathrm{d}y - \int g f_0^a (u_0 u_0^{\mathrm{t}} + i_0)\,\mathrm{d}y.$$

Check with the J_a matrix of Ex. 5.18 that in fact $M_a = J_a/(1+a)$, so that

$$\mathrm{IF}_a(G,y) = (1+a)J_a^{-1}\{f(y,\theta_0)^a u(y,\theta_0) - \xi_a\}.$$

Explain that as $a \to 0$, the limit is $J_0^{-1}u(y,\theta_0)$, the influence function for the ML estimation method. Show also that under conditions, M_a simplifies to $\int f_0^{1+a} u_0 u_0^{\mathrm{t}}\,\mathrm{d}y$.

(b) Consider the case of the normal $\mathrm{N}(\xi,\sigma^2)$. Under model conditions, find formulae for and graph the influence functions $\mathrm{IF}_{1,a}(y)$ and $\mathrm{IF}_{2,a}(y)$, for the ξ and the σ, at the standard position $(\xi,\sigma) = (0,1)$, for a equal to 0, 0.05, 0.10, 0.15, 0.20. Construct a version of Figure 5.2. Note that the BHHJ influence functions mimic those for the ML (i.e. for $a = 0$) for the main expected data range, but that they sensibly redescend down to zero for values far away, via density downweighting. Show in particular that the influence functions are bounded, for each $a > 0$. For illustration of how the BHHJ successfully deals with outliers (without needing to identify them), and for seeing that very little efficiency is lost for small a, see Story vii.5. Formulae needed here include these, which you should prove, in addition to $\xi_{1,a} = 0$:

$$\begin{aligned}
\xi_{2,a} &= -\sigma^{-a}(2\pi)^{-a/2}a/(1+a)^{3/2},\\
M_{11,a} &= (1/\sigma^2)\sigma^{-a}(2\pi)^{-a/2}/(1+a)^{3/2},\\
M_{22,a} &= (1/\sigma^2)\sigma^{-a}(2\pi)^{-a/2}\{3/(1+a)^2 - 2/(1+a) + 1\}.
\end{aligned}$$

(c) Carry out similar analysis for the Gamma (α,β) model.

(d) Consider then the MWL method of Ex. 5.11 and 5.20. For a given weight function $w(y)$, the parameter aimed at for the MWL estimator is the maximiser $\theta_0 = \theta_{0,w}$ of $\int w(g\log f_\theta - f_\theta)$. We need $J_w = \int w f_0 u_0 u_0^{\mathrm{t}}\,\mathrm{d}y + \int w(f_0 - g)i_0\,\mathrm{d}y$ and $\xi_w = \int w f_0 u_0\,\mathrm{d}y = \int w g u_0\,\mathrm{d}y$, computed at θ_0. Show that the influence function becomes

$$\mathrm{IF}(G,y) = J_w^{-1}\{w(y)u(y,\theta_0) - \xi_w\}.$$

Notes and pointers

(xx some pointers to more, including Empirical Likelihood, Hjort et al. (2009, 2018). xx)

(xx make sure we have something on board with BHHJ for regression models, placed after ML for regression. xx)

[xx CR bound: In its simplest form, the inequality goes back to Cramér (1946) and Rao (1945). xx]

(xx M-estimators, Z-estimators, Huber, minimum divergence, point to Basu et al. (2011), more on minimum divergence. two-stage estimators, WalkerHjort24. xx)

[xx least false: a term invented by Hjort, Hjort believes, see Hjort (1986b, 1992), and now used somewhat frequently in the literature. xx]

Read more about risk functions in DeGroot (1970).

For Ex. 5.11, point to Hjort and Jones (1996), Schweder and Hjort (2016), and WalkerHjort24.

[xx check and calibrate what's here and what's in Ch. 7, regarding CD things. xx]

(xx we see that ML matches CR bounds, under model conditions. point to Hajek convolution theorem, and other characterisations. perhaps Hodges superefficient thing too. xx)

(xx point to Hjort (2008), re ML and least false etc. xx)

I.6

Bayesian inference and computation

In frequentist parametric inference, there is a fixed underlying true parameter value, say θ_0, and methods aim at estimating this value, perhaps along with confidence regions or testing. Bayesian inference is radically different, conceptually and operationally. It starts with a prior distribution for the model parameter θ, and proceeds via Bayes theorems to produce the posterior distribution, of the full θ or of relevant focus parameters. Thus 'not knowing θ well' is expressed in terms of probability distributions. This chapter goes through these concepts and operations, including also computational schemes to simulate outcomes from the posterior distributions. One prominent class of such schemes amounts to setting up a Markov Chain Monte Carlo algorithm to the given problem, where the stationary distribution for the chain is precisely the required posterior. This makes Bayesian inference possible in a host of compicated setups, without needing to rely on mathematically feasible formulae.

Key words: Bayes solutions, Bernshteĭn–von Mises theorems, conjugate priors, Jeffreys prior, loss functions, MCMC, prior to posterior distributions

The Bayesian paradigm is to formulate uncertainty about model parameters through probability distributions. If the pre-data uncertainty is a prior density $\pi(\theta)$, this is updated to the post-data posterior density $\pi(\theta \,|\, \text{data})$, via the Bayes theorems.

Consider for illustration and clarification the classical coin flipping experiment, with θ the probability of 'head up'. With n independent flips we have $Y \sim \text{binom}(n, \theta)$. The frequentist postulates that there is an underlying true θ_0, uses perhaps the estimator $\widehat{\theta} = Y/n$, reaches the 95 percent interval $I_n = \widehat{\theta} \pm 1.96\,\{\widehat{\theta}(1-\widehat{\theta})\}/\sqrt{n}$, etc. The key property here us $\Pr_{\theta_0}(\theta_0 \in I_n) = 0.95$; so I_n is a random interval, covering the true θ_0 in 95 percent of actual cases. The Bayesian viewpoint is strikingly different, starting with a prior density $\pi(\theta)$ to reflect what might be considered understanding of θ before the first flip. Post flipping, the Bayesian has reached $\pi(\theta \,|\, y) \propto \pi(\theta) f(y, n, \theta)$, with the binomial likelihood. This may e.g. be used to construct a 95 percent posterior interval J_n for θ, with $\Pr(\theta \in J_n \,|\, y) = 0.95$. The Bayesian is then not interested in 'independent repeated experiments', but just in the data at hand. She is also allowed the statistical luxury of putting prior knowledge into the analysis; if it can be considered known that θ

must be close to 0.50, with values outside $[0.40, 0.60]$ less likely than 2 percent, that can effectively be utilised in Bayesian analysis, but not so easily in frequentist analysis.

In this chapter we go through the basics of such constructions and methods, conceptually and operationally. We also uncover conditions under which the frequentist and Bayesian might actually (approximately) agree, in their final inference statements. The two 95 percent intervals I_n and J_n in the previous paragraph will e.g. tend to be very similar, at least with increasing n.

An attractive feature of Bayesian analysis is that the answer, to a sufficiently well-posed inference question, is crystal clear, without having to study competing methods, carrying out performance and comparison analyses, etc. Essentially, if you give a Bayesian (i) a model, (ii) data, (iii) a list of possible actions, and (iv) a loss function, there is a Master Recipe for the very best action.

Modern Bayesian statistics has flourished since around 1980, partly through computer power and algorithms, making calculations possible that would have been too hard for previous generations. The operational goal is often to be able to generate samples from the posterior distribution, and we give methods for accomplishing this, including Markov Chain Monte Carlo (MCMC).

The Bayesian Master Recipe, with examples and applications

Ex. 6.1 *Poisson data with gamma priors.* This exercise illustrates the basic prior to posterior updating mechanism in a simple Poisson setting. Suppose $Y_1, Y_2, \ldots$ are i.i.d. Poisson with unknown mean θ.

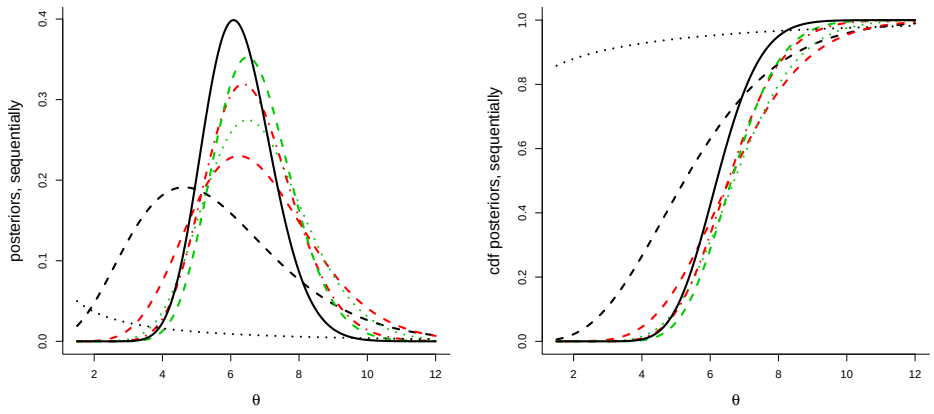


Figure 6.1: *Seven curves are displayed, corresponding to the* Gam$(0.1, 0.1)$ *intial prior for the Poisson parameter θ, along with the six first updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3. The distributions become tighter as more data come in. The smooth black curve is after six data points, the most tight distribution so far. Left panel: the densities; right panel: the c.d.f.s.*

(a) Recall definition and properties of the Gamma distribution from Ex. 1.9. In the present Bayesian context, let $\theta \sim \text{Gam}(a, b)$. The prior mean and variance are $a/b = \theta_0$ and $a/b^2 = \theta_0/b$. In particular, low and high values of b signify high and low variability, respectively. Explain how (a, b) may be set from values of prior mean and prior variance. To exemplify, if these are $(5.5, 7.7)$, find (a, b).

(b) With a single observation Y which is $\text{Pois}(\theta)$ given θ, show that $\theta \,|\, y \sim \text{Gam}(a + y, b + 1)$.

(c) Then suppose there are repeated observations $y_1, \ldots, y_n$, being i.i.d. $\sim \text{Pois}(\theta)$ for given θ. Use the above result repeatedly, e.g. interpreting $p(\theta \,|\, y_1)$ as the new prior before observing y_2, etc., to show that $\theta \,|\, y_1, \ldots, y_n \sim \text{Gam}(a + y_1 + \cdots + y_n, b + n)$. Also derive this result directly, i.e. without necessarily thinking about the data having emerged sequentially.

(d) Suppose the prior used is a rather flat $\text{Gam}(0.1, 0.1)$ and that the Poisson data are 6, 8, 7, 6, 7, 4, 11, 8, 6, 3. Reconstruct a version of Figure 6.1 in your computer, plotting the six first posterior densities $p(\theta \,|\, \text{data}_j)$ (left panel), where data_j is $y_1, \ldots, y_j$, along with the prior density; in the right panel we have the corresponding posterior cumulatives. Complement with another figure also including updated densities 7, 8, 9, 10, for the four last observations, and comment. Also compute the ten Bayes estimates $\widehat{\theta}_j = \text{E}(\theta \,|\, \text{data}_j)$ and the posterior standard deviations, for $j = 1, \ldots, 10$.

(e) The mathematics turned out to be rather uncomplicated in this situation, since the Gamma continuous density matches the Poisson discrete density so nicely. Suppose instead that the initial prior for θ is a uniform over $[0.5, 50]$. Try to compute posterior distributions, Bayes estimates and posterior standard deviations also in this case, and compare with what you found above. Note also that we in this exercise cared about and managed to reach answers for the posterior distributions $\theta \,|\, y_1, \ldots, y_n$, without needing to deal with the also implied marginal distribution for $(y_1, \ldots, y_m)$. These are not Poisson any longer, with the variability in θ is taken into account; see Ex. 6.20.

Ex. 6.2 *The Bayesian Master Recipe.* The general setup is as follows. We have data y_{obs}, seen as the outcome of a random Y over the sample space $\mathcal{Y}$, generated from a model $f_\theta(y) = f(y, \theta)$, with θ having a prior distribution $\pi(\theta)\,\text{d}\theta$ over its parameter space Θ. There is a decision a to be reached, with a belonging to an appropriate action space $\mathcal{A}$, along with a loss function $L(\theta, a)$, a measure of the consequences of decision a if the truth is θ. Whereas the frequentist can attempt different methods for deciding $\widehat{a} = \widehat{a}(Y)$, then compare risk functions, etc., there is a unique optimal strategy for the Bayesian.

(a) Show that the posterior density of θ, that is, the distribution of the parameter given the data, takes the form

$$\pi(\theta \,|\, y) = f_\theta(y)\pi(\theta)/m(y),$$

where $m(y)$ is the required integration constant $\int_\Theta f_\theta(y)\pi(\theta)\,\text{d}\theta$. This is *Bayes' theorem*, and we typically write $\pi(\theta \,|\, y) \propto \pi(\theta) f_\theta(y)$, which reads 'posterior is proportional to prior times likelihood'. Show also that the *marginal distribution* of the data y is $m(y)$.

(b) A decision $a = \widehat{a}(y)$ needs to be reached, as a function of the data. The *Bayes risk* for such a decision is the associated expected loss, $\mathrm{BR}(a,\pi) = \mathrm{E}\, L(\theta, \widehat{a}(Y))$, involving randomness on two levels; θ has a prior, and $\widehat{a}(Y) \,|\, \theta$ is random. Show that it may be expressed in two informative ways:

$$\mathrm{BR}(a,\pi) = \begin{cases} \mathrm{E}\,\{\mathrm{E}\, L(\theta,\widehat{a}(Y)) \,|\, \theta\} = \int R(\widehat{a},\theta)\pi(\theta)\,\mathrm{d}\theta, \\ \mathrm{E}\,\{\mathrm{E}\, L(\theta,\widehat{a}(Y)) \,|\, Y\} = \int_{\mathcal{Y}} \{\int_\Theta L(\widehat{a}(y),\theta)\pi(\theta \,|\, y)\,\mathrm{d}\theta\}\, m(y)\,\mathrm{d}y. \end{cases}$$

The first expression involves the frequentist risk function $R(\widehat{a},\theta) = \mathrm{E}_\theta\, L(\widehat{a}(Y),\theta)$, then averaged with respect to the prior. The 'inner expectation' of the second expression is $\mathrm{E}_\pi\,\{L(\theta,\widehat{a}(y)) \,|\, Y = y\}$, that is, the expected loss given data.

(c) Show then that the optimal Bayes strategy, the one minimising the Bayes risk, is achieved by using

$$\widehat{a} = \operatorname{argmin} g = \text{the value minimising } g,$$

where $g = g(a) = \mathrm{E}_\theta\,\{L(\theta,a) \,|\, y_{\mathrm{obs}}\}$ is the expected posterior loss. The function g is evaluated and minimised over all a, for the given data $y = y_{\mathrm{obs}}$. This is the Bayes recipe. Note that using the recipe in practice only concerns the observed data y_{obs}, and that one does not need to evaluate its risk function.

(d) Above results have been presented and reached in terms of prior and posterior densities, $\pi(\theta)$ and $\pi(\theta \,|\, y$, partly for notational convenience. Show that the arguments go through also for more general priors; these may in particular be mixtures over continuous and discrete measures.

Ex. 6.3 *Some loss functions and their associated Bayes rules.* The Master Recipe of Ex. 6.2 is completely general, and can be applied in new and complicated situations, as long as we have data, a model, a prior for the unknowns, and a loss function. In the Bayesian setup finding or evaluation the posterior distribution of the parameters is always important, carrying separate weight, but if clear decisions are needed one needs also the loss function, say $L(\theta,a)$. Here we go through a short list of commonly used loss functions.

(a) For estimating a one-dimensional θ, with squared error loss $L(\theta,a) = (a-\theta)^2$, show that the Bayes estimator is $\widehat{\theta}_B = \mathrm{E}\,(\theta \,|\, y)$, the posterior mean.

(b) If the loss function is $L(\theta,a) = w(\theta)(a-\theta)^2$, show that the Bayes estimator is

$$\widehat{\theta}_B = \frac{\mathrm{E}\,\{w(\theta)\theta \,|\, \mathrm{data}\}}{\mathrm{E}\,\{w(\theta) \,|\, \mathrm{data}\}}.$$

In particular, when estimating a positive θ using loss $(a-\theta)^2/\theta$, show that the Bayes estimator is $1/\mathrm{E}\,\{(1/\theta) \,|\, \mathrm{data}\}$.

(c) Consider the natural absolute loss function, $L(\theta,a) = |a-\theta|$. Show that the Bayes solution becomes the posterior median, i.e. $\widehat{\theta}_B = G^{-1}(\frac{1}{2} \,|\, \mathrm{data})$, where $G(\theta \,|\, \mathrm{data})$ is the posterior cumulative distribution function.

(d) Suppose one needs the joint estimation of several parameters, say all of $\theta = (\theta_1, \ldots, \theta_p)$, via the loss function $L(\theta, a) = (a-\theta)^{\rm t} M(a-\theta)$, for an appropriate full-rank symmetric matrix M. Show that the Bayes solution again is the posterior mean, but now for the full vector, i.e. $\mathrm{E}\,(\theta \,|\, \text{data})$. In particular, the Bayes solution does not depend on the M matrix, though the actual posterior expected loss, and the Bayes risk, do.

(e) In the previous subquestions the framework has been that of estimating a one-dimensional θ. Check that you understand how these results and insights, for the Bayes solutions, change when the situation is changed to that of estimating a *focus parameter*, say $\phi = g(\theta_1, \ldots, \theta_p)$, a function of the full model parameter.

Ex. 6.4 *How many streetcars in San Francisco?* The streetcars in this city are numbered $1, \ldots, N$. You observe $Y = 203$ and wonder what N is.

(a) Supposing Y has the uniform distribution on $1, \ldots, N$, set up the likelihood, and identify the ML estimate. With the prior $\pi(N)$ proportional to $1/N$, say for $N = 1, 2, \ldots, N_{\max}$ for a suitably high $N_{\max}$, find the posterior distribution, along with its mean and median.

(b) Suppose you after having seen no. 203 also see nos. 157, 222. Update your posterior, and again find the mean and median, qua updated estimates.

(c) The Bayesian setup allows any choice for the prior, so think for a minute and construct your own $\pi_{\rm you}(N)$. Do the updating, for the data 203, 157, 222, and compare with the posterior found above. What is your best estimate of N, if your loss function is of 0-1 type, with $L(N, \widehat{N})$ being 0 if $\widehat{N} = N$ and 1 if $\widehat{N} \neq N$?

Ex. 6.5 *The linex loss function.* When estimating a one-dimensional θ with a $\widetilde{\theta}$, the most traditional loss function is that of squared error, $(\widetilde{\theta} - \theta)^2$, which in particular is symmetric, treating over- and underestimation as equally important. A more flexible loss function is the so-called linex loss, with

$$L_c(\theta, \widetilde{\theta}) = \exp\{c(\widetilde{\theta} - \theta)\} - 1 - c(\widetilde{\theta} - \theta).$$

The c is fine-tuning loss parameter, for the statistician to set, balancing over- against underestimation. Note that both positive and negative values of c are allowed here.

(a) Show that the L_c is always nonnegative. Show that $c > 0$ means penalising overestimation more than underestimation, and vice versa for $c < 0$. For small $|c|$, show that $L_c(\theta, \widetilde{\theta}) \doteq \frac{1}{2} c^2 (\widetilde{\theta} - \theta)^2$, getting back to squared error loss. The constant in front is immaterial for evaluating and comparing loss and risk, and one may use $L_c^*(\theta, \widetilde{\theta}) = L_c(\theta, \widetilde{\theta}) / (\frac{1}{2} c^2)$ to have a smoother transition to the $c = 0$ case of squared error loss.

(b) Show that the expected loss given data can be expressed as

$$\begin{aligned} \mathrm{E}\,\{L_c(\theta, t) \,|\, \text{data}\} &= \mathrm{E}\,[\exp\{c(t-\theta)\} - 1 - c(t-\theta) \,|\, \text{data}] \\ &= \exp(ct) M(-c) - 1 - c(t - \widehat{\xi}), \end{aligned}$$

where $\widehat{\xi}$ is the posterior mean and $M(-c) = \mathrm{E}\,\{\exp(-c\theta) \,|\, \text{data}\}$, the moment-generating function of θ given data, computed at $-c$.

(c) Show that this is minimised for the t_0 where $\exp(ct_0)M(-c) = 1$, or $ct_0 + \log M(-c) = 0$, so that the Bayes estimator becomes $\widehat{\theta}_B = -(1/c)\log M(-c)$. This may be computed numerically, perhaps by simulation, in cases where no clear formula exists for $M(-c)$. Show also that the expected posterior loss, using the Bayes solution, is

$$\min \mathrm{E}\,\{L_c(\theta, t) \,|\, \mathrm{data}\} = -c(t_0 - \widehat{\xi}) = \log M(-c) + c\widehat{\xi}.$$

(d) Using approximations for m.g.f.s close to zero to show that $M(-c) \doteq 1 - c\widehat{\xi} + \frac{1}{2}c^2(\widehat{\xi}^2 + \widehat{\sigma})$, with $\widehat{\xi}$ and $\widehat{\sigma}^2$ the posterior mean and variance. Deduce that $\widehat{\theta}_B \doteq \widehat{\xi} - \frac{1}{2}c\widehat{\sigma}^2$.

(e) In situations where the posterior is based on a sample of size n, the posterior mean $\widehat{\xi}_n$ stays stable wheras the posterior variance $\widehat{\sigma}^2$ goes down with speed $1/n$, i.e. as $\widehat{\sigma}_0^2/n$, for the relevant $\widehat{\sigma}_0^2$. In such cases, $\widehat{\xi}_n - \frac{1}{2}c\widehat{\sigma}_0^2/n$ becomes the approximation to the Bayes linex estimator $\widehat{\theta}_B$. Find in fact the exact Bayes linex estimator, for the case of $Y_1, \ldots, Y_n$ being i.i.d. $\mathrm{N}(\theta, 1)$, with a $\mathrm{N}(0, \tau^2)$ prior for θ; use the updating result from Ex. 6.13(b).

(f) (xx rounding off for now; point to Ex. 6.24, 6.25. xx)

Ex. 6.6 *A Bayesian take on hypothesis testing.* Assume the model parameter θ is either in Ω_0, which we may call the null hypothesis, or not, i.e. in its complement Ω_0^c. Suppose also that the statistician needs to make a decision, either to reject the null, or to accept it. This is the basic framework of hypothesis testing, see Ch. 4, but we now consider the problem from a Bayesian viewpoint.

(a) The decision space is $\{\mathrm{accept}, \mathrm{reject}\}$. For the loss function, take $L(\theta, \mathrm{accept})$ equal to 0 or L_0, if θ is inside or outside Ω_0, and $L(\theta, \mathrm{reject})$ equal to 0 or L_1, if θ is outside or inside Ω_0. Show that

$$\mathrm{E}\,\{L(\theta, \mathrm{accept}) \,|\, \mathrm{data}\} = L_0\, p(\mathrm{data}), \quad \mathrm{E}\,\{L(\theta, \mathrm{reject}) \,|\, \mathrm{data}\} = L_1\,\{1 - p(\mathrm{data})\},$$

where $p(\mathrm{data}) = \Pr(\theta \in \Omega_0^c \,|\, \mathrm{data})$, the probability that the null is wrong, as measured by the Bayesian posterior distribution.

(b) Deduce that one should reject the null when the probability $p(\mathrm{data})$ for its falseness is sufficiently overwhelming, namely when $p(\mathrm{data}) \geq L_1/(L_0 + L_1)$. – If this threshold is 0.95, for example, show that this corresponds to $L_1/L_0 = 19$. Briefly discuss ways of assigning losses L_0 and L_1.

(c) (xx complete this: decision space $\{\mathrm{accept}, \mathrm{reject}, \mathrm{doubt}\}$, with a certain fixed cost L_d for the doubt option, associated with further efforts for getting more data. expected losses given data are $L_0 p$, $L_1(1-p)$, L_d. which is smallest? xx)

Ex. 6.7 *Which subset does the model parameter belong to?* Consider a setup with data from a model with model parameter θ inside its region Ω. Suppose you need to take one of five different possible decisions, $D_1, \ldots, D_5$, and that these are related to where the underlying parameter θ is positioned; if $\theta \in \Omega_j$ the best decision would be D_j, for $j = 1, \ldots, 5$. Here the Ω_j are disjoint and their union is the full parameter region.

(a) Suppose the loss function $L(\theta, D_j)$ is 0 if $\theta \in \Omega_j$ and 100 if $\theta \notin \Omega_j$. Show that $\mathrm{E}\{L(\theta, D_j) \,|\, \text{data}\} = 100\{1 - p_j(\text{data})\}$, where $p_j(\text{data}) = \Pr(\theta \in \Omega_j \,|\, \text{data})$. Hence show that the optimal Bayes strategy is to take the decision associated with the highest posterior probability $p_j(\text{data})$.

(b) Assume there in addition is a 'doubt option', associated with a doubt cost $L_d = 10$; this could e.g. mean planning for getting further data. With decision space $\{D_1, \ldots, D_5, \text{doubt}\}$, what is now the Bayesian strategy?

(c) Generalise the previous setup, and results, to the case where the costs associated with reaching the wrong decision are not equally balanced, say $L(\theta, D_j) = c_{i,j}$, if $\theta \in \Omega_i$, for $i = 1, \ldots, 5$, with $c_{i,i} = 0$ but the other $c_{i,j}$ positive.

Binomial and beta, multinomial and Dirichlet

Ex. 6.8 *The binomial-beta setup.* Let Y given θ be a binomial (n, θ), and for θ take a Beta(a, b) prior, see Ex. 1.21. There we worked with the marginal distribution of Y, and looked at certain properties, but here our aims are Bayesian.

(a) Show that $\theta \,|\, y \sim \text{Beta}(a + y, b + n - y)$. – This is the main and always crucial updating step, getting from the prior to the posterior. In the present case the step is an easy one, since there is only one unknown parameter, and since the product of the prior and the likelihood takes an easy form. Give a description of the posterior also for the not quite so standard case where the prior for θ is uniform on $[0.30, 0.70]$.

(b) Going back to the Beta(a, b) prior again, show that the Bayes estimator, under squared error loss, is

$$\widehat{\theta}_B = \frac{a + y}{a + b + n} = (1 - w_n)\theta_0 + w_n y/n,$$

where $\theta_0 = a/(a + b)$ is the prior mean and $w_n = n/(a + b + n)$. For the case of a uniform prior, show that this leads to $(y + 1)/(n + 2)$. Compute the risk functions $r(\theta) = \mathrm{E}_\theta\,(\widehat{\theta} - \theta)^2$, for the classic frequentist Y/n and for the this $(Y + 1)/(n + 2)$, and find the interval where the latter is better than the former.

(c) Show that the posterior variance becomes

$$\text{Var}\,(\theta \,|\, y) = \frac{\widehat{\theta}_B(1 - \widehat{\theta}_B)}{n + a + b + 1}.$$

(d) If Y is from the binomial $(n, \theta_{\text{true}})$ model, show that Y/n and the Bayes estimator $\widehat{\theta}_B$ are large-sample equivalent, with $\sqrt{n}(Y/n - \widehat{\theta}_B) \to_{\text{pr}} 0$. Deduce that they have the same limit distribution.

Ex. 6.9 *The multinomial-Dirichlet setup.* Here we extend the setup and result of binomial-Beta, to the case of three or more categories. We start with $Y = (Y_1, \ldots, Y_k)$ which for given $p = (p_1, \ldots, p_k)$ is a multinomial $(n, p_1, \ldots, p_k)$. For p we take the Dir$(a_1, \ldots, a_k)$ prior. For details regarding the multinomial and the Dirichlet, see Ex. 1.5 and 1.19.

(a) Show the important and useful result that $(p_1,\ldots,p_k)\,|\,(y_1,\ldots,y_k) \sim \mathrm{Dir}(a_1+y_1,\ldots,a_k+y_k)$.

(b) Show that the Bayes estimator under squared error loss becomes

$$\widehat{p}_{i,B} = \mathrm{E}\,(p_i\,|\,\mathrm{data}) = \frac{a_i+y_i}{a+n} = (1-w_n)p_{0,i} + w_n(y_i/n)$$

for $i=1,\ldots,k$, with prior means $p_{0,i}=a_i/a$, with $a=a_1+\cdots+a_k$, and weight $w_n=n/(a+n)$. Find also the posterior variance and posterior correlation between p_i and p_j.

(c) (xx just a bit more. xx)

Ex. 6.10 *Gott würfelt nicht.* For the multinomial-Dirichlet setup of Ex. 6.9, we reached the posterior characterisation $p\,|\,\mathrm{data} \sim \mathrm{Dir}(a_1+y_1,\ldots,a_k+y_k)$. The importance of this lies in the easy usefulness of simulations, where the posterior distribution of any functions of $(p_1,\ldots,p_k)$ may be read off.

(a) Explain how you may simulate e.g. 10^5 vectors $p=(p_1,\ldots,p_k)$ from the posterior distribution, using the characterisation from Ex. 1.19. Concretely, show that one may use $p_1=G_1/G,\ldots,p_k=G_k/G$, with independent $G_1\sim\mathrm{Gam}(a_1+y_1),\ldots,G_k\sim\mathrm{Gam}(a_k+y_k)$, and sum $G=G_1+\cdots+G_k$.

(b) Suppose you throw a certain and perhaps not entirely standard die 30 times and have counts $(2,5,3,7,5,8)$ of outcomes 1, 2, 3, 4, 5, 6. Use either of the priors (i) 'flat', $\mathrm{Dir}(1,1,1,1,1,1)$; (ii) 'symmetric and more confident', $\mathrm{Dir}(3,3,3,3,3,3)$; (iii) 'unwilling to guess', $\mathrm{Dir}(0.1,0.1,0.1,0.1,0.1,0.1)$, for the probabilities $(p_1,\ldots,p_6)$, to assess the posterior distribution of each of the following quantities:

$$\alpha=p_6/p_1,\ \beta=(1/6)\sum_{j=1}^6(p_j-1/6)^2,\ \gamma=(1/6)\sum_{j=1}^6|p_j-1/6|,\ \delta=\frac{(p_4p_5p_6)^{1/3}}{(p_1p_2p_3)^{1/3}}.$$

For each of $\alpha,\beta,\gamma,\delta$, and for each of the priors, give the 0.05, 0.50, 0.95 quantile points, from 10^5 simulations from the posterior distributions. You should also plot the posterior densities, for each of the four quantities noting the extent to which the prior influences the results.

(c) For the case of $\alpha=p_6/p_1$, exact numerical simulation is possible, without simulation. Do this, and compare with the answers reached via simulation.

(d) The above priors are slightly artificial in this context, since they do not allow the explicit possibility that the die in question is plain boring utterly simply a correct one, i.e. that $p=p_0=(1/6,\ldots,1/6)$. The priors used hence do not give us the possibility to admit that perhaps $\rho=1$, $\alpha=0$, $\beta=0$, $\gamma=1$, after all. This motivates using a mixture prior which allows a positive chance for $p=p_0$. Redo therefore the Bayesian analysis above, with the same $(2,5,3,7,5,8)$ data, for the prior $\frac12\,\delta(p_0)+\frac12\,\mathrm{Dir}(1,1,1,1,1,1)$. Here $\delta(p_0)$ is the 'degenerate prior' that puts unit point mass at position p_0. Compute in particular the posterior probability that $p=p_0$, and display the posterior distributions of ρ,α,β,γ.

Sampling from the posterior via Markov Chain Monte Carlo

Ex. 6.11 *MCMC, I: simulating from a given distribution.* (xx the MCMC basics. Metropolis algorithm. simulating from a couple of distributions. xx) basics: suppose we need to simulat realisations from a given and possibly complicated density $\pi(x)$ on some domain. The Metropolis algorithm works as follows. The task is to form a long chain $x_1, x_2, \ldots$ in your computer, where its limiting distribution is precisely the given π. After having generated $x_{\rm old}$, decide on $x_{\rm new}$ by (i) deciding on a proposal, $x_{\rm prop}$, drawn from a suitable distribution symmetric in $(x_{\rm old}, x_{\rm prop})$, and then (ii) accepting this proposal with probability $p_{\rm accept} = \min(1, \pi(x_{\rm prop})/\pi(x_{\rm old}))$. In algorithmic terms,

$$x_{\rm new} = (1 - {\rm ok})\, x_{\rm old} + {\rm ok}\, x_{\rm new}, \quad \text{where ok} = I(\text{accept}).$$

Markov chain theory (xx point to Ch. 12 xx) secures that this scheme works, in the sense that the chain converges in distribution to that of π; $\Pr(x_n \in A) \to \int_A \pi(x)\,{\rm d}x$ for all continuity sets A.

(a) Set up such a scheme to generate outcomes from the standard normal (ignoring the existence of simpler direct algorithms). Start at any x_0, then draw proposals $x_{\rm prop} \sim {\rm unif}[x_{\rm old} - a, x_{\rm old} + a]$, with acceptance probabilities set up via the general scheme above. Run the chain for a suitably long time and check with a fine histogram that the distribution matches the normal. Keep track of the acceptance probabilities and the overall acceptance rate. The theory works and says that the chain will converge to the standard normal, for any positive fine-turning parameter a; explain however in which ways too small or too large values of a will be ineffective.

(b) Consider a somewhat harder challenge, simulating realisations from the density $\pi = 0.05\,{\rm N}(-2,1) + 0.90\,{\rm N}(0,1) + 0.05\,{\rm N}(2,1)$ via MCMC. Set up a chain that converges to this f in distribution; check that the output produces a fine histogram matches this π.

(c) The algorithm works also in higher dimension. Set up a chain that produces outcomes (x, y, z) from the model on the unit cube with density $\pi(x, y, z) = 1 + \theta(x - \frac{1}{2})(y - \frac{1}{2})(z - \frac{1}{2})$, with $\theta = 3.45$. From the output, read off the correlations, and the probability that $Z \geq (XY)^{1/2}$.

Ex. 6.12 *MCMC, II: simulating from the posterior.* (xx do this. using the strategy above for the prototypical Bayesian task, sampling from $\pi(\theta \,|\, \text{data})$.

(a) In the Poisson-gamma setup of Ex. 6.1 we could find the posterior distribution directly. Suppose however that the prior is outside the gammas, say uniform on $[4.0, 8.0]$. Set up an MCMC to sample from the posterior, given data 6, 8, 7, 6, 7, 4, 11. Compute also the mean, median, standard deviation.

(b) (xx one more, a bit more complex, and then minimising posterior expected loss in the end. xx)

Bayesian analysis for normal models

Ex. 6.13 *The normal prior and posterior with normal data.* Here we go through the basic steps and results for situations with normal data and normal priors for unknown

mean parameters. More elaborate constructions and technical issues are needed when there in addition are unknown parameters in the variance and covariance structure, to be pursued in Ex. 6.14, 6.21.

(a) There are things to think through and to learn from, by working through this very simple setup first. (i) For a single observation Y assume it comes from the $\mathrm{N}(\xi, \sigma^2)$, and take σ as known; (ii) for the unknown mean ξ assume it comes from the prior $\mathrm{N}(\xi_0, \tau_0^2)$, with specified prior parameters ξ_0, τ_0. Show that this leads to a binormal joint distribution for parameter and observation,

$$\begin{pmatrix} \xi \\ Y \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} \xi_0 \\ \xi_0 \end{pmatrix}, \begin{pmatrix} \tau_0^2, & \tau_0^2 \\ \tau_0^2, & \tau_0^2 + \sigma^2 \end{pmatrix}\Big).$$

(b) Use general conditioning results from Ex. 1.41 to infer that

$$\xi \,|\, y \sim \mathrm{N}(\xi_0 + w(y - \xi_0), w\sigma^2), \quad \text{with } w = \tau^2/(\tau^2 + \sigma^2).$$

So w and $1 - w$ are the weights given to the data-based estimate y and the prior guess ξ_0, respectively. Also, w is the reduction factor with which the variance of the prior-free estimator Y, from σ^2 to $w\sigma^2$.

(c) An easy but important extension is to the case of a full sample $Y_1, \ldots, Y_n$ from the $\mathrm{N}(\xi, \sigma^2)$ distribution, independent given the ξ, again with σ taken known and the normal prior $\mathrm{N}(\xi_0, \tau_0^2)$ for the unknown mean. Show that $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$ is sufficient (xx xref here xx), and that

$$\begin{pmatrix} \xi \\ \bar{Y} \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} \xi_0 \\ \xi_0 \end{pmatrix}, \begin{pmatrix} \tau_0^2, & \tau_0^2 \\ \tau_0^2, & \tau_0^2 + \sigma^2/n \end{pmatrix}\Big).$$

Show from this that

$$\xi \,|\, \text{data} \sim \mathrm{N}(\xi_0 + w_n(\bar{y} - \xi_0), w_n\sigma^2/n), \quad \text{with } w_n = \frac{\tau^2}{\tau^2 + \sigma^2/n} = \frac{n\tau^2}{n\tau^2 + \sigma^2}.$$

Again, w_n is both the weight given to the data-based estimate and the factor with which the neutral estimator's variance is reduced, from σ^2/n to $w_n\sigma^2/n$. Note that $w_n \to 1$; discuss how this may be seen as 'the data wash out the prior'.

(d) Discuss the case of a 'flat prior', where τ is taken large, for $\xi \sim \mathrm{N}(\xi_0, \tau^2)$.

(e) In addition to having a coherent updating machine, changing the prior to the posterior, for each new data point, the Bayesian structure implies positive dependence among the observations. From $\mathrm{E}\,(Y_i \,|\, \xi) = \xi$, $\mathrm{Var}\,(Y_i \,|\, \xi) = \sigma^2$, $\mathrm{E}\,(Y_iY_j \,|\, \xi) = \xi^2$, show that

$$\mathrm{E}\,Y_i = \xi_0, \quad \mathrm{Var}\,Y_i = \sigma^2 + \tau_0^2, \quad \mathrm{cov}(Y_i, Y_j) = \tau_0^2, \quad \mathrm{corr}(Y_i, Y_j) = \frac{\tau_0^2}{\sigma^2 + \tau_0^2},$$

Show also that $\mathrm{Var}\,\bar{Y}_n = \sigma^2/n + \tau_0^2$, and discuss what this means for large n.

(f) Prove first the convenient formula

$$v(\xi-\xi_0)^2+n(\xi-\bar y)^2=(v+n)(\xi-\xi^*)^2+d_n(\bar y-\xi_0)^2,$$

where

$$\xi^*=\frac{v\xi_0+n\bar y}{v+n}\quad\text{and}\quad d_n=\frac{vn}{v+n}=(v^{-1}+n^{-1})^{-1},$$

which also may be written and interpreted via $1/d_n=1/v+1/n$.

(g) Show that with any prior $p(\xi)$ for ξ, the marginal density of $(Y_1,\ldots,Y_n)$ can be written

$$\bar f(y_1,\ldots,y_n)=\int(2\pi)^{-n/2}\sigma^{-n}\exp\Big\{-\tfrac12\frac{1}{\sigma^2}\sum_{i=1}^n(y_i-\xi)^2\Big\}p(\xi)\,\mathrm{d}\xi.$$

check all of this with care

For the case of $\mathrm{N}(\xi_0,\tau_0^2)$ worked with above, let first $Q_0=\sum_{i=1}^n(y_i-\bar y)^2$, and verify that $\sum_{i=1}^n(y_i-\xi)^2=Q_0+n(\xi-\bar y)^2$. Writing for mathematical convenience $1/\tau^2=v/\sigma^2$, and $1/d_n=1/n+1/v$, show that

$$\bar f(y_1,\ldots,y_n)=\frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp(-\tfrac12Q_0/\sigma^2)\Big(\frac{v}{v+n}\Big)^{1/2}\exp\{-\tfrac12d_n(\bar y-\xi_0)^2/\sigma^2\}.$$

(h) With the marginal density seen as a function of the two fine-tuning parameters (ξ_0,τ_0^2), find the maximum marginal likelihood estimators $\widehat{\xi_0}$ and $\widehat\sigma$.

(i) We record two matrix identities here, as they will come in handy both here and on later occasions. For a square and invertible A, show that

$$(A+xx^{\mathrm t})^{-1}=A^{-1}-cA^{-1}xx^{\mathrm t}A^{-1},\quad\text{with } c=1/(1+x^{\mathrm t}A^{-1}x);$$

also, that $|A+xx^{\mathrm t}|=|A|(1+x^{\mathrm t}A^{-1}x)$.

(j) Argue directly that $Y=(Y_1,\ldots,Y_n)^{\mathrm t}$ must be multinormal, with $Y\sim\mathrm{N}_n(\xi_0\mathbf{1},\sigma^2I_n+\tau_0^2\mathbf{1}\mathbf{1}^{\mathrm t})$, with $\mathbf{1}$ the vector $(1,\ldots,1)^{\mathrm t}$; the variance matrix has $\sigma^2+\tau_0^2$ on the diagonal and τ_0^2 outside. Show that this agrees with the marginal density formula reached above.

Ex. 6.14 *The gamma-normal prior and posterior.* Let data $y_1,\ldots,y_n$ for given parameters ξ and σ be i.i.d. $\mathrm{N}(\xi,\sigma^2)$. We have seen in Ex. 6.13 that when σ may be taken as a known quantity, then the canonical class of priors for ξ is the normal one. When both parameters are unknown, however, as in most practical encounters, a more elaborate analysis is called for.

(a) Show that the likelihood function may be written as being proportional to

$$L_n(\xi,\sigma)=\exp\Big[-n\log\sigma-\tfrac12\frac{1}{\sigma^2}\{Q_0+n(\xi-\bar y)^2\}\Big],$$

where $\bar y=(1/n)\sum_{i=1}^n y_i$ and $Q_0=\sum_{i=1}^n(y_i-\bar y)^2$.

(b) With *any* given prior $p(\xi, \sigma)$, explain how you may set up a Metropolis type MCMC to draw samples from the posterior distribution. Try this out in practice, using the prior that takes ξ and $\log \sigma$ independent and uniform on say $[-5, 5]$ and $[-10, 10]$, with data that you simulate for the occasion from a $\mathrm{N}(2.345, 1.234^2)$, with $n = 25$. Note that this approach does not need more mathematical algebra as such, apart from the likelihood function above.

(c) There is however a popular and convenient conjugate class of priors for which posterior distributions become particularly clear, with the appropriate algebraic efforts. These in particular involve placing a Gamma prior on the inverse variance $\lambda = 1/\sigma^2$. Say that (λ, ξ) has the gamma-normal distribution with parameters (a, b, ξ_0, v), and write this as $(\lambda, \xi) \sim \mathrm{GN}(a, b, \xi_0, v)$, provided $\lambda = 1/\sigma^2 \sim \mathrm{Gam}(a, b)$ and $\xi \,|\, \sigma \sim \mathrm{N}(\xi_0, \sigma^2/v)$. Show that the prior can be expressed as

$$p(\lambda, \xi) \propto \lambda^{a-1} \lambda^{1/2} \exp\big[-\lambda\{b + \tfrac{1}{2} v(\xi - \xi_0)^2\}\big].$$

What is the unconditional prior variance of ξ?

(d) Using the identity from Ex. 6.13(f), show that if the prior is $(\lambda, \xi) \sim \mathrm{GN}(a, b, \xi_0, v)$, then

$$(\lambda, \xi) \,|\, \text{data} \sim \mathrm{GN}(a + \tfrac{1}{2} n, b + \tfrac{1}{2} Q_0 + \tfrac{1}{2} d_n (\bar{y} - \xi_0)^2, \xi^*, v + n).$$

(e) The special case of a 'flat prior' for ξ, corresponding to letting $v \to 0$ above, is particularly easy to deal with. Show that then

$$(\lambda, \xi) \,|\, \text{data} \sim \mathrm{GN}(a + \tfrac{1}{2} n, b + \tfrac{1}{2} Q_0, \bar{y}, n).$$

Find the posterior mean of σ^2 under this prior.

(f) For an illustration, consider the cigarette consumption data $x_1, \ldots, x_n$, from 2.B, for $n = 44$ US states (actually, 43 states plus the District of Columbia), from the early 1960ies, taken here to form a sample from $\mathrm{N}(\xi, \sigma^2)$. The x_i is the consumption in state i, in hundreds per year, ranging then from 14.0 to 42.4 (which translates to from 3.84 per day to 11.61 per day, in the adult population). Fix the prior that takes (i) $\lambda = 1/\sigma^2 \sim \mathrm{Gam}(a, b)$, with (a, b) taken to correspond to 0.10 and 0.90 prior quantiles for σ being 1.0 and 10.0; (ii) given σ, ξ is a $\mathrm{N}(20.0, \sigma^2/v)$ with $v = 0.50$. Find the precise posterior distribution for (λ, ξ). Give posterior means and 90 percent credibility intervals for ξ, σ, and for the probability $p = \Pr(X \geq 40)$.

(g) For the same data and setup as in the previous point, carry out a second Bayesian analysis, with the simpler prior that corresponds to $v \to 0$. Comment briefly on the differences in results.

Ex. 6.15 *The gamma-multinormal prior for linear regression models.* The aim of the present exercise is to generalise the Gamma-Normal conjugate prior class above to the linear-normal regression model. The model is the very classical one of Ex. 3.31, where

$$y_i = x_{i,1}\beta_1 + \cdots + x_{i,k}\beta_k + \varepsilon_i = x_i^{\mathrm{t}}\beta + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with the ε_i taken i.i.d. $\mathrm{N}(0,\sigma^2)$. Write X for the $n\times k$ matrix of covariates (explanatory variables), with $x_i=(x_{i,1},\ldots,x_{i,k})$ as its ith row, and use y and ε to indicate the vectors of y_i and ε_i. Then $y=X\beta+\varepsilon\sim \mathrm{N}_n(X\beta,\sigma^2 I_n)$ is a concise way to write the full model.

(a) Show that the likelihood function may be written as being proportional to

$$L_n(\beta,\sigma)=\sigma^{-n}\exp\Big[-\tfrac12\frac{1}{\sigma^2}\{Q_0+n(\beta-\widehat\beta)^{\mathrm t}M_n(\beta-\widehat\beta)\}\Big],$$

in which

$$M_n=(1/n)X^{\mathrm t}X=n^{-1}\sum_{i=1}^n x_ix_i^{\mathrm t}\ \text{ and }\ \widehat\beta=(X^{\mathrm t}X)^{-1}X^{\mathrm t}y=M_n^{-1}n^{-1}\sum_{i=1}^n x_iy_i.$$

Also,

$$Q(\beta)=\|y-X\beta\|^2=Q_0+n(\beta-\widehat\beta)^{\mathrm t}M_n(\beta-\widehat\beta),$$

with $Q_0=\sum_{i=1}^n(y_i-x_i^{\mathrm t}\widehat\beta)^2$ the minimum value of Q over all β. Note that $\widehat\beta$ is the classical least squares estimator (and the ML estimator), which in the frequentist framework is unbiased with variance matrix equal to $\sigma^2(X^{\mathrm t}X)^{-1}=(\sigma^2/n)M_n^{-1}$. This is the basis of all classical methods related to the widely popular linear regression model.

(b) Let $p(\beta,\sigma)$ be *any* prior for the $(k+1)$-dimensional parameter of the model. Set up formulae for a Metropolis type MCMC algorithm for drawing samples from the posterior distribution of (β,σ).

(c) In spite of the possibility of solving problems via MCMC (or perhaps acceptance-rejection sampling), as with the previous exercise it is very much worthwhile setting up explicit formulae for the case of a certain canonical prior class. Write $(\lambda,\beta)\sim \mathrm{GN}_k(a,b,\beta_0,M_0)$ to indicate the gamma-normal prior where

$$\lambda=1/\sigma^2\sim\mathrm{Gam}(a,b)\quad\text{and}\quad\beta\,|\,\sigma\sim\mathrm{N}_k(\beta_0,\sigma^2M_0^{-1}).$$

Show that this prior may be expressed as

$$p(\lambda,\beta)\propto\lambda^{a-1}\lambda^{k/2}\exp\big[-\lambda\{b+\tfrac12(\beta-\beta_0)^{\mathrm t}M_0(\beta-\beta_0)\}\big].$$

(d) When multiplying the prior with the likelihood it is convenient to use the following linear algebra identity about quadratic forms, which you should prove first. For symmetric and invertible matrices A and B, and for any vectors a,b,x of the appropriate dimension,

$$(x-a)^{\mathrm t}A(x-a)+(x-b)^{\mathrm t}B(x-b)=(x-\xi)^{\mathrm t}(A+B)(x-\xi)+(b-a)^{\mathrm t}D(b-a),$$

where $\xi=(A+B)^{-1}(Aa+Bb)$ (a weighted average of a and b) and D is a matrix for which several equivalent formulae may be used:

$$\begin{aligned}D&=A(A+B)^{-1}B=B(A+B)^{-1}A\\&=A-A(A+B)^{-1}A=B-B(A+B)^{-1}B=(A^{-1}+B^{-1})^{-1}.\end{aligned}$$

(e) Prove that if (λ, β) has the $\mathrm{GN}_k(a, b, \beta_0, M_0)$ prior, then

$$(\lambda, \beta) \,|\, \text{data} \sim \mathrm{GN}_k(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_0 + \tfrac{1}{2}(\widehat{\beta} - \beta_0)^{\mathrm{t}} D_n (\widehat{\beta} - \beta_0), \beta^*, M_0 + nM_n),$$

where

$$\beta^* = (M_0 + nM_n)^{-1}(M_0\beta_0 + nM_n\widehat{\beta}) \quad \text{and} \quad D_n = M_0(M_0 + nM_n)^{-1}nM_n.$$

This characterisation makes it easy to simulate a large number of (β, σ) from the posterior distribution and hence to carry out Bayesian inference for any parameter of quantity of interest.

(f) Note the algebraic simplifications that result when the M_0 in the prior is chosen as being proportional to the covariate sample variance matrix, i.e. $M_0 = c_0 M_n$. Show that then

$$\beta^* = \frac{c_0\beta_0 + n\widehat{\beta}}{c_0 + n} \quad \text{and} \quad D_n = \frac{c_0 n}{c_0 + n}.$$

In this connection c_0 has a natural interpretation as 'prior sample size'.

(g) A special case of the above, leading to simpler results, is that where β has a flat, non-informative prior, corresponding to very large prior variances, i.e. to $M_0 \to 0$. Show that with such a prior,

$$(\lambda, \beta) \,|\, \text{data} \sim \mathrm{GN}_k(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_0, \widehat{\beta}, nM_n).$$

The prior is improper (infinite integral), but the posterior is proper as long as $\widehat{\beta}$ exists, which requires $X^{\mathrm{t}}X$ to have full rank, which again means at least k linearly independent covariate vectors, and, in particular, $n \geq k$.

(h) (xx repair this. do bladder cancer rates, since linear is ok, whereas lung cancer is more quadratic, done in Story i.15. xx) Go again to the dataset 2.B, for illustration and for flexing your operational muscles. For y use the bladder cancer column of deaths per 100,000 inhabitants and for x use the number of cigarettes sold per capita. Your task is to carry out Bayesian analysis within the linear regression model $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$ for $i = 1, \ldots, 44$, with ε_i taken i.i.d. $\mathrm{N}(0, \sigma^2)$. Specifically, we wish point estimates along with 95 percent credibility intervals for (i) each of the three parameters β_0, β_1, σ; (ii) the probability that $y \geq 25.0$, for a country with cigarette consumption $x = 35.0$; (iii) the bladder cancer death rates y_{45} and y_{46}, per 100,000 inhabitants, for states with cigarette consumption rates $x_{45} = 10.0$ (low) and $x_{46} = 50.0$ (high). You are to carry out such inference with two priors (xx nils repairs this xx): (1) First, the informative one which takes $1/\sigma^2$ a gamma with 0.10 and 0.90 quantiles for σ equal to 1.0 and 5.0, and β_0 and β_1 as independent normals $(5.0, 0.5\,\sigma^2)$ and $(0.0, (2.0\,\sigma)^2)$, given σ. (2) Then, the simpler and partly non-informative one that takes a flat prior for (β_0, β_1) and the less informative one for σ that uses 0.10 and 0.90 prior quantiles 0.5 and 10.0. Finally, compare your results from those arrived at using classical frequentist methods.

Ex. 6.16 *Mixture priors.* Suppose data Y come from a model $f(y, \theta)$, where different priors $\pi_1(\theta), \ldots, \pi_k(\theta)$ can be used, each leading to posterior distributions $\pi_1(\theta \,|\, y), \ldots, \pi_k(\theta \,|\, y)$.

(a) For each of these possible priors (and hence possible posteriors), show that there is a representation $f_j(y,\theta) = \pi_j(\theta) f(y,\theta) = \pi_j(\theta \,|\, y) \bar{f}_j(y)$, where $\bar{f}_j(y) = \int f(y,\theta)\pi_j(\theta)\,\mathrm{d}\theta$ is the marginal density of Y, associated with the $\pi_j(\theta)$ prior.

