Statistical Inference:
666 Exercises, 66 Stories, and Solutions

Nils Lid Hjort
University of Oslo

Emil Aas Stoltenberg
BI Norwegian Business School

– This version, with 0.90 versions of Chapters 1-8 plus Appendix,
last touched by Nils and by Emil, 13-Aug-2023 –

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 several crisp 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, also 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, for simulation etc.)

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. Many 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, Aaron Clauset, Dennis Cristensen, Ingrid Dæhlen, Åsa Engelstad, Arnoldo Frigessi, Ingrid Glad, Håvard Hegre, Aliaksandr Hubin, Ingrid Hobæk Haff, Kristoffer