(b) Suppose now that a full mixture prior is assigned to θ, of the type $\pi(\theta) = p_1\pi_1(\theta) + \cdots + p_k\pi_k(\theta)$, with probabilities $p_1, \ldots, p_k$ summing to 1. Show that this can be interpreted as θ is drawn from prior j with probability p_j. Show also that the marginal distribution of Y can be expressed as $\bar{f}(y) = \sum_{j=1}^k p_j \bar{f}_j(y)$.

(c) Then show that the posterior distribution for θ becomes

$$\pi(\theta \,|\, y) = p_1^* \pi_1(\theta \,|\, y) + \cdots + p_k^* \pi_k(\theta \,|\, y),$$

with revised prior probabilities $p_j^* = p_j \bar{f}_j(y) / \sum_{j'=1}^k p_{j'} \bar{f}_{j'}(y)$ for the different types of priors.

(d) Suppose $Y \sim \text{binom}(n,\theta)$, and that the prior used for θ is $0.15\,\text{Beta}(2,10) + 0.70\,\text{Beta}(15,15) + 0.15\,\text{Beta}(10,2)$. Draw this prior in a plot. With $n = 100$, compute the posterior probabilities p_1^*, p_2^*, p_3^*, and draw the posterior distribution for θ, along with the prior, for each of the cases $y = 12$, $y = 48$, $y = 91$.

(e) (xx revisit the Gott würfelt nicht, Ex. 6.10. do a mixture prior, perhaps $0.50\,\text{Dir}(1,1,1,1,1,1) + 0.50\,\text{Dir}(s,s,s,s,s,s)$, where s is quite big, reflecting the possibility that the die is perfectly fair with probabilities equal to or very close to $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. xx)

(f) Generalise the above to the situation where $\pi(\theta) = \int \pi_\alpha(\theta)\,\mathrm{d}G(\alpha)$ is a mixture of $\pi_\alpha(\theta)$ priors, with $\mathrm{d}G(\alpha)$ a fully general probability measure over the space of hyperparameter α. The α coud be a parameter belonging to a finite set, matching the setup above, or a full continuous mixture. Show that the posterior can be represented as $\pi(\theta \,|\, y) = \int \pi_\alpha(\theta \,|\, y)\,\mathrm{d}G(\alpha \,|\, \text{data})$, where $\mathrm{d}G(\alpha \,|\, \text{data})$ is the posterior for the hyperparameter, and $\pi_\alpha(\theta \,|\, y)$ is the posterior for θ in the setup where α is fixed and known.

(g) xx

The Jeffreys prior

Ex. 6.17 *The Jeffreys prior: the basics.* (xx here we spell out the invariance arguments leading to $\pi_0(\theta) \propto |J(\theta)|^{1/2}$. often improper, but with proper posteriors. xx)

Ex. 6.18 *The Jeffreys prior in certain models.* (xx to come, examples, illustrations, some models. xx)

(a) For binomial (n,p) model, show that the Jeffreys prior is the Beta$(\frac{1}{2},\frac{1}{2})$. Compare this with the geometric, with $f(y,p) = (1-p)^{y-1}p$. (xx different priors, even though the likelihoods are proportional. xx)

(b) Then consider the trinomial (X,Y,Z), with probabilities proportional to $p^x q^y (1-p-q)^z$, with $x+y+z = n$. Find the 2×2 Fisher information matrix, its determinant, and show that the Jeffreys prior is proportional to $p^{-1/2} q^{-1/2} (1-p-q)^{-1/2}$, which is the Dirichlet $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$ for (p,q,r). Generalise to the multinomial $(n, p_1, \ldots, p_k)$ case, with $p_1 + \cdots + p_k = 1$.

(c) For the normal ...

(d) For the Poisson with parameter θ, show that the Jeffreys prior is $1/\theta^{1/2}$.

(e) For the Gamma (a, b), show that the Jeffreys prior takes the form $\pi(a, b) = \pi_1(a)\pi_2(b)$, with $\pi_1(a) \propto \{a\psi'(a) - 1\}^{1/2}$ and $\pi_2(b) \propto 1/b$.

Ex. 6.19 *A simple model for deviations from uniformity.* This exercise illustrates how we can carry out Bayesian analysis for almost any given one-parameter model, via simple numerical techniques; bigger models need bigger tools, as we come back to (xx where xx). Consider the model $f(y, \theta) = 1 + \theta(y - \frac{1}{2})$ for $y \in [0, 1]$.

(a) Show that this indeed defines a bona fide model, for $\theta \in [-2, 2]$, and with c.d.f. $F(y, \theta) = y + \frac{1}{2}\theta(y^2 - y)$. Show that the Fisher information is

$$J(\theta) = \int_{-1/2}^{1/2} \frac{x^2}{1 + \theta x}\,\mathrm{d}x.$$

(xx perhaps more, a formula. xx) Compute and display the Jeffreys prior.

(b) Take $\theta_{\rm true} = 0.333$, simulate say $n = 100$ points from the model, and give a graph for the log-likelihood function. Compute the ML and an approximate 90 percent interval for θ via the methods of Chapter 5.

(c) Then, with a uniform prior on $[-2, 2]$, compute and display the posterior distribution for θ.

(d) Using a fine grid, e.g. with grid length 0.0001, sample say 10^5 points from the posterior distribution. From these, provide $0.05, 0.50, 0.95$ quantiles. (xx just a bit more; round off; point away. xx)

Marginal distributions and Bernshteĭn–von Mises theorems

Ex. 6.20 *The marginal distribution.* Suppose we have data $y_1, \ldots, y_n$ from a model $f(y, \theta)$, with a prior $\pi(\theta)$. Most of the time Bayesians care about the posterior distribution, but on occasion, also in connection with bigger setups, one needs the marginal distribution, which is $\bar f(y_1, \ldots, y_n) = \int L_n(\theta)\pi(\theta)\,\mathrm{d}\theta$, in terms of the likelihood function $L_n(\theta)$. In various setups there will be a clear formula for this $\bar f$, se seen below; see Ex. 6.22 for a very useful approximation method for more complex cases.

(a) Let $y_1, \ldots, y_n$ be independent Bernoulli variables with $\Pr(y_i = 1 \,|\, \theta) = \theta$, and let $\theta \sim \mathrm{Beta}(a, b)$. Writing $z = \sum_{i=1}^n y_i$ for the number of 1s, show that

$$\bar f(y_1, \ldots, y_n) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a + z)\Gamma(b + n - z)}{\Gamma(a + b + n)}.$$

(b) Let then $y_1, \ldots, y_n$ be independent $\mathrm{Pois}(\theta)$, with a $\mathrm{Gam}(a, b)$ prior, as with Ex. 6.1. Show that

$$\bar f(y_1, \ldots, y_n) = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + n\bar y)}{(b + n)^{a+n\bar y}} \frac{1}{y_1! \cdots y_n!}.$$

(c) Consider i.i.d. data $y_i \sim \mathrm{N}(\xi, \sigma^2)$, with known σ and a normal prior $\xi \sim \mathrm{N}(\xi_0, \sigma_0^2)$ for ξ. Find the marginal distribution (xx give a formula here xx).

(d) (xx do also $\mathrm{N}(x_i^{\mathrm{t}}\beta, \sigma^2)$ with $\beta \sim \mathrm{N}(\beta_0, \Sigma_0)$. find the marginal. check with other exercises. xx)

(e) (xx then also for the gamma-normal; give a formula for $\bar{f}(y_1, \ldots, y_n)$. xx)

(f) (xx something to point to empirical Bayes. calibrate carefully with loss-risk Ch. **??**. can already point to Stein things. and to mixtures, where these $\bar{f}(y)$ turn up as ingredients. xx)

Ex. 6.21 *The gamma-normal induced marginal model.* (xx edit intro sentences. $Y_1, \ldots, Y_n$ are i.i.d. from the $\mathrm{N}(\xi, \sigma^2)$, given these two parameters. xx) For the direct Bayesian use one only needs the prior to posterior computation, in this case from the initial $\mathrm{GN}(a, b, \xi_0, v)$ to the updated GN given in Ex. 6.14, and one somehow bypasses the marginal density $\bar{f}(y_1, \ldots, y_n)$ of the data, the likelihood with the parameters (ξ, σ) integrated out according to the prior. On occasion this marginal distribution is important, however, and also finds use as a model in its own right, for positively dependent data.

(a) (xx first things with $\xi \sim \mathrm{N}(\xi_0, \tau_0^2)$, known σ. two ways of computing and seeing $\bar{f}(y_1, \ldots, y_n)$. do marginal moments and correlations. xx)

(b) xx

(c) Then the $\mathrm{GN}(a, b, \xi_0, v)$ gamma-normal prior for (λ, ξ), as with Ex. 6.14. Show first that the likelihood times the prior, $L_n(\lambda, \xi)p(\lambda, \xi)$, can be expressed as

$$\frac{\lambda^{n/2}}{(2\pi)^{n/2}} \exp[-\tfrac{1}{2}\lambda\{Q_0 + n(\xi - \bar{y})^2\}] \frac{b^a}{\Gamma(a)} \lambda^{a-1}(v\lambda)^{1/2} \exp[-\lambda\{b + \tfrac{1}{2}v(\xi - \xi_0)^2\}].$$

Then integrate out the ξ to get

$$\frac{1}{(2\pi)^{n/2}} \frac{b^a}{\Gamma(a)} \lambda^{a+n/2-1} \Big(\frac{v}{v+n}\Big)^{1/2} \exp\{-\lambda(b + \tfrac{1}{2}Q_0 + \tfrac{1}{2}d_n(\bar{y} - \xi_0)^2\},$$

with $d_n = (1/v + 1/n)^{-1}$ as per Ex. 6.14. Show then that this leads to the marginal density being

$$\bar{f}(y_1, \ldots, y_n) = (2\pi)^{-n/2} \Big(\frac{v}{v+n}\Big)^{1/2} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + n/2)}{\{b + \frac{1}{2}Q_0 + \frac{1}{2}d_n(\bar{y} - \xi_0)^2\}^{a+n/2}}.$$

Ex. 6.22 *Approximating the marginal distribution.* In the setup of Ex. 6.20, we go through a useful type of Laplace approximation for the marginal.

(a) Writing as usual $\ell_n(\theta)$ for the log-likelihood, with maximum value $\ell_{n,\max} = \ell_n(\widehat{\theta})$, in terms of the ML estimator, show that

$$\bar{f}(y_1, \ldots, y_n) = \exp(\ell_{n,\max}) \int \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta})\}\pi(\theta)\,\mathrm{d}\theta.$$

With $J_n = -(1/n)\partial^2 \ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ the normalised observed information matrix, of dimension say $p \times p$, show that the marginal can be approximated with

$$\begin{aligned} \bar{f} &\doteq \exp(\ell_{n,\max}) \int \exp\{-\tfrac{1}{2} n(\theta - \widehat{\theta})^{\mathrm{t}} J_n (\theta - \widehat{\theta})\} \pi(\theta)\, \mathrm{d}\theta \\ &= \exp(\ell_{n,\max}) \int \exp(-\tfrac{1}{2} s^{\mathrm{t}} J_n s) \pi(\widehat{\theta} + s/\sqrt{n})\, \mathrm{d}s/n^{p/2} \\ &\doteq \exp(\ell_{n,\max}) \pi(\widehat{\theta}) (2\pi)^{p/2} |J_n|^{-1/2} / n^{p/2}. \end{aligned}$$

(b) (xx a couple of things here. check how successful the approximation is in two setups. the formula

$$\log \bar{f} \doteq \ell_{n,\max} - \tfrac{1}{2} p \log n + \log \pi(\widehat{\theta}) - \tfrac{1}{2} \log |J_n| + \tfrac{1}{2} p \log(2\pi),$$

with its two leading terms, lead to the BIC in Ch. 11. xx)

Ex. 6.23 *Bernshteĭn–von Mises approximations.* Suppose observations $Y_1, \ldots, Y_n$ are i.i.d. from a density $f(y, \theta)$, with $\pi(\theta)$ a prior for the model parameter, of dimension say p. The posterior density can of course be quite complicated, perhaps necessitating numerical efforts, or simulation, for its evaluation. Remarkably, there are generic and simple normal approximations, however.

(a) Show that the posterior density $\pi_n(\theta) = \pi(\theta \,|\, \text{data})$ is proportional to $\pi(\theta) \exp\{\ell_n(\theta)\}$, with $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i, \theta)$ the log-likelihood function.

(b) Let $\widehat{\theta}$ be the ML estimator, and $J_n = -(1/n)\partial^2 \ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ the normalised observed information, as per likelihood theory of Ch. 5. Show that the density of $Z_n = \sqrt{n}(\theta - \widehat{\theta})$ is $g_n(z) = \pi_n(\widehat{\theta} + z/\sqrt{n})(1/n^{p/2})$, and that it can be approximated as

$$g_n(z) \propto \pi(\widehat{\theta} + z/\sqrt{n}) \exp\{\ell_n(\widehat{\theta} + z/\sqrt{n}) - \ell_n(\widehat{\theta})\} \doteq \pi(\widehat{\theta} + z/\sqrt{n}) \exp(-\tfrac{1}{2} z^{\mathrm{t}} J_n z).$$

(c) Suppose then that the data really are i.i.d. from the model, with an underlying θ_{true}. In particular, then $\widehat{\theta} \to_{\mathrm{pr}} \theta_{\mathrm{true}}$ and $J_n \to_{\mathrm{pr}} J = J(\theta_{\mathrm{true}})$, by (xx point to Ch 4 exercises xx). If $\pi(\theta)$ is continuous in a neighbourhood around θ_{true}, show that $g_n(z)$ tends to the density of a $\mathrm{N}_p(0, J^{-1})$, in probability. In concrete terms, show

$$D_n = \int |g_n(z) - \phi_p(z, 0, J^{-1})|\, \mathrm{d}z \to_{\mathrm{pr}} 0.$$

This is one of several versions of *Bernshteĭn–von Mises theorems*. These are Bayesian mirror versions of the classical maximum likelihood asymptotics results in the frequentist camp:

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_{\mathrm{true}}) &\to_d \mathrm{N}(0, J^{-1}), \\ \sqrt{n}(\theta - \widehat{\theta}) \,|\, \text{data} &\to_d \mathrm{N}(0, J^{-1}), \text{ in probability.} \end{aligned}$$

(d) Check two clear situations in detail, comparing the exact posterior density $\pi(\theta \,|\, \text{data})$ with the normal approximation: (i) where $Y \,|\, \theta \sim \mathrm{binom}(n, \theta)$, and $\theta \sim \mathrm{Beta}(a_0, b_0)$; (ii) where $Y_1, \ldots, Y_n \,|\, \theta$ are i.i.d. $\mathrm{Pois}(\theta)$, and $\theta \sim \mathrm{Gam}(a_0, b_0)$. Choose n and (a_0, b_0), and also the true θ_{true}, for your brief investigations.

(e) (xx just a bit more. lazy Bayesian. prior disappears. different Bayesians agree with each other, and also with the frequentist. xx)

Ex. 6.24 *Bayes and minimax normal estimation with the linex loss.* (xx perhaps to be moved to Ch 8. xx) We worked out some basic properties of the linex loss function $\exp\{c(t-\theta)\} - 1 - c(t-\theta)$ in Ex. 6.5. Here we use the Bayesian machinery to find a minimax estimator for the normal mean.

(a) Consider the simple prototype setup where a single X has the $\mathrm{N}(\theta, 1)$ distribution. Show that the estimator $X + d$ has risk function

$$r_c(\theta) = \mathrm{E}_\theta\,[\exp\{c(X+d-\theta)\} - 1 - c(X+d-\theta)] = \exp(cd + \tfrac{1}{2}c^2) - 1 - cd,$$

constant in θ, and that the best estimator of this sort is $\theta^* = X - \frac{1}{2}c$. Show also that the risk achieved, by this estimator, is $\frac{1}{2}c^2$.

(b) Now consider Bayes estimation, with the prior $\theta \sim \mathrm{N}(0, \tau^2)$. Show via Ex. 6.13 that $(\theta \,|\, x) \sim \mathrm{N}(wx, w)$, with $w = \tau^2/(\tau^2+1)$. Show, perhaps via expressing $\theta \,|\, x$ as $wx + w^{1/2}N$ with N a standard normal, that the posterior expected loss is

$$\mathrm{E}\,\{L_c(\theta, t) \,|\, x\} = \exp\{c(t - wx) + \tfrac{1}{2}wc^2\} - 1 - c(t - wx).$$

Deduce that the Bayes estimator is $\widehat{\theta}_B = wx - \frac{1}{2}wc$, and that the posterior expected loss is $\mathrm{E}\,\{L_c(\theta, \widehat{\theta}_B) \,|\, x\} = \frac{1}{2}wc^2$, independent of x.

(c) Show that $\theta^* = X - \frac{1}{2}c$ is minimax. Show also, via Blyth's method, that it is in fact admissible.

(d) Generalise the above to the case of a full sample $X_1, \ldots, X_n$ from $\mathrm{N}(\theta, 1)$. Find the Bayes estimator and associated minimum Bayes risk, for the $\mathrm{N}(0, \tau^2)$ prior, and prove that $\theta^* = \bar{X} - \frac{1}{2}c/n$ is minimax. What is its minimax risk?

(e) Find the distribution of $Z_n = \sqrt{n}(\theta^* - \theta)$, and comment on its limit, (i) when the loss-skewness parameter c is fixed, (ii) when $c = \sqrt{n}$.

(f) (xx perhaps another example with linex loss. and something where we see certain arguments lead to a choice of c. xx)

Ex. 6.25 *More on the linex loss.* For the linex loss, studied initially in Ex. 6.5 and then in Ex. 6.24 for the normal case, we now find out more.

(a) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the $\mathrm{Pois}(\theta)$, with prior $\theta \sim \mathrm{Gam}(a, b)$, as with Ex. 6.1. Show that the Bayes estimator with the linex loss is $\widehat{\theta}_B = (1/c)(a + n\bar{y}) \log\{1 + c/(b+n)\}$. Verify that when $c \to 0$, we retrieve the posterior means of Ex. 6.1.

(b) (xx one more case with a clear formula. perhaps σ in normal. xx)

(c) As we know from Ex. 6.23, an approximation to the posterior distribution is $\theta \,|\, \text{data} \sim \mathrm{N}(\widehat{\theta}_{\mathrm{ml}}, \widehat{\sigma}^2/n)$, in terms of the maximum likelihood estimate and estimated inverse Fisher information. Deduce that $M_n(-c) \doteq \exp(-c\widehat{\theta}_{\mathrm{ml}} + \frac{1}{2}c^2\widehat{\sigma}^2/n)$, in the notation above, and that this leads to the approximation $\widehat{\theta}_B = \widehat{\theta}_{\mathrm{ml}} - \frac{1}{2}c\widehat{\sigma}^2/n$ for the Bayes estimator under linex loss. Show also that the posterior expected loss is approximately $\frac{1}{2}c^2\widehat{\sigma}^2/n$.

(d) (xx an example where we can check the approximation with the exact Bayes estimator, e.g. with Poisson and gamma. xx)

Ex. 6.26 *Multiparameter inference.* (xx something here. joint estimation of $(\theta_1, \ldots, \theta_k)$, say for normal or Poisson or binomial, with Bayes and empirical Bayes. calibrate with Ch8. xx) Consider a setup with $Y_1, \ldots, Y_k$ being independent Poisson counts with parameters $\theta_1, \ldots, \theta_k$, and where the object is to estimate all of the parameters, jointly, with loss function $L(\theta, a) = \sum_{j=1}^k (a_j - \theta_j)^2/\theta_j$. If the parameters somehow are related, as in not too different, this might be built into a Bayesian scheme.

(a) The default estimator would be $\widehat{\theta}_j = Y_j$, for $j = 1, \ldots, k$. Show that its risk function is constant, equal to k.

(b) Consider then the prior which takes $\theta_1, \ldots, \theta_k$ independent from a Gamma (a, b). Show that $\theta_j \,|\, \text{data}$ is a $\text{Gam}(a_j + y_j, b+1)$, find the Bayes estimator, and its risk function. (xx jotting down details here. xx)

$$\widehat{\theta}_j = \widehat{\theta}_j(a,b) = 1/(\mathrm{E}\,(1/\theta_j) \,|\, y_j) = \frac{a + y_j - 1}{b+1}.$$

empirical Bayes: $\mathrm{E}\,(Y_j \,|\, \theta_j) = \theta_j$, $\mathrm{Var}\,(Y_j \,|\, \theta_j)) = \theta_j$, leading to $\mathrm{E}\, Y_j = a/b$, $\mathrm{Var}\, Y_j = a/b + a/b^2$; can then use moment matching to find $\widetilde{a}, \widetilde{b}$. this leads to empirical Bayes estimators

$$\theta_j^* = \widehat{\theta}_j(\widetilde{a}, \widetilde{b}) = \frac{\widetilde{a} + y_j - 1}{\widetilde{b} + 1}.$$

also:

$$\bar{f}(y_j, a, b) = \int_0^\infty \exp(-\theta)\theta^{y_j}/y_j! \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta)\, \mathrm{d}\theta = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+y_j)}{(b+1)^{a+y_j}} \frac{1}{y_j!}$$

Full Bayes, a prior for (a, b):

$$\pi_0(a,b) \prod_{j=1}^k g(\theta_j, a, b) f(y_j, \theta_j) = \pi_0(a,b) \prod_{j=1}^k g(\theta_j, a+y_j, b+1) \bar{f}(y_j, a, b)$$

which leads to $\pi(a, b \,|\, \text{data}) \propto \pi_0(a,b) \prod_{j=1}^k \bar{f}(y_j, a, b)$. then mcmc for sampling from $(a, b, \theta_1, \ldots, \theta_k)$ given data.

(c) (xx then empirical Bayes. Clevenson–Zidek, Stoltenberg–Hjort. xx)

Notes and pointers

(xx mention Varian (1975); Zellner (1986); Claeskens and Hjort (2008a) for the linex loss; but this should perhaps be in Ch. 8. mention Stigler (1983), who uses Bayes's Theorem to help him spot a candidate for an earlier discoverer of Bayes' Theorem. mention MCMC revolution since the 1990ies. point to Story ii.5. xx)

I.7

Confidence distributions, confidence curves, combining information sources

With ϕ a focus parameter, a function of the full parameter vector θ, the Bayesian setup gives a posterior distribution. This requires the conceptually and practically difficult task of defining a prior for the full θ, however. Confidence distributions (CDs) are a frequentist parallel, yielding post-data distributions for such focus parameters, without any prior. In this chapter we develop theory for CDs and confidence curves, and also find ways of combining CDs across different information sources. Computing CDs is not an easy or automatic task, but we develop and illustrate several recipes. For the exponential family class, we derive optimal CDs, with their own clear recipes.

Key words: boundary constraints, combining CDs, confidence curves, confidence distributions, exponential family, meta-analysis, t-bootstrapping

Confidence distributions and confidence curves are fruitful statistical inference summaries. Suppose in general terms that data y stem from a model $f(y, \theta)$, with model parameter $\theta = (\theta_1, \ldots, \theta_p)$, and that $\phi = \phi(\theta_1, \ldots, \theta_p)$ is a parameter of particular interest. A *confidence distribution* for ϕ, a CD, for short, is a function $C(\phi, y)$, such that (i) it is a c.d.f. in ϕ, for each dataset y, and (ii) the distribution of $U = C(\phi_0, Y)$ is uniform at the true value $\phi_0 = \phi(\theta_0)$. In other words,

$$\Pr_{\theta_0}\{a \le C(\phi_0, Y) \le b\} = b - a \quad \text{for each } a, b \in [0, 1].$$

Assuming this random c.d.f. has a unique inverse, then, we have

$$\Pr_{\theta}\{C^{-1}(0.05, Y) \le \phi \le C^{-1}(0.95, Y)\} = 0.90, \tag{7.1}$$

and of course similarly for other choices of quantiles. This is by definition making $[C^{-1}(0.05, y_{\rm obs}), C^{-1}(0.95, y_{\rm obs})]$ a 90 percent confidence interval for the focus parameter ϕ. The CD concept is hence related to and an extension of the confidence intervals, see Ch. 4. The *confidence curve* is a related summary graph, most often computed from the CD via $\mathrm{cc}(\phi, y) = |1 - 2\,C(\phi, y)|$. It has the practical property that $\{\phi\colon \mathrm{cc}(\phi, y) \le 0.90\}$ give the 90 percent interval directly; similarly, all intervals at any desired confidence

level can be read off from the confidence curve. We sometimes write simply $C(\phi)$ and $\mathrm{cc}(\phi)$, omitting the data argument, when clear from the context what the data are, in $C(\phi) = C(\phi, \mathrm{data})$.

Construction a CD is not always an easy or automatic task, but we develop several practical recipes, some of which are based on approximate normality, or on more general methods of likelihood theory. Just as tests have detection power, also CDs have power, and theory is developed below to find optimal CDs in classes of situations. This is partly parallelling the optimal testing methodology of Ch. 4. All in all we develop and illustrate the following recipes: (i) Via the c.d.f. of an estimator; (ii) normal approximation; (iii) based on a pivot; (iv) deviance and Wilks theorem; (v) t-bootstrapping; (vi) the optimal CD via conditional distributions, if inside the exponential family.

The CDs are post-data graphical summaries of the level of uncertainty for any focus parameter, and can be seen as frequentist parallels to the Bayesian posterior distributions; here there is no prior, however. We illustrate this 'clear data-only based posteriors without priors' aspect of the CDs through theory and applications (xx perhaps point to a few Stories xx).

Combining different information sources is a broad statistical theme, going back to the first meta-analysis concepts and methods of Karl Pearson just after 1900 (Simpson and Pearson, 1904). The more familiar meta-analysis methods aim at combining independent estimators for the same quantity, or for providing a broader population assessment of similar but not identical parameters. CDs are useful for such endeavours, and we provide methods for combining sources more general than the traditional ones.

Recipes for constructing CDs

Ex. 7.1 *The probability transform.* Some of the following facts are related to various operations for confidence distributions and confidence curves

(a) Suppose X has a continuous and increasing cumulative distribution function F, i.e. $F(x) = \Pr(X \le x)$. Show that $U = F(X)$ is uniform on the unit interval. Any continuously distributed random variable can hence be transformed to uniformity, via this *probability transform*.

(b) Show that also $U_2 = 1 - F(X)$ and $U_3 = |1 - 2\,F(X)|$ have uniform distributions.

(c) Simulate a million copies of $x_i \sim \mathrm{N}(0,1)$, and check the histogram of $\Gamma_1(x_i^2)$, where Γ_ν is the cumulative distribution function of a χ^2_ν. Comment on what you find.

Ex. 7.2 *Recipe One: via the c.d.f. of an estimator.* Suppose θ is a one-dimensional parameter, for which we need a CD, after having observed data y_{obs}. If there is an estimator $\widehat\theta$, with a distribution depending only on this θ, there is a clear recipe.

(a) Assume therefore that $\widehat\theta$ has a continuous distribution function $K_\theta(x) = \Pr_\theta(\widehat\theta \le x)$; its distribution is here required to depend only on θ, not on other aspects of the underlying model employed. Consider Recipe One, the construction

$$C(\theta, y_{\mathrm{obs}}) = \Pr_\theta(\widehat\theta \ge \widehat\theta_{\mathrm{obs}}) = 1 - K_\theta(\widehat\theta_{\mathrm{obs}}),$$

a curve that can be computed and plotted post-data, where $\widehat{\theta}_{\rm obs} = \widehat{\theta}(y_{\rm obs})$ is the observed estimate. Show that it has the property that the random $C(\theta, Y)$ is uniformly distributed, for each fixed θ.

(b) To illustrate, go through the details for the case of using $\widehat{\theta} = 1/\bar{Y}$, with i.i.d. observations $Y_1, \ldots, Y_n$ from the exponential $\theta \exp(-\theta y)$. Show first that $2\theta Y_i \sim \chi^2_2$, and derive $K_\theta(x) = 1 - \Gamma_{2n}(2n\theta/x)$, with Γ_{2n} the c.d.f. of the χ^2_{2n}. Simulate data and plot the CD $C(\theta, y_{\rm obs}) = \Gamma_{2n}(2n\theta/\widehat{\theta}_{\rm obs})$. From the CD, find a 95 percent interval for θ.

(c) Assume $X_1, \ldots, X_m$ are i.i.d. Expo(θ_1) and that $Y_1, \ldots, Y_n$ are i.i.d. Expo(θ_2). Find the distribution of the estimator $\widehat{\rho} = \widehat{\theta}_1/\widehat{\theta}_2$ for the ratio $\rho = \theta_1/\theta_2$, and derive the associated CD.

(d) Generate $n = 25$ datapoints from the double exponential density $f(y, \theta) = \frac{1}{2}\exp(-|y-\theta)$, using your favourite true θ_0. Compute and display the CD for θ based on the median M_n.

(e) For a simpler and more fundamental illustration, suppose $\widehat{\theta}$ has a normal distribution centred at θ, with a known variance, say $\widehat{\theta} \sim {\rm N}(\theta, \kappa^2)$. Show that Recipe One gives $C(\theta) = \Phi((\theta - \widehat{\theta})/\kappa)$. Check that the famous 95 percent interval $\widehat{\theta} \pm 1.96\,\kappa$ agrees with this.

Ex. 7.3 *Confidence distribution and confidence curve for the normal standard deviation.* The confidence distribution C and the confidence curve cc are close cousins, and they do not need to be both displayed for each new statistical application. Here is a simple illustration. You observe the $n = 6$ data points 4.09, 6.37, 6.87, 7.86, 8.28, 13.13 from a normal distribution and wish to assess the underlying spread parameter, the standard deviation σ.

(a) For the empirical variance, use $\widehat{\sigma}^2 \sim \sigma^2\chi^2_m/m$, with $m = n - 1$, to build the CD

$$C(\sigma, y_{\rm obs}) = {\rm Pr}_\sigma(\widehat{\sigma} \geq \widehat{\sigma}_{\rm obs}) = 1 - \Gamma_m(m\widehat{\sigma}^2_{\rm obs}/\sigma^2).$$

Here $y_{\rm obs}$ represents the observed data, and $\widehat{\sigma}_{\rm obs}$ the observed point estimate. Show that $C(\sigma, Y) \sim$ unif, where Y represents a random data set $Y_1, \ldots, Y_n$, from the σ in question. In particular, the distribution of $C(\sigma, Y)$ does not depend on σ. Make a graph, also of the associated confidence curve

the confidence curve

$$\text{cc}(\sigma, y_{\rm obs}) = |1 - 2\,C(\sigma, y_{\rm obs})| = |1 - 2\,\Gamma_m(m\widehat{\sigma}^2_{\rm obs}/\sigma^2)|.$$

Compute the *median confidence estimate* $\widehat{\sigma}_{0.50} = C^{-1}(0.50, y_{\rm obs})$ and the natural 90 percent confidence interval $[C^{-1}(0.05, y_{\rm obs}), C^{-1}(0.95, y_{\rm obs})]$. Find and display also the *confidence density* $c(\sigma, y_{\rm obs})$, the derivative of the CD.

(b) Compute also the *confidence density* $c(\sigma, y_{\rm obs})$ associated with the CD. Compute furthermore its mode, say σ^*, and briefly assess its properties as an estimator of σ.

(c) A Bayesian approach to the same problem, i.e. finding a posterior distribution for σ, is to start with a prior $\pi(\sigma)$ and then compute $\pi(\sigma \,|\, y_{\rm obs}) \propto \pi(\sigma)g(\widehat{\sigma}, \sigma)$, where $g(\widehat{\sigma}, \sigma)$ is the likelihood, here the density function for $\widehat{\sigma}$ as a function of σ. When does such a Bayesian approach agree with the confidence density?

(d) Suppose there are two independent normal samples, with standard deviations σ_1 and σ_2. Construct a CD for $\rho = \sigma_1/\sigma_2$. Invent a second simple small dataset, to complement the first dataset given above, and then compute and display the confidence curve $\mathrm{cc}(\rho, \mathrm{data})$.

Ex. 7.4 *Computing a CD with simulation and isotonic repair.* (xx to be polished. we use this is Story iii.10 and perhaps in other places, where simulations are expensive. xx) Suppose one observes $y_1, \ldots, y_n$ from the one-parameter Weibull distribution with c.d.f. $F(y, b) = 1 - \exp(-y^b)$, with sample size $n = 25$, and computes the data mean $\bar{y}_{\mathrm{obs}} = 1.313$.

(a) Though we do not actually need this in the CD computations here, find an estimate of b based on $\mathrm{E}\, Y_i = \Gamma(1 + 1/b)$; see Ex. 1.54. Show that $C(b) = \Pr_b(\bar{Y} \le \bar{y}_{\mathrm{obs}})$ is a CD for b.

(b) The practical obstacle here is that $\bar{Y}$ does not have a simple distribution. But we're saved by simulation. Show that the simulation recipe $Y_i^* = V_i^{1/b}$ produces outcomes from the weibull $F(y, b)$, where the V_i are unit exponential. For a grid of b values, e.g. from 0.20 to 1.20, compute the simulation based $C^*(b)$, the proportion of B cases where the simulated $\bar{Y}^*$ is below $\bar{y}_{\mathrm{obs}}$. Compute also the confidence curve $\mathrm{cc}^*(b) = |1 - 2\,C^*(b)|$. For this simple example it is easy to accomplish this with a high B, say 10^5, to make $C^*(b)$ and $\mathrm{cc}^*(b)$ smooth and very close to the real $C(b)$ and $\mathrm{cc}(b)$; for this illustration, however, make the simulation size as relatively small as $B = 100$, and plot the curves, as in Figure 7.1.

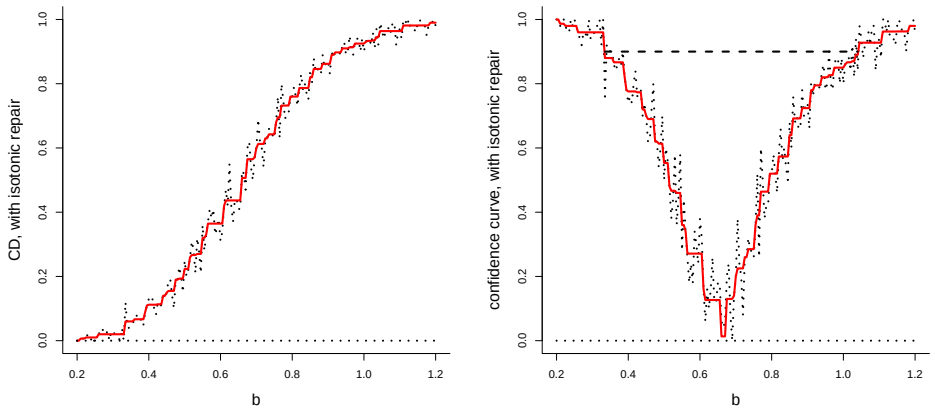


Figure 7.1: *Simulation based confidence distribution* $C^*(b)$ *and confidence curve* $\mathrm{cc}^*(b)$ *for the Weibull parameter* b*, based on the observed sample mean* $\bar{y}_{\mathrm{obs}} = 1.313$ *for* $n = 25$ *data points, along with isotonic repairs. The simulation size here is the low* $B = 100$.

(c) We learn that with a low or moderate simulation size B, the $C^*(b)$ and $\mathrm{cc}^*(b)$ will be wiggly. We can do better, using the prior knowledge that $C(b)$ is increasing. There

isotonic regression

are several repair mechanisms, which from the potentially wiggly $C^*(b)$ create a monotonically increasing curve. A simple scheme is so-called *isotonic regression*, the details of which we do need to get into here. Supposing you have first created `bval` and `Cval` in your `R` session, you may use `Cvaliso=isoreg(bval,Cval)$yf`, which repairs your $C^*(b)$ and $\text{cc}^*(b)$ to ensure monotonicity. Produce versions of Figure 7.1, left and right panels.

(d) (xx round off. explain salient points about generalisability. we need reduction to one-parameter situation. xx)

Ex. 7.5 *An extension of Recipe One.* In Ex. 7.2 we saw that the simple construction $C(\theta, y) = \text{Pr}_\theta(\widehat{\theta} \geq \widehat{\theta}_{\text{obs}})$ gives a CD, in the case of one-dimensional setups with a well-defined estimator $\widehat{\theta}$.

(a) When working with estimators, finetuning efforts are often exuded to trim away biases, getting the scaling right, etc. In a sense this is not needed here, when constructing the CD. Show that if $\widehat{\alpha} = g(\widehat{\theta})$, with any smooth increasing g, the recipe $C^*(\theta) = \text{Pr}_\theta(\widehat{\alpha} \geq \widehat{\alpha}_{\text{obs}})$ gives precisely the same CD as without the g transformation.

(b) So this CD recipe relies merely on having an informative statistic, say Z, with a distribution stochastically increasing in θ; it does not really have to be an estimator for that parameter. Show that $C(\theta, y) = \text{Pr}_\theta(Z \geq z_{\text{obs}})$ is a bona fide CD.

(c) Show also that the construction works, if there are other parameters at play too, as long as the distribution of the chosen Z only depends on θ. Go through the details for the case of the Y_i being $\text{N}(\mu, \sigma^2)$, with $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2$, and also for $Z' = \sum_{i=1}^n |Y_i - M_n|$, where M_n is the empirical median. Compute, display, compare both CDs, based on Z and on Z', for the simple dataset of Ex. 7.3 (with $n = 6$). For the Z case, there is a formula, but for the Z' case you would need simulation, for a grid of σ values; see Ex. 7.4.

(d) (xx one more example, where there is a Z carrying information, but not qua estimator. xx)

Ex. 7.6 *Recipe Two: the normal approximation CD.* Applying Recipe One of Ex. 7.2 to the case of the estimator having a normal distribution leads as we saw there to a clear CD, provided the variance is known. But this is at least approximately so, for large classes of situations, as we've seen in Chs. 2 and 5.

(a) Suppose in general terms that $\widehat{\theta}$ estimates θ, and that its distribution is approximately a $\text{N}(\theta, \kappa^2)$. Explain that $C(\theta) = \Phi((\theta - \widehat{\theta})/\kappa)$ then is an approximate CD for θ. More formally, if $(\widehat{\theta} - \theta_0)/\widehat{\kappa} \to_d \text{N}(0,1)$, at the true parameter θ_0, show that $C(\theta, Y) = \Phi((\theta - \widehat{\theta})/\widehat{\kappa})$ has the property that it converges in distribution to the uniform, at θ_0. In typical applications of these arguments, there is a $\sqrt{n}$ scaling in terms of an underlying sample size, with $\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \text{N}(0, \tau^2)$, say, and $\widehat{\kappa} = \widehat{\tau}/\sqrt{n}$, with $\widehat{\tau} \to_{\text{pr}} \tau$. So this is Recipe Two, the normal approximation CD, most typically of this type $\Phi(\sqrt{n}(\theta - \widehat{\theta})/\widehat{\tau})$.

(b) Simulate a moderate or small dataset from a normal distribution. Compute and display two (approximate) CDs for the mean parameter ξ, (i) using the data mean, (ii) using the data median.

(c) We have seen in Chs. 2 and 5 that approximate normality is highly common, for large classes of estimators, typically along with consistent estimators for the variances. In particular, the delta method implies approximate normality of smooth functions of background estimators (see Ex. 2.47, **??**), making in its turn approximate normality CDs easily available. For a simple illustration, suppose you throw your nearest die, which has probability p of giving a '6', until you get your first '6'. You carry out this geometric experiment $n = 10$ times, giving you the counts $Y_1, \ldots, Y_n$ equal to 1, 2, 17, 18, 20, 4, 3, 1, 15, 3. Use the normal approximation for $\bar{Y}$ to give an approximate CD for p. You may also compare this to what one achieves working with the exact distribution of $\bar{Y}$.

(d) (xx point to logistic and poisson regression, with delta method. estimate β and also $p = H(x_0^{\mathrm{t}}\beta)$. xx)

Ex. 7.7 *Recipe Three: from a pivot to a CD.* (xx check that we're not repetitive regarding pivot. xx) Suppose in general terms that ϕ is some parameter of interest, in a model for observations Y, and that a function $A = \mathrm{piv}(\phi, Y)$ of the parameter and the data has the property that its distribution does not depend on the model parameters (in particular, therefore, not on ϕ, which might itself be a function of other model parameters). We call A a pivot, in more pedantic detail a pivot for the parameter ϕ.

(a) With $Y_1, \ldots, Y_n$ independent from the normal (μ, σ^2), let $R_n = \sum_{i=1}^n |Y_i - \bar{Y}|$ with the sample mean $\bar{Y}$. Show that $(\bar{Y} - \mu)/R_n$ is a pivot. Invent yet another pivot involving μ, with a different denominator.

(b) With two normal samples, say $X_1, \ldots, X_m$ from $\mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n$ from $\mathrm{N}(\mu_2, \sigma_2^2)$, suppose $\rho = \sigma_1/\sigma_2$ is in focus. Show that $(V_1/V_2)/\rho$ is a pivot for ρ, where V_1 and V_2 are the interquartile ranges for the two datasets.

(c) Consider $Y_1, \ldots, Y_n$ from the Cauchy model with density $(1/\pi)/\{1 + (y - \theta)^2\}$. Show that $R_n - \theta$ is a pivot, where $R_n = \frac{1}{2}(Q_{n,0.10} + Q_{n,0.90})$ is the average of the 0.10 and 0.90 quantiles.

(d) Back to the generalities, consider a pivot $A = \mathrm{piv}(\phi, Y)$ for ϕ in some model, increasing in ϕ. Assume the situation is continuous, not discrete, so that the pivot's distribution function K is continuous. Show that $C(\phi, y_{\mathrm{obs}}) = 1 - K(\mathrm{piv}(\phi, y_{\mathrm{obs}}))$ is a proper CD for ϕ.

(e) In clean cases we may derive the precise distribution for the pivot in question, but the CD recipe given above may be used also in more complicated setups, as long as $A = \mathrm{piv}(\phi, Y)$ may be simulated. Make an illustration of this, with the ratio of standard deviations above. Suppose two normal datasets, both of size $n = 100$, lead to interquartile ranges $V_{1,\mathrm{obs}} = 4.44$ and $V_{2,\mathrm{obs}} = 3.33$. Construct and display $C(\rho)$ and $\mathrm{cc}(\rho)$.

(f) (xx make the point that various constructions, involving large-sample approximations to the normal and to chisquares, lead to *approximate pivots*, and then again to approximate CDs and ccs. in particular, Method One, with $\Phi((\phi - \widehat{\phi})/\widehat{\kappa})$ and Method Two, with $\Gamma_1(D(\phi))$, can be seen via approximate pivots. also Method Three, construction of a t type ratio and then bootstrapping. xx)

Ex. 7.8 *CDs from the t pivot.* In Ex. 7.2 we saw that the simple construction $C(\theta, y) = \Pr_\theta(\widehat{\theta} \geq \widehat{\theta}_{\rm obs})$ gives a CD, in the case of one-dimensional setups with a well-defined estimator $\widehat{\theta}$.

(a) For a normal sample from $\mathrm{N}(\mu, \sigma^2)$, we see that several $\Pr_{\mu,\sigma}(Z \geq z_{\rm obs})$ schemes work, in that the Z in question has a distribution depending on σ, but not μ. Attempt to work with $C^*(\mu, y) = \Pr_{\mu,\sigma}(\bar{Y} \geq \bar{y}_{\rm obs})$ – and explain that it will not really work (unless σ is known).

(b) But of course there *are* natural CD constructions for μ here. What is needed is a *pivot*, say $A = \mathrm{piv}(\mu, y)$, a function binding the focus parameter and data together in a way which makes its distribution not depend on the parameters. Study indeed

$$t_n = t_n(\mu, Y) = (\bar{Y} - \mu)/(\widehat{\sigma}/\sqrt{n}),$$

with $\widehat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ the classical empirical variance. Pretend that you in all your cleverness have not seen this t_n before, and are unaware of its relation to a t distribution – but show that the distribution of t_n, call it K_n, does not depend on (μ, σ).

(c) Then show that $C(\mu, y_{\rm obs}) = K_n(t_n(\mu, y_{\rm obs}))$ is a CD for μ. Even if you do not see the connection to the classic t of Student (1908), see Ex. 1.46, you may still carry through this, by simulating $B = 10^5$ realisations of t_n, and use

$$C(\mu, y_{\rm obs}) = K_n^*(t_n(\mu, y_{\rm obs})) = \frac{1}{B} \sum_{j=1}^B I\{t_{n,j} \leq t_n(\mu, y_{\rm obs})\}.$$

Show however that by all means K_n is a t_m, with $m = n - 1$, so the canonical CD for μ is and remains $C(\mu, y_{\rm obs}) = G_m(\sqrt{n}(\mu - \bar{y}_{\rm obs})/\widehat{\sigma}_{\rm obs})$, with G_m the c.d.f. for the t_m.

Ex. 7.9 *Recipe Four: confidence curves via Wilks theorems.* Consider data from a parametric model, leading to the log-likelihood function $\ell_n(\theta)$, and that there is a focus parameter $\phi = g(\theta)$. We have seen likelihood profiling and Wilks theorems in Ch. 5, and know that the deviance $D_n(\phi) = 2\{\ell_{\max} - \ell_{\rm prof}(\phi)\}$ has the property that $D_n(\phi_0) \to_d \chi^2_1$ at the true value $\phi_0 = g(\theta_0)$; see Ex. 5.28.

(a) Recipe Four, utilising the Wilks theorems, is to form $\mathrm{cc}(\phi, y) = \Gamma_1(D_n(\phi))$, with Γ_1 the c.d.f. for the χ^2_1. Show that $\Pr_{\theta_0}(\mathrm{cc}(\phi_0, Y) \leq \alpha) \to \alpha$ for each α, and explain that this makes $\mathrm{cc}(\phi, y)$ and approximate confidence curve.

(b) For an illustration, consider the model $F(y, \theta) = y^\theta$ for observations on $[0, 1]$, where θ is an unknown positive parameter. Write down the log-likelihood function and find a formula for the maximum likelihood (ML) estimator $\widehat{\theta}$. Use also theory of Ch. 5 to write down a normal approximation to the distribution of $\widehat{\theta}$.

(c) Consider the data set

```
0.013 0.054 0.234 0.286 0.332 0.507 0.703 0.763 0.772 0.920
```

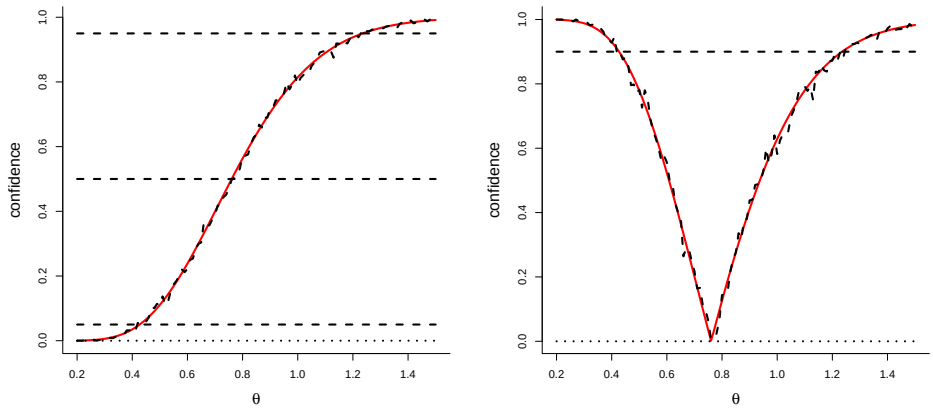


Figure 7.2: *For the simple data example of Ex. 7.9: Left panel: confidence distribution $C(\theta)$, via simulations (black and wiggly curve) and via exact calculations (red and smooth curve); right panel: the two versions of the associated confidence curve* $\mathrm{cc}(\theta)$. *From these we read off the median confidence estimate* $\widehat{\theta}_{0.50} = 0.76$, *and the 90 percent confidence interval* $[0.43, 1.24]$.

Estimate θ and compute the CD $C(\theta) = \Pr_\theta(\widehat{\theta} \geq \widehat{\theta}_{\rm obs})$, along with the confidence curve $\mathrm{cc}(\theta) = |1-2\,C(\theta)|$, (i) using simulations, (ii) using exact probability calculus. Reproduce a version of Figure 7.2.

(d) Supplement these two curves with approximations based (i) on the normal approximation for $\widehat{\theta}$ and (ii) on the chi-squared approximation for the deviance.

(e) (xx somewhere, if not here, then separately: from CD to cc, and from cc to CD, with $C(\phi) = \frac{1}{2} - \frac{1}{2}\mathrm{cc}(\phi)$ for $\phi \leq \widehat{\phi}_{0.50}$ and $\frac{1}{2} + \frac{1}{2}\mathrm{cc}(\phi)$ for $\phi \geq \widehat{\phi}_{0.50}$. particularly useful with these deviance based ccs. xx)

Ex. 7.10 *Median age for men and women in Roman Era Egypt.* In Story ii.11 we work with a rare dataset of lifetimes for 82 men and 59 women in Roman Era Egypt, a century B.C. For the present illustration of constructing confidence curves via log-likelihood profiling, take these lifetimes T_i to have arisen from Weibull distributions, with c.d.f.s $F_m(t) = 1 - \exp\{-(t/a_m)^{b_m}\}$ for men and $F_w(t) = 1 - \exp\{-(t/a_w)^{b_w}\}$ for women.

(a) For a Weibull, with parameters (a, b), show that the q level quantile becomes $\mu(q) = F^{-1}(q) = ac(q)^{1/b}$, with $c(q) = -\log(1-q)$. In particular, the median is $a(\log 2)^{1/b}$.

(b) For the men and women separately, write down the log-likelihood functions $\ell_m(a_m, b_m)$ and $\ell_w(a_w, b_w)$, and then carry out profiling to compute the deviance functions, for the medians μ_m and μ_w. Use Recipe Four, from Ex. 7.9, to construct the confidence curves, and make a version of Figure 7.3, left panel; also, read off 90 percent confidence intervals.

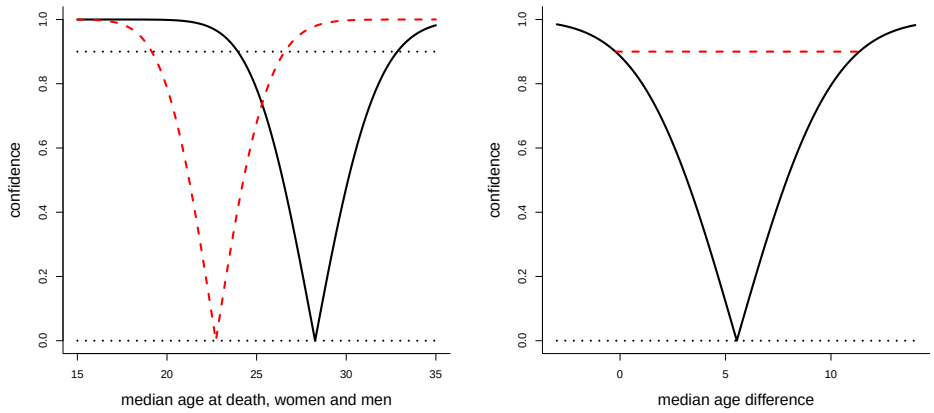


Figure 7.3: *Lifetimes in Roman Era Egypt, a century B.C.: Left panel: confidence curves for the median age μ_w and μ_m, for women and men, via log-likelihood-profiling from fitting separate Weibull distributions, with 90 percent confidence intervals* $[19.95, 26.56]$ *and* $[23.95, 32.84]$. *Right panel: confidence curve for the median age difference* $d = \mu_m - \mu_w$, *via log-likelihood profiling, and 90 percent confidence interval* $[-0.21, 11.34]$.

(c) It appears indeed that men had longer lives than women, in Egypt 2100 years ago; see Story ii.11 for more discussion and details. For this exercise, carry out log-likelihood profiling, for the median difference $d = \mu_m - \mu_w$, from the four-parameter model with the two Weibulls. Compute the deviance function and then the confidence curve $\mathrm{cc}(\cdot)$, leading to Figure 7.3, right panel. Read off a 90 percent interval for d.

(d) Constructing the confidence curves above involved the somewhat laborious computation of log-likelihood profiles. Compare these curves with the easier ones, computationally speaking, based on normal approximations.

Ex. 7.11 *Recipe Five: CDs via approximate pivots and t-bootstrapping.* Consider a parametric model $f(y, \theta)$ for data, with a model parameter of length say p. Suppose there is a focus parameter $\phi = g(\theta)$, estimated as $\widehat{\phi} = g(\widehat{\theta})$, for which we need a CD.

(a) Suppose first that $\widehat{\phi} - \phi$ has some distribution, say G_0, not depending on θ. Under this simple pivotal assumption for 'estimator minus estimand', show that

$$C(\phi) = \Pr_\theta(\widehat{\phi} \geq \widehat{\phi}_{\mathrm{obs}}) = 1 - G_0(\widehat{\phi}_{\mathrm{obs}} - \phi).$$

As indicated in e.g. Ex. 7.4, this can be computed even without knowing the form of G_0, through simulation of many realisations of $\widehat{\phi}^* - \widehat{\phi}$, where the $\widehat{\phi}^*$ is computed from a dataset drawn from the estimated distribution at $\widehat{\theta}$. Explain that this recipe works, even if G_0 is nonsymmetric and not centred well at zero. Simulating the distribution of $\widehat{\phi} - \phi$ at different points in the θ parameter space may also be helpful for checking the assumption needed for the $C(\phi)$ constructed here to be a clear CD.

(b) For a special and simpler case, if $\widehat{\phi} \sim \mathrm{N}(\phi, \kappa^2)$, to a good approximation, with known or well estimated κ, show that the general recipe above leads to

$$C(\phi) = \Pr_\theta(\phi + \kappa N \geq \widehat{\phi}_{\rm obs}) = \Pr(N \geq (\widehat{\phi} - \phi)/\kappa) = \Phi((\phi - \widehat{\phi})/\kappa),$$

and argue that this is really a CD. Note that this requires $Z = (\widehat{\phi} - \phi)/\kappa$ having distribution equal to or close to a standard normal, regardless of where θ is in its parameter space.

(c) Often the standard deviation in the normal approximation is not that sharply estimated from the available data. Consider a Student type ratio $t = (\widehat{\phi} - \phi)/\widehat{\kappa}$, with some appropriate scale estimator $\widehat{\kappa}$. Assume first that t really is a pivot, i.e. that its distribution G is independent or nearly independent of where θ is in its parameter space. Show that

$$C(\phi) = \Pr_\theta((\widehat{\phi} - \phi)/\widehat{\kappa} \geq (\widehat{\phi}_{\rm obs} - \phi)/\widehat{\kappa}_{\rm obs}) = 1 - G((\widehat{\phi}_{\rm obs} - \phi)/\widehat{\kappa}_{\rm obs})$$

is a CD. If G is not known, or too difficult to derive, use simulations, of $t^* = (\widehat{\phi}^* - \widehat{\phi}_{\rm obs})/\widehat{\kappa}^*$, via datasets simulated at position $\widehat{\theta}_{\rm obs}$. We call this a CD computed from t-bootstrapping.

(d) The previous recipe works well if $t = (\widehat{\phi} - \phi)/\widehat{\kappa}$ is close to pivotal, i.e. its distribution is nearly constant over the parameter region. In other cases we may take the t-bootstrapping argument one step further. Write for emphasis $\theta = (\phi, \gamma)$, perhaps in a reparametrisation, where ϕ is in focus and γ is of length $p-1$. The t has some distribution, depending on θ, and we write $\Pr_\theta(t \leq u) = G(u, \phi, \gamma)$. Show that

$$H(\phi, \gamma) = \Pr_{\phi,\gamma}((\widehat{\phi} - \phi)/\widehat{\kappa} \geq (\widehat{\phi}_{\rm obs} - \phi)/\widehat{\kappa}_{\rm obs}) = 1 - G((\widehat{\phi}_{\rm obs} - \phi)/\widehat{\kappa}_{\rm obs}, \phi, \gamma).$$

This is not necessarily a CD, in the strict sense, as this probability may depend not only on ϕ but also on aspects of γ. Often the distribution of t is approximately the same, though, in a neighbourhood around the true value. Argue that this leads to

$$C^*(\phi) = 1 - \widehat{G}((\widehat{\phi}_{\rm obs} - \phi)/\widehat{\kappa}_{\rm obs}, \phi) \quad \text{where } \widehat{G}(u, \phi) = G(u, \phi, \widehat{\gamma}).$$

Such an estimated distribution can be computed via bootstrapping, i.e. simulated datasets at position $\widehat{\theta}$ in the parameter space. With B such simulated datasets, leading to simulated values $\widehat{\theta}^*, \widehat{\phi}^*, \widehat{\kappa}^*$, and hence $t^* = (\widehat{\phi}^* - \widehat{\phi}_{\rm obs})/\widehat{\kappa}^*$.

Ex. 7.12 *Recipe Six: CDs in exponential families.* We have worked with the general exponential family in previous chapters, see Ex. 1.50. In particular we learned in Ex. 4.32 that there are uniformly optimal tests, for individual parameters in such models. The same holds in the present framework of CDs. Suppose data stem from a model of the form $f(y, a, b) = \exp\{aU(y) + b^{\rm t}V(y)\}h(y)$, with U one-dimensional and V of dimension say p. The optimal recipe for a is

$$C^*(a) = \Pr_a\{U(Y) \geq u_{\rm obs} \,|\, V(Y) = v_{\rm obs}\}.$$

This construction is actually optimal, in a power risk function sense we come back to in Ex. 7.29, but we can already start working with this definition and see how it applies in various situations.

(a) Verify from arguments in Ex. 4.32 that $C^*(a)$ indeed depends only on a, not on b.

(b) For an illustration, consider the pair of exponentials of Ex. 4.27. To avoid confusion with the parametrisation, use now $X \sim \text{Expo}(\theta)$, $Y \sim \text{Expo}(\theta+\delta)$, with sum $Z = X+Y$. Show that the joint density is indeed of exponential form, and that the recipe leads to

$$C^*(\delta) = \Pr_\delta(Y \le y_{\text{obs}} \,|\, Z = z_{\text{obs}}) = \frac{1-\exp(-\delta y_{\text{obs}})}{1-\exp(-\delta z_{\text{obs}})}.$$

Find the positive confidence pointmass at $\delta = 0$.

(c) Suppose there are m independent pairs of such exponentials, with $X_i \sim \text{Expo}(\theta_i)$, $Y_i \sim \text{Expo}(\theta_i+\delta)$, and sums $Z_i = X_i + Y_i$. We need a CD for the difference parameter δ. Show that the joint density of the $2m$ variables is on the exponential form, and that the resulting CD must be of the form

$$C^*(\delta) = \Pr_\delta(U \le u_{\text{obs}} \,|\, Z_1 = z_{1,\text{obs}}, \ldots, z_{m,\text{obs}}),$$

with $U = \sum_{i=1}^m Y_i$. There is no clear formula for this conditional distribution, but show that $Y_i \,|\, z_i$ has density $\delta\exp(-\delta y_i)/\{1-\exp(-\delta z_i)\}$ for $y_i \in [0, z_i]$. To show how the CD can be computed, via simulations, suppose as in Ex. 4.27 that the data are the three pairs $(0.927, 0.819), (1.479, 0.408), (3.780, 1.311)$. In that exercise we worked with the optimal test for $\delta = 0$ vs. $\delta > 0$, and needed only the null distribution of U given the three sums z_1, z_2, z_3, i.e. where $\delta = 0$. Now we need to tabulate this conditional distribution also for each $\delta > 0$, however.

(d)

Ex. 7.13 *Optimal CD for a bivariate model.* (xx spell out. we do get the natural answer $C(a) = \Pr_a(\widehat{a} \ge \widehat{a}_{\text{obs}} \,|\, \widehat{b})$. then multivariate. xx)

Ex. 7.14 *Bayesian posteriors as approximate CDs.* (xx to come here: Consider a setup with data y from a model with parameter $\theta = (\theta_1, \ldots, \theta_p)$, and with $\phi = \phi(\theta_1, \ldots, \theta_p)$ a focus parameter. A CD for ϕ has the property $\Pr_\theta\{C^{-1}(0.05, Y) \le \phi \le C^{-1}(0.95, Y)\} = 0.90$, etc., as with (7.1), thus delivering confidence intervals with the right coverage. This is also akin to how Bayesian posterior distributions are used. If a Bayesian prior for θ leads to a posterior for θ, and hence for a cumulative $B(\phi \,|\, y_{\text{obs}})$, then the Bayesian can read off $[B^{-1}(0.05, y_{\text{obs}}) \le \phi \le B^{-1}(0.95, y_{\text{obs}})]$. A question of interest and relevance also in Bayesian contexts is whether such intervals make sense also in the frequentist sense. point to Bernshteĭn–von Mises things in Ex. 6.23. the answer is 'ok' under such conditions, but not outside. xx)

(a) For a simple start example, consider $Y_1, \ldots, Y_n$ which given θ are i.i.d. from the Pois(θ), and with a prior $\theta \sim \text{Gam}(a, b)$; see Ex. (xx suitable exercise Ch 6 xx). Show that the posterior cumulative for θ becomes $B_n(\theta \,|\, \text{data}) = G(\theta, a+n\bar{y}_{\text{obs}}, b+n)$, in terms for the cumulative Gamma, and with $\bar{y}_{\text{obs}}$ the observed data average. Let $\widehat{\theta}_B$ and $\widehat{\tau}_B$ be the posterior mean and standard deviation. Assume now that data $Y_1, Y_2, \ldots$ come from the Pois(θ_0), for a certain θ_0. Show that

$$\begin{aligned} B_n(\theta_0 \,|\, Y_1, \ldots, Y_n) &= \Pr(\theta \le \theta_0 \,|\, Y_1, \ldots, Y_n) = G(\theta_0, a+n\bar{y}, b+n) \\ &\doteq \Phi((\theta_0 - \widehat{\theta}_B)/\widehat{\tau}_B) \to_d \Phi(\text{N}(0,1)) \sim \text{unif}, \end{aligned}$$

with probability 1.

(b) For a similar adventure, start with the Beta(a,b) prior for a binomial probability θ. Show that the posterior cumulative for θ becomes $B_n(\theta\,|\,\text{data}) = \text{Be}(\theta, a+y, b+n-y)$, in terms of the Beta cumulative. Assuming that Y_n is really from a binomial with some true θ_0, show that

$$\begin{aligned} B_n(\theta_0\,|\,Y_n) &= \Pr(\theta \le \theta_0\,|\,Y_n) = \text{Be}(\theta_0, a+Y_b, b+n-Y_n) \\ &\doteq \Phi((\theta_0-\widehat{\theta}_B)/\widehat{\tau}_B) \to_d \Phi(\text{N}(0,1)) \sim \text{unif}, \end{aligned}$$

with probability 1.

(c) Consider then the general parametric situation, supposing $Y_1,\ldots,Y_n$ being i.i.d. from the density $f(y,\theta)$, with $\phi = \phi(\theta)$ a one-dimensional focus parameter. Let θ_0 and $\phi_0 = \phi(\theta_0)$ denote the true parameters. As with Recipe Two, see Ex. 7.6, there is convergence to the standard normal of the standardised $\sqrt{n}(\widehat{\phi}_{\text{ml}}-\phi_0)/\widehat{\kappa}$, say, with the ML estimator for ϕ and $\widehat{\kappa}^2$ consistently estimating $\kappa^2 = c^{\text{t}}J^{-1}c$ (point back to delta method things in Ch2 xx). Now use the Bernshteĭn–von Mises theorem setup of Ex. 6.23 to explain that $\sqrt{n}(\phi-\widehat{\phi}_{\text{ml}})\,|\,\text{data}$ tends in distribution to $\text{N}(0,\kappa^2)$. The Bayesian posterior has a c.d.f. $B_n(\phi\,|\,\text{data})$. Show that at the true value,

$$\begin{aligned} B_n(\phi_0\,|\,\text{data}) &= \Pr\big(\sqrt{n}(\phi-\widehat{\phi}_{\text{ml}})/\widehat{\kappa} \le \sqrt{n}(\phi_0-\widehat{\phi}_{\text{ml}})/\widehat{\kappa}\,|\,\text{data}\big) \\ &= \Phi(\sqrt{n}(\phi_0-\widehat{\phi}_{\text{ml}})/\widehat{\kappa}) + \varepsilon_n, \end{aligned}$$

with $\varepsilon_n \to_{\text{pr}} 0$. Deduce that $B_n(\phi)$ is a CD in the large-sample sense. In this sense we may think of any sensible Bayesian posterior distribution, in regular parametric models, as Recipe Seven for creating a CD. (xx a little more here. xx)

Ex. 7.15 *Confidence distribution for the ratio of explained variation.* (xx to be polished. used in Story i.6. xx) For the classic linear regression model $Y_i = x_i^{\text{t}}\beta + \varepsilon_i$, with i.i.d. $\text{N}(0,\sigma^2)$ noise terms, theory was developed in Ex. 4.37 to estimate the fraction of the variance explained via the covariates. The statistic R^2 given there can be seen as an estimator of the explained variation ratio $\rho = \beta^{\text{t}}\Sigma_n\beta/(\beta^{\text{t}}\Sigma_n\beta+\sigma^2)$, with notation from that exercise. Here we construct full CDs for such parameters.

(a) (xx one covariate at a time. note that σ changes value and interpretaion, depending on whichi covariates are used in the linear regression equation. xx) $\rho = M_n\beta^2/(M_n\beta^2+\sigma^2)$, with $M_n = (1/n)\sum_{i=1}^n (x_i-\bar{x})^2$. then $F = nM_n\widehat{\beta}^2/\widehat{\sigma}^2 \sim F(1,m,n\rho/(1-\rho))$.

(b) (xx then general. xx) Explain that

$$C(\rho) = \Pr_\rho(\widehat{\lambda} \ge \widehat{\lambda}_{\text{obs}}) = \Pr_\rho(F \ge n\widehat{\lambda}_{\text{obs}}/p) = 1 - F(n\widehat{\lambda}_{\text{obs}}/p, p, m, n\rho/(1-\rho))$$

becomes a CD for ρ.

(c) (xx one more thing. xx)

CDs for quantiles

Ex. 7.16 *CDs for quantiles.* Let $Y_1, \ldots, Y_n$ be independent observations from a smooth density f, with c.d.f. F. How can we construct CDs for its quantiles, the $\mu_q = F^{-1}(q)$? We wish such a CD to be nonparametric, without further assumptions on the f. We go through the main ideas for the case of the median $\mu = F^{-1}(\frac{1}{2})$, before extending methods and results to a general quantile $q \in (0,1)$. (xx a bit more prose here; several methods; some better than others in terms of precision and coverage; we draw but briefly on density estimators from Ch. 13. nils needs to check Price and Bonett (2001, 2002). xx)

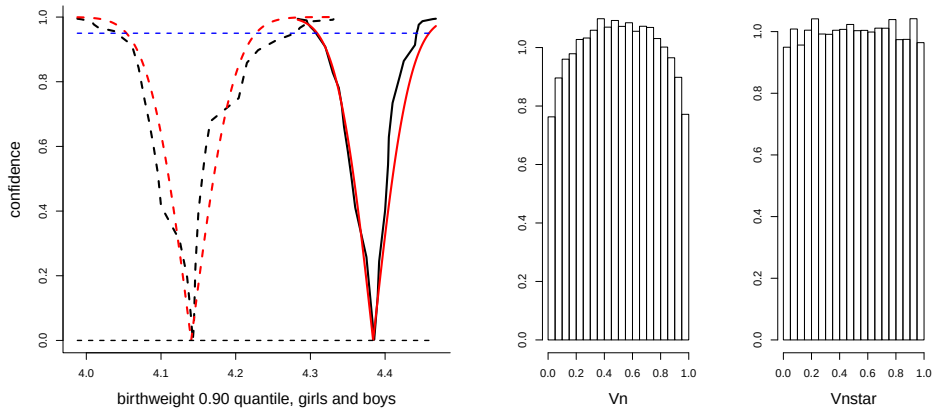


Figure 7.4: Left panel: For Ex. 7.16, confidence curves $\mathrm{cc}(\mu_{0.90})$, for the 0.90 quantiles of the birthweight distributions for girls (to the left) and boys (to the right). The black curves use the Beta method $\mathrm{cc}_n^*(\mu, \mathrm{data})$, with linear interpolation, whereas the slanted curves use the large-sample approximation. The former yields more accurate coverage than the latter. 95 percent intervals for the two 0.90 quaniles are indicated via the blue horizontal line. Right panel: For Ex. 7.17, for the case of f being standard normal, $n = 100$, $q = 0.50$: histograms of V_n and V_n^* based on $\mathrm{sim} = 10^5$ simulations. Also in other cases the V_n^* is much closer to uniformity than is V_n.

(a) It is not difficult to construct a first-order correct CD via large-sample results reached in Chapter 3, see in particular Ex. 3.18. With M_n the sample median, we have $\sqrt{n}(M_n - \mu) \to_d \mathrm{N}(0, \frac{1}{4}/f(\mu)^2)$. Show that as long as $\widehat{\tau}$ is a consistent estimator for $f(\mu)$, then $\sqrt{n}(M_n - \mu)/(\frac{1}{2}/\widehat{\tau}) \to_d \mathrm{N}(0,1)$, and that this leads to the approximate CD

approximate CD for quantiles

$$C_n(\mu, y) = \Phi\big(\sqrt{n}(\mu - M_n)/(\tfrac{1}{2}/\widehat{\tau})\big).$$

One of several choices is to take $\widehat{\tau} = \widehat{f}(M_n)$, with $\widehat{f}(y) = n^{-1}\sum_{i=1}^n h^{-1}K(h^{-1}(y_i - y))$ a kernel density estimator, with some kernel function K and bandwidth h. The best size for this fine-tuning parameter is of the type $h = c/n^{1/5}$, as seen in Chapter 13, and a classic rule of thumb which we typically might resort to here is to take $h = 1.059\,\widehat{\sigma}/n^{1/2}$,

with $\widehat{\sigma}$ the empirical standard deviation. Show that the confidence intervals from this CD take the form $M_n \pm z_0(\frac{1}{2}/\widehat{\tau})/\sqrt{n}$, with z_0 the relevant normal quantile, like 1.96 for intended 95 percent intervals.

(b) A different idea starts out as follows. For the ordered observations $Y_{(1)} < \cdots < Y_{(n)}$, show that

$$\Pr_f(\mu \le Y_{(i)}) = \Pr(\tfrac{1}{2} \le U_{(i)}) = 1 - \mathrm{Be}(\tfrac{1}{2}, i, n-i+1) \quad \text{for } i = 1, \ldots, n.$$

Here $U_{(i)} = F(Y_{(i)})$; these form an ordered sample from the standard uniform, and we saw in Ex. 3.18 that they have Beta distributions. The $\mathrm{Be}(x, a, b)$ is the c.d.f. of a $\mathrm{Beta}(a, b)$. Define a full CD for μ, say $C_n^*(\mu, \mathrm{data})$, via linear interpolation between the $C_n^*(y_{(i)}, \mathrm{data}) = 1 - \mathrm{Be}(\frac{1}{2}, i, n-i+1)$ points. This also yields a confidence curve $\mathrm{cc}_n^*(\mu, \mathrm{data}) = |1 - 2\,C_n^*(\mu, \mathrm{data})|$.

(c) Extend the two methods above, constructed there to deal with the median, to a general quantile $\mu_q = F^{-1}(q)$. For the first CD, use $\sqrt{n}(Q_{n,q} - \mu_q) \to_d \mathrm{N}(0, q(1-q)/f(\mu_q)^2)$, with $Q_{n,q} = F_n^{-1}(q)$ the empirical q quantile, and estimate $\tau_q = f(\mu_q)$ via $\widehat{f}(F_n^{-1}(q))$. For the second CD, show first that $\Pr_f(\mu_q \le Y_{(i)}) = 1 - \mathrm{Be}(q, i, n-i+1)$, and use linear interpolation:

$$C_n^*(\mu_q, \mathrm{data}) = \text{interpolation with } 1 - \mathrm{Be}(q, i, n-i+1) \text{ at } y_{(i)}, \tag{7.2}$$

for $i = 1, \ldots, n$. For μ_q inside $(y_{(i)}, y_{(i+1)})$, therefore, the CD value is interpolation between $1 - \mathrm{Be}(q, i, n-i+1)$ and $1 - \mathrm{Be}(\mu_1, i+1, n-i)$. We call this *the Beta method CD* for quantiles.

the Beta method CD for quantiles

(d) (xx a bit more. for birthweights oslo boys and girls, compute, display, and interpret the confidence curves for the 0.90 quantile, using both of the CD methods. Reproduce a version of Figure 7.4. point to Story i.5. xx)

Ex. 7.17 *CDs for quantiles: how well do they work?* In Ex. 7.16 we found two non-parametric CD recipes, for any quantile $\mu_q = F^{-1}(q)$. Here we investigate how well they work, in terms of actual coverage probabilities for confidence intervals. For the methods $C_n(\mu_q, \mathrm{data})$ and $C_n^*(\mu_q, \mathrm{data})$, define

$$V_n = C_n(\mu_{q,\mathrm{true}}, Y_1, \ldots, Y_n) \quad \text{and} \quad V_n^* = C_n^*(\mu_{q,\mathrm{true}}, Y_1, \ldots, Y_n),$$

with $Y_1, \ldots, Y_n$ drawn from the density in question. Accurate coverage, at all levels, means that the distribution of these two random CDs, at the true value, should be close to the uniform.

(a) To check the precision of these two CDs, carry out a simple simulation experiment. Take f equal to the standard normal, with $\mu_q = \Phi^{-1}(q)$ to be estimated with uncertainty; use $\widehat{\tau} = \widehat{f}(Q_{n,q})$ as above, with K the standard normal kernel and bandwidth $h = 1.059\,\widehat{\sigma}/n^{1/5}$ (which is optimal for the normal case), using the ordinary standard deviation from the data; and then simulate say $\mathrm{sim} = 10^5$ values of $V_n = C_n(\mu_{q,\mathrm{true}}, Y_1, \ldots, Y_n)$ and $V_n^* = C_n^*(\mu_{q,\mathrm{true}}, Y_1, \ldots, Y_n)$. Check, perhaps for $n = 50, 100, 500, 1000$, how close

the distributions of V_n and V_n^* are to the uniform. – For computing the C_n^*, and hence for executing that part of the simulation experiment, the `approx` algorithm of `R` is handy, carrying out linear approximation between the two values $1 - \text{Be}(\frac{1}{2}, i, n-i+1)$ and $1 - \text{Be}(\frac{1}{2}, i+1, n-i)$, for any μ inside the $[Y_{(i)}, Y_{(i+1)})]$ interval.

(b) Conduct a few similar simulation experiments, to see how close V_n and $V_m n^*$ are to uniformity, with different density f, quantile μ_q, sample size n.

(c) To assess 'closeness to uniformity' more accurately, use the monitoring processes of Ex. 3.9. For each of your simulation experiments, in addition to displaying histograms of V_n and V_n^*, compute and display the functions $Z_{\text{sim}}(t) = (\text{sim})^{1/2}\{G_{\text{sim}}(t) - t\}$ and $Z_{\text{sim}}^*(t) = (\text{sim})^{1/2}\{G_{\text{sim}}^*(t) - t\}$, where G_{sim} and G_{sim}^* are the empirical distribution functions of the V_n and V_n^*. Compute also $D_{\text{sim}} = \max_t |Z_{\text{sim}}(t)|$ and $D_{\text{sim}}^* = \max_t |Z_{\text{sim}}^*(t)|$. It will transpire (i) that V_n often is not particularly close to uniformity, unless n is rather large; (ii) that V_n^* is often so close to uniformity, even for moderate n and q near 0 or 1, that we cannot see that the distribution is not uniform, even with 10^5 simulated values.

(d) xx

Ex. 7.18 *Large-sample equivalence for two CDs for quantiles.* (xx to come. details for why the two CDs are large-sample equivalent. harder to show clearly that the 2nd is better than the 1st. nils thinks that it is, though, as of 12-August-2024. xx)

CDs in some nonregular setups (change title in a while)

Ex. 7.19 *Aboriginals and invaders in Watership Down.* Suppose a population of rabbits has been living for a long time on an island, in Hardy–Weinberg equilibrium $(p_0^2, 2p_0q_0, q_0^2)$, which means that pairs of alleles aa, Aa, AA occur with these frequencies, with $q_0 = 1 - p_0$. Suppose next that there's an invading population of new rabbits, with their separate Hardy–Weinberg equilibrium $(p, 2pq, q^2)$, with $q = 1 - p$. We assume that the two populations do not mix, but live on, on the same island, and that rabbitologists don't see the difference. One is interested in learning the fraction λ of newcomers (so the fraction of aboriginals is $1 - \lambda$).

(a) Explain that when one samples n rabbits independently, and find their allele pairs aa, Aa, AA, then these numbers (X, Y, Z) have a trinomial distribution with parameters

$$\text{pr}_1 = (1-\lambda)p_0^2 + \lambda p^2, \quad \text{pr}_2 = (1-\lambda)2p_0q_0 + \lambda 2pq, \quad \text{pr}_3 = (1-\lambda)q_0^2 + \lambda q^2.$$

Note that $\text{pr}_1 + \text{pr}_2 + \text{pr}_3 = 1$.

(b) For the case of $(X, Y, Z) = (118, 438, 444)$, and assuming not only $(p_0, q_0) = (0.25, 0.75)$ known, but also $(p, q) = (0.40, 0.60)$ known, find an estimate and construct a confidence curve $\text{cc}_1(\lambda)$, as with the black smooth Figure 7.5, left panel. Assume next, with the same counts (X, Y, Z), that the home population parameters $(p_0, q_0) = (0.25, 0.75)$ are known, but that the HW parameters $(p, q) = (p, 1-q)$ for the new population are unknown. Again, estimate λ and find a confidence curves $\text{cc}_2(\lambda)$, as for the red slanted curve of Figure 7.5, left panel. Comment on your findings. For your computer script,

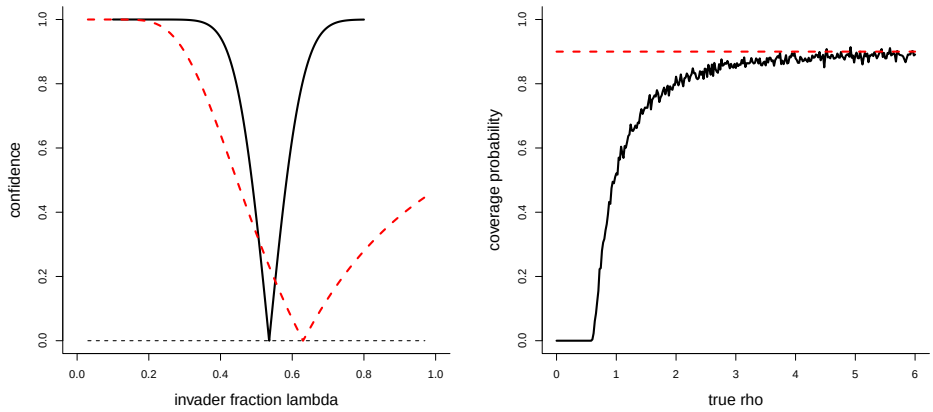


Figure 7.5: *Left panel: For Ex. 7.19, confidence curves for the unknown fraction λ of newcomers, after having counted $(X, Y, Z) = (118, 438, 444)$ of allele pairs aa, Aa, AA. The start population has HW parameters $(p_0, q_0) = (0.25, 0.75)$. (i) The black smooth $\mathrm{cc}_1(\lambda)$ is computed using the knowledge that the new population has HW parameters $(p, q) = (0.40, 0.60)$. (ii) The red slanted $\mathrm{cc}_2(\lambda)$ is computed using only knowledge about (p_0, q_0), i.e. both p, $q = 1-p$, and λ are unknown. Right panel: For Ex. 7.20, as a function of the true ρ, for dimension $p = 3$, the figure shows the actual coverage probability of the Bayesian 90 percent credibility interval, based on the posterior stemming from a flat prior for the θ.*

play a bit with different sample sizes, and with different degrees of difference between (p_0, q_0) and (p, q).

(c) Explain why it is not possible to estimate all (p_0, p, λ) from (X, Y, Z).

Ex. 7.20 *The length problem.* (xx might make a Satellite Collision Story, based on Cunen et al. (2020b). contrasting CD with Bayes. xx) There are several situations, of varying degrees of complexity, where the heart of the matter is, or can be transformed to, the following: with Y having the $\mathrm{N}_p(\theta, \Sigma)$ distribution, with unknown mean vector and known or partly known variance matrix, reach inference for the length $\rho = \|\theta\| = (\theta_1^2 + \cdots + \theta_p^2)^{1/2}$. See e.g. Cunen et al. (2020b) for an application involving the computation and real-time monitoring of the probability that two satellites will collide.

(a) Take first $\Sigma = I_p$, so $y \sim \mathrm{N}_p(\theta, I_p)$, which means independent $Y_i \sim \mathrm{N}(\theta_i, 1)$ for $i = 1, \ldots, p$. Show that the ML estimator of ρ is $\widehat{\rho} = \|Y\|$. Show also that $\widehat{\rho}^2 \sim \chi^2_p(\rho^2)$, the noncentral chi-squared.

(b) Deduce that $\widehat{\rho}^2$ is overshooting its target ρ^2, with mean and variance $p + \rho^2$ and $2p + 4\rho^2$. Find also an expression for $\mathrm{E}\,\widehat{\rho}$, and show that it overshoots ρ.

(c) Show that the natural CD becomes $C(\rho, y) = 1 - \Gamma_p(\widehat{\rho}^2, \rho^2)$.

(d) A typical Bayesian analysis would start with a flat prior for $\theta_1, \ldots, \theta_p$ (xx calibrate and xref Ch 5 for this detail xx). Show that $\theta \,|\, y \sim \mathrm{N}_p(y, I)$, and that this entails $\rho^2 \,|\, y \sim \chi^2_p(\widehat{\rho}^2)$.

(e) For $p = 5$ and $\widehat{\rho} = 7.77$, compute and draw both the CD and the Bayesian posterior distribution,

$$C(\rho, y) = 1 - \Gamma_p(\widehat{\rho}^2, \rho^2) \quad \text{and} \quad B(\rho, y) = \Gamma_p(\rho^2, \widehat{\rho}^2).$$

Comment on what you find.

(f) (xx simulate to illustrate that the CD by construction works, producing confidence intervals with the correct coverage; $U_C = C(\rho_0, Y) \sim$ unif when data stem from the model, at position ρ_0. show however that the Bayesian posterior distribution here risks being very far from producing intervals with the right coverage; $U_B = B(\rho_0, Y)$ is very far from being uniform. point to Figure 7.5, right panel, for the too low coverage probability of the Bayesian 90 percent credibility interval. the CD based intervals have exact coverage. link to Bernshteĭn–von Mises things in Ch. 6; here we're outside BvM terrain. more on why and how. xx)

(g) Generalise the above to the case where $Y \sim \mathrm{N}_p(\theta, \sigma^2 I_p)$.

(h) More generally, with $Y_1, \ldots, Y_n$ being i.i.d. from the $\mathrm{N}_p(\theta, \sigma^2 I_p)$, with σ known, show first that $\bar{Y} \sim \mathrm{N}_p(\theta, (\sigma^2/n) I_p)$. Then show that $\widehat{\rho} = \|\bar{y}\|$ is the ML estimator, with distribution given by $n\widehat{\rho}^2/\sigma^2 \sim \chi^2_p(n\rho^2/\sigma^2)$. On the Bayesian side, show that a flat prior for θ leads to $\theta \,|\, \text{data} \sim \mathrm{N}_p(y, (\sigma^2/n) I_p)$. Show that these statements lead to these generalisations of the above

$$\begin{aligned} C_n(\rho, y) &= 1 - \Gamma_p(n\widehat{\rho}^2/\sigma^2, n\rho^2/\sigma^2), \\ B_n(\rho \,|\, y) &= \Gamma_p(n\rho^2/\sigma^2, (n/\sigma^2)\widehat{\rho}^2). \end{aligned}$$

(i) (xx a bit more, regarding BvM, which holds for fixed p and ρ, with growing n. but misleading picture for finite n. do something to see interplay with n and p. xx)

(j) (xx also, briefly, to the case of $Y \sim \mathrm{N}_p(\theta, \sigma^2 I_p)$, with σ estimated via an independent $\widehat{\sigma}^2 \sim \sigma^2 \chi^2_m/m$. xx)

Ex. 7.21 *Ratio of normal means.* (xx edit and clean. about two exercises here on this. mention Fieller name. start with $x_0 = -a/b$, the point at which a regression type equation $a + bx = 0$. then application to bioassay or similar. xx)

(a) Consider the prototype setup for such questions, where $\widehat{a} \sim \mathrm{N}(a, 1)$ and $\widehat{b} \sim \mathrm{N}(b, 1)$ are independent. Show first that the log-likelihood is a simple $\ell(a, b) = -\frac{1}{2} Q(a, b)$, with $Q(a, b) = (a - \widehat{a})^2 + (b - \widehat{b})^2$, and find a formula for $\ell_{\mathrm{prof}}(x_0) = -\frac{1}{2} Q_{\mathrm{prof}}(x_0)$, where $Q_{\mathrm{prof}}(x_0) = \min\{Q(a, b) \colon x_0 = -a/b\}$.

(b) Show that $(\widehat{a} + \widehat{b} x_0)/(1 + x_0^2) \sim \chi^2_1$, at the true x_0, and hence that

$$\mathrm{cc}(x_0) = \Gamma_1((\widehat{a} + \widehat{b} x_0)/(1 + x_0^2))$$

is a clear confidence curve for x_0. (xx illustrate, with the mildly peculiar confidence regions. find max confidence level. more. xx)

(c) (xx with $\widehat{\sigma}$ on top. with dependence. things fine as long as $(\widehat{a}, \widehat{b})$ is binormal. xx)

(d) (xx bioassay. xx)

Ex. 7.22 *CDs and posterior distributions with boundary constraints.* Here we learn about construction of CDs when there is a boundary condition on the focus parameter. This is sometimes an easy task, involving a natural positive post-data probability on the boundary point. We also compare with Bayesian procedures. Matters may of course be extended and generalised in several directions here, but for simplicity and conciseness we study a very simple prototype situation: y is $\mathrm{N}(\theta, 1)$, and $\theta \geq 0$ a priori.

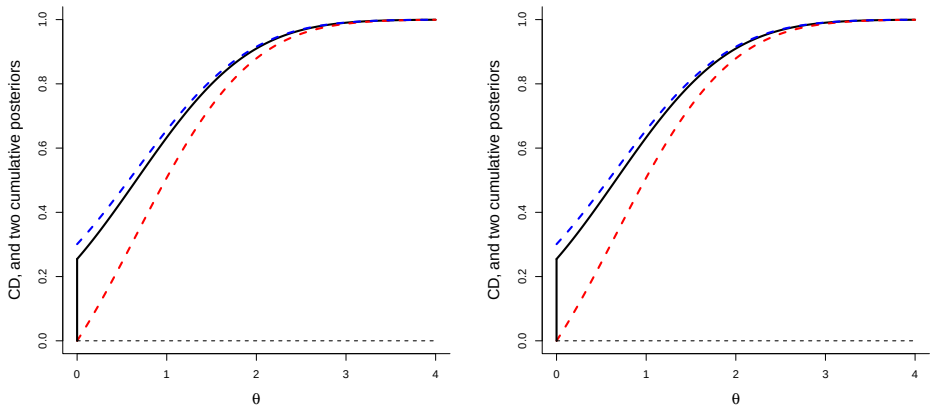


Figure 7.6: Left panel: For Ex. 7.22, with $y_{\mathrm{obs}} = 0.66$ for the $\mathrm{N}(\theta, 1)$ model, the black curve is the natural CD, with positive point mass 0.255 at zero. The red and the blue curves are Bayesian posterior distributions, for the flat prior on the halfline, and for the mixture prior with $\frac{1}{2}$ at zero and $\frac{1}{2}$ flat on the halfline, respectively. Right panel: For Ex. 7.23, (xx to come xx).

(a) Before we come to the parameter constraint, we deal with the more normal situation where there is no a priori constraint. The classical CD is then $C(\theta, y) = \Phi(\theta - y)$. Show that the Bayesian starting with a flat prior for θ finds the posterior distribution $\theta \,|\, y \sim \mathrm{N}(y, 1)$, with cumulative $B(\theta \,|\, y) = \Phi(\theta - y)$, i.e. identical to the canonical CD. – The point below will partly be that this is *not* the same for the constrained problem.

(b) For the remaining points here, assume indeed that $\theta \geq 0$ a priori. Argue that the canonical CD should be $C(\theta, y) = \Phi(\theta - y)$ for $\theta \geq 0$. Its point mass at zero is $\Phi(-y)$. Graph the CD, for the three cases y_{obs} equal to $-0.22, 0.66, 1.99$.

(c) One Bayesian approach in this situation, where $\theta \geq 0$ a priori, is to let θ be flat on $[0, \infty)$. Show that then

$$\theta \,|\, y \sim \frac{\phi(\theta - y)}{\int_0^\infty \phi(\theta - y)\,\mathrm{d}\theta} = \frac{\phi(\theta - y)}{\Phi(y)} \quad \text{for } \theta \geq 0,$$

and that the cumulative posterior distribution becomes

$$B(\theta \,|\, y) = \frac{\Phi(\theta - y) - \Phi(-y)}{1 - \Phi(-y)} = \frac{\Phi(\theta - y) - \Phi(-y)}{\Phi(y)} \quad \text{for } \theta \geq 0.$$

For the three cases of $y_{\rm obs}$ given above, graph the CD along with the Bayesian $B(\theta \,|\, y_{\rm obs})$, and comment on what you find.

(d) In general terms, for the case of $y \,|\, \theta \sim {\rm N}(\theta, 1)$, let θ have the mixture prior distribution $p_0\pi_0 + p_1\pi_1$, with the sub-priors π_0 and π_1 having their individual posteriors $\pi_0(\theta \,|\, y)$ and $\pi_1(\theta \,|\, y)$. Show that the posterior has a natural mixture form,

$$\theta \,|\, y \sim p_0^*(y)\pi_0(\theta \,|\, y) + p_1^*(y)\pi_1(\theta \,|\, y),$$

where

$$p_0(y) = \frac{p_0 f_0(y)}{p_0 f_0(y) + p_1 f_1(y)} \quad \text{and} \quad p_1(y) = \frac{p_1 f_1(y)}{p_0 f_0(y) + p_1 f_1(y)},$$

and with $f_0(y) = \int \phi(y-\theta)\pi_0(\theta)\,{\rm d}\theta$ and $f_1(y) = \int \phi(y-\theta)\pi_1(\theta)\,{\rm d}\theta$ the marginal densities following from the two priors. (This structure generalises to general mixture priors in general models, though that does not concern us just now.)

(e) For the prior $p_0\pi_0 + p_1\pi_1$, with π_0 a unit pointmass at zero and π_1 a flat prior on the halfline, show that $f_0(y) = \phi(y)$ and $f_1(y) = \Phi(y)$. With a 50-50 mixture, show hence that

$$p_0(y) = \frac{\phi(y)}{\phi(y) + \Phi(y)} \quad \text{and} \quad p_1(y) = \frac{\Phi(y)}{\phi(y) + \Phi(y)}.$$

Draw curves of these two posterior probabilities, one for the zero-point and the other for the halfline-based part, as y goes from say -5 to 5. Show that the posterior cumulative distribution becomes $B^*(\theta \,|\, y) = p_0(y) + p_1(y)B(\theta \,|\, y)$ for $\theta \geq 0$. In particular, there is a pointmass $p_0(y)$ at zero. Construct a version of Figure 7.6.

(f) Show that there is no choice of (p_0, p_1) which makes the Bayesian cumulative posterior $B^*(\theta \,|\, y)$ agree with the CD $C(\theta, y)$. Devise a method for selection (p_0, p_1) such that the distance between $B^*(\theta \,|\, y)$ and $C(\theta, y)$ is small, for a relevant range of θ and possible observed $y_{\rm obs}$.

(g) Generalise the formulae above to the case of $y_1, \ldots, y_n$ i.i.d. ${\rm N}(\theta, \sigma^2)$, with known σ.

Ex. 7.23 *CDs for regression parameters with boundary constraints.* (xx to come here: more on boundary parameters, now in simple regression models. pointer to Story iii.10. xx) (xx the point we wish to convey is that the Tore-Sims phenomenon is a general one, easier to understand and analyse in simpler models, separately. so we can have separate points for a model like $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$, where one has prior knowledge $\beta_2 \geq 0$. There is a clear and exact CD for β_2, of the type $C(\beta_2) = G_{\rm df}((\beta_2 - \widehat{\beta}_2)/\widehat{\kappa}_2)$ for $\beta_2 \geq 0$, with a pointmass $G_{\rm df}(-\widehat{\beta}_2/\widehat{\kappa}_2)$ at zero. the Bayesian with flat priors on $\beta_0, \beta_1, \log\sigma$ and a flat prior on $(0, \infty)$ for β_2, a la Sims, will not be able to detect that $\beta_2 = 0$; there's a clear discrepancy between the CD and the Bayesian posterior for that parameter. xx)

Ex. 7.24 *CDs in the truncated exponential model.* Here we consider a model sometimes called the truncated exponential model. We start with its simplest form, with data $Y_1, \ldots, Y_n$ i.i.d. from the density $\exp\{-(y-a)\}$ for $y \geq a$. The a is the unknown start point for the distribution.

(a) Show that the ML estimator is equal to $U_n = \min_{i \leq n} Y_i$, the smallest data point. Show that $n(U_n - a)$ has a unit exponential distribution. Build from this a natural CD for a.

(b) Construct a predictive CD for the next sample point Y_{n+1}. Illustrate by computing and displaying the confidence curve for the text sample point, after having observed the six data points `3.735, 3.338, 10.634, 3.839, 5.667, 5.808`.

(c) Then consider the more realistic two-parameter version of the model, with density

$$f(y_i, a, b) = (1/b)\exp\{-(y_i - a)/b\} \quad \text{for } y_i \geq a,$$

with a being the unknown start-point and b a scale parameter. Show that the ML estimators become $\widehat{a} = U_n$ and $\widehat{b} = (1/n)\sum_{i=1}^n (Y_i - U_n)$, again with U_n being the smallest observation.

(d) Construct accurate CDs and confidence curves for a, for b, and for the next datapoint Y_{n+1}. If some of your formulae cannot be given very explicit mathematical forms, this is ok, as long as numerical solutions can be found via numerical integration or simulation. Give approximations for these CDs for large sample sizes n.

(e) Ignoring these large-sample approximations, compute and display confidence curves for a, b, Y_{n+1} with the simple $n = 6$ dataset above.

Ex. 7.25 *CD inference for the exponential rate, with censored data.* The lifelength distribution for a certain type of technical components is considered exponential, i.e. with density $\theta\exp(-\theta t)$ for $t > 0$, on a priori grounds. To arrive at a point estimate and a confidence curve for θ, the firm producing these components sets in motion the simple experiment where n such items are set to work, under controlled natural conditions. One cannot wait until all components have died out, however, and the firm needs to report what can be said about the lifelength distribution, via θ, a certain time t_0 after project start.

(a) With data of the form observed t_i for the N of the items which have died within t_0, and the information $t_i > t_0$ for the $n - N$ which are still alive and well, show that the combined likelihood function may be expressed as

$$\theta^N \exp\Big[-\theta\Big\{\sum_{t_i \leq t_0} t_i + (n-N)t_0\Big\}\Big].$$

(b) Show that the ML estimator is

$$\widehat{\theta} = N/R = N/\Big\{\sum_{t_i \leq t_0} t_i + (n-N)t_0\Big\}.$$

With increasing sample size, and fixed t_0, find expressions for the probability limits of N/n and R/n, and show that $\widehat{\theta}$ is consistent.

(c) Show in fact that there is a limiting normal distribution here, with $\sqrt{n}(\widehat{\theta}-\theta) \to_d \mathrm{N}(0, \tau(t_0,\theta)^2)$, and attempt to find an explicit (though not necessarily quick and simple) formula for the limit variance.

(d) Explain why the construction $C_n(\theta) = \mathrm{Pr}_\theta(\widehat{\theta} \geq \widehat{\theta}_{\mathrm{obs}})$ yields a CD, and also how it can be computed in practice.

(e) Suppose the experiment described involves $n = 20$ such items, and that the lifelengths for the $N = 11$ of these that conk out before the deadline of $t_0 = 2.00$ years are
`0.528 0.743 0.869 1.180 0.602 0.133 0.327 1.115 0.117 0.208 1.808`
Compute and display perhaps as many as three (exact or approximate) confidence curves for θ, for this little experiment: the one described in (c); one based on the normal approximation to the distribution of the ML estimator; and a t-bootstrap based version. Comment on your findings.

Ex. 7.26 *Estimating n based on observing the first r.* Suppose $Y_1, \ldots, Y_n$ are i.i.d., from some known distribution with density f and cumulative F, but that one only observes the first r order statistics, $Y_{(1)} < \cdots < Y_{(r)}$. Can we estimate n? Such nonstandard problems turn up in various context, from estimating the size of a vocabulary to the number of unseen species. In this exercise we consider the special case of the unit exponential distribution, where the Y_i can be seen as waiting times, so the question may be phrased as how long time do we need to wait, until we've seen all items, when we have used a certain time to observe the first r.

(a) Let then $Y_1, \ldots, Y_n$ be i.i.d. from the unit exponential, and assume $Y_{(1)} < \cdots < Y_{(r)}$ are observed, with unknown n. Observing these r first data points is equivalent to observing the spacings $D_1 = Y_{(1)}$, $D_2 = Y_{(2)} - Y_{(1)}$, up to $D_r = Y_{(r)} - Y_{(r-1)}$. Use Ex. 1.13 to show that the joint distribution of these f first spacings may be written

$$\begin{aligned} g_r(d_1,\ldots,d_r) &= n(n-1)\cdots(n-r+1) \\ &\quad \exp\bigl[-\{n(d_1+\cdots+d_r) - d_2 - 2d_3 - \cdots (r-1)d_r\}\bigr], \end{aligned}$$

and deduce from this that $Y_{(r)}$ is sufficient for n.

(b) With $F(x) = 1-\exp(-x)$, show that $F(Y_{(r)})$ has a Beta distribution with parameters $(r, n-r+1)$.

(c) Show that the optimal CD for n, based on having observed the smallest r datapoints, is

$$C_r(n) = \mathrm{Pr}_n(Y_{(r)} \leq Y_{(r),\mathrm{obs}}) = \mathrm{Be}(F(Y_{(r),\mathrm{obs}}), r, n-r+1).$$

(d) (xx an example or two. suppose $Y_{(r),\mathrm{obs}} = 0.348$ with $r = 33$. estimate n. see nils com87a or thereabouts. give normal approximation. but these are not good for r/n close to zero or one. can we characterise ML estimator. xx)

Ex. 7.27 *(xx another discrete model thing. xx)* (xx to come. xx)

Risk functions for CDs

Ex. 7.28 *Risk functions for CDs.* This exercise looks into risk functions for and hence comparisons between CDs, in simple prototype situations where calculations are easier than for general cases. We start out with $Y_1, \ldots, Y_n$ being i.i.d. from the $\mathrm{N}(\theta, 1)$ model. For a CD $C_n(\theta, y)$, where y denotes the full dataset, the risk function used is

$$\mathrm{risk}_n(C_n, \theta) = \mathrm{E}_\theta \int (\theta' - \theta)^2 \, \mathrm{d}C_n(\theta', Y) = \mathrm{E}_\theta(\theta_{\mathrm{cd}} - \theta)^2,$$

where θ_{cd} is the result of a two-stage random process: data Y lead to the CD $C_n(\theta, Y)$, and then θ_{cd} is drawn from this distribution.

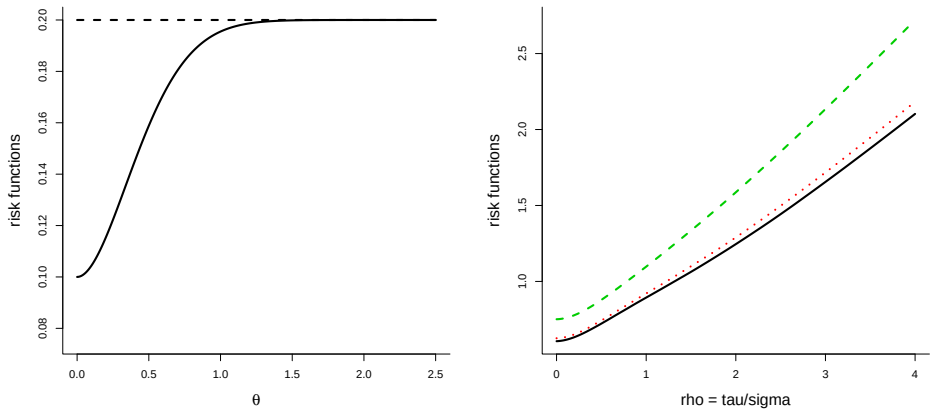


Figure 7.7: *Left panel: For Ex. 7.28, risk function for the CD, in the setup with $Y_1, \ldots, Y_n$ being i.i.d. from $\mathrm{N}(\theta, 1)$, with $n = 10$, but with the restriction $\theta \geq 0$. It starts out at $1/n$ and then grows to the $2/n$ risk of the unrestricted case as θ grows. Right panel: For Ex. 7.30, for the variance component model, with $p = 4$ and $\sigma = 1$, risk functions $r(C, \tau)$ for three CDs for τ. The one based on $Z = \sum_{i=1}^p Y_i^2$ is best, closely followed by the one using $A = \sum_{i=1}^p |Y_i|$, whereas the one using the range $R = \max Y_i - \min Y_i$ does worse.*

(a) Show that the natural CD based on the observed sample mean $\bar{y}_{\mathrm{obs}} = n^{-1} \sum_{i=1}^n y_i$ is $C_n(\theta, y_{\mathrm{obs}}) = \Phi(\sqrt{n}(\theta - \bar{y}_{\mathrm{obs}}))$. Prove that its risk function is $\mathrm{risk}_n(C_n, \theta) = 2/n$.

(b) More generally, assume θ^* is some unbiased estimator of θ, with finite variance τ_n^2, with the property that $\widehat{\theta}^* - \theta$ has a distribution H_n symmetric around zero. Show that the associated CD becomes $C_n^*(\theta, y_{\mathrm{obs}}) = H_n(\theta - \theta^*_{\mathrm{obs}})$, and show that its risk function becomes $2\tau_n^2$. The case of $\bar{Y}$ corresponds to $2/n$. Find the risk function for the case of the median based CD, with say $n = 10$, as for Figure 7.7, left panel.

(c) (xx fix this: Relate the above results to the optimality theorem for CDs, in certain situations, from CLP's Chapter 5. xx)

(d) Now we change gears a bit, by putting the a priori assumption $\theta \geq 0$ on the table. Show that the ML estimator becomes $\widehat{\theta} = \max(0, \bar{y})$, i.e. the sample mean truncated, if necessary, to zero. Argue that this leads to the natural CD

$$\widetilde{C}_n(\theta, y) = \Phi(\sqrt{n}(\theta - \bar{y}_{\rm obs})), \quad \text{for } \theta \geq 0,$$

in particular having a positive point-mass at zero.

(e) With $\theta_{\rm cd}$ drawn from this CD, for given data, show that it may be expressed as $\max(0, \bar{y} + N/\sqrt{n})$, with N a standard normal. Show next that in the two-stage setup, with random data followed by $\theta_{\rm cd}$ drawn from the $\widetilde{C}_n$ CD, we have $\theta_{\rm cd} - \theta = \max(0, \theta + (N + N')/\sqrt{n}) - \theta$, with N' another and independent standard normal. Use this to show that the $\text{risk}_n(\widetilde{C}_n, \theta)$ can be expressed as

$$\begin{aligned} r_n &= \int \big[\{\max(0, \theta + (2/n)^{1/2}x)\} - \theta\big]^2 \phi(x)\,{\rm d}x \\ &= \theta^2\Phi(-(\tfrac{1}{2}n)^{1/2}\theta) + (2/n)\{-(\tfrac{1}{2}n)^{1/2}\theta\phi((\tfrac{1}{2}n)^{1/2}\theta) + 1 - \Phi(-(\tfrac{1}{2}n)^{1/2}\theta)\}. \end{aligned}$$

Compute and display the risk functions for $\widetilde{C}_n$ and C_n, for say $n = 10$, constructing a version of Figure 7.7, left panel. Comment on what we learn from this.

(f) (xx not fully sure about this one. xx) There are various other estimators and CDs worth considering in this $\theta \geq 0$ setting. To simplify matters, take $n = 1$, and consider the Bayes estimator $\widehat{\theta}_B$, the conditional mean of $\theta \,|\, y$, with a flat prior on $(0, \infty)$. Show in fact that $\widehat{\theta}_B = y + \phi(y)/\Phi(y)$, and verify that this is positive even when y is negative. Work out an expression for the naturally associated CD $C_B(\theta) = \Pr_\theta(\widehat{\theta}_B \geq \widehat{\theta}_{B,\rm obs})$, and comment.

Ex. 7.29 *Optimal CDs in exponential family models.* (xx relate to Ex. 7.12. spell out NP things to demonstrate optimality. xx)

Ex. 7.30 *Risk functions for three CDs in a variance components model.* Consider the simple variance component model with independent observations $y_i \sim \text{N}(0, \sigma^2 + \tau^2)$ for $i = 1, \ldots, p$, with σ known and τ the unknown parameter of interest; see Schweder and Hjort (2016, Example 4.1 and Exercise 5.8). The aim here is first to construct CDs based on (i) $Z = \sum_{i=1}^p y_i^2$, (ii) $A = \sum_{i=1}^p |y_i|$, and (iii) the range $R = \max y_i - \min y_i$; and then to compute and compare their risk functions. These are defined as

$$\text{risk}(C, \tau) = \text{E}_\tau\, |\tau_{\rm cd} - \tau| = \text{E}_\tau \int |\tau_{\rm cd} - \tau|\,{\rm d}C(\tau_{\rm cd}, Y),$$

with $\tau_{\rm cd}$ a random draw from the $C(\tau, Y)$ distribution, and with Y itself denoting a dataset drawn from the distribution indexed by τ.

(a) Show that the natural CDs, based on Z, A, R respectively, are

$$\begin{aligned} C_Z(\tau, \text{data}) &= 1 - \Gamma_p(Z_{\rm obs}/(\sigma^2 + \tau^2)), \\ C_A(\tau, \text{data}) &= 1 - G_p(A_{\rm obs}/(\sigma^2 + \tau^2)^{1/2}), \\ C_R(\tau, \text{data}) &= 1 - H_p(R_{\rm obs}/(\sigma^2 + \tau^2)^{1/2}). \end{aligned}$$

Here Γ_p is the cumulative distribution function of $Z_0 = \sum_{i=1}^p N_i^2$, with the N_i being i.i.d. and standard normal, which means $Z_0 \sim \chi^2_p$. Similarly, G_p and H_p are the cumulative distribution functions of $A_0 = \sum_{i=1}^p |N_i|$ and of $R_0 = \max N_i - \min N_i$, respectively.

(b) Show that a random draw $\tau_{\rm cd}$ from the first of these, i.e. C_Z, for a given dataset, can be represented as $\tau_{\rm cd} = (Z_{\rm obs}/K - \sigma^2)_+^{1/2}$, where x_+ is notation for the truncated-to-zero quantity $\max(x, 0)$, and where $K \sim \chi^2_p$. In the situation where data are random, from the model at position τ, deduce that

$$\tau_{\rm cd} - \tau = \{(\sigma^2 + \tau^2)K_0/K - \sigma^2\}_+^{1/2} - \tau = \sigma\big[\{(1+\rho^2)K_0/K - 1\}_+^{1/2} - \rho\big],$$

where $\rho = \tau/\sigma$, and K_0, K are two independent draws from the χ^2_p. In other words, $F = K_0/K \sim F_{p,p}$, a F distribution with degrees of freedom (p, p). Use this to compute the risk function $\text{risk}(C_Z, \tau)$, for $p = 4$ and $\sigma = 1$; this is the lowest of the three risk functions of Figure 7.7, right panel.

(c) Then consider the C_A option. Show that a random draw from an observed $C_A(\tau, \text{data})$ can be written $\tau_{\rm cd} = \{(A_{\rm obs}/A)^2 - \sigma^2\}_+^{1/2}$. Deduce that for random data behind the CD, we have the representation

$$\tau_{\rm cd} - \tau = \{(\sigma^2 + \tau^2)(A_0/A)^2 - \sigma^2\}_+^{1/2} - \tau = \sigma\big[\{(1+\rho^2)(A_0/A)^2 - 1\}_+^{1/2} - \rho\big],$$

with A and A_0 two independent draws from the G_p distribution. Use this to compute $\text{risk}(C_A, \tau)$. There is no simple expression for the density of A_0/A, so use simulation.

(d) Carry out similar analysis for the third CD, based on the range R. Construct a version of Figure 7.7, right panel.

(e) Use your programme to explore the three risk functions for other values of p.

Meta-analysis, combining confidence distributions

(xx need more sorting and polish. xx)

Ex. 7.31 *CD and cc for binomial probabilities.* Suppose y is observed from a binomial (n, θ). The task is to construct a CD and a cc for θ.

(a) Show that the standard normal approximations for y (xx give pointer here to large-sample chapter) lead to

$$C_a(\theta, y) = \Phi\Big(\frac{n\theta - y}{\{n\theta(1-\theta)\}^{1/2}}\Big) \quad \text{and} \quad C_b(\theta, y) = \Phi\Big(\frac{\sqrt{n}(\theta - \widehat\theta)}{\{\widehat\theta(1-\widehat\theta)\}^{1/2}}\Big),$$

with $\widehat\theta = y/n$ the standard estimator for θ.

(b) The recipe of Ex. 7.5 does not quite work here since y has a discrete distribution. This invites the half-correction method

$$C(\theta, y) = \Pr_\theta(y > y_{\rm obs}) + \tfrac12 \Pr_\theta(y = y_{\rm obs}).$$

For say $n = 20$ and $y = 12$, compute and display this CD, along with (i) the same CD but without the half-correction, (ii) the two simple normal approximations above. Try other combinations of (n, y), and demonstrate that they are approximately equal for moderate to large n.

(c) To investigate the basic CD property, take $n = 20$ and $\theta_{\text{true}} = 0.33$. Simulate a large number of $C(\theta_0, y)$, $C_a(\theta_0, y)$, $C_b(\theta_0, y)$, to check for their approximate uniform distribution. Try other values of (n, θ_0), and summarise your findings.

Ex. 7.32 *CD and cc for comparing binomials.* In Ex. 7.31 we learn how to construct CDs for separate binomial parameters. Consider now the 2×2 table setup with two binomials, as with Ex. 4.30, say $y_0 \sim \text{binom}(m_0, p_0)$ and $y_1 \sim \text{binom}(m_1, p_1)$. How do we reach precise inference for the extent to which p_0 and p_1 differ?

(a) We first use the logistic transform $p_0 = H(\theta_0)$ and $p_1 = H(\theta_0 + \gamma)$, with $H(u) = \exp(u)/\{1 + \exp(u)\}$. Show that

$$\gamma = \log \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \log \frac{p_1}{1-p_1} - \log \frac{p_0}{1-p_0},$$

the log-odds difference. Write up the likelihood function for the observed (Y_0, Y_1) to deduce (xx via the optimal CD exercise xx) that the optimal CD for γ takes the form

$$C(\gamma) = \Pr_\gamma(Y_1 > y_{1,\text{obs}} \,|\, Z = z_{\text{obs}}) + \tfrac{1}{2}\Pr_\gamma(Y_1 = y_{1,\text{obs}} \,|\, Z = z_{\text{obs}}),$$

with $Z = Y_0 + Y_1$. The conditional distribution in question is the eccentric hypergeometric, found in Ex. 4.30. (xx do simple example here. this CD is used in both Stories i.1 and i.10. we use Ex. 4.30. xx)

Ex. 7.33 *Meta-analysis for Lidocain data.* The following data table is from Normand (1999), and pertains to prophylactic use of lidocaine after a heart attack. The aim is to evaluate mortality from prophylactic use of lidocaine in acute myocardial infarction. We view the data here as pairs of binomials, with $y_{1,i} \sim \text{binom}(m_{i,1}, p_{1,i})$ and $y_{1,0} \sim \text{binom}(m_{i,0}, p_{1,0})$.

m1	m0	y1	y0
39	43	2	1
44	44	4	4
107	110	6	4
103	100	7	5
110	106	7	3
154	146	11	4

(a) Write the probabilities in logistic fashion, i.e. $p_{i,0} = H(\theta_{i,0})$ and $p_{i,1} = H(\theta_{i,0} + \gamma_i)$, with $H(u) = \exp(u)/\{1 + \exp(u)\}$. Show that

$$\gamma_i = H^{-1}(p_{i,1}) - H^{-1}(p_{0,i}) = \log \frac{p_{i,1}}{1-p_{i,1}} \Big/ \frac{p_{0,i}}{1-p_{i,0}},$$

the log-odds difference. Construct and display the optimal CD for the γ_i, and also for the odds ratio $\rho_i = \exp(\gamma_i)$, for each of the six studies.

(b) Assume then that the log-odds parameter γ is the same, across studies, so that the six binomial data pairs relate to seven parameters. Find the optimal CD for this γ, and for the common odds ratio $\rho = \exp(\gamma)$. Translate the CDs to confidence curves, and display the six + one curves in a diagram. How would you conclude?

Ex. 7.34 *Comparing Poisson parameters.* (xx ranting on a bit, to be edited. point to Story i.11, application to sucide attempts rates. xx) Suppose $Y_0 \sim \text{Pois}(m_0\theta_0)$ and $Y_1 \sim \text{Pois}(m_1\theta_1)$. In what precise way is θ_1 different from θ_0? Writing $\gamma = \theta_1/\theta_0$, show that the likelihood is proportional to $\exp\{-\theta_0(m_0+m_1\gamma)\}\theta_0^{y_0+y_1}\gamma^{y_1}$. Explain that the optimality recipe tells us inference should be made based on the distribution of $Y_1 \,|\, (Z=z)$, where $Z = Y_0 + Y_1$. Show that $Y_1 \,|\, (Z=z)$ has the binomial distribution $(z, m_1\gamma/(m_0+m_1\gamma))$. Show how this leads to the optimal CD

$$\begin{aligned} C(\gamma) &= \text{Pr}_\gamma(Y_1 > y_{1,\text{obs}} \,|\, z) + \tfrac{1}{2}\text{Pr}_\gamma(Y_1 = y_{1,\text{obs}} \,|\, z) \\ &= 1 - B_z(y_{1,\text{obs}}, m_1\gamma/(m_0+m_1\gamma)) + \tfrac{1}{2}b_z(y_{1,\text{obs}}, m_1\gamma/(m_0+m_1\gamma)). \end{aligned}$$

(xx point to Story i.11, for $y_0 = 1$, $y_1 = 7$, for the patient years m_0, m_1, in Aursnes et al. (2005). compare with their Bayes gamma priors, both informative and less informative. xx)

Ex. 7.35 *Basic meta-analysis.* (xx to come, and calibrated with later stuff. xx) There is a very wide literature on combining information, with different names and labels, including meta-analysis, data fusion, etc. This exercise looks into some of the more basic versions, and where CDs will be helpful in later extensions below.

(a) Suppose $y_j \sim \text{N}(\phi, \sigma_j^2)$, for $j = 1, \ldots, k$ independent sources, with the same focus parameter ϕ, and with variances taken to be known or well estimated. Consider the linear combination estimator $\widehat{\phi} = \sum_{j=1}^k a_j y_j$. Show that it is unbiased, provided $\sum_{j=1}^k a_j = 1$, and find its variance. Show that the best choice, yielding minimal variance among the unbiased ones, is $a_j \propto 1/\sigma_j^2$, leading to

$$\widehat{\phi} = \frac{\sum_{j=1}^k y_j/\sigma_j^2}{\sum_{j=1}^k 1/\sigma_j^2}.$$

Show indeed that $\widehat{\phi} \sim \text{N}(\phi, \kappa^2)$, with this minimal variance being $\kappa^2 = (\sum_{j=1}^k 1/\sigma_j^2)^{-1}$. Comment on what this leads to for the case where the σ_j are equal.

(b) In various settings there is a need to generalise the setting above to one where $y_j \,|\, \phi_j \sim \text{N}(\phi_j, \sigma_j^2)$, with these individual mean parameters not being equal, but having their own distribution, say $\phi_j \sim \text{N}(\phi_0, \tau^2)$. The task is then to reach inference for both the overall mean ϕ_0 and the spread τ among the ϕ_j. Show that $y_j \sim \text{N}(\phi_0, \sigma_j^2 + \tau^2)$, and that the log-likelihood function becomes

$$\ell(\phi_0, \tau) = -\tfrac{1}{2}\sum_{j=1}^k \Big\{\log(\sigma_j^2 + \tau^2) + \frac{(y_j - \phi_0)^2}{\sigma_j^2 + \tau^2}\Big\}.$$

(c) Considering the spread parameter τ first, show that the profiled log-likelihood can be written

$$\ell_{\text{prof}}(\tau) = -\tfrac{1}{2}\sum_{j=1}^{k}\Big[\log(\sigma_j^2+\tau^2) + \frac{\{y_j - \widehat{\phi}_0(\tau)\}^2}{\sigma_j^2+\tau^2}\Big], \quad \text{for } \tau \geq 0,$$

in which

$$\widehat{\phi}_0(\tau) = \frac{\sum_{j=1}^k y_j/(\sigma_j^2+\tau^2)}{\sum_{j=1}^k 1/(\sigma_j^2+\tau^2)}$$

is the best linear combination estimator for ϕ_0, for the fixed τ under inspection.

(d) (xx fix this, needs to be clearer. xx) Consider this profiled log-likelihood as a function of $\gamma = \tau^2$, rather than of τ, and show that its derivative at zero is

$$D = \tfrac{1}{2}\sum_{j=1}^{k}\frac{1}{\sigma_j^2}\Big\{\frac{(y_j-\widetilde{\phi}_0)^2}{\sigma_j^2} - 1\Big\}.$$

Here $\widetilde{\phi}_0 = \widehat{\phi}(0)$. A small D means that the y_j data have a low spread, and vice versa. Show that if $D \leq 0$, then $\widehat{\tau}_{\text{ml}} = 0$, and if that $D > 0$, then $\widehat{\tau}_{\text{ml}}$ is positive.

(e) (xx fix this, needs to be clearer. xx) Show next that

$$Q(\tau) = \sum_{j=1}^{k}\frac{\{y_j - \widehat{\phi}(\tau)\}^2}{\sigma_j^2+\tau^2} \sim \chi^2_{k-1}.$$

(f) (xx work with $\ell_{\text{prof}}(\tau)$. partly from CLP. derivative at zero. xx) By maximising over ϕ_0, for each given τ, show that

$$\ell_{\text{prof}}(\tau) = -\tfrac{1}{2}\sum_{j=1}^{k}\Big[\log(\sigma_j^2+\tau^2) + \frac{\{y_j - \widehat{\phi}(\tau)\}^2}{\sigma_j^2+\tau^2}\Big].$$

Ex. 7.36 *Combining CDs for the same parameter.* (xx a few exercises here. first for the same parameter, basic. then to CLP Ch 13 settings, then CDs to likelihood; then II-CC-FF. xx) Suppose that $C_1(\phi), \ldots, C_k(\phi)$ are independent CDs for the same parameter ϕ, perhaps based on different sets of data. How can these be properly combined?

(a) Show that $N_j = \phi^{-1}(C_j(\phi_{\text{true}}))$ is standard normal, at the true position in the parameter space underlying the C_j. With $w_1, \ldots, w_k$ numbers such that $\sum_{j=1}^k w_j^2 = 1$, show that

$$\bar{C}(\phi) = \Phi\Big(\sum_{j=1}^{k} w_j \Phi^{-1}(C_j(\phi))\Big)$$

is a proper combination CD for ϕ.

(b) (xx point back to Ex. 7.35. xx)

(c)

Ex. 7.37 *The problem of the Nile.* (xx starting a rant and will see how it goes. xx) Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent and exponential, the X_i with parameter θ, the Y_i with parameter $1/\theta$.

(a) Show that the log-likelihoods, for the X_i and for the Y_i parts, become

$$\ell_1(\theta) = n(\log\theta - \theta\bar{X}), \quad \ell_2(\theta) = n(-\log\theta - \bar{Y}/\theta),$$

with separate ML estimators $\widehat{\theta}_1 = 1/\bar{X}$ and $\widehat{\theta}_2 = \bar{Y}$. Show further that we may represent these as $\widehat{\theta}_1 \sim \theta A$, $\widehat{\theta}_2 \sim \theta B$, where $A = 2n/K$ and $B = L/(2n)$, with K and L being χ^2_{2n}.

(b) Show that the canonical CDs for θ, based on the two parts, are

$$C_1(\theta) = \Gamma_{2n}(2n\theta/\widehat{\theta}_1), \quad C_2(\theta) = 1 - \Gamma_{2n}(2n\widehat{\theta}_2/\theta).$$

The combination recipe hence leads to

$$\bar{C}(\theta) = \Phi\big((1/\sqrt{2})\{\Phi^{-1}(C_1(\theta)) + \Phi^{-1}(C_2(\theta))\}\big).$$

(c) The full log-likelihood becomes $\ell(\theta) = -n(\theta\bar{X} + \bar{Y}/\theta)$. Show that this is maximised for

$$\widehat{\theta}_{\rm ml} = (\bar{Y}/\bar{X})^{1/2} = (\widehat{\theta}_1\widehat{\theta}_2)^{1/2} \sim \theta F^{1/2},$$

where $F = A/B = L/K \sim F_{2n,2n}$. Show that the associated with this ML estimator is $C_{\rm ml}(\theta) = 1 - F_{2n,2n}(\widehat{\theta}^2_{\rm ml}/\theta^2)$.

(d) Run some simple experiments, where you for a few values of n simulate data and then plot the four CDs, and the four confidence curves (the two separate ones, the combination one, and the ML based one).

(e) (xx a bit more. ML etc. work, even though the situation is slightly irregular. $(\bar{X}, \bar{Y})$ is sufficient, but not complete. we need two informants for the single parameter. xx)

Ex. 7.38 *From CD to likelihood.* (xx to come. with illustrations. normal conversion. $\ell(\phi) = -\frac{1}{2}\Gamma_1^{-1}(\mathrm{cc}_j(\phi))$. xx)

Ex. 7.39 *II-CC-FF: Independent Inspection, Confidence Conversion, Focused Fusion.* (xx to come. using Cunen and Hjort (2022). point to Bayesian updating being part of this, but allows user keeping only prior for the focus parameter. aim to demonstrate iiccff in Story i.11. xx)

Ex. 7.40 *Private attributes.* (xx to be checked with care. xx) The probability ψ of cheating at exams might be hard to estimate, but a bit of randomisation might grant anonymity and yield valid estimates. Suppose there are three cards, two with the statement 'I did cheat', and 'I did not cheat' on the third. Students are asked to draw one of the three cards randomly and answer either *true* or *false* to the drawn statement, without revealing it.

(a) Show that the probability of *true* is $(1+\psi)/3$. Assume a binomial model for the number of students answering *true*, and devise a CD for ψ.

(b) Assume 1000 students go through the simple post-exam exercise above (anonymously). Find and display CDs for ψ for the cases of respectively 300, 350, 400 out of the 1000 answered *true*.

Notes and pointers

[xx A few remarks. xx]

we point to Schweder and Hjort (2016, Example 3.11), Fisher (1930), Xie and Singh (2013), Hjort and Schweder (2018), Cunen and Hjort (2022), Singh et al. (2005), ...

(xx repair. we point to Story iii.10. xx) There's a notable discrepancy between the frequentist Schweder-Hjort CD and the Bayesian posterior distribution associated with a flat prior on the $[0,\infty)$ interval, in cases where the $y_{\rm obs}$ is close to, or perhaps even to the left of, the boundary point.

I.8

Loss, risk, performance, optimality

Statistics is a mathematical formalisation of how to make good decisions under uncertainty. One source of uncertainty is that the future or the true state of nature, say θ, is not known when we are to make our decisions, and since the utility or loss of a decision depends on θ, we need to be clear about how bad it is when our decision is off. This is the role of loss functions, of which the square error is the most well known example. Decisions are based on data, and it is not anodyne how we use the data: Some procedures for going from data to a decision are better than others. Therefore, it makes sense to see how a certain procedure performs on average. This is the role played by risk functions, of which the mean square error is the most well known example. Risk functions average out the data, but they still depend on θ, so the risk function of various decision procedures are often difficult to order. Some might be preferable for certain values of θ, while others might be better for other values of θ. This chapter introduces criteria that let us, nevertheless, say something about how good a decision procedure is. A decision procedure is said to be *admissible* if there is no other that does better, in terms of the risk function, whatever the truth or future may be. One says that a decision procedure is *minimax* if it is the the best one in the most unfortunate situation. In proving that certain decision procedures are admissible or minimax, Bayesian thinking is an essential tool. This includes the concepte of Bayes risk, where not only the data is average out, but also the various possible states of nature are averaged out.

Key words: admissibility, Rao–Blackwell, loss functions, minimaxity, multiparameter estimation, UMVU estimators, risk functions

A statistical decision, as most decisions in life when you think about it, is a function of what we observe to the space of all possible decisions we can make in a given setting: I look out the window and see grey clouds, and choose to take my umbrella with me when I go out. I see that you are smiling and I think that you are happy. Formally, an action is a function $a\colon \mathcal{X} \to \mathcal{A}$, where $\mathcal{X}$ is the space in which the data take its values, the *sample space*; while $\mathcal{A}$ is the *action space*, that is the collection of all possible actions we might take. How wise our choice of $a \in \mathcal{A}$ is or turns out to be, depends on the true *state of nature* θ. This θ lives in a *parameter space* Θ, and is an unknown *parameter* governing the probability distribution P_θ from which the data $X \in \mathcal{X}$ are generated. A

Loss function

loss function $L(\theta, a)$ measures 'how much' we lose by choosing action a when the true state of nature is θ. We assume that

$$L\colon \Theta \times \mathcal{A} \to [0, \infty),$$

so that the best possible loss is zero (in general, loss functions do not need to be nonnegative, see e.g. Schervish (1995, Chapter 3.1) for a more general introduction). To be concrete, consider the point estimation problem with data $X_1, \ldots, X_n$ independent $\mathrm{N}(\theta, 1)$, where θ is an unknown parameter to be estimated under the loss function $L(\theta, a) = (a - \theta)^2$. Here, the amount lost is the squared distance between a and θ. If we wish to test $H_0\colon \theta \leq 0$ versus $H_A\colon \theta > 0$, the action space is $\mathcal{A} = \{\text{keep } H_0, \text{ reject } H_0\}$, and a natural loss function can be described by

$$L(\theta, a) = \begin{array}{r|cc} & \theta \leq 0 & \theta > 0 \\ \hline \text{keep } H_0 & 0 & 1 \\ \text{reject } H_0 & 1 & 0 \end{array}.$$

[xx see comment about 0-1 loss in Schervish (1995, p. 215) xx] With this loss function, we lose the same amount when rejecting a true null hypothesis as when failing to reject a null hypothesis that is false. In statistical jargon that you probably already know, failing to reject a null hypothesis is called a Type I error, while when we fail to reject a null hypothesis that is false, we commit a Type II error.

In the classical or frequentist setup, different decision procedures are compared by the loss they incur for each value of θ, that is, by their *risk function*

$$R(\theta, a) = \mathrm{E}_\theta\, L(\theta, a(X)) = \int_{\mathcal{X}} L(\theta, a(x))\, \mathrm{d}P_\theta(x).$$

Notice that for each decision procedure a, the risk function $R(\theta, a)$ is a function of θ, so that for two estimators a_1 and a_2 their respective risk functions may cross, that is $R(\theta, a_1) < R(\theta, a_2)$ for some θ, while $R(\theta, a_1) > R(\theta, a_2)$ for other values of θ. Thus, comparison of risk functions only provides a partial ordering of decision procedures, and as such does not point clearly at a best decision procedure. What is clear, however, is that if $R(\theta, a_1) \leq R(\theta, a_2)$ for all θ, with strict inequality for at least one value of θ, then a_2 should be discarded from the competition: We say that a_1 *dominates* a_2, and a_2 is said to be *inadmissible*. A decision procedure that is not dominated by any other decision procedure is *admissible*. In and of itself admissibility does not tell us much about an estimator. Consider for example the estimator $a'(X) = \theta'$ that returns the value θ' whatever the data. Clearly, no estimator can perform better than a' in θ', but that does not, for obvious reasons, make it an estimator we would like to use. One (not very principled) fix to the problem of comparing estimators is to limit the search of a best estimator to the class of estimators that are unbiased: An estimator $\widehat{\theta}$ is unbiased for the parameter θ if $\mathrm{E}_\theta\, \widehat{\theta} = \theta$ for all θ. Another principle by which to compare decision procedures is the the *minimax principle*. According to this principle, the estimator with the best performance in the worst possible scenario ought to be chosen. We say that a decision rule $a^\star$ is *minimax* if it minimises the maximum risk, that is if

Admissibility

unbiased estimator

Minimax

$$\inf_{a \in \mathcal{A}} \sup_{\theta \in \Theta} R(a, \theta) = \sup_{\theta \in \Theta} R(a^\star, \theta).$$

If you are a Bayesian and venture into the business of constructing prior distributions $\pi(\theta)$ over the parameter space Θ, then the problem of risk functions only being partially ordered can be circumvented. What you are interested in then is the *Bayes risk* of the decision procedures you are comparing, that is

Bayes risk

$$\mathrm{BR}(a, \pi) = \mathrm{E}_\pi \, R(\theta, a) = \int_\Theta R(\theta, a) \, \pi(\theta) \, \mathrm{d}\theta.$$

For a given a and prior π, the Bayes risk is a number, and under certain conditions that we will explore in the exercises to come, there will be a unique decision procedure a that, for a given prior π, minimises $\mathrm{BR}(a, \pi)$. This decision procedure is the *Bayes solution*. As we will see, Bayes solutions are, despite the name, extremely important for Bayesians and frequentists alike.

Bayes solution

UMVU estimator, Rao–Blackwell, and Lehmann–Scheffé

Ex. 8.1 *Coin tossing.* To get a feeling for some of the basic challenges and concepts concerning the comparison of various estimator, we start out with the emblematic problem of estimating the probability of heads in $n = 10$ independent tosses of a coin. Let $Y_1, \ldots, Y_n$ be independent Bernoulli random variables with expectation θ.

(a) Sketch the risk function of the maximum likelihood estimator $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ under squared error loss $L(\delta, \theta) = (\delta - \theta)^2$. Recall that $n = 10$.

(b) Suppose we have some intuition about where on the unit interval the expectation θ might be located, close to a value $0 < \theta_0 < 1$, say. One way in which such a prior hunch might be employed is by taking as our estimate a convex combination of the maximum likelihood estimator and θ_0, that is

$$\delta_a(Y_1, \ldots, Y_n) = a\bar{X}_n + (1 - a)\theta_0,$$

for some $0 \leq a \leq 1$. For $a = 1/2$ and $\theta_0 = 1/2$, sketch the risk function of this estimator. Suppose that your task, as was Pierre-Simon Laplace's in 1781 or so, is to estimate the probability of giving birth to a boy. Which of the two above estimators do you prefer, the maximum likelihood estimator or δ_a with $a = 1/2$ and $\theta_0 = 1/2$?

(c) Based on the risk functions you sketched in (a) and (b), we see that the two estimators are difficult to compare. The maximum likelihood estimator has lower risk than δ_a for certain values of θ, while δ_a performs better for other values of θ. The risk functions cross and none of the two is uniformly better than the other. An easy fix to this problem of comparison, is to limit our search for an estimator to the class of estimators that are unbiased for what we are estimating. Look back at Ex. 5.15 and explain why, when the search is restricted to the class of unbiased estimators, the maximum likelihood estimator is the clear winner.

(d) The risk of the estimators from (a) and (b) vary widely with what the true θ is. Choosing a best estimator, when the yardsstick is the squared error loss function, seems therefore to require some prior hunch about where θ really is. A criterion for risk function

comparison that does not require such a prior hunch, is minimaxity: An estimator is minimax if it minimises the maximum risk. Estimators whose risk functions are constant are, as we will soon see, good candidates for being minimax. Consider the estimator from Ex. (b) with $\theta_0 = 1/2$. Find a function $a = a(n)$ such that the risk function $R(\theta, \delta_{a(n)})$ is constant. For $n = 10$, sketch the risk function of your estimator (that is, draw a line). Suppose you have absolutely no idea whatsoever about where in the unit interval θ may be located. Which of your three estimators of θ do you prefer? In Ex. 8.9 we learn that $\delta_{a(n)}$ is indeed minimax.

Ex. 8.2 *Uniformly minimum variance unbiased estimators.* As the sketch in Ex. 8.1 (hopefully) illustrates, comparing risk functions is not always straightforward, and a somewhat *ad hoc* way of making the problem of finding a best estimator tractable is by limiting the search for a best estimator to the class of unbiased estimators. What is variably called a best unbiased estimator, the uniformly minimum variance unbiased estimator, the UMVU estimator, is defined as follows: An estimator $\delta^\star(Y)$ is the uniformly minimum variance unbiased estimator for $g(\theta)$ if it is unbiased for $g(\theta)$, and for any other estimator $\delta(Y)$ that is unbiased for $g(\theta)$, it holds that $\mathrm{Var}_\theta\, \delta^\star(Y) \leq \mathrm{Var}_\theta\, \delta(Y)$ for all θ. When feasible (see Ex. 5.14 and 5.15–5.16), the easiest way of establishing that an unbiased estimator is an uniformly minimum variance unbiased estimator, is to verify that it achieves the Cramér–Rao lower bound.

UMVU estimator

(a) Let us look at a few examples of the Cramér–Rao approach: (i) Suppose $X \sim \mathrm{N}(\theta, 1)$, and show that X is uniformly minimum variance unbiased for θ. (ii) Suppose Y has an exponential distribution with mean θ. Show that Y is uniformly minimum variance unbiased for θ. (iii) Let $Y_i = \beta_0 + \beta_1 x_i + \sigma\varepsilon_i$ for $i = 1, \ldots, n$, where the covariates $x_1, \ldots, x_n$ are fixed numbers, and the $\varepsilon_1, \ldots, \varepsilon_n$ are independent standard normal random variables. Show that the least squares estimator for (β_0, β_1) is the uniformly minimum variance unbiased estimator.

(b) Let $Y_1, \ldots, Y_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$. It is immediate from (a) that $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ is uniformly minimum variance unbiased for μ. Show that the estimator $\widehat{\sigma}_n^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2/(n-1)$ is *not* uniformly minimum variance unbiased for σ^2. In Exercise 8.3 we will see that there is no unbiased estimator of σ^2 attaining the Cramér-Rao lower bound.

Ex. 8.3 *Cramér–Rao and Cauchy–Schwarz.* The proof of the Cramér–Rao inequality that we met in Ex. 5.15–5.16, is a clever application of the Cauchy–Schwarz inequality.

(a) Let X and Y be two square integrable random variables with expectation zero. Show that $|\mathrm{E}\, XY| = \{\mathrm{Var}\,(X)\mathrm{Var}\,(Y)\}^{1/2}$ if and only if X and Y are linearly related, $Y = a + bX$, for example.

(b) Explain why the Cramér–Rao inequality is an equality if and only if the estimator $\delta(y)$ and the score function are linearly related, that is

$$\frac{\partial}{\partial\theta} \log f(Y;\theta) = a(\theta) + b(\theta)\delta(Y), \quad \text{for all } \theta,$$

for some function $a(\theta)$ and $b(\theta)$. Solve the differential equation above for $f(y;\theta)$, and state what this entails for estimators and distributions when it comes to possible attainment of the Cramér–Rao lower bound. You may have a look back at Ex. 4.19 and 4.20.

(c) In Ex. 8.2(b) we saw that with $Y_1, \ldots, Y_n$ i.i.d. $\mathrm{N}(\mu, \sigma^2)$, the unbiased estimator $\widehat{\sigma}_n^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$ for σ^2 does not attain the Cramér–Rao lower bound. Use the result from (b) to argue that the Cramér–Rao lower bound may only be attained when μ is known.

Ex. 8.4 *Sufficiency and Rao–Blackwell.* Suppose that Y has a distribution from a family $\{P_\theta \colon \theta \in \Theta\}$ of distributions. Recall from Ex. 4.16 that $T = T(Y)$ is a sufficient statistic for this family distributions if the conditional distribution of Y given T does not depend on θ. The Rao–Blackwell theorem says that any estimator can be improved upon by conditioning on a sufficient statistic. This is an important result, as it tells us that in our search for a best estimator, we need only consider those estimators that are functions of a sufficient statistic.

Rao–Blackwell theorem

(a) Let $\delta'(Y)$ be an unbiased estimator for $g(\theta)$, and suppose that $T(Y)$ is sufficient for θ. Consider the estimator given by $\delta(Y) = \mathrm{E}_\theta\{\delta'(Y) \mid T\}$. Then $\delta(Y)$ is better than $\delta'(Y)$. The proof proceeds in three steps. First, explain why $\delta(Y)$ is an estimator; second, show that $\delta(Y)$ is unbiased, and third, show that $\mathrm{Var}_\theta\, \delta(Y) \leq \mathrm{Var}_\theta\, \delta'(Y)$ for all θ. You have now proven the Rao–Blackwell theorem for unbiased estimators. In Ex. 8.8 we look at this theorem in a more general decision theoretic framework.

(b) Suppose that $Y_1, \ldots, Y_n$ are i.i.d. uniforms on $[0, \theta]$, with $\theta > 0$ an unknown parameter. Show that the estimators

$$\delta_1 = \frac{2}{n} \sum_{i=1}^n Y_i, \quad \text{and} \quad \delta_2 = \frac{n+1}{n} \max_{i \leq n} Y_i,$$

are both unbiased for θ. There are (at least) two ways of showing that δ_2 is a better estimator than δ_1. Try them both. First, compute the variances of both estimators. Second, appeal to the Rao–Blackwell theorem, that is, more concretely, use the results from Ex. 3.17 to establish that $\mathrm{E}\{\delta_1 \mid \max_{i \leq n} Y_i\} = \delta_2$ almost surely.

(c) Let $Y_1, \ldots, Y_n$ be i.i.d. Pois(λ). We seek to estimate $\theta = \Pr(Y_1 = 0) = \exp(-\lambda)$. Find the maximum likelihood estimator for θ, say $\widehat{\theta}$, and show that $\mathrm{E}\,(\widehat{\theta}) = \gamma\{1 + O(1/n)\}$, meaning the the maximum likelihood estimator for θ is biased. Next, find the Cramér–Rao lower bound for unbiased estimators of θ. Look back at Ex. 8.3 and consider whether this lower bound can be attained.

Another estimation strategy is to estimate $\theta = \Pr(Y_1 = 0)$ by the share of zero counts, that is $\widetilde{\theta} = n^{-1} \sum_{i=1}^n I\{Y_i = 0\}$. This estimator is clearly unbiased for θ, why is it not best unbiased? Finally, with the aid of the sufficient statistic $T(Y) = \sum_{i=1}^n Y_i$ we Rao–Blackwellise the estimator $\widetilde{\theta}$. Derive an expression for this Rao–Blackwellised estimator, $\widetilde{\theta}_{\mathrm{rb}}$, say. Does $\widetilde{\theta}_{\mathrm{rb}}$ attain the Cramér–Rao lower bound?

Ex. 8.5 *Best unbiased, completeness, and Lehmann–Scheffé.* From the Rao–Blackwell theorem we know that any candidate for being an uniformly minimum variance unbiased estimator must be a function of a sufficient statistic. This limits our search. In this exercise we establish that in our search for the UMVU estimator, we are only looking for one estimator. Thereafter, it is shown that an estimator is best unbiased if and only if it

is uncorrelated with all unbiased estimators of zero. This yields a characterisation of the best unbiased estimators, albeit one of limited utility as it is, without further conditions, hard to describe all unbiased estimators of zero. Finally, completeness – a condition on the distribution of the data – is introduced, ensuring that the only unbiased estimator of zero is zero itself.

(a) Suppose that δ is uniformly minimum variance unbiased for $g(\theta)$, and that so is δ'. Let $\delta'' = \delta/2 + \delta'/2$, and use the Cauchy–Schwarz inequality to show that $\mathrm{Var}_\theta\, \delta'' \leq \mathrm{Var}_\theta\, \delta$. But since δ is an UMVU estimator and δ'' is unbiased (check it), it must be the case that $\mathrm{Var}_\theta\, \delta'' = \mathrm{Var}_\theta\, \delta$. Look back at the results in Ex. 8.3, and use this to establish that if δ and δ' are both uniformly minimum variance unbiased, then $\delta = \delta'$ almost surely, for all θ.

UMVU estimator is unique

(b) [xx rewrite xx]Suppose that δ is an unbiased estimator for $g(\theta)$ and a function of a sufficient statistic. How may we improve on δ? Well, the family of estimators $\delta_a = \delta + a\varepsilon$ as a ranges of the real numbers, and ε is some mean zero random variable, constitute a class of unbiased estimators. Show that if $\mathrm{cov}_\theta(\delta, \varepsilon) \neq 0$ for some θ, then a may be chosen so that $\mathrm{Var}_\theta(\delta_a) < \mathrm{Var}_\theta(\delta)$ for some value(s) of θ, which entails that δ is not best unbiased. Prove the converse, namely that if δ is unbiased and $\mathrm{cov}_\theta(\delta, \varepsilon)$ for all θ and all mean zero random variables ε, then δ is uniformly minimum variance unbiased.

characterisation of the UMVU estimator

(c) It is in general no easy task to show that an unbiased estimator, or more generally, a statistic $T = T(Y)$ say, is uncorrelated with all unbiased estimators of zero. Since the correlation between any random variable and zero is zero, the task would be much easier if we knew of the distribution of T that the only unbiased estimator of zero, is zero itself. That is, if for any measurable function h, $\mathrm{E}_\theta\, h(T) = 0$ implies $\Pr_\theta\{h(T) = 0\} = 1$, for all θ. We recall from Ex. 4.23 that a family of distributions with this property is called *complete*. Alternatively, we just say that the statistic $T(Y)$ is complete.

Suppose that T is sufficient and complete for θ. Let $\delta = \delta(T)$ be unbiased for $g(\theta)$. Prove the Lehmann–Scheffé theorem, that is, show that the estimator δ is the unique uniformly minimum variance unbiased estimator for $g(\theta)$.

Lehmann–Scheffé theorem

(d) Look back at Ex. 8.4(b). Show that the estimator δ_2 is the uniformly minimum variance unbiased estimator.

(e) The completeness requirement in the Lehmann–Scheffé theorem was motivated by the characterisation in (b), saying that an estimator is uniformly minimum variance unbiased if and only if it is uncorrelated with all unbiased estimators of zero. A perhaps more illuminating motivation comes from the fact, proven in Ex. 8.4(a), that a best estimator must be based on a sufficient statistic. Intuitively, by getting rid of information irrelevant to the estimation problem at hand, we reduce the variance of our estimation procedure. Taking this intuition to its logical conclusion, we deduce that a best estimator must be based on a statistic achieving the maximum amount of data compression, while still retaining all the information in the data about the parameter we seek to estimate. In other words, a best estimator must be based on a minimal sufficient statistic. Recall from Ex. 4.21 that a statistic S is minimal sufficient if for any sufficient statistic T there

exists a measurable function g so that $S = g(T)$. Show that if δ is an unbiased estimator, we form the estimators $\delta' = \mathrm{E}(\delta \mid T)$ and $\delta'' = \mathrm{E}(\delta \mid S)$, where T is sufficient and S is minimal sufficient, then $\mathrm{Var}(\delta'') \leq \mathrm{Var}(\delta')$. Now, suppose that T is sufficient *and* complete. Use the Lehmann–Scheffé theorem and (a) to conclude that δ' and δ'' must be almost surely equal. In view of this equality, it may not come as a surprise that if T is sufficient and complete, then T is minimal sufficient. A fact we will prove in (g).

(f) Here is a toy example illustrating some of the points made in (e). Let X_1 and X_2 be independent Bernoulli(θ) random variables and consider the estimator $\widehat{\theta} = (X_1 + X_2)/2$ and the estimator $\delta = \delta(X_1, X_2)$ given by

$$\delta(x_1, x_2) = \begin{cases} 1, & (x_1, x_2) = (1, 1), \\ 2/3, & (x_1, x_2) = (1, 0), \\ 1/3, & (x_1, x_2) = (0, 1), \\ 0, & (x_1, x_2) = (0, 0). \end{cases}$$

Explain why both $\widehat{\theta}$ and δ are sufficient for θ. Show that δ is unbiased for θ, and show that the variance of δ exceeds the variance of $\widehat{\theta}$ for all values of $\theta \in (0, 1)$.

Bahadur's theorem

(g) The results quoted at the end of (e) is Bahadur's theorem: If T is sufficient and complete, then T is minimal sufficient.

To prove this, let W be another sufficient statistic, and assume, with out loss of generality, that T and W are real valued. We must show that there is a function g such that $T = g(W)$. If $T = \mathrm{E}_\theta(T \mid W)$, then we have found a function g, it is $g(w) = \mathrm{E}_\theta(T \mid W = w)$, and g does not depend on θ since W is sufficient. Let us therefore prove that T equals $\mathrm{E}_\theta(T \mid W)$ almost surely for all θ. To this end, assume that T has finite variance, and define $g(W) = \mathrm{E}_\theta(T \,|\, W)$ and $h(T) = \mathrm{E}_\theta\{g(W) \,|\, T\}$. Now, use the tower property of conditional expectation a couple of times and that T is complete to show that $T = h(T)$ almost surely, for all θ. Next, combine the above with the variance decomposition formula to obtain

$$\mathrm{Var}_\theta g(W) = \mathrm{E}_\theta \mathrm{Var}_\theta(g(W) \,|\, T) + \mathrm{E}_\theta \mathrm{Var}_\theta(T \,|\, W) + \mathrm{Var}_\theta g(W),$$

from which we conclude that $T = \mathrm{E}_\theta(T \,|\, W)$ almost surely for all θ. To get rid of the finite variance assumption on T, replace T by $f(T) = 1/\{1 + \exp(-T)\}$ (which clearly has finite variance) throughout the proof, to conclude that $T = f^{-1}(g(W))$.

Ex. 8.6 *A uniform mean.* Let $Y_1, \ldots, Y_n$ be i.i.d. unif(a, b). We are to estimate the mean $\mu = (a + b)/2$.

(a) Show that $(\min_{i \leq n} Y_i, \max_{i \leq n} Y_i)$ is sufficient and complete.

(b) Propose an unbiased estimator for μ, and find its variance.

Ex. 8.7 *A weird unbiased estimator.* Limiting our search for a best estimator to the class of unbiased estimators lacks the decision theoretic foundation that the principle of minimising expected loss enjoys. More on this in the Notes and Pointers section.

Sometimes, the search for unbiasedness might lead us astray. Let Y be a random variable with density

$$f(y,\theta) = \frac{\theta^y \exp(-\theta)}{y!\{1-\exp(-\theta)\}}, \quad \text{for } y = 1, 2, \ldots,$$

with $\theta > 0$. This is a Poisson distribution truncated at zero, and the probability of being truncated is $\exp(-\theta)$.

(a) Show that $\delta_u(Y) = (-1)^{Y+1}$ is the unique unbiased estimator for $\exp(-\theta)$.

(b) Find an expression for the risk function of $\delta_1(Y)$.

(c) Propose an estimator with uniformly smaller risk than $\delta_1(Y)$.

Ex. 8.8 *More Rao–Blackwellisation.* Recall that a function is convex if g is a convex if for all x, y in its domain, and all $\lambda \in [0,1]$,

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y).$$

Jensen's inequality states that if g is convex, then

$$g(\mathrm{E}\, X) \leq \mathrm{E}\, g(X),$$

with equality only if $g(x) = a + bx$. Jensen's inequality also holds conditional expectations, see Ex. A.24(g). In this exercise we will look at loss functions $L(\delta, \theta)$ that are convex in δ for all θ. Think of your favourite loss function, and you will realise that this is a quite natural requirement.

(a) Let $\delta = \delta(X)$ be an estimator of θ. Suppose that $T = T(X)$ is sufficient for θ and define $\delta^\star(T) = \mathrm{E}\,\{\delta(X) \mid T\}$. Explain why $\delta^\star(T)$ is an estimator.

(b) Suppose $L(\delta, \theta)$ is convex in δ for all θ. Show that $R(\delta^\star, \theta) \leq R(\delta, \theta)$ for all θ.

(c) When will the inequality in (b) be strict for all θ?

(d) Suppose that $\mathrm{E}_\theta\, \delta^\star(T) = h(\theta)$, and that T is complete. Show that, in the class of estimators $\{\delta \colon \mathrm{E}_\theta\, \delta = h(\theta)\}$, the estimator $\delta^\star$ is the unique estimator minimising the risk. [xx check this for the general case here presented xx].

(e) [xx Let $L(\delta, \theta) = (\delta - \theta)^2$ and specialise to UMVU estimator xx]

(f) Suppose $L(\delta, \theta)$ is convex in δ for all θ, and that δ_π is the unique Bayes solution under the prior π. Show that δ_π must be a function of a sufficient statistic.

Minimaxity

Ex. 8.9 *Tools for minimaxity.* In Exercise. 8.1 we compared three different estimators, which is fine, but what we ultimately want to say something about is the performance of an estimator compared to *all* other estimators. To do so, we need some more tools. We start out with convenient tools for establishing minimaxity, from which we will see that the estimator in Exercise 8.1(d) is minimax.

(a) Let δ_π be a Bayes solution with respect to the prior distribution π, and suppose that

$$\mathrm{BR}(\delta_\pi, \pi) = \sup_\theta R(\theta, \delta_\pi). \tag{8.1}$$

Show that δ_π is minimax.

(b) Show that if δ_π satisfies (8.1) and is the unique Bayes solution with respect to π, then δ_π is the unique minimax procedure.

(c) Show that if a Bayes solution has constant risk, then it is minimax.

(d) Show that if an estimator has constant risk and is admissible, it is minimax.

(e) Show that if an estimator is unique minimax, it is admissible.

Ex. 8.10 *The minimax estimator in Bernoulli problem.* Let $Y_1, \ldots, Y_n$ be independent Bernoulli with success probability θ.

(a) Give θ a $\mathrm{Beta}(a\theta_0, a(1-\theta_0))$ prior distribution, and find an expression for the posterior expectation.

(b) Find an expression for the risk function under squared error loss when $a = n^{3/2}$ and $\theta_0 = 1/2$, and conclude. See Ex. 8.1(d).

Ex. 8.11 *Minimaxity and sequences of priors.* In Exercise 8.9(b) we assumed that the equality in (8.1) is attained. A prior distribution that succeeds in attaining this equality is, for natural reasons, called a *least favourable* prior distribution. If no such prior distribution exists, we cannot use the conclusion of the exercise to prove minimaxity. Consider independent $X_1, \ldots, X_n \mid \theta$ from $\mathrm{N}(\theta, 1)$. It seems reasonable that a least favourable prior for θ should spread its mass evenly out on the real line, that is

$$\int_a^{a+c} \pi(\theta)\,\mathrm{d}\theta = \int_b^{b+c} \pi(\theta)\,\mathrm{d}\theta, \qquad \text{for all } a, b \in \mathbb{R} \text{ and } c > 0.$$

This distribution is Lebesgue measure on $\mathbb{R}$, and is not a proper probability distribution. This hints at the result above not being applicable. To fix this, the idea is to approximate an improper distributions with proper ones. In the case of the normals, one may try $\theta \sim \pi_k(\theta)$, where $\pi_k(\theta)$ is the density of a uniform distribution over $[-k, k]$, then let k grow.

(a) Suppose that δ is an estimator and $(\pi_k)_{k\geq 1}$ a sequence of prior distributions such that

$$\sup_\theta R(\delta, \theta) = \lim_{k\to\infty} \mathrm{BR}(\delta_{\pi_k}, \pi_k),$$

with δ_{π_k} being the Bayes solution for π_k. Show that δ is minimax.

(b) Let $X_1, \ldots, X_n$ be independent $\mathrm{N}(\mu, \sigma^2)$. We are to estimate μ under the squared error loss $L(\widehat{\mu}, \mu) = (\widehat{\mu} - \mu)^2$. You may consider the sequence of priors $\mu \sim \mathrm{N}(0, \tau_k)$ for $k = 1, 2, \ldots$ to show that the estimator $\widehat{\mu} = \bar{X}_n$ is minimax.

(c) Let $X_1, \ldots, X_n$ be independent $\mathrm{Poisson}(\theta)$. We want to estimate θ under the weighted loss function $L(\widehat{\theta}, \theta) = \theta^{-1}(\widehat{\theta} - \theta)^2$. Use Gamma priors to show that $\widehat{\theta} = \bar{X}_n$ is minimax.

Admissibility and Bayes

Ex. 8.12 *Some Bayes and some admissibility.* If we have at hand an estimator δ, the most convenient way of showing that δ is admissible is to show that it is Bayes. In fact, it is almost true that an estimator is admissible if and only if it is Bayes. We'll get to the cases where this implication fails, but as a rule of thumb it is pretty safe.

(a) Suppose that $X \sim f_\theta$, where $\theta \in \Theta = \{\theta_1, \ldots, \theta_k\}$ for some finite $k \geq 2$. Consider the estimator δ_π that is Bayes for the prior $\pi = \{\pi_1, \ldots, \pi_k\}$, where π_j is the prior mass given to θ_j. Show that if $\pi_j > 0$ for $j = 1, \ldots, k$, then δ_π is admissible.

(b) Why does the conclusion of Ex. 8.12(a) fail if $\pi_j = 0$ for one or more j?

(c) Show that if a Bayes solution is *unique*, then it is admissible. Or, equivalently, if every Bayes rule with respect to a prior π has the same risk function, then they are all admissible.

(d) To clarify what the uniqueness of the Bayes solution refers to, let's look at an example where there are several Bayes solutions. Let $X \sim \text{unif}(0, \theta)$ and suppose that we want to estimate θ under the squared error loss function $L(\delta, \theta) = (\delta - \theta)^2$. Suppose θ is given the prior distribution that is uniform on $(0, c)$. Find the posterior distribution $\theta \mid (X = x)$ and derive at least two different Bayes solutions (there are uncountably many). [xx comment on this exercise, more relevant in BNP, point to Nils' 1976-proof and the mistake made by Lehmann. Could also mention Lindley and his so-called Cromwell's rule xx].

(e) Recall that a parameter value θ is in the *support* of π (a probability density in our notation) if it is contained in the set $\{\theta \in \Theta \colon \pi(\theta) > 0\}$. Let $\{f_\theta \colon \theta \in \Theta\}$ be a model, and suppose that (i) the support of the prior π is Θ; and that (ii) the risk function $R(\theta, \delta)$ is continuous in θ for all estimators δ. Show that if δ_π is Bayes with respect to π and have finite Bayes risk, then δ_π is admissible.

Ex. 8.13 *Generalised Bayes.* Suppose that in some experiment involving data from a normal distribution with expectation θ, you have no idea whatsoever about where on the real line θ might be located. A natural 'prior' is therefore $\pi(\theta) \propto 1$ (or just take it equal to one) that spreads the 'probability' mass uniformly over the real line. Now, $\pi(\theta) \propto 1$ corresponds to Lebesgue measure on the real line, and is not a probability measure because

$$\int_{\mathbb{R}} \pi(\theta)\, \mathrm{d}\theta = \infty.$$

The fact that $\pi(\theta)$ does not integrate to one does not, however, stop us from using it to derive estimators using 'Bayes' theorem. Priors that are not probability distributions are called improper priors. improper priors

(a) Suppose $X_1, \ldots, X_n$ are independent $\text{N}(\theta, \sigma^2)$. Suppose θ is given the improper prior $\pi(\theta) = 1$. Show that $\pi(\theta \mid x_1, \ldots, x_n) = \text{N}(\bar{X}_n, \sigma^2/n)$. A generalised Bayes estimator is the estimator δ minimising the posterior expected loss $\text{E}\{L(\delta, \theta) \mid \text{data}\}$. Let $L(\delta, \theta) = (\delta - \theta)^2$, and find the generalised Bayes estimator for θ.

(b) The estimator you found in (a) is generalised Bayes, why is this not enough to conclude that it is admissible?

(c) (xx xx)

Ex. 8.14 *Blyth's method.* We call δ a *limiting Bayes* estimator if there is a sequence $\{\pi_k\}_k$ of possibly improper priors such that the corresponding Bayes estimators δ_{π_k} converge almost surely to δ. Blyth's methods can be paraphrased as saying that limits of Bayes's estimators are admissible. We start, in (a) by proving Blyth's method, as is clear by now, when it comes to admissibility proofs by contradiction is the way to go.

(a) Let $\delta^\star$ be an estimator. Suppose $\Theta \subset \mathbb{R}^p$ is open, and that $R(\delta, \theta)$ is continuous in θ for all estimators δ. Let $(\pi_k)_{k\geq 1}$ be a sequence of (possibly improper) prior distributions such that $\mathrm{BR}(\delta^\star, \pi_k) < \infty$ for all k, and for any open set $\Theta_0 \subset \Theta$,

$$\frac{\mathrm{BR}(\delta^\star, \pi_k) - \mathrm{BR}(\delta_{\pi_k}, \pi_k)}{\int_{\Theta_0} \pi_k(\theta)\,\mathrm{d}\theta} \to 0, \quad \text{as } k \to \infty;$$

Then $\delta^\star$ is admissible.

Normal mean is admissible

(b) Let $X_1, \ldots, X_n$ be i.i.d. from a $\mathrm{N}(\theta, 1)$, where θ is an unknown parameter to be estimated under under the squared error loss function $L(\delta, \theta) = (\delta - \theta)^2$. Consider the sequence of prior distributions $\pi_k(\theta) = \mathrm{N}(0, \tau_k^2)$, and show that the Bayes solution is

$$\delta_{\pi_k}(X) = \frac{n\bar{X}_n}{n + 1/\tau_k^2},$$

Take it from here and show that $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is admissible.

Ex. 8.15 *Bernoulli mean with weighted risk.* Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(θ). We wish to estimate θ, and we are particularly interested in precise estimates of very small and very large values of θ. Therefore, we'll work with the loss function

$$L(\delta, \theta) = \frac{(\delta - \theta)^2}{\theta(1-\theta)}.$$

(a) Compute the risk function of the maximum likelihood estimator. What's noticeable about this risk function?

(b) We now take a Bayesian point of view and give θ a Beta($a\theta', a(1-\theta')$) prior distribution. Compute the expectation and variance of this prior.

(c) With the prior introduced in (b), find the posterior distribution $\pi(\theta \mid x_1, \ldots, x_n)$. Find also the Bayes solution δ_π, i.e., the minimiser of the Bayes risk $\mathrm{BR}(\delta, \theta) = \int R(\delta, \theta)\pi(\theta)\,\mathrm{d}\theta$. [xx introduce Bayes risk earlier xx].

(d) Tweak the parameters of the Beta prior distribution, so that the Bayes solution you found above equals the maximum likelihood estimator from (a). What desirable properties does the maximum likelihood estimator possess?

Ex. 8.16 *Estimating the standard deviation.* Suppose $X_1, \ldots, X_n$ are i.i.d. from $\mathrm{N}(0, \sigma^2)$. We are to estimate σ under the loss function

$$L(\delta, \sigma) = \frac{(\delta - \sigma)^2}{\sigma}. \tag{8.2}$$

(a) Find the maximum likelihood estimator, say $\widehat{\sigma}_{\mathrm{ml}}$, and show that its risk function is

$$R(\sigma, \widehat{\sigma}) = \sigma\{(n-1)/n + b_n^2 + (b_n - 1)^2\}, \quad \text{where} \quad b_n = \sqrt{\frac{2}{n}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.$$

You may now use Stirling's formula $\Gamma(z) = (2\pi/z)^{1/2}(z/e)^z$ to show that $b_n \to 1$, and $R(\sigma, \widehat{\sigma}) \to 2\sigma$ as $n \to \infty$, as we already knew from ML-theory (see Ex. xx in Chapter 5).

(b) Consider the prior distribution $\sigma \sim \pi(\sigma)$, whose density is

$$\pi(\sigma) \propto (1/\sigma)^{a+1} \exp(-b/\sigma^2).$$

with $a > 1$ and $b > 0$. Find the prior expectation of σ. Find also the prior expectation of $1/\sigma$.

(c) Find the posterior distribution $\sigma \mid x_1, \ldots, x_n$, and derive the Bayes solution under the loss function given in (8.2).

(d) Show that the maximum likelihood estimator is inadmissible by exhibiting an estimator, say $\delta^\star$, with uniformly smaller risk. *Hint*: Consider $\delta_\alpha = \alpha\widehat{\sigma}_{\mathrm{ml}}$.

(e) Is $\delta^\star$ admissible? *Hint*: Use Blyth's method.

Stein's phenomenon

Ex. 8.17 *The James–Stein estimator.* When estimating the price of apples in Oslo, the height of women in Bergen, and the unemployment rate in Trondheim, it is sometimes advantageous to use information about apples is Oslo and women in Bergen to say something about the unemployment rate in Trondheim. The point is that when estimating an ensemble of unrelated things, we can sometimes do better in the estimation by borrowing information across unrelated things. This phenomenon is known as Stein's paradox or the Stein effect. See Stein (1956); James and Stein (1961) for the original articles, and, for example Efron and Morris (1977) and Stigler (1990) for lucid presentations. In the present exercise we'll look at Stein's 1956–1961 result, a result that initiated a whole field of statistical research known as shrinkage estimation.

Let $Y_i \sim \mathrm{N}(\theta_i, 1)$ be independent for $i = 1, \ldots, p$ with $p \geq 3$. We are to estimate $\theta_1, \ldots, \theta_p$ under the combined loss function

$$L(\delta, \theta) = \sum_{i=1}^{p} (\delta_i - \theta_i)^2.$$

The standard approach is to use Y_i as an estimator of θ_i. The estimator Y_i is the maximum likelihood estimator, it is admissible under $(\delta_i - \theta_i)^2$, it is the uniformly minimum variance unbiased estimator, etc.

(a) For obvious reasons, we call $Y = (Y_1, \ldots, Y_p)$ the standard or the natural estimator. Compute its risk function.

(b) For a single $Y \sim \mathrm{N}(\theta, 1)$, show that under very mild conditions on the function $b(y)$, one has

$$E_\theta\,(Y-\theta)b(Y) = E_\theta\, b'(Y),$$

where b' is the derivative of b. *Hint*: Use integration by parts.

(c) Let now $b(y) = (b_1(y), \ldots, b_p(y))$. Generalise what you found in (b) to

$$\mathrm{E}_\theta\,(Y_i-\theta_i)b_i(Y) = \mathrm{E}_\theta\, b_{i,i}(Y),$$

where $b_{i,i}(y) = \partial b_i(y)/\partial y_i$.

(d) What you found in (b) and (c) is known as Stein's lemma. We are now going to use Stein's lemma to construct an estimator that uniformly dominates Y. Consider a general competitor to Y of the form $\delta(Y) = (\delta_1(Y), \ldots, \delta_p(Y))$, with

$$\delta_i(Y) = Y_i - b_i(Y). \tag{8.3}$$

Show that the difference in risk between Y and estimators of the form (8.3) can be expressed as

$$R(\delta,\theta) - R(Y,\theta) = \mathrm{E}_\theta\, D(Y),$$

where

$$D(y) = \sum_{i=1}^{p} \{b_i(y)^2 - 2b_{i,i}(y)\}.$$

Then $R(\delta,\theta) = p + \mathrm{E}_\theta\, D(Y)$. The fabulous thing about such a simple lemma as Stein's, is that $D(y)$ does not depend on the unknown $\theta_1, \ldots, \theta_p$. We can therefore try to find a data dependent function $b(y)$ such that $D(y) < 0$ for all y, and consequently an estimator that uniformly dominates the standard estimator. It turns out to be impossible to find such functions $b(y)$ when $p \le 2$, but it is possible for $p \ge 3$.

(e) Try $b_i(y) = ay_i/\|y\|^2$, with $\|y\|^2$ being the squared Euclidian norm $\sum_{i=1}^p y_i^2$, corresponding to

$$\delta(y) = y - b(y) = \big(1 - \frac{a}{\|y\|^2}\big)y.$$

With this choice of $b(y)$, show that

$$D(y) = \frac{1}{\|y\|^2}\{a^2 - 2a(p-2)\}.$$

Show that this is negative for a range of a values provided $p \ge 3$. Demonstrate that the optimal a is $a = p-2$, corresponding to the estimator

$$\delta_{\mathrm{JS}}(Y) = \big(1 - \frac{p-2}{\|Y\|^2}\big)Y. \tag{8.4}$$

This estimator is known as the James–Stein estimator. Show that the risk function of this estimator can be expressed as

$$R(\delta_{\text{JS}}, \theta) = p - (p-2)^2 \text{E}_\theta \frac{1}{\|Y\|^2}.$$

Show that the greatest reduction in risk from using δ_{JS} instead of Y takes place when $\theta_1 = \cdots = \theta_p = 0$, and compute the risk $R(\delta_{\text{JS}}, 0)$ in this point.

(f) We'll now make a connection to empirical Bayes procedures. Start with a prior that takes $\theta_1, \ldots, \theta_p$ independent from $\text{N}(0, \tau^2)$. Show that the Bayes solution is $\delta^B = (\delta_1^B, \ldots, \delta_p^B)$, with

$$\delta_i^B(Y) = \alpha\, Y_i, \quad i = 1, \ldots, p, \quad \text{where} \quad \alpha = \frac{\tau^2}{\tau^2 + 1}. \tag{8.5}$$

(g) The empirical Bayes approach consists of estimating hyperparameters from data. Hyperparameters are those parameters set by the statistician in a pure Bayesian approach. Show that the marginal distribution of $y_1, \ldots, y_p$ is a product of $\text{N}(0, 1 + \tau^2)$ distributions. Find the maximum likelihood estimator of α. Use the maximum likelihood estimator to find an unbiased estimator, say $\tilde{\alpha}$, of α. The empirical Bayes estimator is then $\delta_{\text{EB}}(Y) = \tilde{\alpha} Y$. What's noticeable about this estimator?

Ex. 8.18 *Resolving the paradox.* [xx make an exercise based on insights from Stigler (1990), perhaps?]

Ex. 8.19 *Poisson means and inadmissibility of ML-estimator.* To show that an estimator is inadmissible it suffices to showcase one estimator that dominates it. Let $Y_1, \ldots, Y_p$ be independent Poisson with means $\theta_1, \ldots, \theta_p$. We are to estimate the $\theta = (\theta_1, \ldots, \theta_p)$ under the loss function

$$L(\theta, \delta) = \sum_{i=1}^{p} \frac{(\delta_i - \theta_i)^2}{\theta_i},$$

where $\delta = (\delta_1, \ldots, \delta_p)$. The maximum likelihood estimator δ_{ml} takes $\delta_{\text{ml},i}(Y) = Y_i$ for $i = 1, \ldots, p$. Clevenson and Zidek (1975) showed that δ_{ml} is inadmissible by constructing an estimator, say δ_{CZ}, such that $R(\theta, \delta_{\text{CZ}}) < R(\theta, \delta_{\text{ml}})$ for all θ. In this exercise we derive this estimator [xx and try to show that it is admissible. xx]

(a) Let $Z = \sum_{i=1}^p Y_i$ be the sum of the p independent Poisson observations, write $\gamma = \sum_{i=1}^p \theta_i$ for the sum of the p Poisson means, and define $\pi_i = \theta_i/\gamma$ for $i = 1, \ldots, p$. Show that

$$(Y_1, \ldots, Y_n) \mid (Z = z) \sim \frac{z}{y_1! \cdots y_p!} \pi_1^{y_1} \cdots \pi_p^{y_p}.$$

This establishes that $\text{E}\,(Y_i \mid Z) = Z\pi_i$ and $\text{Var}(Y_i \mid Z) = Z\pi_i(1 - \pi)$ for $i = 1, \ldots, p$.

(b) Prove the following little lemma. If $X \sim \text{Poisson}(\theta)$ and g is a function such that $g(0) = 0$, then

$$\text{E}\, g(X)/\theta = \text{E}\, g(X+1)/(X+1).$$

(c) Consider the estimator $\delta^\star$ whose components are given by

$$\delta^\star(Y) = (1 - \phi(Z))Y_i, \quad \text{for } i = 1, \ldots, p.$$

The game to be played now (as with the James–Stein estimator of Exercise xx), is to find an expression for the risk difference $R(\theta, \delta^\star) - R(\theta, \delta_{\rm ml})$ that is independent of the unknown parameters. Using the results from (a) and (b) it is indeed the case that the risk difference $D(Z) = R(\theta, \delta^\star) - R(\theta, \delta_{\rm ml})$ can be expressed as

$$D(Z) = \mathrm{E}_\gamma \{[\phi(Z+1)^2 - 2\phi(Z+1)][(Z+1) + (p-1)] + 2\phi(Z)Z\}.$$

Derive this expression for $D(Z)$.

(d) Suppose that the function ϕ is such that $\phi(z)z$ is increasing. Under this assumption, find a function ϕ that ensures that $D(z) < 0$ for all $z \in \{0, 1, 2, \ldots\}$. The estimator $\delta(Y) = (1 - \phi(Z))Y$ with this function ϕ inserted is the estimator $\delta_{\rm CZ}$ of Clevenson and Zidek (1975) [xx fix, this is a class of estimators xx]. Conclude that the maximum likelihood estimator is inadmissible.

(e) We have shown that $\delta_{\rm CZ}$ uniformly [xx nytt begrep xx] dominates the maximum likelihood estimator, however, we do not yet now whether or not there exists and estimator that dominates $\delta_{\rm CZ}$. Show that $\delta_{\rm CZ}$ is admissible.

Hypothesis testing

Ex. 8.20 *Testing a simple hypothesis.* Let $X \sim f_\theta(x)$ and consider the simple hypothesis $H_0 \colon \theta = \theta_0$ versus the simple alternative $\theta = \theta_1$. The statistical tests ϕ, with $\phi(x) = 1$ meaning 'reject H_0, and $\phi(x) = 0$ 'keep H_0', are to be evaluated under the loss function

$$L(\phi, \theta_0) = \begin{cases} 0, & \text{if } \phi(x) = 0, \\ K_1, & \text{if } \phi(x) = 1, \end{cases} \qquad L(\phi, \theta_1) = \begin{cases} K_2, & \text{if } \phi(x) = 0, \\ 0, & \text{if } \phi(x) = 1. \end{cases}$$

(a) Let $0 < \pi_0 < 1$ be your prior probability of H_0 being true. Derive an expression for the posterior expected loss, and show that the Bayes solution ϕ_π is of the likelihood ratio type

$$\phi_\pi(x) = \begin{cases} 1, & \text{if } f(x \mid \theta_1) > k_\pi f(x \mid \theta_0), \\ 0, & \text{if } f(x \mid \theta_1) < k_\pi f(x \mid \theta_0). \end{cases}$$

Find k_π and relate this quantity to the level of a test.

(b) Let now $X \mid \theta$ be $\mathrm{N}(\theta, 1)$. We want to test $H_0 \colon \theta = 0$ versus $\theta_1 = 1/2$ using the Bayes solution when the prior is $\pi_0 = 1/2$. Find K_1 and K_2 such that $\mathrm{E}_{\theta_0} \phi_\pi(X) = 0.05$.

(c) Show that any Bayesian test with a prior giving weight to both the null- and the alternative hypothesis, is the most powerful test of its size. *Hint*: Use what you know about Bayes solutions and admissibility.

Density estimation

Ex. 8.21 *Unbiased estimation of a parametric density.* (xx earlier nils exercise from Ch3, not pushed to this Ch8. needs to be connected to sufficiency and completeness, perhaps to exponential family. xx) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from a parametric density $f(y, \theta)$, like the normal or the Gamma or the Beta. How can we construct an unbiased estimator of the density function itself? Assume there is a sufficient statistic here, say $T = T(Y_1, \ldots, Y_n)$.

(a) A very simple estimator for the window probability

$$p(\theta) = \Pr(Y \in [a, b]) = \int_a^b f(y, \theta)\, \mathrm{d}y$$

is $\widehat{p} = I(Y_1 \in [a, b])$, using very simply a single data point. Show that it is unbiased.

(b) This also invites the somewhat more intelligent estimator $\bar{p} = n^{-1} \sum_{i=1}^n I(Y_i \in [a, b])$, the binomial proportion of data points inside the $[a, b]$ window. Show that it is unbiased and find a formula for its variance.

(c) Typically this estimator can be beaten, however. Consider indeed

$$p^* = \mathrm{E}\,(\widehat{p}\,|\,T) = \Pr(Y_1 \in [a, b]\,|\,T).$$

Explain why this is actually an estimator, i.e. that it does not depend on the parameter θ, and that it is unbiased. Show also that the construction $\mathrm{E}\,(\bar{p}\,|\,T)$ leads to the very same p^*.

(d) Let $f_n(y\,|\,T)$ be the density of a Y_i given T. Explain why it does not depend on the parameter, and that

$$p^* = \int_a^b f_n(y\,|\,T)\, \mathrm{d}y, \quad \text{for all windows } [a, b].$$

(e) Show that $f_n(y\,|\,T)$ is unbiased, and also the minimum variance estimator among all siuch unbiased estimators.

(f) For each of the following parametric densities, find a formula for this minimum variance unbiased estimator for the density. (i) The $\mathrm{N}(\mu, 1)$. (ii) The $\mathrm{N}(0, \sigma^2)$. (iii) The two-parameter normal $\mathrm{N}(\mu, \sigma^2)$. (iv) The exponential $\theta \exp(-\theta y)$.

(g) (xx give them 25 data points from a normal, perhaps even a tiny real dataset. plot the different estimates of both $f(y)$ and of $\log f(y)$. convey the point that smallish differences and nuances are better picked up and seen on the log scale. xx)

Notes and pointers

[xx some notes and pointers here xx]

I.14

Bootstrapping

Part II

Stories

Part III

Appendix

III.A

Mini-primer on measure and integration theory

[xx Mini-primer on measures, probabilities on spaces, integration theory. Background for rest of the book. xx]

Chapter introduction

(xx mini intro to measure theory and integration, background for probability measures, distributions, densities, models, etc. we also explain in one paragraph that yes, these things matter, and without them we cannot work properly; on the other hand, for most of the work we do, also in later chapters, we do not need to think too much about it. it's also a matter of becoming *basically literate* in the probability language underlying theoretical and also applied statistics. xx)

Various aspects of probability theory, and hence statistics methodology, rest on the general theory of measure and integration. If all random variables we meet have nice distributions and densities, on regular domains, like an interval, the real line, or open subsets of Euclidean spaces, we can get pretty far without this underlying measure and integration theory. To formulate concepts in natural generality, and to develop tools and demonstrate basic properties for these, however, one needs this more general theory. In particular, the business of defining probabilities, for perhaps complicated events in not-so-standard spaces, demands theory beyond 'ordinary' integration.

Essentials of measure, integration, and probability

Ex. A.1 *Some set theory.* Let Ω be a set, and $A, B, A_1, A_2, \ldots$ subsets thereof. The *union* $A \cup B$ consists of the points $\omega \in \Omega$ that are in A or in B (or in both A and B). So $\cup_{n=1}^{\infty} A_n = \{\omega \in \Omega\colon \omega \in A_n \text{ for at least one } n\}$. The *intersection* $A \cap B$ consits of the points $\omega \in \Omega$ that are in A and in B, so $\cap_{n=1}^{\infty} A_n = \{\omega \in \Omega\colon \omega \in A_n \text{ for all } n\}$. The *complement* A^c consists of all the points $\omega \in \Omega$ that are not in A, $A^c = \{\omega \in \Omega\colon \omega \notin A\}$. The *set difference* $A \setminus B = A \cap B^c$, and the *symmetric difference* of A and B is $A \Delta B = (A \setminus B) \cup (B \setminus A)$. We say that A is a *subset* of B if all the elements of A are also elements of B, and denote this $A \subset B$. If $A \subset B$ and $B \subset A$, then $A = B$. The collection of all subsets of Ω, is called the *power set* and is denoted 2^{Ω}.

(a) Prove the *distributive laws* distributive laws

$$B \cap (\cup_{n=1}^{\infty} A_n) = \cup_{n=1}^{\infty} (B \cap A_n), \quad \text{and} \quad B \cup (\cap_{n=1}^{\infty} A_n) = \cap_{n=1}^{\infty} (B \cup A_n),$$

and *de Morgan's laws*, de Morgan's laws

$$(\cup_{n=1}^{\infty} A_n)^c = \cap_{n=1}^{\infty} A_n^c, \quad \text{and} \quad (\cap_{n=1}^{\infty} A_n)^c = \cup_{n=1}^{\infty} A_n^c.$$

(b) The *empty set*, denoted $\emptyset$, is the set with no element, think of it as $\emptyset = \{\}$. Write down a truth table with the columns $P\colon \omega \in \emptyset$, $Q\colon \omega \in A$, and 'if P then Q', to prove the vacuous truth that the empty set is a subset of any set.

(c) To any function $f\colon \Omega \to \mathcal{X}$, there is an associated inverse image. The inverse image, denoted f^{-1}, is a set mapping $f^{-1}\colon 2^{\mathcal{X}} \to 2^{\Omega}$, defined by

$$f^{-1}(B) = \{\omega \in \Omega\colon f(\omega) \in B\}, \quad \text{for } B \in \mathcal{X}.$$

Show that that for subsets $B, B_1, B_2, \ldots$ of $\mathcal{X}$, the inverse image f^{-1} preserves the set operations complement, union, and intersection in the sense that $f^{-1}(B^c) = (f^{-1}(B))^c$, $f^{-1}(\cup_{n=1}^{\infty} B_n) = \cup_{n=1}^{\infty} f^{-1}(B_n)$, and $f^{-1}(\cap_{n=1}^{\infty} B_n) = \cap_{n=1}^{\infty} f^{-1}(B_n)$.

(d) The sets Ω or $\mathcal{X}$ will often be the real line or a subset thereof. For $a < b$, the open interval is $(a, b) = \{x \in \mathbb{R}\colon a < x < b\}$, the closed interval is $[a, b] = \{x \in \mathbb{R}\colon a \le x \le b\}$, and so on. If $a = b$, then $(a, b) = \emptyset$ and $[a, b] = a$. Show that $(a, b) = \cup_{n=1}^{\infty} (a, b - 1/n]$, $[a, b] = \cap_{n=1}^{\infty} (a - 1/n, b]$, and that $\{a\} = \cap_{n=1}^{\infty} (a - 1/n, a] = \cap_{n=1}^{\infty} (a - 1/n, a + 1/n)$.

(e) A subset B of $\mathbb{R}$ is open if for any $x \in B$ you can fit a little ε-interval $(x - \varepsilon, x + \varepsilon)$ around x. Show that any open set in $\mathbb{R}$ is a countable union of open intervals. To do so, you may consider open intervals with rational endpoints, and recall that a countable union of countable sets, is countable.

(f) The *Cartesian product* of two sets A and B is the set $A \times B = \{(a, b)\colon a \in A, b \in B\}$, of three sets A, B and C it is $A \times B \times C = \{(a, b, c)\colon a \in A, b \in B, c \in C\}$, and so on. If $A = \{1, 2, 3\}$ and $B = \{4, 5\}$ what is $A \times B$? Make small sketches of the Cartesian products $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, and also of $\mathbb{R} \times \mathbb{N}$, and $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$. Cartesian product

Ex. A.2 *Measurable spaces.* Underlying all of statistics and probability is a mathematical model for randomness. This model consists of three things: a set Ω, called the *sample space*, containing all possible outcomes of the random phenomenon we are interested in; a family $\mathcal{A}$ of subsets of Ω, whose members are called *events*; and a function Pr having $\mathcal{A}$ as its domain and the unit interval as its range, called a *probability measure*.

(a) We start with a *measurable space*, say $(\Omega, \mathcal{A})$, consisting of a non-empty set Ω and a collection $\mathcal{A}$ of subsets of Ω. The subsets in $\mathcal{A}$ are later to be given values, perhaps probabilities, in terms of a measure. For $\mathcal{A}$, we demand that a σ-algebra of sets

(i) if $A \in \mathcal{A}$, then $A^c = \Omega \setminus A \in \mathcal{A}$;

(ii) if $A_1, A_2, \ldots$ are in $\mathcal{A}$, then $\cup_{j=1}^{\infty} A_j \in \mathcal{A}$.

A family of subsets $\mathcal{A}$ with these properties is called a σ-algebra. Thus, a σ-algebra is family of subsets that is closed under complements and countable unions. Show that if $\mathcal{A}$ is a σ-algebra, then Ω and the emptyset $\emptyset$ are in $\mathcal{A}$. Show that the power set 2^Ω is a σ-algebra. At the other extreme, show that the trivial σ-algebra $\{\emptyset, \Omega\}$ is a σ-algebra. And somewhat intermediately, show that if $A \subset \mathcal{A}$ is some nonempty set, then $\{\emptyset, A, A^c, \Omega\}$ is an σ-algebra. For a different type of example, consider $\mathcal{A}$, all subsets of $\mathbb{R}$ which are either empty, finite, or countably infinite, or whose complements are either empty, finite, or countably infinite. Show that $\mathcal{A}$ is a σ-algebra.

(b) Show that if $\mathcal{A}$ is a σ-algebra, with $B, A_1, A_2, \ldots \in \mathcal{A}$, then also $A_1 \cap A_2$, $A_1 \cap A_2 \cap A_3$, and even $\cap_{i=1}^\infty A_i$, is in $\mathcal{A}$. Show that sets like $A_1 \cap (A_2 \cup A_3 \cup A_4)^c \cap A_5$, $B \cap (\cup_{n=1}^\infty A_n)$, and $A \setminus B$ are in $\mathcal{A}$.

(c) Show that an intersection of σ-algebras must be a σ-algebra. Hence we may start by identifying a list of basis events, finite or infinite, say $\mathcal{B}_0$, and then define $\mathcal{B} = \sigma(\mathcal{B}_0)$, as *the smallest σ-algebra* containing all sets in $\mathcal{B}_0$. The σ-algebra $\mathcal{B}$ is said to be *generated* by $\mathcal{B}_0$, and $\mathcal{B}_0$ is called a generating family. Working with basis events that generate a σ-algebra is much more convenient than trying to somehow list all types of subsets of the σ-algebra.

the Borel-σ-algebra

(d) A famous and important example of a generated σ-algebra is the *Borel-σ-algebra on the real line*, denoted $\mathcal{B}(\mathbb{R})$, defined as the σ-algebra generated by all open intervals (a, b). Show that sets $\{a\}$, $[a, b]$, $[a, b)$, $(a, b]$, $(-\infty, b)$, $(-\infty, b]$, (a, ∞), $[a, \infty)$, as well as all countable unions and intersections of these, are then also in $\mathcal{B}(\mathbb{R})$. Show also that any of the families $\{(a, \infty)\colon a \in \mathbb{R}\}$, $\{[a, \infty)\colon a \in \mathbb{R}\}$, $\{(-\infty, a)\colon a \in \mathbb{R}\}$, and $\{(-\infty, a]\colon a \in \mathbb{R}\}$ also generates $\mathcal{B}(\mathbb{R})$.

(e) Similarly to the real case, we define $\mathcal{B}(\mathbb{R}^k)$, the Borel-$\sigma$-algebra on of $\mathbb{R}^k$, as the σ-algebra generated by all rectangles $(a_1, b_1) \times \cdots \times (a_k, b_k)$. Show that $\mathcal{B}^k$ then also must contain all closed rectangles $[a_1, b_1] \times \cdots \times [a_k, b_k]$, all rectangles of half-open sets $(a_1, b_1] \times \cdots \times (a_k, b_k]$, etc., and also all open sets of $\mathbb{R}^k$.

(f) Let $\mathcal{B}_0$ be the collection of all intervals of the form $(-\infty, q]$, with $q \in \mathbb{Q}$. Show that $\sigma(\mathcal{B}_0) = \mathcal{B}(\mathbb{R})$. A σ-algebra generated by a countable collection of sets is said to be separable. Let (Ω, d) be a separable metric space, i.e., Ω contains a countable and dense subset. Show that the smallest σ-algebra on Ω that contains all the open sets is a separable σ-algebra.

(g) If $(\Omega_1, \mathcal{F})$ and $(\Omega_2, \mathcal{G})$ are two measurable spaces, then we can construct the product space $(\Omega_1 \times \Omega_2, \mathcal{F} \otimes \mathcal{G})$, where the *product σ-algebra* $\mathcal{F} \otimes \mathcal{G}$ is the smallest σ-algebra on the Cartesian product $\Omega_1 \times \Omega_2$ that contains all sets of the form $F \times G$, where $F \in \mathcal{F}$ and $G \in \mathcal{G}$. Show that if $\Omega_1 = \Omega_2 = \mathbb{R}$ and $\mathcal{F} = \mathcal{G} = \mathcal{B}(\mathbb{R})$, then $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$. Generalise to dimension $k > 2$.

extended real numbers

(h) Plus and minus infinity will occur naturally as limits of sequences of real functions. Consider, for example, $f_n(x) = (\sqrt{n/2\pi}) \exp(-\frac{1}{2} n x^2)$ for $n = 1, 2, \ldots$, i.e., the density of the mean of n independent standard normals. Define the *extended real numbers* $\bar{\mathbb{R}} =$

$\mathbb{R} \cup \{-\infty, \infty\}$. Similarly, $\bar{\mathbb{R}}_+ = \mathbb{R} \cup \infty$. If we allow for functions taking values in $\bar{\mathbb{R}}$, then $f_n(x)$ converges pointwise to $f(x)$, where $f(x) = 0$ if $x \neq 0$, and $f(x) = \infty$ if $x = 0$. Show that the Borel-σ-algebra on $\bar{\mathbb{R}}$ is $\mathcal{B}(\bar{\mathbb{R}}) = \sigma(\mathcal{B}(\mathbb{R}), -\infty, \infty)$; or, equivalently, the σ-algebra generated by $\mathcal{B}_0 = \{(a, b), [-\infty, b), (a, \infty] \colon a, b \in \mathbb{R}\}$. Show similarly, that the Borel-σ-algebra on $\bar{\mathbb{R}}_+$ is $\sigma(\mathcal{B}(\mathbb{R}_+), \infty)$.

(i) In general, the Borel-σ-algebra on a set $\mathcal{X}$ is defined as the σ-algebra generated by all the open sets in $\mathcal{X}$, where open is defined in terms of a metric or a topology on $\mathcal{X}$. Consider the space $C[0, 1] = \{\text{all continuous functions } f \colon [0, 1] \to \mathbb{R}\}$, equipped with $d(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|$. The space $(C[0, 1], d)$ is a separable metric space (as we will see in Ex. 9.5). The open ε-ball around $g \in C[0, 1]$ is $B_\varepsilon(g) = \{f \in C[0, 1] \colon d(f, g) < \varepsilon\}$, and the Borel-$\sigma$-algebra on $C[0, 1]$ is the smallest σ-algebra containing all the open balls. Show that this σ-algebra is generated by $\cup_{\varepsilon \in \mathbb{Q} \cap (0,\infty)} \{B_\varepsilon(g_1), B_\varepsilon(g_2), \ldots\}$, where $\{g_1, g_2, \ldots\}$ is dense in $C[0, 1]$; and also by the collection of sets of the form $\{f \in C[0, 1] \colon (f(x_1), \ldots, f(x_n)) \in B_1 \times \cdots B_n\}$ where $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$ and $0 \leq x_1 < x_2 < \cdots < x_n \leq 1$.

Ex. A.3 *Measurable functions.* Let $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$ be two measurable spaces. A function $f \colon \Omega \to \mathcal{X}$ is *$\mathcal{A}/\mathcal{B}$-measurable* if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$. When it is clear what σ-algebras are involved, we simply say that f is measurable. When $\mathcal{X}$ is $\mathbb{R}$, $\bar{\mathbb{R}}$, or $\mathbb{C}$, $\mathcal{B}$ will always be the Borel-σ-algebra.

(a) In view of the efforts in Ex. A.2, the following lemma, tells much of the story of this exercise. Let $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$ be two measurable spaces, and suppose that $\mathcal{B}_0$ is a collection of subsets of $\mathcal{X}$ that generates $\mathcal{B}$. Then f is measurable if and only if $f^{-1}(B_0) \in \mathcal{A}$ for all $B_0 \in \mathcal{B}_0$. Look back at Ex. A.1(c), and prove this lemma.

(b) On a measurable space $(\Omega, \mathcal{A})$, let $f \colon \Omega \to \mathbb{R}$ be a measurable function. Show that $f^{-1}(a, b) \in \mathcal{A}$ for all open intervals (a, b), and that this might as well be taken as our definition of measurability for real valued functions. In particular, show that the sets $f^{-1}\{a\}$, $f^{-1}[a, b)$, $f^{-1}(a, b]$, $f^{-1}(-\infty, b)$ $f^{-1}(-\infty, b]$, $f^{-1}(a, \infty)$, and $f^{-1}[a, \infty)$ are all in $\mathcal{A}$. Generalise to extended real valued measurable functions $f \colon \Omega \to \bar{\mathbb{R}}$.

(c) Suppose that (Ω, d) is a a metric space equipped with its Borel-σ-algebra, and let $f \colon \Omega \to \bar{\mathbb{R}}$ be a continuous function. Show that f is measurable.

(d) We will frequently use the characteristic function, a function taking values in the complex plane. An open subset of $\mathbb{C}$ can be identified with the open subsets of $\mathbb{R}^2$. For example, and open box in $\mathbb{C}$ is $\{x + iy \colon a < x < b, c < y < d\}$. Therefore, any collection of subsets of $\mathbb{R}^2$ that generates $\mathcal{B}(\mathbb{R}^2)$ can, by making the appropriate identifications, be used to generate $\mathcal{B}(\mathbb{C})$. Let $f = g + ih$ be a complex valued function on the measurable space $(\Omega, \mathcal{A})$, for measurable functions $g, h \colon \Omega \to \mathbb{R}$. Show that f is measurable if and only if g and h are measurable. In particular, use (c) to show that function $\exp(ix) = \cos(x) + i \sin(x)$ is measurable.

(e) Let $f \colon \Omega \to \mathcal{X}$ be a measurable function between $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$. The σ-algebra generated by f is the smallest σ-algebra on Ω such that f is measurable, i.e., the intersection of all σ-algebras with respect to which f is measurable. We denote it $\sigma(f)$, and

clearly, $\sigma(f) \subset \mathcal{A}$. Show that $\sigma(f) = \{f^{-1}(B)\colon B \in \mathcal{B}\}$, and that if $\mathcal{X} = \bar{\mathbb{R}}$ and $\mathcal{B} = \mathcal{B}(\bar{\mathbb{R}})$, then $\sigma(f)$ may be generated by the collection of sets $\pi(f) = \{\{\omega \in \Omega\colon f(\omega) \le x\}\colon x \in \mathbb{R}\}$.

(f) First, let $f_1, \ldots, f_n$ be finitely many measurable functions. Show that $\max(f_1, \ldots, f_n)$ and $\min(f_1, \ldots, f_n)$ are measurable. Next, let $f_1, f_2, \ldots$ be a sequence of extended real valued measurable functions. Show that $\sup_{n\ge1} f_n$ and $\inf_{n\ge1} f_n$ are also measurable functions. Show that if $0 \le g_1 \le g_2 \le \cdots$ is a sequence of functions where $g_n(\omega)$ is nondecreasing (in n) for each ω, then the limit function g, with $g(\omega) = \lim_{n\to\infty} g_n(\omega)$, is also measurable. Next, show that $\limsup_{n\to\infty} f_n$ and $\liminf_{n\to\infty} f_n$ are measurable. Finally, show that, if it exists, the limit function $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ is a measurable function.

Simple functions

(g) A simple function is a function taking on only finitely many values, meaning that if g is a simple function it can be written $g = \sum_{j=1}^{k} c_j I_{A_j}$ for sets $A_1, \ldots, A_k \in \mathcal{A}$ such that $\cup_{j=1}^{k} A_j = \Omega$ and real constants $c_1, \ldots, c_k$. If the $A_1, \ldots, A_k$ are disjoint we say that g is a simple function on standard form. With f any nonnegative measurable function on $(\Omega, \mathcal{A})$, show that the sequence of simple functions

Approximation by simple functions

$$f_n(\omega) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{A_{n,k}}(\omega) + n I_{A_n}(\omega),$$

with $A_{n,k} = \{\omega \in \Omega\colon (k-1)/2^n \le f(\omega) < m/2^n\}$ and $A_n = \{\omega \in \Omega\colon f(\omega) \ge n\}$, is measurable, and is such that $0 \le f_1 \le f_2 \le \cdots$ and $f(\omega) = \lim_n f_n(\omega)$ for each ω.

(h) On $(\Omega, \mathcal{A})$, let $f, g\colon \Omega \to \mathbb{R}$ be measurable functions, and let a, b be constants. Show that $af + bg$, fg, and f/g if $g \neq 0$ are measurable.

(i) Let $f\colon \Omega \to \mathcal{X}$ and $g\colon \mathcal{X} \to \bar{\mathbb{R}}$ be measurable functions, on the measurable spaces $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{C})$, respectively. Show that the composition $g(f(\cdot))\colon \Omega \to \bar{\mathbb{R}}$ is measurable.

(j) In (i), it is important that g is $\mathcal{C}$-measurable. Let $\mathcal{C}'$ be a σ-algebra on $\mathcal{X}$, containing at least one set, for example C', that is not in $\mathcal{C}$. Define $g = I_{C'}$, and let $h = g(f(\cdot))$, with f the $\mathcal{A}/\mathcal{C}$-measurable function from (i). Show that h is not measurable. [xx can make this more concrete, see, e.g. Romano and Siegel (1986, p. 36) xx]

Ex. A.4 *Measure and measure spaces.*. A *measure* μ on a measurable space $(\Omega, \mathcal{A})$ is a function $\mu\colon \mathcal{A} \to [0, \infty]$, giving values to all sets A in $\mathcal{A}$, with the following properties:

(i) $\mu(\emptyset) = 0$, for the empty set;

(ii) $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$, for $A_1, A_2, \ldots$ disjoint sets in $\mathcal{A}$, .

The resulting triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*. We say that μ is a *finite measure* if $\mu(\Omega)$ is finite. Prime examples of finite measures are those with $\mu(\Omega) = 1$, such measures are *probability measures*, to be returned to in the rest of this book. If Ω can be represented as a countable union $\cup_{i=1}^{\infty} A_i$ of measurable sets, with each $\mu(A_i)$ finite, we say that μ is a σ-finite measure.

σ-finite measure

(a) Show that countable additivity implies finite additivity, namely $\mu(A_1 \cup \cdots \cup A_n) = \mu(A_1) + \cdots \mu(A_n)$ for any finite sequence of disjoint measurable sets. Show also that if $A, B \in \mathcal{A}$ and $A \subset B$, then $\mu(A) \leq \mu(B)$. Deduce that $\mu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$, for any sequence of sets $A_1, A_2, \ldots$ in $\mathcal{A}$.

(b) Suppose that $A_1 \subset A_2 \subset \cdots$ is a nondecreasing sequence of sets in $\mathcal{A}$. Show that $\nu(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \nu(A_n)$. Suppose now that $A_1 \supset A_2 \supset \cdots$ is a decreasing sequence in $\mathcal{A}$. Show that $\nu(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \nu(A_n)$, provided $\nu(A_n)$ is finite for some n. Continuity of measure

(c) Let ν be a *finitely additive* measure. That is, a function $\nu\colon \mathcal{A} \to [0, \infty]$, such that (i) $\nu(\emptyset) = 0$, and (ii) $\nu(A \cup B) = \nu(A) + \nu(B)$ for all disjoint sets $A, B \in \mathcal{A}$. Suppose that for any nondecreasing sequence $A_1 \subset A_2 \subset \cdots$ in $\mathcal{A}$ it is the case that $\nu(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \nu(A_n)$. Show that ν is a measure. Suppose that for any nonincreasing sequence $B_1 \supset B_2 \supset \cdots$ in $\mathcal{A}$ such that $\cap_n^{\infty} A_n = \emptyset$, we have $\lim_{n\to\infty} \nu(B_n) = 0$. Show that ν is a measure, provided $\nu(\Omega)$ is finite [xx check xx]. In other words, countably additivity is a continuity property in disguise.

(d) Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with no atoms, that is, $\mu(\{\omega\}) = 0$ for all $\omega \in \Omega$. Show that $\mu(A) = 0$ for any countable set $A \in \mathcal{A}$.

(e) Let μ be a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\mu((a, b)) = b - a$. Show that $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ has no atoms, and that $\mu([a, b]) = \mu((a, b]) = \mu([a, b)) = b - a$. This measure, called Lebesgue measure, is constructed in Ex. A.7.

(f) Consider the σ-algebra $\mathcal{A}$ of those subsets of $\mathbb{R}$ which are empty, or finite, or countably infinite, or complements of such sets. Let $\nu(A)$ be the simple measure which counts the number of elements in A. Show that it defines a measure, called the *counting measure*, and that it is not finite nor σ-finite. counting measure

(g) Consider the σ-algebra $2^{\mathbb{N}}$ of all subsets of $\mathbb{N} = \{1, 2, \ldots\}$. Let ν be the counting measure. Show that ν is σ-finite.

(h) Let $(\Omega, \mathcal{A}, \nu)$ be a measure space, $(\mathcal{X}, \mathcal{B})$ a measurable space, and $f\colon \Omega \to \mathcal{X}$ a measurable function. Define the set function $\nu f^{-1}(B) = \mu(f^{-1}(B))$ for $B \in \mathcal{B}$. Show that νf^{-1} is a measure on $\mathcal{B}$.

Ex. A.5 *π-systems, d-systems, monotone classes.* We often need to prove that a certain property of a measure, an integral, or the like, holds for all sets in a σ-algebra. Without further tools this is a tall order, simply because it can be exceedingly hard to give 'closed form' characterisations of all the elements of a σ-algebra. Think, for example, of the Borel-σ-algebra on the real line, how are you to describe all its elements without recourse to the family of sets that generates it? This is the motivation for the definitions and results we introduce in this exercise. We start with the definitions we need: Let $\mathcal{C}$ be a collection of subsets of a set Ω, then $\mathcal{C}$ is

- an algebra if it is closed under complements and finite unions;
- a π-system if $A \cap B \in \mathcal{C}$ for every $A, B \in \mathcal{C}$;

- a d-system if (i) $\Omega \in \mathcal{C}$; (ii) $A \subset B$ then $B \setminus A \in \mathcal{C}$ for $A, B \in \mathcal{C}$; and (iii) if $A_1 \subset A_2 \subset \cdots$ are in $\mathcal{C}$ then $\cup_{n=1}^{\infty} A_n \in \mathcal{C}$.
- a monotone class if (i) $A_1 \subset A_2 \subset \cdots$ are in $\mathcal{C}$ then $\cup_{n=1}^{\infty} A_n \in \mathcal{C}$, and (ii) if $A_1 \supset A_2 \supset \cdots$ are in $\mathcal{C}$ then $\cap_{n=1}^{\infty} A_n \in \mathcal{C}$;

(a) Show that the collection of sets $\pi(f)$ defined in A.3(e) is a π-system.

(b) Show that (i) an algebra is a π-system; (ii) a d-system is a monotone class; (iii) a σ-algebra is a π-system, a d-system, and a monotone class; (iv) a collection of sets that is both an algebra and a monotone class is a σ-algebra; (v) a collection of sets that is both a π-system and a d-system is a σ-algebra.

Dynkin's lemma

(c) If Π is a π-system, and $\mathcal{D}$ a d-system that contains Π, then $\sigma(\Pi) \subset \mathcal{D}$. In particular $\sigma(\Pi) = d(\Pi)$, where $d(\Pi)$ is the smallest d-system containing Π. In (d) we will get a glimpse of the power of this lemma. Here, we prove it, step by step. Step 1: Explain why it is enough to prove that $d(\Pi)$ is a π-system. Step 2: To prove that $d(\Pi)$ is a π-system, form the set $d(\Pi)_A = \{B \in d(\Pi) \colon A \cap B \in d(\Pi)\}$, and show that if $A \in d(\Pi)$, then $d(\Pi)_A$ is a d-system. Step 3: Show that if $A \in \Pi$, then $d(\Pi) \subset d(\Pi)_A$. Step 4: Show that if $A \in d(\Pi)$, then we still have that $d(\Pi) \subset d(\Pi)_A$. Step 4: Conclude from the above that $d(\Pi)$ is a π-system.

(d) Let ν and μ be measures on $(\Omega, \mathcal{A})$ such that $\nu(\Omega) = \mu(\Omega)$ is finite, and suppose ν and μ agree on a π-system that generates $\mathcal{A}$, i.e., $\mathcal{A} = \sigma(\Pi)$. Consider the collection of sets given by

$$\mathcal{D} = \{A \in \mathcal{A} \colon \nu(A) = \mu(A)\}.$$

Since $\Pi \subset \sigma(\Pi) = \mathcal{A}$ and $\nu(A) = \mu(A)$ for all $A \in \Pi$, it is clearly the case that $\Pi \subset \mathcal{D}$. Show that $\mathcal{D}$ is a d-system and conclude from (c) that $\nu = \mu$.

Monotone class theorem

(e) We now turn to a theorem that is quite similar to that proved in (c), the difference being that our basis events form an algebra, and an algebra is a π-system, but the reverse need not hold. Here is the theorem: Let Ω be a set, and $\mathcal{A}_0$ an algebra of subsets of Ω. Suppose that $\mathcal{M}$ is a monotone class of subsets of Ω and that $\mathcal{A}_0 \subset \mathcal{M}$. Then $\sigma(\mathcal{A}_0) \subset \mathcal{M}$. To prove this, we proceed in steps. Step 1: Let $m(\mathcal{A}_0)$ be the smallest montone class containing $\mathcal{A}_0$, and argue that it suffices to prove that $m(\mathcal{A}_0)$ is a σ-algebra. Step 2: Show that $m(\mathcal{A}_0)$ is closed under countable unions. Step 3: Form the set $\mathcal{M}_c = \{B \in m(\mathcal{A}_0) \colon B^c \in m(\mathcal{A}_0)\}$, argue why it is now sufficient to prove that $\mathcal{M}_c$ is a monotone class, and prove that $\mathcal{M}_c$ is indeed a monotone class.

(f) If $\mathcal{A}_0$ is an algebra, a natural question is how much bigger than $\mathcal{A}_0$ the σ-algebra it generates is. A partial answer to this question can be given by an application of the monotone class theorem from which we conclude that if $\mathcal{A}_0$ is an algebra, then every set in the σ-algebra generated by $\mathcal{A}_0$ can be approximated arbitrarily well by sets from $\mathcal{A}_0$. To make this precise, let $(\Omega, \mathcal{A}, \nu)$ be a finite measure space. Let $\mathcal{A}_0$ be an algebra contained in $\mathcal{A}$, and form the family of subsets

$$\mathcal{M} = \{A \in \mathcal{A} \colon \text{for any } \varepsilon > 0 \text{ there is an } A_0 \in \mathcal{A}_0 \text{ such that } \nu(A \Delta A_0) < \varepsilon\}.$$

Here, $A\Delta B = (A \setminus B) \cup (B \setminus A)$ is the symmetric difference of the sets A and B. Show that $\nu(\cup_{j=1}^{\infty} A_j \Delta \cup_{j=1}^{\infty} B_j) = \nu(\cap_{j=1}^{\infty} A_j^c \Delta \cap_{j=1}^{\infty} B_j^c) \leq \sum_{j=1}^{\infty} \nu(A_j \Delta B_j)$, which you may use to show that $\mathcal{M}$ is a monotone class, and appeal the monotone class theorem to conclude that given $A \in \sigma(\mathcal{A}_0)$ and $\varepsilon > 0$, we can find $A_0 \in \mathcal{A}_0$ so that $\nu(A \Delta A_0) < \varepsilon$.

(g) Let ν and μ be two finite measures on $(\Omega, \mathcal{A})$ such that $\nu = \mu$ on an algebra $\mathcal{A}_0$ contained in $\mathcal{A}$, and consider the collection of sets $\mathcal{M} = \{A \in \mathcal{A} \colon \mu(A) = \nu(A)\}$. Use a monotone class argument to show that $\nu = \mu$ on the σ-algebra generated by $\mathcal{A}_0$.

(h) The measures in (g) need not be finite. Suppose that $\Omega = \cup_{n=1}^{\infty} A_n$ for sets $A_1, A_2, \ldots$ in the algebra $\mathcal{A}_0$ on which μ and ν agree, and that $\mu(A_n)$ and $\nu(A_n)$ are finite for $n = 1, 2, \ldots$, then $\mu = \nu$ on $\sigma(\mathcal{A}_0)$. To prove this, fix C_n, show that collection of sets $\mathcal{M}_n = \{A \in \mathcal{A} \colon \mu(A \cap C_n) = \nu(A \cap C_n)\}$ forms a monotone class, and take it from there.

Ex. A.6 *Lifting good candidates to bona fide measures.* An important and useful general result from measure theory is Carathéodory's Extension Theorem. The point of this theorem is to lift a set function defined on simpler subsets of a set than those contained in a σ-algebra, to bona fide measures defined on the full σ-algebra. We start out with a set function $\nu_0 \colon \mathcal{A}_0 \to [0, \infty]$ working on subsets A_0 of an algebra $\mathcal{A}_0$ of subsets of a set Ω. Suppose that ν_0 has the properties (i) $\nu_0(\emptyset) = 0$; and (ii) if $A_1, A_2, \ldots$ is a disjoint sequence in $\mathcal{A}_0$ whose union happens to be in $\mathcal{A}_0$, then $\mu_0(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_0(A_n)$. Under these conditions, Carathéodory's Extension Theorem says that ν_0 on $\mathcal{A}_0$ can be lifted to a full measure ν on $\sigma(\mathcal{A}_0)$, with $\nu(A) = \nu_0(A)$ for all $A \in \mathcal{A}_0$. Also, if $\Omega = \cup_{n=1}^{n} A_n$ for sets $A_1, A_2, \ldots$ in $\mathcal{A}_0$ and $\mu_0(A_n) < \infty$ for all n, then this extension is unique.

Carathéodory's Extension Theorem

Notice that if $\mathcal{A}_0$ had been a σ-algebra, and not merely an algebra, then ν_0 would have been a measure. This is what distinguishes the present theorem from the results of Ex. A.5(d) and Ex. A.5(g), where ν and μ were assumed to be defined on σ-algebras, that is, they were assumed to be measures. In the extension theorem, by contrast, ν_0 has measure-like properties, but is only defined on an algebra.

Let us also mention the variations of Carathéodory's Extension Theorem where the function μ_0, instead of being defined on an algebra, as here, is defined on a semialgebra or on a semi-ring. A *semialgebra*, say $\mathcal{S}$, is a collection of subsets of a set Ω such that (i) if $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$; and (ii) if $A \in \mathcal{S}$, then $A^c = \cup_{j=1}^{n} B_j$ for $B_1, \ldots, B_n \in \mathcal{S}$. A *semi-ring*, say $\mathcal{R}$, is a collection of subsets of a set Ω such that (i) $\emptyset \in \mathcal{R}$; (ii) if $A, B \in \mathcal{R}$, then $A \cap B \in \mathcal{R}$; and (iii) if $A, B \in \mathcal{R}$, then $A \setminus B = \cup_{j=1}^{n} C_j$ for $C_1, \ldots, C_n \in \mathcal{R}$. The advantage with these two variations of Carathéodory's Extension Theorem is that natural basis events often constitute a semialgebra or a semi-ring, but not an algebra. For example, $\mathcal{S} = \{\text{all intervals on } \mathbb{R}\}$ is a semialgebra, but not an algebra; and $\mathcal{R} = \{(a, b] \colon a, b \in \mathbb{R}\}$ is a semi-ring, but not an algebra. An algebra, on the other hand, is both a semialgebra and a semi-ring. The reader might verify these claims. Assuming that the basis events constitute a semialgebra or a semi-ring, rather than an algebra, therefore amounts to imposing weaker conditions than we do here, and, as a consequence, the proof is more involved than with the algebra version of the extension theorem. We now prove the extension theorem, as stated above, through a string of exercises.

(a) Let $\mathcal{A}_0$ be an algebra of subsets of a set Ω, and let $\lambda\colon \mathcal{A}_0 \to [0,\infty]$ be such that $\lambda(\emptyset) = 0$. Consider the collection of elements of $\mathcal{A}_0$ that splits every element of $\mathcal{A}_0$ 'as it should', namely

$$\mathcal{A}_0^\lambda = \{B \in \mathcal{A}_0\colon \lambda(B\cap A) + \lambda(B^c \cap A) = \lambda(A) \text{ for all } A \in \mathcal{A}_0\}.$$

Show that $\mathcal{A}_0^\lambda$ is an algebra, and that λ is finitely additive on $\mathcal{A}_0^\lambda$.

(b) If $\mathcal{A}$ is a σ-algebra of subsets of Ω, the set function $\lambda\colon \mathcal{A} \to [0,\infty]$ is called an outer measure if

(i) $\lambda(\emptyset) = 0$;

outer-measure

(ii) $A, B \in \mathcal{A}$ with $A \subset B$, then $\lambda(A) \subset \lambda(B)$;

(iii) if $A_1 \subset A_2 \subset \cdots$ are sets in $\mathcal{A}$, then $\lambda(\cup_{j=1}^\infty A_j) \le \sum_{j=1}^\infty \lambda(A_j)$.

Carathéodory's lemma

Carathéodory's lemma (not the theorem yet) says that if λ is an outer measure on the measurable space $(\Omega, \mathcal{A})$, then the collection of elements of the σ-algebra $\mathcal{A}$ that splits every element of $\mathcal{A}$ 'as it should', that is

$$\mathcal{A}^\lambda = \{B \in \mathcal{A}\colon \lambda(B\cap A) + \lambda(B^c \cap A) = \lambda(A) \text{ for all } A \in \mathcal{A}\},$$

form a σ-algebra, and λ is countably additive on $\mathcal{A}^\lambda$, meaning that $(\Omega, \mathcal{A}^\lambda, \lambda)$ is a measure space. Let $B_1, B_2, \ldots$ be a sequence of sets in $\mathcal{A}^\lambda$, and set $B = \cup_{j=1}^\infty B_j$. To show that $\lambda(A) = \lambda(A\cap B) + \lambda(A \cap B^c)$ for any $A \in \mathcal{A}$, show that $\lambda(A) \le \lambda(A\cap B) + \lambda(A\cap B^c)$ and $\lambda(A) \ge \lambda(A\cap B) + \lambda(A\cap B^c)$. For the latter inequality, consider sets $C_n = \sum_{j=1}^n B_j$, and use that from (a), λ is finitely additive.

(c) We are now ready for the extension theorem, as stated in the introduction to this exercise, and prove this theorem in three steps. Step 1: For $G \in 2^\Omega$, define

$$\lambda(G) = \inf\{\sum_{j\ge 1} \mu_0(F_j)\colon F_1, F_2, \ldots \in \mathcal{A}_0 \text{ such that } G \subset \cup_{j\ge 1} F_j\},$$

and prove that λ is an outer measure. By (b), λ is a measure on $(\Omega, \mathcal{A}^\lambda)$, with $\mathcal{A}^\lambda$ defined as in (b) (so $\mathcal{A} = 2^\Omega$). This observation leads to the two final steps. Step 2: Show that $\mathcal{A}_0 \subset \mathcal{A}^\lambda$. Step 3: Show that $\lambda = \mu_0$ on $\mathcal{A}^\lambda$. We can then define μ to be the restriction of λ to $\mathcal{A} = \sigma(\mathcal{A}_0) \subset \mathcal{A}^\lambda$, with the inclusion coming from the already proven fact that $\mathcal{A}^\lambda$ is a σ-algebra.

(d) Point to Ex. A.5(h) to argue that if $\Omega = \cup_{n=1}^\infty A_n$ for sets $A_1, A_2, \ldots \in \mathcal{A}_0$ with $\mu(A_n) < \infty$ for $n = 1, 2, \ldots$, then the extension is unique.

Ex. A.7 *Lebesgue measure.* In this exercise we apply Carathéodory's extension theorem to the construction of Lebesgue measure on $\mathbb{R}$. The basic property of Lebesgue measure, say λ, is that for any interval

$$\lambda((a,b)) = \lambda([a,b]) = \lambda((a,b]) = \lambda([a,b)) = b - a,$$

the length of the interval in question. We want to have λ as a measure on the Borel-σ-algebra $\mathcal{B}(\mathbb{R})$, but it is not at all obvious that such a measure exists. Recall from

Ex. A.2 that the only way we mange to describe the elements of $\mathcal{B}(\mathbb{R})$ are via the various generating classes (e.g. all open intervals), so how are we then to describe $\lambda(B)$ for some arbitray set B in $\mathcal{B}(\mathbb{R})$? The solution is to define λ for a simpler class of subsets of $\mathbb{R}$, and then use Carathéodory's theorem to extend the domain of λ to all of $\mathcal{B}(\mathbb{R})$.

(a) Let $\mathcal{S}$ be the collection of all half-open intervals $(a, b]$ on $\mathbb{R}$ (if $a \geq b$, then $(a, b] = \emptyset$), as well as all infinite intervals of the form $(-\infty, b]$ and (a, ∞). Let $\mathcal{A}_0$ be a collection of sets consisting of all finite disjoint unions of elements of $\mathcal{S}$. Show that $\mathcal{A}_0$ is an algebra, and that $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{A}_0)$.

(b) We define $\lambda\colon \mathcal{A}_0 \to [0, \infty]$ to be such that $\lambda((a, b]) = b - a$ when a, b are finite, and to be ∞ when at least one of them are not. Thus, λ gives what we think of as the length of an interval. For disjoint intervals $I_1, \ldots, I_k \in \mathcal{S}$, define

$$\lambda(\cup_{j=1}^n I_j) = \lambda(I_1) + \cdots + \lambda(I_n),$$

so, in particular, if $I_j = (a_j, b_j]$ with $a_j, b_j \in \mathbb{R}$ for $j = 1, \ldots, n$, then $\lambda(\cup_{j=1}^n (a_j, b_j]) = \sum_{j=1}^n (b_j - a_j)$. Show that this is unambiguous, meaning that if $\cup_{j=1}^n I_j = \cup_{i=1}^m I_i'$, then $\lambda(\cup_{j=1}^n I_j) = \lambda(\cup_{i=1}^m I_i')$.

(c) Show that $\lambda(\emptyset) = 0$, and that λ is finitely additive on $\mathcal{A}_0$. In order to lift λ from the algebra $\mathcal{A}_0$ on which it is currently defined to the σ-algebra $\mathcal{B}(\mathbb{R})$ that $\mathcal{A}_0$ generates, we need to show that λ is *countably* additive on $\mathcal{A}_0$. This is the hard part. We start by showing that if $\cup_{j=1}^\infty (a_j, b_j] = (a, b]$ for disjoint intervals $(a_1, b_1], (a_2, b_2], \ldots$, then $\lambda(a, b] = \sum_{j=1}^n \lambda(a, b]$. To prove this equality, prove the inequality both ways. First, show that if $\cup_{j=1}^n (a_j, b_j] \subset (a, b]$, for disjoint $(a_1, b_1], \ldots, (a_n, b_n]$, then $\sum_{j=1}^n \lambda(a_j, b_j] \leq \lambda(a, b]$. Second, if $(a_1, b_1], (a_2, b_2], \ldots$ are disjoint, and $\cup_{j=1}^\infty (a_j, b_j] \subset (a, b]$, then $\sum_{j=1}^\infty \lambda(a_j, b_j] \leq \lambda(a, b]$. Now we get to the reverse inequality. Third, show that if $(a, b] \subset \cup_{j=1}^n (a_j, b_j]$, then $\lambda(a, b] \leq \sum_{j=1}^n (a_j, b_j]$. Fourth, if $(a, b] \subset \cup_{j=1}^\infty (a_j, b_j]$, then $\lambda(a, b] \leq \sum_{j=1}^\infty (a_j, b_j]$. To prove this fourth claim, note that for any $\varepsilon > 0$ smaller than $b - a$,

$$[a + \varepsilon, b] \subset (a, b] \subset \cup_{j=1}^\infty (a_j, b_j] \subset \cup_{j=1}^\infty (a_j, b_j + \varepsilon/2^j).$$

Thus, the closed and bounded set $[a+\varepsilon, b]$ is covered by the open sets $(a_1, b_1+\varepsilon/2), (a_2, b_2+\varepsilon/4), (a_3, b_3+\varepsilon/8), \ldots$, so by the Heine–Borel theorem it must then have a finite subcover, i.e., $[a + \varepsilon, b]$ is compact. Take it from here.

(d) Show that λ extends to a measure on Borel σ-algebra $\mathcal{B}(\mathbb{R})$, and that this extension, which is Lebesgue measure on the real line, is unique (see Ex. A.6(d)).

(e) Now that we have Lebesgue measure on the real line, we can construct Lebesgue measure on any subinterval of the real line, for example $([0, 1], \mathcal{B}[0, 1])$ or $([0, \infty), \mathcal{B}[0, \infty)$. Construct Lebesgue measure on these two measurable spaces.

(f) Establish similarly Lebesgue measure on $(\mathbb{R}^2, \mathcal{B}^2)$, i.e. on the plane, with its Borel sets, starting from the area of rectangles $\lambda((a_1, b_1) \times (a_2, b_2)) = (b_1 - a_1)(b_2 - a_2)$. Via the Carathéodory lifting, this gives rise to a well-defined way of measuring the area of any Borel subset A on the plane.

(g) Once the fundamental Lebesgue measure has been properly put on the map, it will be easy to define classes of others, via *cumulative distribution functions* and *densities*; see Ex. A.14 and A.20 below. It is nevertheless useful to go through direct arguments, resembling those for the Lebesgue measure itself, for a few concrete instances. Do this for the measures μ and ν on the positive halfline, starting with respectively $\mu(a,b) = \log(b/a)$ and $\nu(a,b) = b^2 - a^2$, for intervals (a,b).

Ex. A.8 *Almost surely and infinitely often.* Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. A property is said to hold μ almost surely, or μ-a.s., or simply a.s. if there is no confusion about the underlying measure, if it holds for for all ω outside of a set of μ-measure zero.

(a) Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ be a measure space with λ Lebesgue measure. Look back at Ex. A.4(e) to show that the indicator function $I_{\mathbb{R}\setminus\mathbb{Q}} = 1$ almost surely.

(b) From Ex. A.2(h), recall the sequence $f_n(x) = (\sqrt{2\pi/n})^{-1} \exp(-\frac{1}{2}nx^2)$ for $n = 1, 2, \ldots$, defined on the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$. Show that $f_n \to 0$ almost surely.

(c) Suppose that the sequence of measurable functions $f_1, f_2, \ldots$ converges to f almost surely. Show that f is measurable. Compare this to what you showed in Ex. A.3(f).

(d) Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $A_1, A_2, \ldots$ a sequence of sets in $\mathcal{A}$. Show that $\limsup_{n\to\infty} A_n = \emptyset$ if and only if $\lim_{n\to\infty} \mu(\cup_{m=n}^{\infty} A_m) = 0$.

(e) Let $A_1, A_2, \ldots$ be a sequence of sets, and $I_{A_1}, I_{A_2}, \ldots$ the corresponding sequence of indicators functions. Show that $\liminf_{n\to\infty} I_{A_n}(\omega) = 1$ if and only if $\omega \in \cup_{n\geq 1} \cap_{m\geq n} A_m$. Show also that $\limsup_{n\to\infty} I_{A_n}(\omega) = 1$ if and only if $\omega \in \cap_{n\geq 1} \cup_{m\geq n} A_m$. These two equivalences motivate the definitions

$$\liminf_{n\to\infty} A_n = \cup_{n\geq 1} \cap_{m\geq n} A_m, \quad \text{and} \quad \limsup_{n\to\infty} A_n = \cap_{n\geq 1} \cup_{m\geq n} A_m.$$

Show that these sets are measurable provided the sets $A_1, A_2, \ldots$ are. In probability and statistics one encounters the notion of something occurring *infinitely often*, or i.o. This notion is defined by

$$\begin{aligned} A_{\text{i.o.}} &= \limsup_{n\to\infty} A_n = \cap_{n\geq 1} \cup_{m\geq n} A_m \\ &= \{\omega \in \Omega\colon \text{for every } n \text{ there is an } m = m(\omega) \geq n \text{ such that } \omega \in A_m\} \\ &= \{\omega \in \Omega\colon \omega \in A_n \text{ for infinitely many } n\}. \end{aligned}$$

Borel–Cantelli lemma

(f) Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and assume that $A_1, A_2, \ldots \in \mathcal{A}$ are such that $\sum_{n=1}^{\infty} \mu(A_n) < \infty$. Show that $\mu(A_{\text{i.o.}}) = 0$. In Ex. A.19 we will study several illustration of the use of this lemma, and also see that it has a partial converse.

Fatou's lemma for sets

(g) Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and let $A_1, A_2, \ldots \in \mathcal{A}$. Show that

$$\mu(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} \mu(A_n).$$

This inequality, known as Fatou's lemma, holds not only for sequences of indicator functions, but, as we will see in Ex. A.11(b), for any sequence of nonnegative measurable functions.

Ex. A.9 *Convergence in measure/probability.* Let $f, f_1, f_2, \dots$ be measurable functions on a measurable space $(\Omega, \mathcal{A}, \mu)$. The sequence $f_1, f_2, \dots$ converges to f *in measure*, if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mu(\{\omega \in \Omega \colon |f_n(\omega) - f(\omega)| \geq \varepsilon\}) = 0.$$

If $\mu(\Omega) = 1$, so that μ is a probability measure, then convergence in measure is called convergence in probability.

convergence in probability

(a) If $f_n \to f$ almost surely, then $f_n \to f$ in measure. To show this, consider the sets $A_n = \{\omega \in \Omega \colon |f_n(\omega) - f(\omega)| \geq \varepsilon\}$ for some $\varepsilon > 0$, show that $\cap_{n\geq 1} \cup_{m\geq n} A_n \subset A$, where A is the set where $f_n(\omega)$ does not converge to $f(\omega)$, and take it from there.

(b) Show that if $f_n \to f$ in measure, then there is a subsequence f_{n_j} such that $f_{n_j} \to f$ almost surely. To construct such a subsequence, you may use the Borel–Cantelli lemma from Ex. A.8(f).

(c) Let $([0,1], \mathcal{B}, \lambda)$ be the unit interval with the Borel-σ-algebra and the Lebesgue measure. Divide the unit interval in $2, 3, \dots$ pieces: $A_1 = [0, 1/2]$, $A_2 = (1/2, 1]$, $A_3 = [0, 1/3]$, $A_4 = (1/3, 2/3]$, $A_5 = (2/3, 1]$, and so on. Define the function $f_n(x) = I(x \in A_n)$. Show that $f_n \to 0$ in measure, but not almost surely.

Ex. A.10 *The Lebesgue integral.* After having defined measures and measurable functions, the next goal is to form a well-defined *integral*, say $\int f \, d\nu = \int f(\omega) \, d\mu(\omega)$, with $(\Omega, \mathcal{A}, \mu)$ a measure space, and with $f \colon \Omega \to \bar{\mathbb{R}}$ a measurable function.

(a) We start with f a nonnegative and simple function on standard form, that is, f is on the form $f = a_1 I_{A_1} + \cdots + a_k I_{A_k}$ for some $n \geq 1$, with disjoint sets $A_1, \dots, A_k \in \mathcal{A}$ such that $\cup_{j=1}^k A_k = \Omega$, and nonnegative real constants $a_1, \dots, a_n$. We then define the integral of f as

$$\int f \, d\mu = a_1 \mu(A_1) + \cdots + a_k \mu(A_k). \tag{A.1}$$

If any of these sets have infinite measure and coefficient zero, we follow the measure theoretical convention that $0 \cdot \infty = 0$. Show that (A.1) is unambiguous, giving the same value for different representations of the same simple function, i.e., show that if $g = b_1 I_{B_1} + \cdots + b_n I_{B_n}$ is another nonnegative simple function on standard form such that $g(\omega) = f(\omega)$ for all $\omega \in \Omega$, then $\int g \, d\mu = \int f \, d\mu$.

(b) Show that $\int f \, d\mu$, as defined in (A.1), is linear and nondecreasing. That is, for nonnegative and simple functions on standard form f and g, and nonnegative real constants a, b, we have $\int (af + bg) \, d\mu = a \int f \, d\mu + b \int g \, d\mu$; and if $f \leq g$, then $\int f \, d\mu \leq \int g \, d\mu$.

(c) Next, we extend the integral to any nonnegative measurable function $f \colon \Omega \to \bar{\mathbb{R}}_+$. To do so, pick a sequence of nonnegative simple functions $0 \leq f_1 < f_2 < \cdots$ such that $f_n \to f$, the existence of which is guaranteed by Ex. A.3(g), and define

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu.$$

We need to prove that $\int f \, d\mu$ is independent of the approximating sequence f_n. To do this, let $0 \leq g_1 < g_2 < \cdots$ be some other sequence of nonnegative simple functions on

standard form converging to f, and show, first, that $\lim_{n\to\infty} \int g_n \, d\mu \le \int f \, d\mu$. Second, consider the sets $A_n = \{\omega \in \Omega \colon g_n(\omega) \ge f_k(\omega)\}$, argue that $\liminf_{n\to\infty} I_{A_n} f_k = f_k$, and use Fatou's lemma (Ex. A.8(g)) to show the reverse inequality, $\lim_{n\to\infty} \int g_n \, d\mu \ge \int f \, d\mu$.

(d) Let's look at some of the properties of the integral defined in (d). Let $f, g \colon \Omega \to \bar{\mathbb{R}}_+$ be measurable, and show that the integral is (i) *linear* $\int (af+bg) \, d\mu = a \int f \, d\mu + b \int g \, d\mu$; and (ii) *nondecreasing*, if $f \le g$ almost surely, then $\int f \, d\mu \le \int g \, d\mu$. Show also that (iii) $\int f \, d\mu = 0$ if and only if $f = 0$ almost surely; (iv) if $f > 0$ almost surely, then $\int f \, d\mu > 0$; (v) if $f = g$ almost surely, then $\int f \, d\mu = \int g \, d\mu$; (vi) if $\int_A f \, d\mu = \int_A g \, d\mu$ for all $A \in \mathcal{A}$ and μ is σ-finite, then $f = g$ almost surely; and (vii) if $\int f \, d\mu < \infty$, then $f < \infty$ almost surely.

(e) Now, we extend the integral to measurable functions $f \colon \Omega \to \bar{\mathbb{R}}$ taking on positive and negative values. This requires some more care, for if a nonnegative function f equals ∞ on a set of positive measure, then we can set $\int f \, d\mu = \infty$, but this does not work for functions taking on really large positive and really small negative values, as we may end up in $\infty - \infty$ type trouble. Therefore, the integral of functions taking values in $\bar{\mathbb{R}}$ is only defined for those functions that are *integrable*. For a fully general measurable f, we may represent it as $f = f_+ - f_-$, where $f_+ = \max(f, 0)$ and $f_- = \max(-f, 0)$ are two nonnegative measurable functions (see Ex. A.3(f)). We say that f is integrable if $\int |f| \, d\mu$ is finite. For integrable functions f, we then define

$$\int f \, d\mu = \int f_+ \, d\mu - \int f_- \, d\mu.$$

Show that f is integrable if and only if $\int f_+ \, d\mu$ and $\int f_- \, d\mu$ are finite. Show also that if μ is a finite measure, and f is bounded, then f is integrable.

(f) If A is measurable and f is a measurable function, show that fI_A is measurable too. We can hence define $\int_A f \, d\mu = \int I_A f \, d\mu$, where, in the case that f takes both positive and negative values, we require that $I_A f$ is integrable.

(g) Assume that $f, g \colon \Omega \to \bar{\mathbb{R}}$ are two integrable, measurable functions, and let $a, b \in \mathbb{R}$ be constants. Show that $af + bf$ is measurable and integrable, and that

(i) $\int (af + bg) \, d\mu = a \int f \, d\mu + b \int g \, d\mu$;
(ii) if $f \le g$, then $\int f \, d\mu \le \int g \, d\mu$;
(iii) $|\int f \, d\mu| \le \int |f| \, d\mu$;

(h) Some more properties of the integral defined in (e). Let $f, g \colon \Omega \to \bar{\mathbb{R}}$ be measurable functions, and $A, B \in \mathcal{A}$. Show that (i) if $A \subset B$ are measurable sets, then $\int_A f \, d\mu \le \int_B f \, d\mu$; (ii) if $\mu(A) = 0$, then $\int_A f \, d\mu = 0$; (iii) if f is integrable and $f = 0$ almost surely, then $\int f \, d\mu = 0$; (iv) if $f = g$ almost surely, then $\int f \, d\mu = \int g \, d\mu$; and (v) if f and g are integrable and $\int_A f \, d\mu = \int_A g \, d\mu$ for all $A \in \mathcal{A}$, then $f = g$ almost surely.

(i) Let $f = g + ih$ be a measurable complex valued function defined on the measurable space $(\Omega, \mathcal{A}, \mu)$, where $g, h \colon \Omega \to \mathbb{R}$ are measurable functions (see Ex. A.3(d)). Since $|f| \le |g| + |h|$, we see that f is integrable, i.e., $\int |f| \, d\mu < \infty$, if g and h are. For an

integrable $f\colon \Omega \to \mathbb{C}$ we define $\int f\,\mathrm{d}\mu = \int g\,\mathrm{d}\mu + i\int h\,\mathrm{d}\mu$. Show that linearity of the integral, (g)(i), extends to the integral of complex valued functions, where now $a, b \in \mathbb{C}$; and that the inequality $|\int f\,\mathrm{d}\mu| \le \int |f|\,\mathrm{d}\mu$ also holds for complex valued functions.

Ex. A.11 *Convergence theorems.* One of the main objectives of the integration theory developed in the preceding exercises is to find general criteria for when $\lim_{n\to\infty}\int f_n\,\mathrm{d}\mu = \int \lim_{n\to\infty} f_n\,\mathrm{d}\mu$. The theorems that give various sets of conditions for when we can pass the limit under the integral sign, are the convergence theorems of measure theory. Remember that the measure μ can be any measure on any measurable space, so the theorems that follow are very general, they will, for example, apply to sums $\sum_{j=1}^{\infty} f_n(j)$, as well as to Riemann integrals $\int f_n(x)\,\mathrm{d}x$. All the functions below are defined on a measure space $(\Omega, \mathcal{A}, \mu)$, and take values in $\bar{\mathbb{R}}$, $\bar{\mathbb{R}}_+$, or $\mathbb{C}$. Note that these theorems are often stated under the assumption that a sequence converges almost surely to some limit function. Here, however, we state these theorems with the weaker and statistically and probabilistically more applicable assumption, we think, that the convergence occurs in measure (or in probability of $\mu(\Omega) = 1$). The reader is free, perhaps even advised, to first prove the theorems under the an convergence a.s. assumption, and then extend the results to its in measure version.

(a) Suppose that $f_1, f_2, \ldots$, is a sequence of measurable functions such that $|f_n(\omega)| \le M$ for all ω and n, that μ is a finite measure, and that $f_n \to f$ in measure. Show that $\lim_{n\to\infty}\int f_n\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$. Bounded convergence

(b) Suppose $f_1, f_2, \ldots$ is a sequence of nonnegative measurable functions. Show that Fatou's lemma

$$\int \liminf_{n\to\infty} f_n\,\mathrm{d}\mu \le \liminf_{n\to\infty} \int f_n\,\mathrm{d}\mu.$$

Assume in addition that $f_n \to f$ in measure. Show that $\int f\,\mathrm{d}\mu \le \liminf_{n\to\infty}\int f_n\,\mathrm{d}\mu$.

(c) Suppose that $f_1, f_2, \ldots$ is an nondecreasing sequence of nonnegative measurable functions such that that $f_n \to f$ in measure. Show that $f_n \to f$ almost surely, and that $\lim_{n\to\infty}\int f_n\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$. Monotone convergence

(d) Let $f_1, f_2, \ldots$ be a sequence of measurable functions such that $f_n \to f$ in measure. Suppose there is a nonnegative integrable function g so that $|f_n(\omega)| \le g(\omega)$ almost surely for each n. Show that the limit function f is integrable, and $\lim_{n\to\infty}\int |f_n - f|\,\mathrm{d}\mu = 0$ and that $\lim_{n\to\infty}\int f_n\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$. Dominated convergence

(e) As a corollary to the Dominated convergence theorem, suppose that the dominating function g, instead of being merely integrable, is so that g^p is integrable for some $p \ge 1$. Show that then f^p is integrable, and $\lim_{n\to\infty}\int |f_n - f|^p\,\mathrm{d}\mu = 0$.

(f) We extend the Dominated convergence theorem to complex valued functions. That is, suppose that the $f_1, f_2, \ldots$ in (d) is a sequence of measurable complex valued functions (see Ex. A.10(i)), satisfying the conditions in (d). Show that the same conclusion holds.

(g) Suppose that $f_1, f_2, \ldots$, is a sequence of measurable complex valued functions such that $|f_n(\omega)| \le M$ for all ω and n, that μ is a finite measure, and that $f_n \to f$ in measure. Show that $\lim_{n\to\infty}\int f_n\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$.

Ex. A.12 *More properties of the integral/Applications of the convergence theorems.* In this exercise we apply the convergence theorems of Ex. A.11 to work out a few more properties of the Lebesgue integral and touch on a proof strategy that appears again and again when working with the Lebesgue integral, so often that the strategy sometimes goes by the name of a bootstrapping argument (not to be confused with bootstrapping in statistics). All the functions in this exercise are defined on a measurable space $(\Omega, \mathcal{A}, \mu)$.

(a) First, let $A_1 \subset A_2 \subset \cdots$ be sets in $\mathcal{A}$, and let $f\colon \Omega \to \bar{\mathbb{R}}$ be integrable over $\cup_{j=1}^{\infty} A_j$. Show that $\lim_{n\to\infty} \int_{A_n} f \,\mathrm{d}\mu = \int_A f \,\mathrm{d}\mu$ where $A = \cup_{j=1}^{\infty} A_j$. Second, let $B_1 \supset B_2 \supset \cdots$ be sets in $\mathcal{A}$, and assume that $f\colon \Omega \to \bar{\mathbb{R}}$ is integrable over B_1. Show that $\lim_{n\to\infty} \int_{B_n} f \,\mathrm{d}\mu = \int_B f \,\mathrm{d}\mu$, where $B = \cap_{n=1}^{\infty} B_n$.

(b) Let $f_1, f_2, \ldots$ be a sequence of extended real valued measurable functions. They are nonnegative only if explicitly mentioned. (i) If $f_n \geq 0$ for each n, show that $\int \sum_{j=1}^{\infty} f_j \,\mathrm{d}\mu = \sum_{j=1}^{\infty} \int f_j \,\mathrm{d}\mu$. Here, both sides are either finite or infinite. (ii) If $\sum_{j=1}^{\infty} f_j < \infty$ almost surely, and each partial sum is bounded by the same integrable function g, then f_n and each partial sum is integrable, and $\int \sum_{j=1}^{\infty} f_j \,\mathrm{d}\mu = \int \sum_{j=1}^{\infty} f_j \,\mathrm{d}\mu$. (iii) If $\sum_{j=1}^{\infty} \int |f_j| \,\mathrm{d}\mu < \infty$, then $\sum_{j=1}^{\infty} |f_j| < \infty$ almost surely, this sum is integrable, and $\int \sum_{j=1}^{\infty} f_j \,\mathrm{d}\mu = \int \sum_{j=1}^{\infty} f_j \,\mathrm{d}\mu$. See Ex. A.31(c) for important applications of these results.

(c) Let $f_1, f_2, \ldots$ be a sequence of measurable complex valued functions. Show that if $\sum_{n=1}^{\infty} \int |f_n| \,\mathrm{d}\mu < \infty$, then $\int \sum_{n=1}^{\infty} f_n \,\mathrm{d}\mu = \sum_{n=1}^{\infty} \int f_n \,\mathrm{d}\mu$.

(d) Deduce from (b) that if f is a nonnegative measurable function, then $\nu(A) = \int_A f \,\mathrm{d}\mu$ for $A \in \mathcal{A}$ defines a measure on $(\Omega, \mathcal{A})$. One says that f is the density of ν with respect to μ, and write $f = \mathrm{d}\nu/\mathrm{d}\mu$. Show that if $\mu(A) = 0$ then $\nu(A) = 0$; one says that ν is absolutely continuous with respect to μ, a property denoted $\nu \ll \mu$. Notice also that since ν is a measure, the properties of measures studied in Ex. A.4 and Ex. A.5 carry over to integrals of nonnegative functions. We treat densities in more details in Ex. A.20.

(e) Let $\nu(A) = \int_A f \,\mathrm{d}\mu$ be the measure introduced in (d). We want to show that that

$$\int g \,\mathrm{d}\nu = \int gf \,\mathrm{d}\mu, \tag{A.2}$$

bootstrapping argument

for any measurable function g. To do so, we emply a *bootstrapping argument*. It goes like this: First, prove (A.2) for indicator functions $g = I_A$. Second, extend (A.2) to simple functions $g = \sum_{j=1}^{k} a_j I_{A_j}$. Third, use Ex. A.2(g) and the monotone convergence theorem, and deduce that (A.2) holds for measurable functions $g\colon \Omega \to \bar{\mathbb{R}}_+$. Finally, using linearity again, show that (A.2) holds for all measurable functions $g\colon \Omega \to \bar{\mathbb{R}}$, provided g is integrable with respect to ν.

(f) Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, $(\mathcal{X}, \mathcal{B})$ a measurable space, and $f\colon \Omega \to \mathcal{X}$ a measurable function. From Ex. A.4(h) we know that $\mu f^{-1}\colon \mathcal{B} \to [0, \infty]$ is a measure.

change of variable

Let $g\colon \mathcal{X} \to \mathbb{R}$ be integrable with respect to μf^{-1}. Show that

$$\int_{f^{-1}(B)} g(f(\omega)) \mathrm{d}\mu(\omega) = \int_B g(x) \,\mathrm{d}(\mu f^{-1})(x).$$

You may once again use a bootstrapping argument to show this. For the measurability of $g(f(\cdot))$, see Ex. A.3(i).

(g) Here is a lemma whose importance in Chapter 5 cannot be exaggerated, and whose proof employs the dominated convergence theorem. Let $(\mathcal{Y}, \mathcal{A})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be measurable spaces. Suppose that $g\colon \mathcal{Y} \times \mathbb{R} \to \bar{\mathbb{R}}$ is such that $y \to g(y, \theta)$ is measurable and integrable for each $\theta \in (a, b) \subset \mathbb{R}$, and that $\partial g(y, \theta)/\partial\theta$ exists for all $\theta \in (a, b)$. Show that $\partial g(y, \theta)/\partial\theta$ is measurable. Suppose further that there is an integrable function $h(y)$ such that $|\partial g(y, \theta)/\partial\theta| \leq h(y)$ for all $\theta \in (a, b)$ and all y, combine the fundamental theorem of calculus and the dominated converge theorem to show that

Derivative under the integral sign

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int g(y, \theta)\, \mathrm{d}\mu(y) = \int \frac{\partial}{\partial\theta}\, g(y, \theta)\, \mathrm{d}\mu(y), \quad \text{for } \theta \in (a, b),$$

that is, we can pass the derivative under the integral sign.

(h) Suppose $f\colon [a, b] \to \mathbb{R}$ is a Riemann integrable function. Show that f is Lebesgue measurable, and that its classical Riemann-definition integral $\int_a^b f(x)\, \mathrm{d}x$ coincides with the more general integral we've worked with in this exercise, $\int_a^b f(x)\, \mathrm{d}\lambda(x) = \int_a^b f\, \mathrm{d}\lambda$, with λ the Lebesgue measure defined on the Lebesgue subsets of $[a, b]$.

Riemann and Lebesgue

Ex. A.13 *Probability spaces.* Mathematically speaking, we are free to define the basics of probabilities, along with axioms these should satisfy, without yet tying these to the so-called real world. So let us define a *probability space* as a triple $(\Omega, \mathcal{A}, P)$, where Ω is a set; $\mathcal{A}$ a σ-algebra of subsets of Ω; and $P\colon \mathcal{A} \to [0, 1]$ a *probability measure*, defined simply to be a measure, in the sense of Ex. A.4, with full measure $P(\Omega) = 1$.

a probability space

a probability measure

We may envisage P as a probability machine, assessing to each A a probability $P(A)$. Such a probability measure on $(\Omega, \mathcal{A})$ has axiomatic properties following those of more general measures, given in Ex. A.2, and for convenience stated again here, for the present case of $P(\Omega) = 1$. We demand

axioms for a probability space

(i) that $P(\emptyset) = 0$;
(ii) that $\Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i)$ for are disjoint sets $A_1, A_2, \ldots$ in $\mathcal{A}$;
(iii) and that $P(\Omega) = 1$.

The subsets A can be given several names, including *events*; the conceptual idea is that we do not yet know whether a certain A occurs or not, but we can give it a probability.

(a) For all events A and B show that $\Pr(A \setminus B) = \Pr(A) - \Pr(A \cap B)$; that $\Pr(A) = 1 - \Pr(A^c)$; and that $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Generalise this latter formula to the union of three events, and then try to generalise it to the union of four or more events.

(b) If A and B have probabilities 0.95, or more, show that $\Pr(A \cap B) \geq 0.90$. Generalise. This simple lower-bounding of certain types of probabilities is sometimes called the Bonferroni method, or Bonferroni correction.

(c) From Ex. A.4(b) we know that if $A_1 \subset A_2 \subset \cdots$, then $\Pr(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \Pr(A_n)$; and, secondly, if $A_1 \supset A_2 \supset \cdots$, then $\Pr(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \Pr(A_n)$. Show that either of these two statements could replace (ii) in the axiom list above.

(d) Above we have been careful to define probability measures P for large collections of events, namely σ-algebras, but also avoiding defining $P(A)$ for *every* subset A. Attempting to do that, in various natural spaces, will lead to difficulties and incoherencies, related to the existence of non-measurable sets. These issues are not present when the full space Ω is finite, however, as one can simply allow *every subset* to be included, in the set of subsets for which a probability is attached. Show indeed that if $\Omega = \{\omega_1, \ldots, \omega_m\}$, with perhaps a large m, and these singletons are attached probabilities $p_1, \ldots, p_m$ (non-negative, with sum 1), then

$$\Pr(A) = \sum_{j\colon \omega_j \in A} p_j, \quad \text{for any subset } A,$$

defines a probability measure on $(\Omega, \mathcal{A})$, where, in this case, $\mathcal{A} = 2^\Omega$ the set of all 2^m subsets (which, from Ex. A.2(a), we know is a σ-algebra). Generalise to the case of countably big spaces, say $\Omega = \{\omega_1, \omega_2, \ldots\}$, with pointmasses $p_1, p_2, \ldots$ summing to 1. In these cases the collection $\mathcal{A}$ of *all* subsets is the natural set of events.

Ex. A.14 *Distribution functions.* Consider the case where the probability space is $(\mathbb{R}, \mathcal{B}, P)$, with $\mathcal{B} = \mathcal{B}(\mathbb{R})$ the Borel σ-algebra on the real line, and P is some probability measure on this measurable space. For such a P, define the *cumulative distribution function* (c.d.f. for short) as

the c.d.f., the cumulative distribution function

$$F(t) = \Pr(A_t), \quad \text{with } A_t = (-\infty, t],$$

where we also allow the simpler notation $P(\infty, t]$ for $P((-\infty, t])$.

(a) Show that F is nondecreasing, right continuous, with $F(t) \to 1$ and $F(t) \to 0$ as $t \to \infty$ and $t \to -\infty$, respectively. Show also that $F(t - 1/n) \to \Pr(-\infty, t)$, and that

$$\Pr(a + 1/n, b - 1/n] = F(b - 1/n) - F(a + 1/n) \to F(b-) - F(a) = \Pr[a, b),$$

for all intervals (a, b). Here $F(b-)$ is notation for the limit of $F(b - \varepsilon)$ as $\varepsilon \to 0_+$, converging to zero from above, and is also the same as $P(-\infty, b)$.

(b) Show that $\Pr(\{t\})$, the probability assigned to the fixed point t, is $F(t) - F(t-)$. This probability is often zero, as is the case for all t if F is continuous. Show that the set D_F of discontinuities for F is at most countably infinite.

(c) Suppose P_1 and P_2 are two probability measures on $(\mathbb{R}, \mathcal{B})$, with the same c.d.f., i.e. $F_1 = F_2$. An important fact (to say the least) is that if $F_1 = F_2$, then indeed $P_1(A) = P_2(A)$ for *all* $A \in \mathcal{B}(\mathbb{R})$. Show this, from Carathéodory's Extension Theorem of Ex. A.6, or, alternatively from one of the theorems of Ex. A.5. Very conveniently, this allows one to define a full probability measure P by giving only its c.d.f., or its values for all intervals. For example, saying that $P(a, b) = \int_a^b (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)\, \mathrm{d}x$, for all intervals (a, b), is a sufficient description of the standard normal distribution; we don't need to give a more laborious recipe for how to compute $P(A)$ for more complicated events A.

(d) Suppose P is a probability measure on $(\mathbb{R}^2, \mathcal{B})$, where $\mathcal{B} = \mathcal{B}(\mathbb{R}^2)$ is the Borel-σ-algebra in the plane (see Ex. A.2(e)). Define the cumulative distribution function corresponding to P by

$$F(t_1, t_2) = P(A_{t_1,t_2}) = P((-\infty, t_1] \times (-\infty, t_2]).$$

Show that for any rectangle,

$$P((a_1, b_1] \times (a_2, b_2]) = F(a_1, a_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_2, b_2).$$

Use again Ex. A.5, or indeed Carathéodory Extension Theorem of Ex. A.6, to prove that if two probability measures are equal for all rectangles, then they are identical, i.e., giving the same probability to *any* Borel set. Thus a probability measure P on $(\mathbb{R}^2, \mathcal{B}^2)$ is fully determined by giving its $F(t_1, t_2)$ function.

(e) Attempt to generalise (d) to dimension k, i.e., to $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$; in particular, the probability attached to a rectangle $(a_1, b_1] \times (a_k, b_k]$ can be expressed as a sum of values of F computed at the 2^k vertices of the rectangle, with ± 1 signs, as seen above for $k = 2$.

(f) Let P be a probability measure such that the distribution function $F(x) = P(-\infty, x]$ (see Ex. A.14(c)) is montonically increasing. Suppose that g is a real valued function such that the Riemann–Stieltjes integral $\int_a^b g \, \mathrm{d}F$ exists. Show that $\int_{[a,b]} g \, \mathrm{d}P = \int_a^b g \, \mathrm{d}F$.

Ex. A.15 *Random variables.* Speaking mathematically, a *random variable* is a measurable function on a probability space. Measurable functions were defined in Ex. A.3, but let us repeat some of the details. With $(\Omega, \mathcal{A}, \Pr)$ a 'background' probability space, we construct random variables as measurable functions $X \colon \Omega \to \mathcal{X}$, where $(\mathcal{X}, \mathcal{B})$ is the measurable space where $X(\omega)$ lands. Measurability means that the inverse images $X^{-1}(B) = \{\omega \in \Omega \colon X(\omega) \in B\}$ are in $\mathcal{A}$ for any $B \in \mathcal{B}$. As is common, we often write $\{X \in B\}$ instead of the more cumbersome $\{\omega \in \Omega \colon X(\omega) \in B\}$, and $\Pr(X \in B)$ instead of $\Pr(\{\omega \in \Omega \colon X(\omega) \in B\})$.

(a) The probability distribution, distribution, or law, of a random variable X, say P, is defined by

probability distribution

$$P(B) = \Pr(X \in B) = \Pr(X^{-1}(B)), \quad \text{for } B \in \mathcal{B}.$$

With pedantic care, we define P via $(\Pr X^{-1})(B) = \Pr(X^{-1}(B))$. Even though this is just repeating Ex. A.4(h) with the extra requirement that $P(\mathcal{X}) = 1$, show that $P = \Pr X^{-1}$ indeed is a probability measure on $(\mathcal{X}, \mathcal{B})$.

(b) Often what matters is the distribution of X, rather than particularities of the background space. Indeed there may be different spaces $(\Omega_j, \mathcal{A}_j, \Pr_j)$ and random variables $X_j \colon \Omega_j \to \mathcal{X}$ inducing precisely the same distribution, i.e., the different $P_j = \Pr_j X_j^{-1}$ might be identical. For a given P on $(\mathcal{X}, \mathcal{B})$, show that the identity map $x \mapsto x$ is one such construction, leading to a random variable X with distribution P. In the case of $X_j \colon \Omega_j \to \mathbb{R}$, we have seen in Ex. A.14 that what matters is the c.d.f. $\Pr_j(X_j \le t) = P_j(-\infty, t] = F_j(t)$; as long as these are equal, the distributions $P_j = \Pr_j X_j^{-1}$ are identical. Give three separate such constructions of the standard normal distribution.

(c) If $X\colon \Omega \to \bar{\mathbb{R}}$ is a random variable, defined on a background probability space $(\Omega, \mathcal{A}, \Pr)$, its *mean*, or *expected value*, is defined as

the expectation

$$\mathrm{E}\, X = \int X \,\mathrm{d}\Pr = \int X(\omega)\,\mathrm{d}\Pr(\omega),$$

as long as this integral is finite, i.e., X is integrable. Since the expectation is a Lebesgue integral (as defined in Ex. A.10), the convergence theorems of Ex. A.11, as well as the properties of the Lebesgue integral derived in Ex. A.10(g)–(h) and Ex. A.12 apply. In particular, with $g\colon \mathbb{R} \to \bar{\mathbb{R}}$ any measurable function, deduce from Ex. A.12(f) that

$$\mathrm{E}\, g(X) = \int g(X(\omega))\,\mathrm{d}\Pr(\omega) = \int g(x)\,\mathrm{d}P(x), \quad \text{with } P = \Pr X^{-1},$$

provided g is P-integrable. In particular, only the distribution P of X matters, not the details associated with the background probability space.

(d) With the mean of a real random variable well defined, we may of course go on to other and higher moments. For a random variable $X\colon \Omega \to \bar{\mathbb{R}}$, as above, with $\xi = \mathrm{E}\, X$,

the variance

show that

$$\mathrm{E}\,(X-\xi)^2 = \int (X-\xi)^2 \,\mathrm{d}\Pr = \int_{-\infty}^{\infty} (x-\xi)^2 \,\mathrm{d}P(x) = \int_0^{\infty} y \,\mathrm{d}Q(y),$$

with Q the distribution of $Y = (X-\xi)^2$; so there's no ambiguity. This quantity is of course *the variance of* X, denoted $\mathrm{Var}\, X = \mathrm{E}\,(X-\xi)^2$. The square root of $\mathrm{Var}\, X$ is called the *standard deviation* of X.

(e) Consider integrable random variables $X, Y\colon \Omega \to \bar{\mathbb{R}}$ defined on the same background probability space $(\Omega, \mathcal{A}, \Pr)$. Show that $\mathrm{E}\,(aX + bY) = a\,\mathrm{E}\, X + b\,\mathrm{E}\, Y$ (see Ex. A.10(g)), and generalise . In particular, for integrable random variables $X_1, \ldots, X_n$, we have $\mathrm{E}\,(X_1 + \cdots + X_n) = \mathrm{E}\, X_1 + \cdots + \mathrm{E}\, X_n$, regardless of any dependencies between these variables.

(f) For variables with finite second moments, show that $\mathrm{Var}\, X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2$. Since the variance must be nonnegative, we get that $\mathrm{E}\, X^2 \geq (\mathrm{E}\, X)^2$. This is a nice reminder of what way the inequality goes in Jensen's inequality: $\mathrm{E}\, g(X) \geq g(\mathrm{E}\, X)$ whenever $g\colon \mathbb{R} \to \mathbb{R}$ is a convex function, i.e., a function such that $g(ax + (1-a)y) \leq ag(x) + (1-a)g(y)$ for all $0 \leq a \leq 1$ and all $x, y \in \mathbb{R}$. Prove it. Show also that $(\mathrm{E}\,|X|^r)^{1/r}$ is increasing in r.

Jensen's inequality

Ex. A.16 *Product measure and iterated integrals.* Let $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ be two σ-finite measure spaces. The measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ consists of the Cartesian product $\mathcal{X} \times \mathcal{Y}$ (see Ex. A.1(f)) and the σ-algebra $\mathcal{A} \otimes \mathcal{B}$ generated by the measurable rectangles $A \times B$, for $A \in \mathcal{A}$ and $B \in \mathcal{B}$. In this exercise we first construct a measure $\mu \times \nu$ on $\mathcal{A} \otimes \mathcal{B}$, such that $(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$ for all measurable rectangles $A \times B$. This measure is called the product measure. Second, we establish conditions under which we can compute double integrals by iterated integration,

product measure

$$\int f \,\mathrm{d}(\mu \times \nu) = \int \int f(x,y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x) = \int \int f(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y). \tag{A.3}$$

for measurable functions $f\colon \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$.

(a) Show that $\Pi = \{A \times B \colon A \in \mathcal{A}, B \in \mathcal{B}\}$ is a π-system generating $\mathcal{A} \otimes \mathcal{B}$.

(b) For (A.3) to make sense, we need a technical lemma. Assume that $f \colon \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is measurable, and that μ is σ-finite. Then (i) $x \mapsto f(x, y)$ is $\mathcal{A}$-measurable for each $y \in \mathcal{Y}$; and (ii) $\int_{\mathcal{X}} f(x, y)\, \mathrm{d}\mu(x)$ is $\mathcal{B}$-measurable. To prove (i) and (ii), assume that μ is finite, check (i) and (ii) for $f = I_C$ for $C \in \Pi$, then use Dynkin's lemma (Ex. A.5(c)) to show that (i) and (ii) holds for for $f = I_C$ for $C \in \mathcal{A} \otimes \mathcal{B}$. Third, use a bootstrapping argument, and, finally, extend what you have to μ being σ-finite.

(c) Due to our efforts in (b), it makes sense to define

$$\lambda(C) = \int \int I_C(x, y)\, \mathrm{d}\nu(y)\, \mathrm{d}\mu(x), \quad \text{for } C \in \mathcal{A} \otimes \mathcal{B}.$$

Show that λ is a measure, and argue that it is the unique measure such that $\lambda(A \times B) = \mu(A)\nu(B)$ on the π-system from (a). Conclude from this that (A.3) holds for indicator functions (so we could have defined λ with the order of integration reversed), and, indeed λ is the product measure $\mu \times \nu$.

(d) Combine what you found in (c) with the monotone convergence theorem (a boostrapping argument) to show that (A.3) holds for all measurable functions $f \colon \mathcal{X} \times \mathcal{Y} \to [0, \infty]$. Tonelli's theorem

(e) Suppose that $f \colon \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is integrable with respect to $\mu \times \nu$. Show that (A.3) holds also in this case. Notice that the sets $\{x \in \mathcal{X} \colon \int_{\mathcal{Y}} |f(x, y)|\, \mathrm{d}\nu(y) = \infty\}$ and $\{y \in \mathcal{Y} \colon \int_{\mathcal{Y}} |f(x, y)|\, \mathrm{d}\mu(x) = \infty\}$ have μ- and ν-measure zero, respectively (see Ex. A.10(d)), so you can modify f on these sets to avoid $\infty - \infty$ type trouble. Fubini's theorem

Ex. A.17 *Convolutions.* Let X and Y be independent real random variables with distributions P_X and P_Y, and cumulative distribution functions $F_X(x) = P_X(-\infty, x]$ and $F_Y(y) = P_Y(-\infty, y]$.

(a) For $B \in \mathcal{B}(\mathbb{R})$ write $B - x = \{b - x \colon b \in B\} = \{y \in \mathbb{R} \colon x + y \in B\}$, and use Ex. A.16(d) to show that

$$\Pr(X + Y \in B) = \int_{\mathbb{R}} P_Y(B - x)\, \mathrm{d}P_X(x) = \int_{\mathbb{R}} P_X(B - y)\, \mathrm{d}P_Y(y).$$

This defines the measure $P_Y * P_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, called the *convolution* of P_X and P_Y, i.e., $(P_X * P_Y)(B) = \int_{\mathbb{R}} P_Y(B - x)\, \mathrm{d}P_X(x)$, for $B \in \mathcal{B}(\mathbb{R})$. Note also that the convolution is commutative, i.e., $P_Y * P_X = P_X * P_Y$.

(b) Show that $Z = X + Y$ has cumulative distribution function

$$H(z) = \int F_Y(z - x)\, \mathrm{d}F_X(x) = \int F_X(z - y)\, \mathrm{d}F_Y(y).$$

Show also that if P_X and P_Y have densities, say f_X and f_Y, with respect to Lebesgue meaures on the real line, then the distribution of Z, namely $P_Y * P_X$, has density

$$h(z) = \int_{\mathbb{R}} f_Y(z - x) f_X(x)\, \mathrm{d}x = \int_{\mathbb{R}} f_X(z - y) f_Y(y)\, \mathrm{d}y,$$

also, as we see, with respect to Lebesgue measure. The density h is often denoted $(f_X * f_Y)(z)$, which is called the convolution of f_X and f_Y.

(c) (xx just a bit more, with X discrete and Y having a density. an illustration. also pointers to mgf things and CLT etc. xx)

(d) If $f, g\colon \mathbb{R} \to \mathbb{R}$ are two functions such that $\int g(z-x)f(x)\,\mathrm{d}x$ is finite for all z, we may also define the convolution $(g * f)(z) = \int g(z-x)f(x)\,\mathrm{d}x = \int f(z-y)g(y)\,\mathrm{d}y$ for $z \in \mathbb{R}$. Show that if g is bounded and continuously differentiable with a bounded derivative g', and $\int f(x)\,\mathrm{d}x$ is finite, then $(f * g)(z)$ is also bounded and continuously differentiable, with derivative

$$(g * f)'(z) = (g' * f)(z) = \int g'(z-x)f(x)\,\mathrm{d}x.$$

You may look at Ex. A.12(g) to find weaker conditions under which the equality above holds. Notice also that $f(x)\,\mathrm{d}x$ may be replaced by any finite measure, i.e., the existence of a density is not used.

(e) As a corollary to (d), suppose that g is k times continuously differentiable function that vanishes outside of a compact set (i.e., g is nonzero only on a closed and bounded set), and that f is as before. Show that $(g*f)$ is also k times continuously differentiable, with derivatives $(g * f)^{(j)} = (g^{(j)} * f)$ for $j \le k$. Moreover, show that if f also vanishes outside of a compact set (which need not be the same as for g), then $g*f$ vanishes outside of a compact set.

(f) Let g be as in (e), and suppose that f is the density of the uniform distribution on $[a, b]$, i.e., $f(x) = 1/(b-a)$ for $x \in [a, b]$, and $f(x) = 0$ elsewhere. Show that $(g * f)$ has one more continuous derivative than g, and that these derivatives take the form

$$(g * f)^{(j+1)}(z) = \frac{g^{(j)}(z-a) - g^{(j)}(z-b)}{b-a}, \quad \text{for } j = 1, \dots, k.$$

(g) We'll use our findings above to prove the existence of a infinitely smooth density function with support $[-1, 1]$. Let $U_1, U_2, \dots$ be i.i.d. uniforms on $[-1, 1]$. Show that

$$Y_3 = \frac{U_1}{2} + \frac{U_2}{4} + \frac{U_3}{8},$$

has a density, say f_3, that is one time continuously differentiable. Proceed by induction to show that the density of $Y_m = \sum_{n=1}^{m} U_n/2^n$ is $m-2$ times continuously differentiable. Finally, argue that the density of

$$Y = \sum_{n=1}^{\infty} U_n/2^n = \sum_{n=1}^{m} U_n/2^n + \sum_{n \ge n+1} U_n/2^n,$$

is infinitely smooth, i.e., has infinitely many continuous derivatives.

Ex. A.18 *Independence.* Here we define and work through basic properties of *independence*, for events and for random variables.

(a) For a probability space $(\Omega, \mathcal{A}, \Pr)$, we start out saying that two events A and B are independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$. Show that then also A and B^c are independent, A^c and B are independent, and A^c and B^c are independent. Show that all events are independent of the emptyset and of the full set Ω.

(b) Try to exhibit an example, with a finite Ω, of events A, B, C such that A and B are independent, A and C are independent, B and C are independent, but where $\Pr(A \cap B \cap C) \neq \Pr(A)\Pr(B)\Pr(C)$. Hence care is needed when defining independence for more than two events. We say that $A_1, \ldots, A_n$ are independent if

$$\Pr(A_{i_1} \cap \cdots A_{i_k}) = \Pr(A_{i_1}) \cdots \Pr(A_{i_k}), \quad \text{for any } \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}.$$

Show that this is equivalent to requiring that $\Pr(B_1 \cap \cdots \cap B_n) = \Pr(B_1) \cdots \Pr(B_n)$ for the 2^n such equations obtained by setting $B_j = A_j$ or $B_j = A_j^c$.

(c) The definition in (b) extends to countably infinite many events. The events $A_1, A_2, \ldots$ are independent if for any finite number of distinct indices $i_1, \ldots, i_n \in \{1, 2, \ldots\}$, the events $A_{i_1}, \ldots, A_{i_n}$ are independent, in the sense defined in (b). We say that the sub-σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots$ of $\mathcal{A}$ are independent if for arbitrary representatives $A_1 \in \mathcal{G}_1$, $A_2 \in \mathcal{G}_2$, and so on, the events $A_1, A_2, \ldots$ are independent. Suppose that Π_1 and Π_2 are two π-systems such that $\Pr(A_1 \cap A_2) = \Pr(A_1)\Pr(A_2)$ whenever $A_1 \in \Pi_1$ and $A_2 \in \Pi_2$. Use Dynkin's lemma (see Ex. A.5(c)) to show that $\sigma(\Pi_1)$ and $\sigma(\Pi_2)$ are independent. Generalise to π-systems $\Pi_1, \Pi_2, \ldots$ and σ-algebras $\sigma(\Pi_1), \sigma(\Pi_2), \ldots$.

(d) Consider random variables $X_1, X_2, \ldots$ defined on the probability space $(\Omega, \mathcal{A}, \Pr)$. We say that $X_1, X_2, \ldots$ are independent if the σ-algebras they generate, $\sigma(X_1), \sigma(X_1), \ldots$, are independent in the sense defined in (c). Suppose that X and Y are two random variables defined $(\Omega, \mathcal{A}, \Pr)$, and that

independent random variables

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x)\Pr(Y \leq y), \quad \text{for all } x, y \in \mathbb{R},$$

i.e., their joint c.d.f. equals the product of the marginal c.d.f.'s. Use the result from (c) to show that X and Y are independent.

(e) Consider the probability space $(\{1, 2, 3, 4, 5, 6\}, 2^\Omega, \Pr)$, where $\Pr$ is the uniform probability measure on this space. Define the random variables $X = I_{\{1,3,5\}}$ and $Y = I_{\{5,6\}}$. Write down the full σ-algebras $\sigma(X)$ and $\sigma(Y)$, and show that, indeed, $\Pr(A \cap B) = \Pr(A)\Pr(B)$ for all $A \in \sigma(X)$ and $B \in \sigma(Y)$.

(f) Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces, and let $X \colon \Omega \to \mathcal{X}$ and $Y \colon \Omega \to \mathcal{Y}$ be random variables on the same underlying probability space, with distributions P_1 and P_2, respectively. From Ex. A.16 we know that the measure $P_1 \times P_2$ on, defined by

$$(P_1 \times P_2)(C) = \int_{\mathcal{X}} \int_{\mathcal{Y}} I_C(x, y) \, \mathrm{d}P_2(y) \, \mathrm{d}P_1(x) = \int_{\mathcal{Y}} \int_{\mathcal{X}} I_C(x, y) \, \mathrm{d}P_1(x) \, \mathrm{d}P_2(y),$$

for $C \in \mathcal{A} \otimes \mathcal{B}$, is the unique probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ such that $(P_1 \times P_2)(A \times B) = P_1(A)P_2(B)$ for all measurable rectangles $A \times B$. Show that X and Y are independent if and only if (X, Y) has distribution $P_1 \times P_2$.

(g) So $(P_1 \times P_2)(C) = \Pr((X, Y) \in C)$ is now properly defined for much more complicated sets than the direct product sets $A \times B$. Let X and Y be independent, both with uniform distributions on $[-1, 1]$, with subintervals of equal length having the same probability. Find the probability that (X, Y) lands inside the unit circle.

(h) Show that if X and Y are independent random variables, and g and h are measurable functions, then $g(X)$ and $h(Y)$ are also independent. Use a bootstrapping argument (see Ex. A.12(e)) to show that

$$\mathrm{E}\, g(X)h(Y) = \mathrm{E}\, g(X)\, \mathrm{E}\, h(Y),$$

covariance

provided $g(X)$ and $h(Y)$ are integrable. The *covariance* of two random variables W and Z with means $\xi_W = \mathrm{E}\, W$ and $\xi_Z = \mathrm{E}\, Z$, respectively, is defined by

$$\mathrm{cov}(W, Z) = \mathrm{E}\,(W - \xi_W)(Z - \xi_Z).$$

Show that $\mathrm{cov}(W, Z) = \mathrm{E}\,(WZ) - \xi_W \xi_Z$, and conclude that if W and Z are independent, then $\mathrm{cov}(W, Z) = 0$.

(i) We must of course extend (f) and (h) to the case of more than two independent random variables. With $X_1, \ldots, X_n$ defined on the same underlying probability space $(\Omega, \mathcal{A}, \Pr)$, their distributions are $P_1 = \Pr X_1^{-1}, \ldots, P_n = \Pr X_n^{-1}$. Suppose that these random variables are independent, i.e., $\sigma(X_1), \ldots, \sigma(X_n)$ are independent σ-algebras (see definition in (d)), and show that this is equivalent to

$$\Pr(X_1 \le x_1, \ldots, X_n \le x_n) = P_1(-\infty, x_1] \cdots P_n(-\infty, x_n], \quad \text{for all } x_1, \ldots, x_n \in \mathbb{R},$$

that is, again, the joint c.d.f. equals the product of the marginal c.d.f.'s. Generalise the construction in (f): For $j = 1, \ldots, n$, suppose that $X_j \colon \Omega \to \mathcal{X}_j$ where $(\mathcal{X}_j, \mathcal{B}_j)$ are measurable spaces, and show that this gives rise to a well-defined product probability measure $Q = P_1 \times \cdots \times P_n$ on the σ-algebra $\mathcal{B}_1 \otimes \cdots \otimes \mathcal{B}_n$, generated by all rectangles $B_1 \times \cdots \times B_n$, where $B_j \in \mathcal{B}_j$ for $j = 1, \ldots, n$. Generalise also the if and only if claim from (f) to this higher dimensional setting.

(j) Show that if $X_1, \ldots, X_n$ are independent, and $g_1, \ldots, g_k$ are measurable functions, then also $g_1(X_1), \ldots, g_k(X_k)$ are independent. Show also that

$$\mathrm{E}\, g_1(X_1) \cdots g_k(X_k) = \mathrm{E}\, g_1(X_1) \cdots \mathrm{E}\, g_k(X_k),$$

when these means exist. The variance of a random variable was defined in Ex. A.15(d). Let $X_1, \ldots, X_n$ be random variables with finite second moments, show that

$$\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{cov}(X_i, X_j),$$

and, consequently, by (h), that $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$ if the $X_1, \ldots, X_n$ are independent. We note that double sums such as that on the right above are often written $\sum_{1 \le i < j \le n} a_i a_j = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j$. (xx point briefly to stigler and seven pillars. xx)

(k) Why product spaces? Because they are an efficient way of constructing probability spaces on which an arbitrary number of independent random variables with arbitrary distributions live. To see what is meant by this, consider the probability space from (e). Is it possible to construct two independent Bernoulli random variables with the same success probability on this space? Is it possible to construct more than two independent Bernoulli random variable on this space?

(l) Suppose that $A_1, A_2, \ldots, B_1, B_2, \ldots$ are independent events. Show that the σ-algebras $\sigma(A_1, A_2, \ldots)$ and $\sigma(B_1, B_2, \ldots)$ are independent. Let $A_1, A_2, \ldots$ be independent events, and consider the tail-σ-algebra $\mathcal{A} = \cap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \ldots)$. Prove Kolmogorov's zero-one law, namely that if $A \in \mathcal{A}$, then $\Pr(A) = 0$ or $\Pr(A) = 1$.

Ex. A.19 *The Borel–Cantelli lemma.* Let $A_1, A_2, \ldots$ be events, in a relevant probability space, with probabilities $p_i = \Pr(A_i)$. Consider $A_{\text{i.o.}} = \cap_{n\geq 1} \cup_{m \geq n} A_m = \limsup_{n\to\infty} A_n$, the full-sequence event corresponding to the A_n occurring infinitely often (see Ex. A.8). In Ex. A.8(f) we proved that of $\sum_{i=1}^{\infty} p_i$ is convergent then $\Pr(A_{\text{i.o.}}) = 0$, so sooner or later, there will be a finite (but random) n, such that none of the A_m will ever occur, for $m > n$. In (b) we prove a partial converse of this result.

(a) Let $A_1, A_2, \ldots$ be a sequence of events, and let N be the total number of occurrences of the A_i. Show that $\mathrm{E}\, N = \sum_{i=1}^{\infty} p_i$.

(b) Assume in addition that the $A_1, A_2, \ldots$ are independent. Show that if $\sum_{i=1}^{\infty} p_i$ is divergent, then $\Pr(A_{\text{i.o.}}) = 1$. In particular, for the case of independent events, there can't be say a 50 percent chance that there will be infinitely many occurrences.

(c) Consider independent Bernoulli 0-1 variables X_i with $\Pr(X_i = 1) = p_i$. What is the probability for having infinitely many $X_i = 1$, for $p_i = 1/i^{0.99}$, for $p_i = 1/i$, for $p_i = 1/i^{1.01}$?

(d) Let $X_1, X_2, \ldots$ be i.i.d. from the unit exponential distribution. Will there be infinitely many cases with $X_i \geq 0.99 \log i$, with $X_i \geq \log i$, with $X_i \geq 1.01 \log i$?

(e) Let $X_1, X_2, \ldots$ be i.i.d. standard normal. Show first that

$$\Pr(X_i \geq a) = 1 - \Phi(a) \doteq \phi(a)/a,$$

in the sense that the ratio between the exact and the approximate quantities tends to 1. (xx this is the Mills ratio. xx) Show that there will be infinitely many cases with $|X_i| \geq (2 \log i)^{1/2}$.

(f) (xx one or two more. new records, $\Pr(R_n = 1) = 1/n$. xx)

Ex. A.20 *Probability densities.* We have seen in Ex. A.14 that probability measures on the real line are fully characterised by the cumulative distribution functions. Very often there is an even more practical and satisfying way of defining a probability distribution, however, via its probability density function. These may be defined not only in familiar situations with continuous distributions, but with discrete data, and with measures having both continuous and discrete components.

probability density function

(a) In various classical situations, the density is simply the derivative of the cumulative distribution function, say $f(x) = F'(x)$, when the random variable X in question has a differentiable c.d.f. F. From the fundamental theorem of calculus,

$$\Pr(X \in [a, b]) = F(b) - F(a) = \int_a^b f(x)\, \mathrm{d}x, \quad \text{for all } [a, b].$$

The general theory of measure and integration allows the clear definition of $\int_A f(x)\,\mathrm{d}x$ for *any* Borel set A. Show that $\Pr(X \in A) = \int_A f(x)\,\mathrm{d}x$, for all such A, i.e. not merely for intervals. Giving $f(x)$, instead of the cumulative $F(x)$, or perhaps more complicated ways of defining a distribution $P(A)$ for all A, is the most convenient (and traditional) way in which to define a probability distribution.

absolute continuity

the Radon–Nikodym theorem

(b) Suppose in general terms that ν and μ are σ-finite measures on a measurable space $(\mathcal{X}, \mathcal{A})$ (see Ex. A.4). Suppose next that the measure ν is dominated by μ, meaning that $\mu(A) = 0$ implies $\nu(A) = 0$; one also says that ν is absolutely continuous with respect to μ. Under these conditions, the Radon–Nikodym theorem gives a converse to what we found in Ex. A.12(d): it says that there is a nonnegative $\mathcal{A}$-measurable function f, such that

$$\nu(A) = \int_A f(x)\,\mathrm{d}\mu(x) \quad \text{for all } A \in \mathcal{A}. \tag{A.4}$$

The function f, called the *density* of ν with respect to μ, is often denoted $\mathrm{d}\nu/\mathrm{d}\mu$, to remind us that this is a density of ν with respect to μ. If $\nu = P$ is a probability measure, then $f = \mathrm{d}P/\mathrm{d}\nu$ is a probability density function. [xx fix xx] We defer the proof of the Radon–Nikodym theorem until Ex. 10.10, when we have developed the appropriate tools for proving it in a nice probabilistic manner.

Look back at Ex. A.10(d) and explain why the density $f = \mathrm{d}\nu/\mathrm{d}\mu$ in (A.4) is only unique μ-almost surely. Explain why (a), where F has a derivative and is the integral of this derivative, matches this more general setup, where μ is the Lebesgue measure, with $\mu(a, b) = b - a$ for all intervals. Many classes of probability distributions, like the normal, the gamma, the Beta, the Weibull, the exponential, the t, the chi-squared, etc., are of this type, where a clear probability density function can be given as here, that is, with respect to standard Lebesgue measure.

(c) The strength of the general $f = \mathrm{d}P/\mathrm{d}\mu$ machinery above is that it can be fruitfully used for large classes of other probability measures too, not only for those which are dominated by the Lebesgue measure. The dominating measure is often chosen by mathematical convenience, to match the situation at hand. For the Poisson and other distributions, with random variables landing in $\mathcal{X} = \{0, 1, 2, \ldots\}$, consider, for any subset of $\mathcal{X}$, $\mu(A)$ equal to the number of numbers $j \in A$, that is, μ is the counting measure on the integers, which, from Ex. A.2(g), we know is a σ-finite measure. Show that with P having a Poisson distribution P, with mean θ, that there is a density $f = \mathrm{d}P/\mathrm{d}\mu$, given by $f(x) = \exp(-\theta)\theta^x/x!$ for $x = 0, 1, 2, \ldots$, in the sense given above.

(d) Consider a probability measure P on $[0, 1]$ with probabilities 0.1 and 0.1 at positions 0 and 1, and which has $P(a, b) = 0.8\,(b - a)$ for (a, b) inside $(0, 1)$. Thus P is not continuous, and not discrete, but a mixture. Show that P is dominated by the measure μ, which has pointmasses 1 and 1 at the points 0 and 1, and is uniform inside $(0, 1)$. Find the probability density $f(x) = \mathrm{d}P(x)/\mathrm{d}\mu$.

(e) Suppose P is dominated by a σ-finite μ, with $f(x) = \mathrm{d}P(x)/\mathrm{d}\mu$ the probability density, as per (A.4). With X having distribution P, and $g(x)$ being a function for which

the mean is finite (with respect to P), we can now take the change of variable formula from Ex. A.15?? one step further, so to speak. With the help of Ex. A.12(e), show that

$$\mathrm{E}\, g(X) = \int g(x)\, \mathrm{d}P(x) = \int g(x)\, \frac{\mathrm{d}P(x)}{\mathrm{d}\mu}\, \mathrm{d}\mu(x) = \int g(x) f(x)\, \mathrm{d}\mu(x).$$

In particular, if μ is Lebesgue measure on the real line, we get from Ex. A.12(h) that $\mathrm{E}\, g(X) = \int g(x) f(x)\, \mathrm{d}x$, and we are back to the expectation formula from introductory statistics courses.

(f) (xx round this off. drive home that this makes it possible and convenient to derive results in a general manner, point to Cramér–Rao, which we need to redo, as of 12-August-2024, and also that we can handle any type of mixed distributions, not merely the classic ones, the continuous and the discrete. ask for the mean and variance of the 0.1, 0.1, 0.8 distribution above. xx)

Ex. A.21 *Radon–Nikodym derivatives* Let P, Q, and μ be σ-finite measures on the measure space $(\Omega, \mathcal{A})$.

(a) Show that if $Q \ll P$ and $P \ll \lambda$, then $Q \ll \lambda$, and that we have the following relation of Radon–Nikodym derivatives,

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu} = \frac{\mathrm{d}Q}{\mathrm{d}P} \frac{\mathrm{d}P}{\mathrm{d}\mu}, \quad \mu\text{-almost surely.}$$

The second part here is, to a certain extent, Ex. A.12(e) in new dressing, and you might use that exercise to prove it.

(b) Show that if $Q \ll P$, then the density $\mathrm{d}Q/\mathrm{d}P$ is positive Q-almost surely. Next, show that if $Q \ll P$ and $P \ll Q$, then

$$\frac{\mathrm{d}P}{\mathrm{d}Q} = I(\mathrm{d}Q/\mathrm{d}P > 0)\Big(\frac{\mathrm{d}Q}{\mathrm{d}P}\Big)^{-1}, \quad \text{almost surely,}$$

with respect to both Q and P. You may again use Ex. A.12(e).

(c) Suppose that $Q \ll \mu$ and $P \ll \mu$. Let $N = \{\mathrm{d}P/\mathrm{d}\mu = 0\}$. Show that there is a measurable function $f \geq 0$ such that

$$Q(A) = \int_A f\, \mathrm{d}P + Q(A \cap N).$$

The pair (f, N) is called the Lebesgue decomposition of Q with respect to P. If P and Q are two σ-finite measures, in particular probability measures, then $Q + P$ is σ-finite and $Q \ll Q + P$ and $P \ll Q + P$, and a Lebesgue decomposition of Q with respect to P exists. Such constructions will1§ play an important role in parts of Chapter 2.

(d) Suppose that $Q \ll \mu$ and $P \ll \mu$. Show that $Q \ll P$ if and only if the set $\{\mathrm{d}Q/\mathrm{d}\mu > 0\} \cap \{\mathrm{d}P/\mathrm{d}\mu = 0\}$ has measure zero under μ. Show also that if this is the case, i.e., if $\mu(\{\mathrm{d}Q/\mathrm{d}\mu > 0\} \cap \{\mathrm{d}P/\mathrm{d}\mu = 0\}) = 0$, then

$$\frac{\mathrm{d}Q}{\mathrm{d}P} = \frac{\mathrm{d}Q/\mathrm{d}\mu}{\mathrm{d}P/\mathrm{d}\mu} I\{\mathrm{d}P/\mathrm{d}\mu = 0\}, \quad P\text{-almost surely.}$$

(e)

(f) Let P_ξ be the $\mathrm{N}(\xi, 1)$ distributions, and $P_{\xi+n^{-1/2}h}$ be the $\mathrm{N}(\xi+n^{-1/2}h, 1)$ distribution, e.g., P_ξ has density $f(x;\xi) = (1/\sqrt{2\pi})\exp(-\frac{1}{2}(x-\xi)^2)$ with respect to Lebesgue measure on the real line. Show that $P_{\xi+n^{-1/2}h} \ll P_\xi$; that

$$\frac{\mathrm{d}P_{\xi+n^{-1/2}h}}{\mathrm{d}P_\xi}(x) = \exp\Big(\frac{h}{\sqrt{n}}(x-\xi) - \frac{1}{2}\frac{h^2}{n}\Big),$$

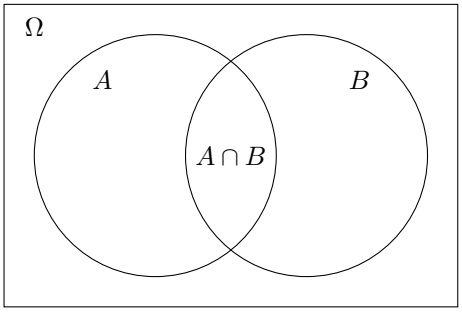
and that $\mathrm{E}_\xi\,(\mathrm{d}P_{\xi+n^{-1/2}h}/\mathrm{d}P_\xi)(X) = 1$, i.e., when the expectation is taken under P_ξ.

(g) Let $(\mathcal{X}, \mathcal{B}_1, \mu)$ and $(\mathcal{Y}, \mathcal{B}_2, \nu)$ be two σ-finite measurable spaces. Suppose that $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$ are independent random variables with distributions P_1 and P_2, and that $P_1 \ll \mu$ and $P_2 \ll \nu$. Denote the densities f_1 and f_2. Show that the distribution of (X, Y) has density $f_1 f_2$ with respect to the product measure $\mu \times \nu$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_1 \otimes \mathcal{B}_2)$ (see Ex. A.18(f)). Generalise to higher dimensions. This result is fundamental to the likelihood theory we cover in Chapter 5.

Ex. A.22 *Basic conditional probability.* Let $(\Omega, \mathcal{A}, \Pr)$ be a probability space. The Venn-diagram below provides the intuition behind the definition of the conditional probability of an event A, given that the event B has occurred. If all we know is that B occurred, then the probability that A also occured is the the size of the area of B that intersects A, relative to the total size of B. Here, size is measured by, well, a probability measure. Thus, the conditional probability of A given B is defined by

$$\Pr(A \,|\, B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \quad \text{provided } \Pr(B) > 0. \tag{A.5}$$

The notation $\Pr(A \,|\, B)$ might be unfortunate, because it may seem that the events A and B are on an equal footing. They are not. The event we condition on, namely B, is fixed, while the event we are computing the conditional probability of, that is A, can change. In fact, $A \mapsto P(A \,|\, B)$ is a probability measure, while $B \mapsto P(A \,|\, B)$ is not.



(a) Show that $\Pr(A \,|\, B)$ is a probability measure $(\Omega, \mathcal{A})$.

(b) The function $L(B) = \Pr(A \,|\, B)$ where the event A is fixed and the event we condition on might change, is called the likelihood function, and $L(B)$ is referred to as the likelihood of B. Show that $L(B)$ is *not* a probability measure.

(c) Suppose that B has positive probability. Show that A and B are independent if and only if $\Pr(A \mid B) = \Pr(A)$.

(d) Let $A_1, \ldots, A_n$ be disjoint events such that $\cup_{j=1}^n A_j = \Omega$. From the definition of conditional probability, deduce the law of total probability, that for any event B, $\Pr(B) = \Pr(B \mid A_1)\Pr(A_1) + \cdots + \Pr(B \mid A_n)\Pr(A_n)$. Deduce also Bayes' theorem,

law of total probability

Bayes theorem

$$\Pr(A_j \mid B) = \frac{\Pr(B \mid A_j)\Pr(A_j)}{\Pr(B \mid A_1)\Pr(A_1) + \cdots + \Pr(B \mid A_n)\Pr(A_n)},$$

provided $\Pr(B) > 0$.

Ex. A.23 *Conditional expectation.* Let X be an integrable real random variable on a probability space $(\Omega, \mathcal{A}, \Pr)$. The probability of X landing in $B \in \mathcal{B}(\mathbb{R})$ given the event $A \in \mathcal{A}$, is, using the definition in (A.5), $\Pr(X \in B \mid A) = \Pr(\{X \in B\} \cap A)/\Pr(A)$, provided $\Pr(A) > 0$. We can then define the conditional expectation of X given A by

$$\mathrm{E}\,(X \mid A) = \frac{\mathrm{E}\, X I_A}{\Pr(A)} = \int_A X \frac{\mathrm{d}\Pr(\omega)}{\Pr(A)} = \int X \mathrm{d}\Pr(\omega \mid A).$$

The point being that conditioning on an event A with positive probability is just a matter of using the definition in (A.5) in the obvious manner.

(a) Let's look at an example. Suppose that $X \sim \mathrm{N}(0,1)$, and set $A = \{X \geq c\}$ for some constant c. Show that $\mathrm{E}\,(X \mid A) = \int x\phi(x)/\{1 - \Phi(c)\}I(x \in [c, \infty))\,\mathrm{d}x$.

(b) In (a), we conditioned on the event $A = \{X \geq c\}$. But what if we want a function $Z \colon \Omega \to \mathbb{R}$ such that $Z(\omega) = \mathrm{E}\,(X \mid A)$ if $\omega \in A$, and $Z(\omega) = \mathrm{E}\,(X \mid A^c)$ if $\omega \in A^c$? That is, Z is the function

$$Z = \mathrm{E}\,(X \mid A^c)I_{A^c} + \mathrm{E}\,(X \mid A)I_A.$$

While $\mathrm{E}\,(X \mid A^c)$ and $\mathrm{E}\,(X \mid A)$ are constants, the indicator functions I_{A^c} and I_A are random variables, and so Z is also a random variable. We define this function Z to be the conditional expectation of X given I_A, denoted $\mathrm{E}\,(X \mid I_A)$. Thus, conditional expectations given a random variables, are themselves random variables. Notice that $\mathrm{E}\,(X \mid I_A)$ is measurable with respect to $\sigma(I_A) = \{A, A^c, \emptyset, \Omega\} \subset \mathcal{A}$, and also that if we modify $\mathrm{E}\,(X \mid I_A)$ on a set of measure zero, e.g., set $Z' = \mathrm{E}\,(X \mid I_A) + aI_{\{X \in \mathbb{Q}\}}$ for $a \in \mathbb{R}$, then $\Pr(Z \neq Z') = 0$. Show that $\mathrm{E}\,\mathrm{E}\,(X \mid I_A) = \mathrm{E}\, X$, and that, indeed $\mathrm{E}\, Z' = \mathrm{E}\, X$.

(c) Let X be some integrable random variable, and let Y be a discrete random variable with values in $\{y_1, y_2, \ldots\}$. Define $A_j = \{Y = y_j\} = \{\omega \in \Omega \colon Y(\omega) = y_j\}$ for $j = 1, 2, \ldots$. As a mild extension of (b), we can define the conditional expectation of X given Y as the random variable

$$\mathrm{E}\,(X \mid Y) = \sum_{j=1}^{\infty} \mathrm{E}\,(X \mid A_j)I_{A_j} \tag{A.6}$$

Show that $\mathrm{E}\,\mathrm{E}\,(X \mid Y) = \mathrm{E}\, X$, and also $\mathrm{E}\, I_{A_j}\mathrm{E}\,(X \mid Y) = \mathrm{E}\, I_{A_j} X$ for any $A_j = \{Y = y_j\}$. In fact, try to show that

$$\mathrm{E}\, I_G \mathrm{E}\,(X \mid Y) = \mathrm{E}\, I_G\, X, \tag{A.7}$$

for any event G in the σ-algebra generated by Y.

(d) So far we have only considered conditioning on events or on discrete random variables. But what if the random variable we want to condition on, say Y, is continuous, so that $\Pr(Y = y) = 0$ for all y. Then the definition of $\mathrm{E}\,(X \,|\, Y)$ given in (A.6) does not make sense, as it would involve division by zero. The solution to this problem is to take (A.7) as the definition of conditional expectation. Here is how. On the probability space $(\Omega, \mathcal{A}, \Pr)$ let X be an integrable random variable, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{A}$. Then any $\mathcal{G}$-measurable random variable Z such that

conditional expectation

$$\mathrm{E}\, I_G Z = \mathrm{E}\, I_G X, \quad \text{for all } G \in \mathcal{G}, \tag{A.8}$$

is called the conditional expectation of X given $\mathcal{G}$, denoted $\mathrm{E}(X \,|\, \mathcal{G})$. This definition entails that the conditional expectation is only defined up to sets of measure zero, and we call any Z satisfying (A.8) a *version* of the conditional expectation. Suppose that Z and Z' are two $\mathcal{G}$-measurable random variables, so that both $\mathrm{E}\, I_G Z = \mathrm{E}\, I_G X$ and $\mathrm{E}\, I_G Z' = \mathrm{E}\, I_G X$ for all $G \in \mathcal{G}$. Show that $Z = Z'$ almost surely.

(e) Prove that conditional expectation exists. More to the point, appeal to the Radon–Nikodym theorem, and show that if X is an integrable random variable on $(\Omega, \mathcal{A}, \Pr)$, and $\mathcal{G} \subset \mathcal{A}$, then there exists a $\mathcal{G}$-measurable random variable Z satisfying (A.8).

(f) There are some very useful results that follow directly from the definition in (A.8). You should convince yourself that they do. First, if X is $\mathcal{G}$-measurable, then $\mathrm{E}\,(X \,|\, \mathcal{G}) = X$, almost surely. Second, for any integrable X, we have that,

$$\mathrm{E}\, X = \mathrm{E}\,\mathrm{E}\,(X \,|\, \mathcal{G}). \tag{A.9}$$

When $\mathcal{G} = \sigma(Y)$ for a random variable Y, we typically write $\mathrm{E}(X \,|\, Y)$ instead of the more cumbersome $\mathrm{E}\,(X \,|\, \mathcal{G})$ or $\mathrm{E}\,\{X \,|\, \sigma(Y)\}$, and get the formula $\mathrm{E}\, X = \mathrm{E}\,\mathrm{E}\,(X \,|\, Y)$. Show that $\mathrm{E}\, X$ equals the conditional expectation of X given the trivial σ-algebra, that is, $\mathrm{E}\, X = \mathrm{E}\,(X \,|\, \{\emptyset, \Omega\})$, almost surely, and deduce that (A.9) is a special case of the tower property of conditional expectation, which says that when $\mathcal{G} \subset \mathcal{B}$ are sub-σ-algebras of $\mathcal{A}$,

tower property of conditional expectation

$$\mathrm{E}\,(X \,|\, \mathcal{B}) = \mathrm{E}(\,\mathrm{E}\,(X \,|\, \mathcal{B}) \,|\, \mathcal{G}), \quad \text{a.s.}.$$

(g) Comparing the definition in (A.8) with the result in (A.7), we see that $\mathcal{G}$ corresponds to the σ-algebra generated by Y. Let's go 'backwards', and see that (A.8) leads back to the definition we started out with. Suppose that $G_1, G_2, \ldots$ are disjoint sets whose union equals Ω, and let $\mathcal{G}$ be the σ-algebra generated by $G_1, G_2, \ldots$. Define

$$Z(\omega) = \begin{cases} \mathrm{E}\,(X \,|\, G_j), & \omega \in G_j \text{ and } \Pr(G_j) > 0, \\ z_j & \omega \in G_j \text{ and } \Pr(G_j) = 0, \end{cases}$$

for arbitrary constant $z_1, z_2, \ldots$. Show that Z is a version of $\mathrm{E}\,(X \,|\, \mathcal{G})$.

(h) Deduce from (g) that if $X = a$ on a set $G \in \mathcal{G}$, then $aI_G + \mathrm{E}\,(X \mid \mathcal{G})I_{G^c}$ is a version of $\mathrm{E}\,(X \,|\, \mathcal{G})$. In (g) you might have already used the fact that if $G \subset \mathcal{G}$ has no nonempty proper subsets, then $\mathrm{E}\,(X \,|\, \mathcal{G})$ must be constant over G. Prove it.

(i) Let $X, Y_1, \dots, Y_n$ be real random variables in $(\Omega, \mathcal{A})$, and set $Y = (Y_1, \dots, Y_n)$. From Ex. A.3(i) we know that if $g\colon \mathbb{R}^n \to \mathbb{R}$ is a measurable function, then $X = g(Y)$ is measurable with respect to $\sigma(Y)$. Combine Ex. A.3(f)&(g) to show that the converse also holds: If X is measurable with respect to $\sigma(Y)$, then there exists a measurable function $g\colon \mathbb{R}^n \to \mathbb{R}$ such that $X = g(Y)$.

(j) The upshot of (i) is that for the conditional expectation $\mathrm{E}(X \,|\, Y)$, there exists a measurable function g so that $g(Y) = \mathrm{E}(X \,|\, Y)$, almost surely. So when we write $\mathrm{E}(X \,|\, Y = y)$, we mean $g(y)$. Show that, with P_Y the distribution of Y,

$$\mathrm{E}(XI_B) = \int_C \mathrm{E}(X \,|\, Y = y)\, \mathrm{d}P_Y(y), \quad \text{for every } B = Y^{-1}(C) \in \sigma(Y).$$

In particular, Show that if (X, Y) has joint density $f_{X,Y}$ with respect to Lebesgue measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$, then the function

$$g(y) = \frac{\int x f_{X,Y}(x, y)\, \mathrm{d}x}{\int f_{X,Y}(x, y)\, \mathrm{d}x}.$$

is such that $g(Y) = \mathrm{E}(X \,|\, Y)$ almost surely.

Ex. A.24 *Properties of conditional expectation.* Even though a conditional expectation is a random variable and an expectation is a constant, many of the same properties and theorems apply. Let $(\Omega, \mathcal{A}, \Pr)$ be a probability space, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{A}$, and let X and Y be integrable random variables.

(a) Use bootstrapping to show that if XY is integrable, and Y is $\mathcal{G}$-measurable, we can 'take out what is known' from the conditional expectation, that is

take out what is known

$$\mathrm{E}(XY \,|\, \mathcal{G}) = Y\mathrm{E}(X \,|\, \mathcal{G}), \quad \text{a.s.}.$$

(b) If $X = a$ a.s., show that $\mathrm{E}(X \,|\, \mathcal{G}) = a$. For constants a and b, show that

$$\mathrm{E}(aX + bY \,|\, \mathcal{G}) = a\,\mathrm{E}(X \,|\, \mathcal{G}) + b\,\mathrm{E}(Y \,|\, \mathcal{G}), \quad \text{a.s.}.$$

Show that if $X \le Y$ almost surely, then $\mathrm{E}(X \,|\, \mathcal{G}) \le \mathrm{E}(Y \,|\, \mathcal{G})$ almost surely.

(c) Show that Fatou's lemma holds for conditional expectation, that is

$$\mathrm{E}(\liminf_{n\to\infty} X_n \,|\, \mathcal{G}) \le \liminf_{n\to\infty} \mathrm{E}(X_n \,|\, \mathcal{G}), \quad \text{a.s.}.$$

(d) Show that if X_n is a monotone increasing (or decreasing) sequence of integrable random variables such that $X_n \to X$ a.s., then $\mathrm{E}(X_n \,|\, \mathcal{G}) \to \mathrm{E}(X \,|\, \mathcal{G})$ a.s..

monotone convergence

(e) Show that if X_n is a sequence of integrable random variables such that $X_n \to X$ a.s. and $|X_n| \le Y$ a.s. for an integrable random variable Y, then $\mathrm{E}(X_n \,|\, \mathcal{G}) \to \mathrm{E}(X \,|\, \mathcal{G})$ a.s..

dominated convergence

(f) Show that in **??** and (e), if the sequence X_n only converges in probability to X (instead of almost surely), then we have $\mathrm{E}(X_n \,|\, \mathcal{G}) \to_p \mathrm{E}(X \,|\, \mathcal{G})$.

(g) Let g be a convex function such that $g(X)$ is integrable.. Show that

$$\mathrm{E}\,\{g(X)\,|\,\mathcal{G}\} \geq g(\mathrm{E}\,\{X\,|\,\mathcal{G}\}), \quad \text{a.s..}$$

Applying this to the convex function $g(x) = |x|^p$ for some $1 \leq p < \infty$, we have that $|\mathrm{E}\,(X \mid \mathcal{G})| \leq \mathrm{E}\,(|X|\,|\,\mathcal{G})$, a.s.. Show also that $(\mathrm{E}\,|\mathrm{E}\,(X \mid \mathcal{G})|^p)^{1/p} \leq (\mathrm{E}\,|X|^p)^{1/p}$

(h) Suppose that X is square integrable. The conditional variance of X given $\mathcal{G}$, denoted $\mathrm{Var}(X\,|\,\mathcal{G})$, is any version of the random variable $\mathrm{E}\,\{(X - \mathrm{E}\,(X\,|\,\mathcal{G}))^2\,|\,\mathcal{G}\}$. Show that if Y and Z are $\mathcal{G}$-measurable random variables, such that XY is square integrable, then

$$\mathrm{Var}(YX + Z\,|\,\mathcal{G}) = Y^2\mathrm{Var}(X\,|\,\mathcal{G}), \quad \text{a.s..}$$

variance decomposition

(i) Suppose that X is square integrable. Show the following very useful variance decomposition formula

$$\mathrm{Var}(X) = \mathrm{E}\,\mathrm{Var}(X\,|\,\mathcal{G}) + \mathrm{Var}(\mathrm{E}\,(X\,|\,\mathcal{G})), \quad \text{a.s..}$$

We'll meet this formula in the proof of the Rao–Blackwell theorem, see Ex. 8.4.

Ex. A.25 *Conditional probability, distributions, and densities.* Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. The *conditional probability* of the event $A \in \mathcal{F}$ given a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ is defined as $\Pr(A\,|\,\mathcal{G}) = \mathrm{E}\,(I_A\,|\,\mathcal{G})$. If $X\colon \Omega \to \mathcal{X}$ is a random variable with values in the measurable space $(\mathcal{X}, \mathcal{B})$, the *conditional distribution* of X given $\mathcal{G}$ is $P_X(B\,|\,\mathcal{G}) = \Pr(X \in B\,|\,\mathcal{G})$ as B ranges over $\mathcal{B}$. Since the conditional expectation is only defined almost surely, so are conditional probabilities and distributions. This means that any function $p(\omega, B)$ such that for each $B \in \mathcal{B}$, $p(\omega, B) = P_X(B\,|\,\mathcal{G})(\omega)$ for Pr-almost all ω, is a called the conditional distribution of X given $\mathcal{G}$. In this book, we deal exclusively with random variables taking values in complete and separable metric spaces, so we can, and will, assume that all the conditonal distributions we encounter are *regular*, i.e., they are such that $B \mapsto P_X(B\,|\,\mathcal{G})(\omega)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$ for Pr-almost all $\omega \in \Omega$.

(a) Let X and Y be real valued random variables. When, for some $B \in \mathcal{B}(\mathbb{R})$, we write $P_X(B\,|\,Y = y)$ or $\Pr(X \in B\,|\,Y = y)$, we mean the function $\mathrm{E}\,(I_B(X)\,|\,Y = y)$ introduced in Ex. A.23(i). A version of the conditional distribution of X given Y ...

(b) Let $P_X(B\,|\,\mathcal{G})$ be the conditional distribution of a real-valued random variable X given $\mathcal{G}$, and suppose that $g\colon \mathbb{R} \to \mathbb{R}$ is a measurable function such that $g(X)$ is integrable. Show that

$$\mathrm{E}\,(g(X) \mid \mathcal{G})(\omega) = \int g(x)\, P_X(\mathrm{d}x\,|\,\mathcal{G})(\omega),$$

for Pr-almost all $\omega \in \Omega$. This is the conditional expectation analogue of the second expression for $\mathrm{E}\,g(X)$ in Ex. A.15**??**.

(c) Suppose that $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ are σ-finite measure spaces, and let $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$ be random variables. Let P_Y be the distribution of Y on $(\mathcal{Y}, \mathcal{B})$, and let $P_{X,Y}$ be their joint distribution on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$. Suppose that $P_{X,Y} \ll \mu \times \nu$ with density $f_{X,Y}$. Show that $P_Y \ll \nu$ with density

$$f_Y(y) = \int_{\mathcal{X}} f_{X,Y}(x, y)\, \mathrm{d}\mu(x),$$

and that the conditional distribution of X given $Y = y$ has densities

$$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

with respect to μ. Convince yourself that the two expressions above also holds with X and Y switching roles. In particular, show that the densities above can be expressed as

$$f_{X \mid Y}(x \mid y) = \frac{f_{Y \mid X}(y \mid x) f_X(x)}{f_Y(y)}.$$

This is Bayes' theorem for densities. In the case that X is a (random) parameter, and Y are the data, $f_{X \mid Y}(x \mid y)$ is called the posterior density of the parameter given the data.

(d) Show that if $g \colon \mathbb{R} \to \mathbb{R}$ is a measurable function such that $g(X)$ is integrable, then

$$\mathrm{E}\,(g(X) \mid \sigma(Y)) = \int g(x) f_{X \mid Y}(x \mid Y)\, \mathrm{d}\mu(x),$$

almost surely. Note that this is the conditional expectation analogue of the expression for $\mathrm{E}\, g(X)$ in Ex. A.20(e).

(e) The results in (c) are in the background of many of the conditional probability calculations to be carried out in this book, but they cannot always be used. In Ex. A.23(a), we had $X \sim \mathrm{N}(0,1)$ and the event $A = \{X \geq c\}$. Consider the random variable $Y = I(X \geq c)$. The distributions of X and Y have, of course, densities with respect to Lebesgue and counting measure, say λ and μ, respectively. Show, however, that the distribution of (X, Y) is *not* dominated by the product measure $\lambda \times \nu$ on $(\mathbb{R} \times \{0,1\}, \mathcal{B}(\mathbb{R}) \otimes 2^{\{0,1\}})$. Define the set function $\rho(C) = \lambda(\{x \in \mathbb{R} \colon (x, I\{x \geq c\}) \in C\})$ on $\mathcal{B}(\mathbb{R}) \otimes 2^{\{0,1\}}$. Show that ρ is a measure, and that $P_{X,Y} \ll \rho$. Find an expression for the density. Introduce the set function $\lambda_{X \mid Y}(B \mid y) = \lambda(B_y)$, where $B_y = \{x \in \mathbb{R} \colon I(x \geq c) = y\}$ for $B \in \mathcal{B}(\mathbb{R})$. Show that $\lambda_{X \mid Y}(B \mid y)$ is a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for $y = 0, 1$; that the conditional distribution of X given $Y = y$ is dominated $\lambda_{X \mid Y}(B \mid y)$; and find an expression for the density.

(f)

(g) Let X and Y be independent random variables on the probabilty space $(\Omega, \mathcal{A}, \Pr)$. Let g be a real-valued function such that $g(X,Y)$ is integrable with respect to $\Pr$. Show that

$$\mathrm{E}\,(g(X,Y) \mid \sigma(Y)) = \int g(x,Y)\, \mathrm{d}P_X(x),$$

where $P_X = \Pr X^{-1}$ is the distribution of X.

(h)

(i) Suppose that $(\mathcal{X}, \mathcal{A}, \mu)$ be a σ-finite measure space, and $(\Theta, \mathcal{B})$ a measurable space. Let $X \colon \Omega \to \mathcal{X}$ and $\theta \colon \Omega \to \Theta$ be random variables on the same underlying probability space $(\Omega, \mathcal{F}, \Pr)$. Suppose that the conditional distribution of X given θ is P_θ, and that $P_\theta \ll \mu$ for all $\theta \in \Theta$, with densities $f_\theta(x)$. Let Π be the distribution of θ. We think of

θ as a random parameter, and, as is the convention, we do not distinguish between the random variable θ and the values it attains. Show that (X, θ) has joint distribution

$$\Pr((X,\theta) \in C) = \int I_C(x,\theta) f_\theta(x) \,\mathrm{d}(\mu \times \Pi)(x,\theta),$$

for $C \in \mathcal{A} \otimes \mathcal{B}$, where $\mu \times \Pi$ is the product measure on $(\mathcal{X} \times \Theta, \mathcal{A} \otimes \mathcal{B})$. Show that the posterior distribution of θ given X is

$$\Pr(\theta \in B \,|\, \sigma(X)) = \int_B \frac{f_\theta(X)\,\mathrm{d}\Pi(\theta)}{\int_{\mathcal{X}} f_\theta(X)\,\mathrm{d}\Pi(\theta)},$$

almost surely. Show also that if ν is a σ-finite measure on $(\Theta, \mathcal{B})$ such that $\Pi \ll \nu$ with density π, then the posterior density of θ given X is

$$\pi(\theta \,|\, x) = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)\pi(\theta)\,\mathrm{d}\nu(\theta)},$$

with respect to ν. This version of Bayes' theorem, with this notation, is the workhorse formula of Chapter 6.

(j) (xx the more general cases, and examples; Bayes' theorem for densities; include also classical transformation formula, with Jacobian etc.)

(k)

(l)

Ex. A.26 *Conditional independence.* [xx some intro text xx]

(a)

(b) Let $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ be σ-algebras. Show that $\mathcal{F} \perp\!\!\!\perp \mathcal{H} \,|\, \mathcal{G}$ if and only if

$$\Pr(H \,|\, \mathcal{F} \vee \mathcal{G}) = \Pr(H \,|\, \mathcal{G}), \text{ a.s., for all } H \in \mathcal{H},$$

where $\mathcal{F} \vee \mathcal{G}$ is the smallest σ-algebra that contains $\mathcal{F}$ and $\mathcal{G}$.

Ex. A.27 *Extension of probability spaces.* An extension of a probability space $(\Omega, \mathcal{F}, \Pr)$ is a product space $(\Omega \times \mathcal{X}, \mathcal{F} \otimes \mathcal{B})$ equipped with a probability measure $\Pr'$ such that $\Pr'(A \times \mathcal{X}) = \Pr(A)$ for all $A \in \mathcal{F}$.

(a) Let Y be a random variable (or vector, or process) on $(\Omega, \mathcal{F}, \Pr)$. We can extend Y to be defined on the extension $(\Omega \times \mathcal{X}, \mathcal{F} \otimes \mathcal{B}, \Pr')$ by setting $Y'(\omega, x) = Y(\omega)$. Show that Y' has the same distribution as Y, and, in particular, that $\mathrm{E}'\, g(Y') = \mathrm{E}\, g(Y)$ for all measurable $g\colon \mathcal{X} \to \mathbb{R}$.

(b) Let Q be a probability kernel (or Markov kernel) between the probability space $(\Omega, \mathcal{G}, \Pr)$ and the measurable space $(\mathcal{X}, \mathcal{B})$. That is $Q\colon \Omega \times \mathcal{X} \to [0, \infty)$ is so that

$\omega \mapsto Q(\omega, B)$ is $\mathcal{G}$-measurable for each fixed $B \in \mathcal{B}$; and
$B \mapsto Q(\omega, B)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$ for each $\omega \in \Omega$.

The regular conditional distributions introduced in Ex. A.25 are probability kernels. Other examples of probability kernels are the mixed normal densities we meet in Ex. 2.56, where $\sigma\colon \Omega \to (0,\infty)$ a random variable, and

$$Q(\omega, B) = \int_B \frac{1}{\sqrt{2\pi}\sigma(\omega)} \exp(-\frac{x^2}{2\sigma(\omega)^2}\,\mathrm{d}x).$$

If we were to simulate a random variable X from $Q(\omega,\cdot)$ we would perhaps run something ressembling the following little script

```
sigma <- rgamma(1,2,2) # for example
X <- rnorm(1,0,sigma)
```

and almost without thinking about it, X would have been taken as conditionally independent of all other random variables, given σ. Since our probability space may not be rich enough to support such as conditionally independent random variable (see example below), we might need enlarge the probability space. Here is how: Suppose we have a probability space $(\Omega,\mathcal{F},\Pr)$ and $\mathcal{G}\subset\mathcal{F}$. Let Q be a probability kernel between $(\Omega,\mathcal{G},\Pr)$ and $(\mathcal{X},\mathcal{B})$. Consider $(\Omega\times\mathcal{X},\mathcal{G}\otimes\mathcal{B},\Pr')$ with

$$\Pr'(A) = \int_{\Omega\times\mathbb{R}} I_A(\omega,x)Q(\omega,\mathrm{d}x)\,\mathrm{d}\Pr(\omega),\ \text{for all } A\in\mathcal{G}\otimes\mathcal{B}.$$

Show that (i) $\Pr'(G\times\mathcal{X}) = \Pr(G)$ for all $G\in\mathcal{G}$, meaning that the product space just defined is indeed an extension; (ii) that the random variable $X(\omega,x)=x$ is so that $\Pr'(X\in B\,|\,\sigma(V)) = Q(\cdot,B)$, $\Pr'$-almost surely; and (iii) that X is conditionally independent of all $\mathcal{F}$-measurable random variables, given $\mathcal{G}$.

(c) Consider the probability space $(\{H,T\},2^{\{H,T\}},\Pr)$ with $\Pr$ a probability measure such that $\Pr(H)=p$. Let $\theta\colon\{H,T\}\to\{0,1\}$ be such that $\Pr(\theta=1)=p$, and consider the probability kernel $Q(\omega,B) = \theta(\omega)\int_B\phi(x)\,\mathrm{d}x + (1-\theta(\omega))\int_B\phi(x/\sqrt{2})/\sqrt{2}\,\mathrm{d}x$, where ϕ is the standard normal density. Construct an extension on which the random variable X has conditional distribution Q, given θ.

Ex. A.28 *Regularity and approximations.* Let $(\mathcal{X},d)$ be a metric space and consider the measurable space $(\mathcal{X},\mathcal{B},\mu)$, where $\mathcal{B}$ is the Borel σ-algebra, and μ is a finite measure. The distance from a point x to a set A is $d(x,A)=\inf\{d(x,y)\colon y\in A\}$. Using this distance, we can express any closed set F as a countable intersection of open sets, as follows $F=\cap_{n=1}^\infty\{x\in\mathcal{X}\colon d(x,F)<1/n\}$. So by De Morgan's laws, any open set G can be expressed as a countable union of closed sets $G=\cup_{n=1}^\infty\{x\in\mathcal{X}\colon d(x,G^c)\geq 1/n\}$. These facts are useful in what follows.

(a) Show that $\mu(B)=\sup_{F\subset B}\mu(F)$ for any open set B, where the supremum is taken over closed sets F. Similarly, show that $\mu(C)=\inf_{G\supset C}\mu(G)$ for any closed set C, with the infinum taken over open sets G. Recall that $\{B\subset\mathcal{X}\colon B \text{ is open}\}$ and $\{C\subset\mathcal{X}\colon C \text{ is closed}\}$ are π-systems, and that they both generate the Borel σ-algebra, facts we use below.

(b) Show that $\mathcal{D}_c = \{B \in \mathcal{B}\colon \mu(B) = \sup_{F\subset B} \mu(F), F \text{ closed}\}$ and $\mathcal{D}_o = \{B \in \mathcal{B}\colon \mu(B) = \inf_{B\subset G} \mu(G), G \text{ open}\}$ are d-systems, and conclude from Dynkin's lemma (see Ex. A.5) that for any $B \in \mathcal{B}$,

$$\mu(B) = \sup_{F\subset B} \mu(F) = \inf_{G\supset B} \mu(G),$$

where the supremum and the infimum are taken over closed sets F and open sets G, respectively.

(c) Show that for any measurable function $f\colon \mathcal{X} \to \mathbb{R}$ such that $\int f^p \,\mathrm{d}\mu < \infty$, i.e., $f \in L^p(\mathcal{X}, \mathcal{B}, \mu)$, there is a sequence $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$ of bounded and continuous functions such that $\int |f_n - f|^p \,\mathrm{d}\mu \to 0$.

Ex. A.29 *The mean via the cumulative distribution function.* Consider a random variable X on $[0, \infty)$, with cumulative function F.

(a) Show that the mean $\mathrm{E}\, X = \int_0^\infty x \,\mathrm{d}F(x)$ also can be expressed as $\int_0^\infty (1 - F)\,\mathrm{d}x$. You may use $x = \int_0^\infty I(x > y)\,\mathrm{d}y$ and then Fubini. Show furthermore that $\mathrm{E}\, X^p = \int_0^\infty \{1 - F(x^{1/p})\}\,\mathrm{d}x$.

(b) As a simple illustration, consider X with density function $f(x) = \theta \exp(-\theta x)$, where θ is a positive parameter. Find the cumulative F, and compute $\mathrm{E}\, X$ in two ways.

(c) (xx something with a discrete distribution too. find the AmStat paper we talked briefly about, for a bit more. xx)

Ex. A.30 *Moment-generating functions: examples.* [xx we have not introduces these distributions yet. xx] (xx nils lifts these to Ch1; then need post-polish here in App. xx) For a random variable Y, with distribution P, its moment generating function is

$$M(t) = \mathrm{E} \exp(tY) = \int \exp(ty) \,\mathrm{d}P(y),$$

defined for each t at which the expectation exists. The moment generating function is useful for finding and characterising distributions, for finding their moments, for handling the distributions of sums of variables, and in connection with distributional limits. When Y has a density $f(y)$ (with respect to Lebesgue measure), we have $M(t) = \int \exp(ty) f(y)\,\mathrm{d}y$, and if it is discrete with pointmasses $f(y)$ for sample space S, say, then $M(t) = \sum_{y\in S} \exp(ty) f(y)$. The expectation operator is more general, however, and $M(t)$ is perfectly defined also for intermediate cases where Y can have both discrete and continuous parts; see Ex. A.15.

(a) For a standard normal $Y \sim \mathrm{N}(0, 1)$, show that $M(t) = \exp(\frac{1}{2}t^2)$. When $Y \sim \mathrm{N}(\mu, \sigma^2)$, derive $M(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.

(b) For $Y \sim \mathrm{Expo}(\theta)$, show that $M(t) = 1/(1 - t/\theta)$, for $t < \theta$.

(c) For $Y \sim \mathrm{Gam}(a, b)$, with density $\{b^a/\Gamma(a)\}y^{a-1} \exp(-by)$, show that $M(t) = \{b/(b - t)\}^a$, for $t < b$. In particular, $M(t) = 1/(1 - t)^a$ for $\mathrm{Gam}(a, 1)$.

(d) Suppose Y is equal to zero with probability 0.90, but a standard normal with probability 0.10. Find the $M(t)$, and generalise.

(e) For the binomial (n, p), show that $M(t) = \{1 - p + p\exp(t)\}^n$.

(f) For $Y \sim \text{Pois}(\theta)$, find $M(t) = \exp\{\theta(e^t - 1)\}$. Use this, with Ex. 1.26, to find $M(t)$ also for the negative binomial (a, p). (xx hm, should give the formula here. xx)

(g) Let $Y = \pm 1$ with probabilities $\frac{1}{2}, \frac{1}{2}$. Show that

$$M(t) = \cosh(t) = \tfrac{1}{2}(e^t + e^{-t}) = 1 + (1/2)t^2 + (1/4!)t^4 + (1/6!)t^6 + \cdots .$$

(h) For the uniform distribution on the unit interval, show that $M(t) = \{\exp(t) - 1\}/t$, for $t \neq 0$, and with $M(0) = 1$.

(i) Let Y have the uniform distribution on the $[-1, 1]$ interval. Show that

$$M(t) = \frac{\exp(t) - \exp(-t)}{2t} = \frac{\sinh t}{t},$$

and that this function may be written as the infinite sum $1 + (1/3!)t^2 + (1/5!)t^4 + \cdots$.

Ex. A.31 *Moment-generating functions: properties.* Among the basic properties of moment-generating functions is that it generates moments. As we will see in this exercise, the rth derivative of $M(t) = \text{E}\exp(tY)$ at the point zero equals $\text{E}\,Y^r$.

(a) Suppose that the moment generating function $M(t)$ of a random variable Y is finite for all $t \in (-t_0, t_0)$, for some $t_0 > 0$. For some t in this interval, the sum $M(-t) + M(t)$ is then also clearly finite. Appeal to Ex. A.12(b) to show that the finiteness of $M(-t) + M(t)$ implies that $\text{E}\,|Y|^{2k}$ is finite for all $k \geq 1$. Now, use that $|x|^{2k-1} \leq 1 + |x|^{2k}$ to fill in the odd gaps. This highlights the restrictiveness of the moment-generating function: by assuming its existence in a neighbourhood of zero, we are effectively assuming that all moments exist.

(b) To see that the converse of (a) does not hold, that is, a distribution with finite moments of all orders may not have a moment-generating function that is finite in some interval around zero, consider the log-normal distribution; see Ex. 1.53.

(c) Provided $M(t)$ is finite in an interval around zero, say $(-t_0, t_0)$, $t_0 > 0$, you may again use Ex. A.12(b) to show that

$$M(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\,\text{E}\,Y^k, \quad \text{for } |t| < t_0.$$

We now see that if we can slip the derivative inside this sum, then we have the property mentioned in the introduction. To see that we can, choose points a and b so that $|t| < a < b < t_0$, and show that

$$\Big|\frac{(t+h)^k - t^k}{h}\Big| \leq 3ka^{k-1},$$

provided $|h| < a - |t|$. Next, show that there is an $M > 0$ so that $3ka^{k-1} \leq 3Mb^k$, and use this to conclude that $\sum_{k=0}^{\infty} 3ka^{k-1}\mathrm{E}\,Y^k < \infty$. Using these results, explain why $M'(t) = \sum_{k=1}^{\infty}(t^{k-1}/(k-1)!)\mathrm{E}\,Y^k$. Finally, by induction, show that

$$M^{(r)}(t) = \sum_{k=r}^{\infty} \frac{t^{k-r}}{(k-r)!}\,\mathrm{E}\,Y^k.$$

This expression shows that moment-generating functions generate moments in the sense that $M^{(r)}(0) = \mathrm{E}\,(Y^r)$.

(d) For $X \sim \mathrm{N}(0,1)$, show that $M(t)$ for $|X|$ becomes $2\exp(\frac{1}{2}t^2)\Phi(t)$, and use this to find its mean and variance.

(e) It Y has mean ξ and standard deviation σ, and moment-generating function $M(t)$, give a formula for that of $Y' = (Y-\xi)/\sigma$. Illustrate this in the case of $Y \sim \mathrm{Pois}(\theta)$, computing and drawing the moment-generating function of $(Y-\theta)/\theta^{1/2}$, alongside $\exp(\frac{1}{2}t^2)$. Comment on what you find.

(f) If Y has a distribution symmetric around zero, such that Y and $-Y$ have the same distribution, then $M(t) = M(-t)$, so it depends on t only via $|t|$.

Ex. A.32 *Uniqueness of the Laplace transform.* For random variables taking values on the positive halfline $[0,\infty)$, it is sometimes more convenient to work with the Laplace transform instead of the moment-generating function. Let X be a random variable with support $[0,\infty)$, and denote its distribution and c.d.f. by P and F, respectively. Define the Laplace transform $L(t) = \mathrm{E}\,\exp(-tX)$, for $t \geq 0$. We here follow Billingsley (1995, pp. 284–286) in proving that $L(t)$ determines the distribution of X.

(a) Show that $L(t)$ is finite for all $t \geq 0$, and deduce from Ex. A.31(c) that the rth derivative of $L(t)$ is $L^{(r)}(t) = (-1)^r\mathrm{E}\,X^r\exp(-tX)$.

(b) To prove that the Laplace transform uniquely determines the distribution of X, we make a detour via the Poisson distribution. Let Y_θ be a Poisson random variable with mean θ, i.e., Y_θ has density $f(r;\theta) = (1/r!)\theta^r\exp(-\theta)$ with respect to counting measure on $\{0,1,2,\ldots\}$. Use the Chebyshov inequality (see Ex. 2.11) to show that $Y_\theta/\theta \to_p 1$ as $\theta \to \infty$. Let $G(t;\theta) = \Pr(Y_\theta/\theta \leq t) = \sum_{r=0}^{\lfloor \theta t\rfloor} f(r;\theta)$ be the c.d.f. of Y_θ/θ, where $\lfloor u\rfloor = \max\{m \in \mathbb{Z}\colon m \leq u\}$. Deduce from the convergence in probability result that as θ tends to infinity, $G(t;\theta) \to 1$ if $t > 1$ and $G(t;\theta) \to 0$ if $t < 1$.

(c) Show that we can write,

$$\sum_{r=0}^{\lfloor ty\rfloor} \frac{(-1)^r}{r!}t^r L^{(r)}(t) = \mathrm{E}\,G(y/X;tX).$$

and show that $G(y/x;tx) \to I\{0 \leq x \leq y\}$ as $t \to \infty$, almost everywhere. Appeal to the right convergence theorem to conclude that $\mathrm{E}\,G(y/X;tX) \to F(y)$, for all continuity points of $F(y) = P(-\infty, y]$.

(d) [xx include continuity theorem for Laplace transform, i.e., that for nonnegative random variables X_n the Laplace transforms converge to a the Laplace transform of a rv X, then $X_n \to_d X$ xx]

Ex. A.33 *Moment-generating functions for sums.* (xx point here to Ex. **??**, and more. xx) If Y_1 and Y_2 are independent, with given distributions, say with densities f_1 and f_2, then their sum $Z = Y_1 + Y_2$ have of course a well-defined distribution, and its density can be expressed as

$$g_2(z) = \int f_1(z - y_2) f_2(y_2)\,\mathrm{d}y_2 = \int f_1(y_1) f_2(z - y_1)\,\mathrm{d}y_1.$$

With algebraic patience this may e.g. be used to show that if $Y_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$, then indeed $Y_1 + Y_2$ is normal too, with parameters $\mu_1 + \mu_2$ and $\sigma_1^2 + \sigma_2^2$; see Ex. 1.2. Such convolutions quickly become convoluted in more general setups, however, and finding the density of say $Y_1 + Y_2 + Y_3 + Y_4$ from given densities f_1, f_2, f_3, f_4 may become too complicated. Pushing the matter to the domain of moment-generating functions instead makes matters simpler.

(a) When X and Y are independent, then $M_{X+Y}(t) = M_X(t) M_Y(t)$, in the obvious notation. This generalises of course to the case of more than two independent variables.

(b) Let $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, for $i = 1, \ldots, n$, with these variables being independent. Find the moment-generating function for the sum $Z = Y_1 + \cdots + Y_n$, and use the characterisation property to establish that indeed $Z \sim \mathrm{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

(c) Let $Y_1, \ldots, Y_k$ be independent Gamma distributed variables, with parameters (a_1, b), $\ldots, (a_k, b)$; see Ex. 1.9. Show that their sum is a Gamma with parameters $(\sum_{i=1}^k a_i, b)$.

(d) Suppose $Z = Y_1 + Y_2$, with these two being independent, and suppose you know that $Y_1 \sim \mathrm{N}(0, 2)$ and $Z \sim \mathrm{N}(0, 7)$. Prove that Y_2 must be a $\mathrm{N}(0, 5)$.

(e) Similarly, suppose $Z = Y_1 + Y_2$, with these two being independent, and assume it is known that $Y_1 \sim \chi^2_{10}$ and that $Z \sim \chi^2_{24}$. Prove that $Y_2 \sim \chi^2_{14}$.

Ex. A.34 *Characteristic functions.* The characteristic function of a random variable X is defined as

$$\varphi(t) = \mathrm{E}\,\exp(itX) = \mathrm{E}\,\cos(tX) + i\,\mathrm{E}\,\sin(tX),$$

with $i = \sqrt{-1}$ the complex unit, and $t \in \mathbb{R}$. As the name suggests, the characteristic function is useful for finding and characterising distributions, for finding their moments, for handling the distributions of sums of variables, and for finding with distributional limits. What distingusihes it from the moment-generating function is that it always exists, i.e., we do not need to make the assumption of all moments being finite (see Ex. A.31(a)).

(a) Show that the characteristic function always exists, in fact $|\varphi(t)| \leq 1$. Establish that $|\varphi(t+h) - \varphi(t)| \leq \mathrm{E}\,|\exp(ihX) - 1|$, and use this inequlity to show that $\varphi(t)$ is uniformly continuous.

(b) Show that if $Z \sim \mathrm{N}(0,1)$, then its characteristic function is $\varphi_Z(t) = \exp(-\frac{1}{2}t^2)$, and that if $X \sim \mathrm{N}(\mu, \sigma^2)$ its characteristic function is $\varphi_X(t) = \exp(it\mu - \frac{1}{2}t^2\sigma^2)$.

(c) Show that the Cauchy distribution with density $f(x) = (1/\pi)(1+x^2)^{-1}$ has characteristic function $\exp(-|t|)$. Note that this function does not have a derivative at zero, corresponding to the fact that the Cauchy distribution does not have a finite mean.

(d) Suppose that X has c.d.f. F. Show that the characteristic function is real-valued if and only if the distribution is symmetric, that is. $F(x) = 1 - F(-x)$ for all x.

(e) For $m = 0, 1, 2, \ldots$ define $r_m(x) = \exp(ix) - \sum_{k=0}^{m}(ix)^k/k!$, and let $r_{-1}(x) = \exp(ix)$. Convince yourself that $r_m(x) = r_{m-1}(x) - (ix)^m/m!$, and that $r_m(x) = i\int_0^x r_{m-1}(y)\,\mathrm{d}y$ for $x > 0$ and $r_m(x) = -i\int_x^0 r_{m-1}(y)\,\mathrm{d}y$ for $x < 0$. Show that $|r_0(x)| \le \min(2, |x|)$, and proceed by induction to show that

$$|\exp(ix) - \sum_{k=0}^{m}\frac{(ix)^k}{k!}| \le \min\big(\frac{2|x|^m}{m!}, \frac{|x|^{m+1}}{(m+1)!}\big), \quad \text{for } m = 0, 1, 2, \ldots.$$

In particular, for $m = 0, 1, 2$,

$$\begin{aligned} |\exp(ix) - 1| &\le \min(|x|, 2), \\ |\exp(ix) - (1+ix)| &\le \min(\tfrac{1}{2}|x|^2, 2|x|), \\ |\exp(ix) - (1+ix-\tfrac{1}{2}x^2)| &\le \min(\tfrac{1}{6}|x|^3, x^2). \end{aligned}$$

(f) Use the inequality in (e) and the dominated convergence theorem to show that if $\mathrm{E}\,X^2$ is finite, then $\varphi(t) = 1 + it\mathrm{E}\,X - \frac{1}{2}t^2\mathrm{E}\,X^2 + o(t^2)$ as $t \to 0$. In particular, if $\mathrm{E}\,X = 0$ and $\mathrm{E}\,X^2 = \sigma^2$, we have

$$\varphi(t) = 1 - \tfrac{1}{2}t^2\sigma^2 + o(t^2), \quad \text{as } t \to 0.$$

Generalise to show that if $\mathrm{E}\,|X|^m < \infty$ for some $m \ge 1$, then

$$\varphi(t) = \sum_{k=0}^{m}((it)^k/k!)\mathrm{E}\,X^k + o(t^m), \quad \text{as } t \to 0.$$

(g) Assume that $\mathrm{E}\,|X|$ is finite. For $h > 0$ write,

$$\frac{\varphi(t+h) - \varphi(t)}{h} = \mathrm{E}\,\exp(itX)\big(\frac{\exp(ihX) - 1}{h}\big),$$

and, taking limits as $h \to 0$, combine the inequality from (e) and the dominated convergence theorem to show that $\varphi'(t) = \mathrm{E}\,\{iX\exp(itX)\}$. Proceed inductively to show that, provided $\mathrm{E}\,|X|^r$ is finite, the rth derivative of the characteristic function is

$$\varphi^{(r)}(t) = \mathrm{E}\,\{(iX)^r\exp(itX)\}.$$

This shows that the moments can be read off from the characteristic function. Use the same proof technique as in (a) to show that $\varphi^{(r)}(t)$ is uniformly continuous.

Ex. A.35 *Uniqueness of characteristic functions.* If two random variables have identical characteristic functions, then their distributions are identical too. This fact is proved via so-called 'inversions theorems', providing a mechanism for finding the distribution of a random variable from its characteristic function.

(a) One such inversion formula is as follows: If the random variable X has characteristic function $\varphi(t)$ that is integrable, i.e., $\int |\varphi(t)|\,\mathrm{d}t < \infty$, then X has density f, for which a formula is

$$f(x) = \frac{1}{2\pi}\int \exp(-itx)\varphi(t)\,\mathrm{d}t.$$

Write down what this means, in the cases of a normal and a Cauchy, and verify the implied formulae. Show that f in each such case of an integrable $\varphi(t)$ necessarily becomes continuous.

(b) As a small digression, show that the characteristic function of the uniform distribution on $[-1,1]$ is $\varphi(t) = \sin(t)/t$. Deduce that

$$\int |\frac{\sin t}{t}|\,\mathrm{d}t = \infty \quad \text{even though} \quad \int \frac{\sin t}{t}\,\mathrm{d}t = \pi.$$

(c) Point (a) above gives a formula for the density f of a random variable, in the case of it having an integrable characteristic function φ. One also needs a more general formula, for the case of random variables that do not have densities, etc. Let X be any random variable, with cumulative distribution function F and characteristic function φ (but with nothing assumed about it having a density), and add a little Gaußian noise to X,

$$U_\sigma = X + \sigma Z, \text{ with } Z \sim \mathrm{N}(0,1),$$

with $\sigma > 0$ and Z is independent of X. Then U_σ has a density, even if X does not have one. Our intention is to let $\sigma \to 0$, to come back to X. Show that U_σ has cumulative distribution function and density of the form

$$F_\sigma(u) = \int \Phi\big(\frac{u-x}{\sigma}\big)\,\mathrm{d}F(x), \quad \text{and} \quad f_\sigma(u) = \int \frac{1}{\sqrt{2\pi}\sigma}\exp\big(-\frac{(u-x)^2}{2\sigma^2}\big)\,\mathrm{d}F(x),$$

where $\Phi(z) = \int_{-\infty}^z (1/\sqrt{2\pi})\exp(-\frac{1}{2}x^2)\,\mathrm{d}x$ is the standard normal distribution function.

(d) Verify that the characteristic function of U_σ is $\varphi(t)\exp(-\frac{1}{2}t^2\sigma^2)$, and that it is integrable. Thus, according to (a) we must have that the density of U_σ is

$$f_\sigma(u) = \frac{1}{2\pi}\int \exp(-itu)\varphi(t)\exp(-\tfrac{1}{2}t^2\sigma^2)\,\mathrm{d}t.$$

To see that this equality is true, start by showing that $\mathrm{E}\int \exp\{it(X-u) - \frac{1}{2}t^2\sigma^2\}\,\mathrm{d}t = 2\pi f_\sigma(u)$, and apply Fubini's theorem.

(e) Deduce from (d) that for $a < b$,

$$\Pr(U_\sigma \in (a,b]) = \frac{1}{2\pi}\int \frac{\exp(-itb) - \exp(-ita)}{-it}\varphi(t)\exp(-\tfrac{1}{2}t^2\sigma^2)\,\mathrm{d}t,$$

and from this, derive the general inversion formula, valid for all continuity points a and b ($a < b$) of F, inversion formula

$$F(b) - F(a) = \lim_{\sigma \to 0} \frac{1}{2\pi} \int \frac{\exp(-itb) - \exp(-ita)}{-it} \varphi(t) \exp(-\tfrac{1}{2} t^2 \sigma^2) \, \mathrm{d}t.$$

(f) Assume that X and Y are random variables with identical characteristic functions. Show that X and Y must be equal in distribution.

(g) Let $X_1, \ldots, X_n$ be independent $\mathrm{N}(\mu_j, \sigma_j^2)$ random variables. Show that $\sum_{j=1}^n X_i$ has a normal distribution.

(h) By the inequality in Ex. A.34(e), and using that the absolute value of the complex exponential is 1, show that for all real x and y,

$$|\exp(iy) - \exp(ix)| \leq |x - y|.$$

characteristic functions with finite integral

Now, let X be a random variable with characteristic function φ, and suppose that $\int |\varphi(t)| \, \mathrm{d}t$ is finite. Show that X has density $f(x) = (2\pi)^{-1} \int \exp(-itx)\varphi(t) \, \mathrm{d}t$, meaning that the claim in (a) is true.

Ex. A.36 *Characteristic functions for vector variables.* With $X = (X_1, \ldots, X_k)^{\mathrm{t}}$ a random vector, in dimension k, we define its characteristic functions as

$$\varphi(t_1, \ldots, t_k) = \mathrm{E} \exp(it^{\mathrm{t}} X) = \mathrm{E} \exp\{i(t_1 X_1 + \cdots + t_k X_k)\}$$

for $t = (t_1, \ldots, t_k)^{\mathrm{t}}$.

(a) Show that the properties from Ex. A.34, with the appropriate amendments, generalises to the k-dimensional case. In particular, show that $|\varphi(t_1, \ldots, t_k)| \leq 1$, and [xx list relevant props here xx].

(b) Show that if the components are independent, then $\varphi(t_1, \ldots, t_k) = \varphi_1(t_1) \cdots \varphi_k(t_k)$, in terms of the individual characteristic functions $\varphi_1, \ldots, \varphi_k$. Thus, the characteristic function of any subset of $(X_1, \ldots, X_k)$ can be retrieved by setting the appropriate subset of $(t_1, \ldots, t_k)$ to zero.

(c) Show for the multinormal case, where $X \sim \mathrm{N}_k(\xi, \Sigma)$, that $\varphi(t) = \exp(it^{\mathrm{t}}\xi - \frac{1}{2} t^{\mathrm{t}} \Sigma t)$.

(d) The inversion formula derived in Ex. A.35 also generalises to the k-dimensional case. In analogy to that exercise, let $Z_1, \ldots, Z_k$ independent standard normal random variables, and define $U_{\sigma,j} = X_j + \sigma Z_j$ for $j = 1, \ldots, k$. Amend the proof from said exercise to show that

$$\begin{aligned} &\Pr\{U_{\sigma,1} \in (a_1, b_1], \ldots, U_{\sigma,k} \in (a_k, b_k]\} \\ &= \frac{1}{(2\pi)^k} \int \cdots \int \prod_{j=1}^k \frac{\exp(-it_j b_j) - \exp(-it_j a_j)}{-it_j} \exp(-\tfrac{1}{2} t_j^2 \sigma^2) \varphi_X(t_1, \ldots, t_k) \, \mathrm{d}t_1 \cdots \mathrm{d}t_k. \end{aligned}$$

Next, take the limit as $\sigma \to 0$ to obtain the general inversion formula for k-dimensional random vectors. Conclude that if $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$ are random vectors with identical characteristic functions, then they are also identical in distribution.

(e) Let $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$ be two random vectors. Show that if $c^{\mathrm{t}}X = c_1X_1 + \cdots c_kX_k = c_1Y_1 + \cdots c_kY_k = c^{\mathrm{t}}Y$ for all vectors $c = (c_1, \ldots, c_k)^{\mathrm{t}}$, then X and Y are identical in distribution.

(f) Let $X = (X_1, \ldots, X_k)^{\mathrm{t}}$ be a random vector, and suppose that its characteristic functions $\varphi(t_1, \ldots, t_k)$ is such that $\int \cdots \int |\varphi(t_1, \ldots, t_k)| \mathrm{d}t_1, \ldots \mathrm{d}t_k < \infty$. Show that X has density

$$f(x_1, \ldots, x_k) = \frac{1}{(2\pi)^k} \int \cdots \int \Big(\prod_{j=1}^{k} \exp(-it_jx_j) \Big) \varphi(t_1, \ldots, t_k) \, \mathrm{d}t_1 \cdots \mathrm{d}t_k.$$

Ex. A.37 *Smoothness of characteristics functions.* In Ex. A.34(g) we have seen that the more moments a random variable X has, the smoother its characteristic function $\varphi(t)$ is. In (a) , which is the Riemann–Lebesgue lemma, the smoothness of $\varphi(t)$ is connected to the behaviour of $\varphi(t)$ as $|t|$ tends to infinity.

(a) Suppose that X has density f with respect to Lebesgue measure. Show that $\varphi(t) \to 0$ as $|t| \to \infty$. Show that if f has $k \geq 1$ integrable derivatives, then $\varphi(t) = o(1/t^k)$ as $|t| \to \infty$.

(b) Suppose that the characteristic function φ of the random variable X is such that $|\varphi(t)| = 1$ for all t. Show that X must be equal to a constant, i.e., it has a degenerate distribution.

(c) Show that $\sup_{|t|>\varepsilon} |\varphi(t)| < 1$ for any $\varepsilon > 0$.

(d) Generalise the above to higher dimensions, i.e., the characteristic functions of random vectors of dimension $k \geq 2$.

Ex. A.38 *Uniqueness of moment-generating functions.* Moment-generating functions characterise distributions: If X and Y are random variables such that $\mathrm{E}\exp(tX) = \mathrm{E}\exp(tY) < \infty$ for all $t \in (-t_0, t_0)$, then X and Y have the same distribution.

(a) Suppose that $\mathrm{E}\, f(X) = \mathrm{E}\, f(Y)$ for all bounded and continuous functions f. Approximate the indicator function $I\{y \leq x\}$ by a sequence of bounded and continuous functions to show that $F(x) = G(x)$ for all x. From Ex. A.14(c), this entails that X and Y have the same distribution.

(b) Let X and Y be random variables with distributions F and G on the unit interval, with identical moment-generating functions, $\int_0^1 \exp(tx) \, \mathrm{d}F(x) = \int_0^1 \exp(tx) \, \mathrm{d}G(x)$ for all $t \in (-t_0, t_0)$, say. Show that then $\int_0^1 p(x) \, \mathrm{d}F(x) = \int_0^1 p(x) \, \mathrm{d}G(x)$ for all polynomials $p(x)$. Use the Weierstraß approximation theorem, see Ex. 2.18, to show that this equality must hold for all continuous functions f. Point to (a) and conclude.

(c) Suppose that X and Y have identical moment-generating functions that are finite on $(-t_0, t_0)$. Let φ_X and φ_Y be their characteristic functions. We follow Billingsley (1995) in showing that $\varphi_X = \varphi_Y$. Appeal to Ex. A.31(c) to argue that

$$\lim_{k\to\infty} \frac{t^{2k} \, \mathrm{E}\, |X|^{2k}}{(2k)!} = 0, \quad \text{for } t \in (-t_0, t_0).$$

Let $0 < s < \min(t_0, 1)$, and fix an $0 < r < s$. Argue that there is an k_0 so that $2kr^{2k-1} < s^{2k}$ for all $k \geq k_0$. Use this and the inequality $|x|^{2k-1} \leq 1 + |x|^{2k}$ to show that $\lim_{k\to\infty} t^{2k-1} \operatorname{E}|X|^{2k-1}/(2k-1)! = 0$ for $t \in (-t_0, t_0)$, as well. Next, use the inequality in Ex. A.34(e) to show that

$$\left|\varphi_X(t+h) - \sum_{k=0}^{m-1} \frac{t^k}{k!}\varphi_X^{(k)}(t)\right| \leq \frac{|h|^m}{m!}\operatorname{E}|X|^m, \quad \text{for } h \in (-r, r),$$

with a similar inequality holding for φ_Y for the same h, where $\varphi_X^{(k)}(t)$, $k = 0, 1, \ldots$ are the derivatives from Ex. A.34(g). Since X and Y have identitical moment sequences (why?), $\varphi_X^{(k)}(0) = \varphi_Y^{(k)}(0)$ for $k = 0, 1, 2, \ldots$, and so $\varphi_X(t) = \varphi_Y(t)$ for $t \in (-r, r)$. For an arbitrary $0 < \varepsilon < r$, consider the inequality above with $t = r - \varepsilon$ and $t = -r + \varepsilon$, and argue that $\varphi_X(t) = \varphi_Y(t)$ for $t \in (-2r, 2r)$. Use the same argument with $t = 2r - \varepsilon$ and $t = -2r + \varepsilon$, to argue that $\varphi_X(t) = \varphi_Y(t)$ for $t \in (-3r, 3r)$, and so on, forevermore.

(d) (xx perhaps other conditions ensuring that identical moment sequences ensure identical distribiutions. Then a few counterexamples! xx)

Ex. A.39 *Posterior distributions without Bayes.* In Chapter 15 we will need to find posterior distributions without densities, i.e., in situations where neither of the formulae of Ex. A.25(i) are applicable.

(a)

(b)

Ex. A.40 *Something More.* (xx not yet an exercise, but a place to jot down a few comments, also as of 12-August-2024. we need the 'double variance' formula too. and show we round off ChZero with a few things to make the readers feel 'aha, so after all of this, we can do familiar things again', with ordinary integrals and sums and means and variances. perhaps a few simple but nonstandard things too. xx)

Notes and pointers

(xx we point to some of the many books on measure theory for probability and statistics, and also to where key ideas originated. Kolmogorov (1933a,b). Billingsley (1968); Royden and Fitzpatrick (2010) also Shiryaev (1996) and Williams (1991) and Kallenberg (2002). also, briefly, to 'what is a statistical model', McCullagh (2002). Cantor set and Cantor function, F is continuous on $[0, 1]$ but not at all absolutely continuous. In connection with Ex. 2.5, mention that $X_n \to_d X$ and X_n uniformly integrable, implies $\operatorname{E} X_n \to \operatorname{E} X$, and that the proof of this employs a theorem of Skorokhod, see Billingsley (1995, p. 338) xx)

Notes to Ex. A.25. A conditional distribution is said to be *regular* if $P_X(B \,|\, \mathcal{G})(\omega)$ is a distribution on $(\mathcal{X}, \mathcal{B})$ for Pr-almost all $\omega \in \Omega$. Not all conditional distributions are regular (see, e.g. Dudley (2002, Problem 6, p. 351)), but, fortunately, if the measurable space $(\mathcal{X}, \mathcal{B})$ is composed of a complete and separable metric space $(\mathcal{X}, d)$, and the Borel-σ-algebra $\mathcal{B}$ on $\mathcal{X}$ (see Ex. A.2(i)), then there exists a regular conditional distribution

of X given $\mathcal{G}$ (see, e.g., Dudley (2002, Theorem 10.2.2, p. 345) or Schervish (1995, Lemma B.40, p. 621)). In this book, we deal exclusively with complete and separable metric spaces, and all conditional distributions will be assumed regular without further mention.

For Ex. A.17, Emil is indebted to the lecture notes for the course Statistics 381: Measure-Theoretic Probability 1, by Steven Lalley, at the University of Chicago. Provide a citation

References

Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, 2:951–972.

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis. A process point of view*. Springer, New York.

Aalen, O. O. and Gjessing, H. K. (2004). Survival models based on the Ornstein–Uhlenbeck process. *Lifetime Data Analysis*, 10:407–423.

Aït-Sahalia, Y. and Jacod, J. (2014). *High-Frequency Financial Econometrics*. Princeton University Press, Princeton.

Aldous, D. J. and Eagleson, G. K. (1978). On mixing and stability of limit theorems. *The Annals of Probability*, pages 325–331.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, Berlin.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24:3–30.

Ashworth, S., Berry, C. R., and de Mesquita, E. B. (2021). *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton University Press, Princeton.

Aursnes, I., Tvete, I. F., Gåsemyr, J., and Natvig, B. (2005). Suicide attempts in clinical trials with paroxetine randomised against placebo. *BMC Medicine*, xx:1–5.

Aursnes, I., Tvete, I. F., Gåsemyr, J., and Natvig, B. (2006). Even more suicide attempts in clinical trials with paroxetine randomised against placebo. *BMC Psychiatry*, xx:1–3.

Ball, P. (1999). *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*. American Academy for the Advancement of Science, Washington.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*.

Barfort, S., Klemmensen, R., and Larsen, E. G. (2020). Longevity returns to political office. *Political Science Research and Methods*, 9:658–664.

Bartolucci, F. and Lupparelli, M. (2008). Focused Information Criterion for capture-recapture models for closed populations. *Scandinavian Journal of Statistis*, 9:658–664.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a densithy power divergence. *Biometrika*, 85:549–559.

Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Press, London.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, 15:377–180.

Billingsley, P. (1961). *Statistical Inference for Markov Processes*. Chicago University Press, Chicago.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Billingsley, P. (1995). *Probability and Measure. Third Edition*. Wiley, New York.

Blower, J. G., Cook, L. M., and Bishop, J. A. (1981). *Estimating the Size of Animal Populations*. Allen & Unwin, Kondon.

Boitsov, V. D., Karsakov, A. L., and Trofimov, A. G. (2012). Atlantic water temperature and climate in the barents sea, 2000–2009. *ICES Journal of Marine Science*, 69:833–840.

Bolt, U. (2013). *Faster Than Lightning: My Autobiography*. HarperSport, London.

Borgan, Ø., Fiaccone, R. L., Henderson, R., and Barreto, M. L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in brazil. *Scandinavian Journal of Statistics*, 34:53–69.

Borgan, Ø. and Keilman, N. (2019). Do Japanese and Italian women live longer than women in Scandinavia? *European Journal of Population*, 35:87–99.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353–360.

Breiman, L. (2001). Statistical modeling: The two cultures [with comments and a rejoinder by the author]. *Statistical Science*, 16:199–231.

Brunborg, H., Lyngstad, T. H., and Urdal, H. (2003). Accounting for genocide: How many were killed in Srebrenica? *European Journal of Population*, 19:229–248.

Candès, E. J., Lei, L., and Ren, Z. (2021). Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, 84:772–793.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *American Statistician*, 46:167–174.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118:e2107794118.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion and a rejoinder]. *Journal of the American Statistical Association*, 98:900–916.

Claeskens, G. and Hjort, N. L. (2008a). Minimizing average risk in regression. *Econometric Theory*, 24:493–527.

Claeskens, G. and Hjort, N. L. (2008b). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances*, 4:1–9.

Clauset, A. (2020). On the frequency and severity of interstate wars. In Gleditsch, N. P., editor, *Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences*, pages 113–128. Springer, Berlin.

Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous estimation of the means of independent Poisson laws. *Journal of the American Statistical Association*, 70:698–705.

Cox, D. R. (1958). Some problems with statistical inference. *The Annals of Mathematical Statistics*, 29:357–372.

Cox, D. R. (1972). Regression models and life-tables [with discussion]. *Journal of the Royal Statistical Society: Series B*, 34:187–202.

Cox, D. R. and Brandwood, L. (1959). On a discriminatory problem connected with the worlks of Plato. *Journal of the Royal Statistical Society Series B*, 21:195–200.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman & Hall, London.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Cramér, H. (1976). Half a century with probability theory: some personal reflections. *Annals of Probability Theory*, 4:509–546.

Cunen, C. (2015). Mortality and Nobility in the Wars of the Roses and Game of Thrones. *FocuStat Blog, University of Oslo*, iv.

Cunen, C., Hermansen, G. H., and Hjort, N. L. (2018). Confidence distributions for change points and regime shifts. *Journal of Statistical Planning and Inference*, 195:14–34.

Cunen, C. and Hjort, N. L. (2015). Optimal inference via confidence distributions for two-by-two tables modelled as Poisson pairs: fixed and random effects. In Nair, V., editor, *Proceedings of the 60th World Statistics Congress, ISI Rio*, pages xx–xx. Springer, Rio.

Cunen, C. and Hjort, N. L. (2022). Combining information from diverse sources: the II-CC-FF paradigm. *Scandinavian Journal of Statistics*, 49:625–656.

Cunen, C. and Hjort, N. L. (2024). Survival and event history models and methods via Gamma processes. Technical report, University of Oslo. Technical report.

Cunen, C., Hjort, N. L., and Nygård, H. M. (2020a). Statistical sightings of better angels. *Journal of Peace Research*, 57:221–234.

Cunen, C., Hjort, N. L., and Schweder, T. (2020b). Confidence in confidence distributions! *Proceedings of the Royal Society, A*, 476:1–5.

Cunen, C., Walløe, L., and Hjort, N. L. (2020c). Focused model selection for linear mixed models, with an application to whale ecology. *Annals of Applied Statistics*, 14:872–904.

Dagsvik, J. K., Fortuna, M., and Moen, S. H. (2020). How does temperature vary over time?: Evidence on the stationary and fractal nature of temperature fluctuations. *Journal of the Royal Statistical Society A*, pages 883–908.

De Blasi, P. and Hjort, N. L. (2007). Bayesian survival analysis in proportional hazard models with logistic relative risk. *Scandinavian Journal of Statistics*, 34:229–257.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. John Wiley & Sons, Hoboken, N.J.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge.

Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.

Efron, B. (2023). *Exponential Families in Theory and Practice*. Cambridge University Press, Cambridge.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236:119–127.

Einmahl, J. H. J. and Smeets, S. G. W. R. (2011). Ultimate 100 m world records through extreme-value theory. *Statistica Neerlandica*, 65:32–42.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, London.

Fagerland, M., Lydersen, S., and Laake, P. (2017). *Statistical Analaysis of Contingency Tables*. Chapman and Hall/CRC, New YOrk.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80:758–770.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.

Franklin, B. (1793). *The Autobiography of Benjamin Franklin*. Dover, New York. Reprinted from Dover, New York, 1996.

Friesinger, A. (2004). *Mein Leben, mein Sport, meine besten Fitness-Tipps*. Goldmann, Berlin.

Frigessi, A. and Hjort, N. L. (2002). Statistical methods for discontiniuous phenomena. *Journal of Nonparametric Statistics*, 14:1–5.

Galton, F. (1889). *Natural Inheritance*. Macmillan, London.

Geißler, A. (1889). Beiträge zur Frage des Geschlechts verhältnisses der Geborenen. *Zeitschrift des königlichen sächsischen statistischen Bureaus*, 35:1–24.

Gelman, A., Hill, J., and Vehtari, A. (2022). *Regression and Other Stories*. Cambridge University Press, Cambridge.

Gelman, A. and Nolan, D. (2002). A probability model for golf putting. *Teaching Statistics*, 24:93–95.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Ghosh, M. (2002). Basu's theorem with applications: a personalistic review. *Sankhya*, 35:721–741. Special issue in memory of D. Basu.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17:295–314.

Gjessing, H. K., Aalen, O. O., and Hjort, N. L. (2003). Frailty models based on Lévy processes. *Advances in Applied Probability*, 35:532–550.

Glad, I. K., Hjort, N. L., and Ushakov, N. G. (2003). Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, 30:415–427.

Gleditsch, N. P. (2020). *Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences (edited book)*. Springer, Berlin.

Goudie, I. B. J. and Goudie, M. (2007). Who captures the marks for the Petersen estimator? *Journal of the Royal Statistical Society, Series A*, 170:825–839.

Gran, J. M. and Stensrud, M. J. (2022). Hva er forventet levealder? *Tidsskrift for Den norske legeforening*, page 245.

Grønneberg, S. and Hjort, N. L. (2012). On the errors committed by sequences of estimator functionals. *Mathematical Methods of Statistics*, 20:327–346.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrics*, 11:1–12.

Hall, P. G. (1983). Large-sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, 11:1156–1174.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon–Nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, 20:225–241.

Hanche-Olsen, H. and Holden, H. (2010). The Kolmogorov-Riesz compactness theorem. *Expositiones Mathematicase*, 28:385–394.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition.* Springer, New York.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 9:97–109.

Haug, K. K. (2019). Focused model selection for Markov chain models, with an application to armed conflict data. Technical report, University of Oslo. Master Thesis.

Heger, A. (2011). Jeg og jordkloden. *Dagsavisen*, Dec. 16.

Hermansen, G. H., Hjort, N. L., and Kjesbu, O. S. (2016). Modern statistical methods applied on extensive historic data: Hjort liver quality time series 1859-2012 and associated influential factors. *Canadian Journal of Fisheries and Aquatic Sciences*, 73:273–295.

Hjort, J. (1914). *Fluctuations in the Great Fisheries of Northern Europe, Viewed in the Light of Biological Research.* Conseil Permanent International Pour l'Exploration de la Mer, Copenhagen.

Hjort, N. L. (1986a). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinaavian Journal of Statistics*, 13:63–85.

Hjort, N. L. (1986b). *Notes on the Theory of Statistical Symbol Recognition.* Norwegian Computing Centre, Oslo.

Hjort, N. L. (1990a). Goodness of fit tests for life history data based on cumulative hazard rates. *Annals of Statistics*, 18:1221–1258.

Hjort, N. L. (1990b). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, 18:1259–1294.

Hjort, N. L. (1992). On inference in parametric survival data models. *International Statisical Review*, xx:355–387.

Hjort, N. L. (1994). The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association*, 89:665–675.

Hjort, N. L. (2007). And quiet does not flow the Don: Statistical analysis of a quarrel between Nobel laureates. In Østreng, W., editor, *Concilience*, pages 134–140. Centre for Advanced Research, Oslo.

Hjort, N. L. (2008). Discussion of P.L. Davies' article 'Approximating data'. *Journal of the Korean Statistical Society*, 37:221–225.

Hjort, N. L. (2014). Discussion of efron's article 'Estimation and accuracy after model selection'. *Journal of the American Statistical Association*, 110:1017–1020.

Hjort, N. L. (2017a). Cooling of Newborns and the Difference Between 0.244 and 0.278. *FocuStat Blog, University of Oslo*, xv.

Hjort, N. L. (2017b). The Semifinals Factor for Skiing Fast in the Finals. *FocuStat Blog, University of Oslo*, xv.

Hjort, N. L. (2018a). Overdispersed Children. *FocuStat Blog, University of Oslo*, xxi.

Hjort, N. L. (2018b). Towards a More Peaceful World [insert '!' or '?' here]. *FocuStat Blog, University of Oslo*, xvii.

Hjort, N. L. (2019a). The Magic Square of 33. *FocuStat Blog, University of Oslo*, xxi.

Hjort, N. L. (2019b). Sudoku Solving by Probability Models and Markov Chains. *FocuStat Blog, University of Oslo*, xxi.

Hjort, N. L. and Claeskens, G. (2003a). Frequentist model averaging [with discussion and a rejoinder]. *Journal of the American Statistical Association*, 98:879–899.

Hjort, N. L. and Claeskens, G. (2003b). Rejoinder to the discussion of the Hjort and Claeskens and Claeskens and Hjort papers. *Journal of the American Statistical Association*, 98:917–925.

Hjort, N. L. and Fenstad, G. (1992). On the last time and the number of times an estimator is more than ε from its target value. *The Annals of Statistics*, 20:469–489.

Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 23:882–904.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24:1619–1647.

Hjort, N. L. and Koning, A. J. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, 14:113–132.

Hjort, N. L. and Lumley, T. (1993). Normalised local hazard plots. Technical report, Department of Statistics, University of Oxford, Oxford.

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics*, 37:1079–1111.

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2018). Hybrid combinations of parametric and empirical likelihoods. *Statistica Sinica*, 27:2389–2407.

Hjort, N. L. and Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In Nair, V., editor, *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell Doksum*, pages 463–492. World Scientific, New Jersey.

Hjort, N. L. and Pollard, D. B. (1993). Asymptotics for minimisers of convex processes. Technical report, Department of Mathematics, University of Oslo.

Hjort, N. L. and Schweder, T. (2018). Confidence distributions and related themes: introduction to the special issue. *Journal of Statistical Planning and Inference*, 195:1–13.

Hjort, N. L. and Stoltenberg, E. A. (2021). The partly parametric and partly nonparametric additive risk model. *Lifetime Data Analysis*, 27:1–31.

Hjort, N. L. and Varin, C. (2008). ML, PL, QL in Markov chain models. *Scandinavian Journal of Statistics*, 35:64–82.

Hjort, N. L. and Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *Annals of Statistics*, 37:105–131.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960.

Holum, D. (1984). *The Complete Handbook of Speed Skating*. High Peaks Cyclery, Lake Pacid.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Logistic Regression*. Wiley, New York.

Hveberg, K. (2019). *Lene din ensomhet langsomt mot min*. Aschehoug, Oslo.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge.

Inlow, M. (2010). A moment generating function proof of the Lindeberg–Lévy central limit theorem. *American Statistician*, 64:228–230.

Jacod, J. and Mémin, J. (1981). Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité. In *Séminaire de Probabilités (Strasbourg), tome 15*, pages 529–546. Springer.

Jacod, J. and Protter, P. (2004). *Probability Essentials. Second Edition*. Springer, Berlin.

Jacod, J. and Shiryaev, A. (2013). *Limit Theorems for Stochastic Processes. Second Edition*. Springer, Berlin.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R. Second Edition*. Springer, New York.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379.

Jamtveit, B., Jacobsen, A. U., and Wyller, T. B. (2018). Utvikling i andel administrativt personale i norske helseforetak. *Samfunnsøkonomen*, 6:17–21.

Jamtveit, B., Jettestuen, E., and Mathiesen, J. (2009). Scaling properties of European research units. *Proceedings of the National Academy of Sciences*, 106:13160–13163.

Jansen, D. (1994). *Full Circle*. Villard Books, New York.

Jones, M. C. (1991). The roles of ISE and MISE in density estimation. *Statistics and Probability Letters*, 12:51–56.

Jones, M. C., Hjort, N. L., Harris, I. R., and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, 88:865–873.

Jullum, M. and Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, 27:951–981.

Jullum, M. and Hjort, N. L. (2019). What price semiparametric Cox regression? *Lifetime Data Analysis*, 25:406–438.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2020). *Noise: A Flaw in Human Judgment*. William Collins, London.

Kallenberg, O. (2002). *Foundations of Modern Probability. Second Edition*. Springer, Berlin.

Kjesbu, O. S., Opdal, A. F., Korsbrekke, K., Devine, J. A., and Skjæraasen, J. E. (2014). Making use of Johan Hjort's 'unknown' legacy: reconstruction of a 150-year coastal time-series on northeast Arctic cod (Gadus morhua) liver data reveals long-term trends in energy allocation patterns. *ICES Journal of Marine Science*, 71:2053–2063.

Kjetsaa, G., Gustavson, S., Beckman, B., and Gil, S. (1984). *The Authorship of The Quiet Don [also published in Russian]*. Solum/Humanities Press, Oslo.

Klein, R., Knudtson, M. D., Lee, K. E., Gangnon, R., and Klein, B. E. (2008). The Wisconsin epidemiologic study of diabetic retinopathy: XXII the twenty-five-year progression of retinopathy in persons with type 1 diabetes. *Ophtalmology*, 115:1859–1868.

Klotz, J. (1972). Markov chain clustering of births by year. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability Theory*, 4:173–185.

Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. *Annals of Statistics*, 1:373–379.

Koehler, J. J. and Conley, C. A. (2003). The "hot hand" myth in professional basketball. *Journal of Sport and Exercise Psychology*, 25:253–259.

Kolmogorov, A. N. (1933a). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer, Berlin.

Kolmogorov, A. N. (1933b). Sulla determinazione empirica di una legge di distribuzione. *Giorn Ist Ital Attuar*, 4:83–91.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer, New York.

Kusolitsch, N. (2010). Why the theorem of Scheffé should be rather called a theorem of Riesz. *Periodica Mathematica Hungarica*, 61:225–229.

Laptook, A. e. a. (2017). Effect of therapeutic hypothermia initiated after 6 hours of age on death and disability among newborns with hypoxic-ischemic encephalopathy: A randomized clinical trial. *Journal of the American Medical Association*, 318:1550–1560.

Larkey, P. D., Smith, R. A., and Kadane, J. B. (1989). It's okay to believe in the "hot hand". *Chance*, 2:22–30.

Le May Doan, C. (2002). *Going For Gold*. McClelland & Stewart Publisher, Toronto.

LeCam, L. (1986). The Central Limit Theorem around 1935. *Statistical Science*, 1:78–91.

Lehmann, E. L. (1950). *Notes on the Theory of Estimation*. Berkeley University Press, Berkeley. Notes recorded by Colin Blyth.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111.

Leike, A. (2001). Demonstration of the exponential decay law using beer froth. *European Journal of Physics*, 23:1–21.

Lessing, D. (1997). *Walking in the Shade: Volume Two of My Autobiography, 1949 to 1962*. xx, xx.

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15:211–225.

Lindqvist, B. H. (1978). A note on Bernoulli trials with dependence. *Scandinavian Journal of Statistics*, 5:205–208.

Loader, C. (1996). Local likelihood density estimation. *Annals of Statistics*, 67:1602–1618.

Lum, K., Price, M. E., and Banks, D. (2013). Applications of multiple systems estimation in human rights research. *American Statistician*, 24:191–200.

Markov, A. A. (1906). Распространение закона больших чисел на величины, зависящие друг от друга [Extending the law of large numbers for variables that are dependent of each other]. Известия Физико-математического общества при Казанском университете (2-я серия), 15:124–156.

Markov, A. A. (1913). Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь [Example of a statistical investigation illustrating the transitions in the chain for the 'Evgenii Onegin' text]. Известия Академии Наук, Санкт-Петербург (6-я серия), 7:153–162.

Marron, S. and Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics*, 20:712–736.

McCloskey, R. (1943). *Homer Price*. Scholastic Inc., New York.

McCullagh, P. (2002). What is a statistical model? [with discussion]. *Annals of Statistics*, 30:1225–1310.

Miller, J. B. and Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, 86:2019–2047.

Miller, J. B. and Sanjurjo, A. (2021). Is it a fallacy to believe in the hot hand in the NBA three-point contest? *European Economic Review*, 138:103771.

Mykland, P. A. and Zhang, L. (2012). The econometrics of high frequency data. In Kessler, M., Lindner, A., and Sørensen, M., editors, *Statistical Methods for Stochastic Differential Equations*, pages 109–190. CRC Press.

Mykland, P. A., Zhang, L., and Chen, D. (2019). The algebra of two scales estimation, and the S-TSRV: High frequency estimation that is robust to sampling times. *Journal of Econometrics*, 208:101–119.

Neyman, J. and Pearson, E. (1933). On the problem of the most efficient statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 68:289–337.

Normand, S.-L. T. (1999). Tutorial in biostatistics: Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18:321–359.

O'Neill, B. (2014). Some useful moment results in sampling problems. *American Statistician*, A 231:282–296.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series*, 5(302):157–175.

Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika*, 1:261–264.

Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, 6:5–84.

Peterson, A. V. (1975). Nonparametric estimation in the competing risks problem. Technical report, Department of Statistics, Stanford University.

Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Books, Toronto.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

Price, R. M. and Bonett, D. G. (2001). Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation*, 68:xx–xx.

Price, R. M. and Bonett, D. G. (2002). Distribution-free confidence intervals for difference and ratio of medians. *Journal of Statistical Computation and Simulation*, 72:xx–xx.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletins of the Calcutta Mathematical Society*, pages 81–91.

Reeves, R. V. (2022a). *Of Boys and Men: Why the Modern Male is Struggling, Why it Matters, and What to Do About It*. Brookings Institution Press, Washington, D.C.

Reeves, R. V. (2022b). Redshirt the boys. *The Atlantic*, October.

Romano, J. P. and Siegel, A. F. (1986). *Counterexamples in Probability and Statistics*. Wadsworth & Brooks/Cole, Belmont.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Royden, H. L. and Fitzpatrick, P. M. (2010). *Real Analysis [4th ed.]*. Pearson Education Asia, Beijin.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimation. *Scandinavian Journal of Statistics*, 9:65–78.

Rydén, J. (2020). On features of fugue subjects: A comparison of J.S. Bach and later composers. *Journal of Mathematics and Music*, pages 1–20.

Saleh, J. H. (2019). Statistical reliability analysis for a most dangerous occupation: Roman emperor. *Palgrave Communication*, 5:1–7.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43:142–1542.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18:434–438.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

Schervish, M. J. (1995). *Theory of Statistics*. Springer, New York.

Schömig, A., Mehili, J., de Waha, A., Seyfarth, M., Pahce, J., and Kastrati, A. (2008). A meta-analysis of 17 randomized trials of a percutaneous coronary intervention-based strategy in patients with stable coronary artery disease. *Journal of the American College of Cardiology*, 52:894–904.

Schweder, T. (1980). Scandinavian statistics, some early lines of development. *Scandinavian Journal of Statistics*, 7:113–129.

Schweder, T. (1999). Early statistics in the Nordic countries – when did the Scandinavians slip behind the British? *Bulletin of the International Statistical Institute*, 58:1–4.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, London.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9.

Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica*, 1:185–202.

Shiryaev, A. N. (1996). *Probability. Second edition*. Springer, Berlin.

Shumway, R. H. and Stoffer, D. S. (2016). *Time Seires Analysis and Its Applications [4th ed.]*. Springer, Heidelberg.

Silver, N. (2012). *The Signal and the Noise: Why so Many Predictions Fail, but Some Don't*. Penguin.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Simpson, R. J. S. and Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3:1243–1246.

Sims, C. A. (2012a). Appendix: inference for the Haavelmo model. Technical report, Puplic Policy & Finance, Princeton University, Princeton, NJ.

Sims, C. A. (2012b). Statistical modeling of monetary policy and its effects [Sveriges Riksbank Prize in Memory of Alfred Nobel lecture]. *American Economic Review*, xx:1–22.

Singh, K., Xie, M., and Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Annals of Statistics*, 33:159–183.

Slud, E. (1989). Clipped Gaussian processes are never M-step Markov. *Journal of Multivariate Analysis*, 29:1–14.

Smith, T. D. (1994). *Scaling Fisheries: The Science of Measuring the Effects of Fishing 1855–1955*. Cambridge University Press, Cambridge.

Spiegelberg, W. (1901). *Aegyptische und Griechische Eigennamen aus Mumientiketten der Römischen Kaiserzeit*. Greek Inscriptions, Cairo.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206.

Stigler, S. M. (1973). Studies in the history of probability and statistics. xxxii: Laplace, Fisher and the discovery of the concept of sufficiency. *Biometrika*, 60:439–445.

Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, 27:1055–1098.

Stigler, S. M. (1983). Who discovered Bayes's Theorem? *American Statistician*, 37:290–296.

Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155.

Stigler, S. M. (2006). How Ronald Fisher became a mathematical statistician. *Mathematics and Social Sciences*, 44:23–30.

Stoltenberg, E. A. (2019). An MGF proof of the Lindeberg theorem. Technical report, Department of Mathematics, University of Oslo.

Stoltenberg, E. A. and Hjort, N. L. (2021). Models and inference for on-off data via clipped Ornsten–Uhlenbeck processes. *Scandinavian Journal of Statistics*, 48:908–929.

Stout, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York.

Student (1908). The probable error of a mean. *Biometrika*, 6:1–25.

Swensen, A. R. (1983). A note on convergence of distributions of conditional moments. *Scandinavian Journal of Statistics*, 10:41–44.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

Tversky, A. and Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. *Chance*, 2:16–21.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Varian, H. R. (1975). Distributive justice, welfare economics, and the theory of fairness. *Philosophy and Public Affairs*, 4:223–247.

Voldner, B., Frøslie, K. F., Haakstad, L., Hoff, C., and Godang, K. (2008). Modifiable determinants of fetal macrosomia: role of lifestyle-related factors. *Acta Obstetricia et Gynecologica Scandinavica*, 87:423–429.

von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Annals of Mathematical Statistics*, xx:808–818.

von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*. B.G. Teubner, Berlin.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media, Berlin/Heidelberg.

Walløe, L., Hjort, N. L., and Thoresen, M. (2019a). Major concerns about late hypothermia study. *Acta Paediatrica*, 108:588–589.

Walløe, L., Hjort, N. L., and Thoresen, M. (2019b). Why results from Bayesian statistical analyses of clinical trials with a strong prior and small sample sizes may be misleading: The case of the NICHD Neonatal Research Network Late Hypothermia Trial. *Acta Paediatrica*, 108:1190–1191.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wardrop, R. L. (1995). Simpson's paradox and the hot hand in basketball. *The American Statistician*, 49:24–28.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge.

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D., Riffe, T., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., C, W., and M, B. (2021). Methods protocol for the Human Mortality Database. University of California, Berkeley, US, and Max Planck Institute for Demographic Research, Rostock, Germany. `https://www.mortality.org/` [Version 6. Last revised January 26, 2021].

Wissner-Gross, Z. (2020). Can you feed the hot hand? https://fivethirtyeight.com/features/can-you-feed-the-hot-hand/. Accessed: December 12, 2020.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review [with discussion and a rejoinder]. *International Statistical Review*, 81:3–39.

Zabriskie, B. N., Corcoran, C., and Senchaudhuri, P. (2021). A comparison of confidence distribution approaches for rare event meta-analysis. *Statistics in Medicine*, 40:5276–5297.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81:446–451.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100:1394–1411.

Name index

Subject index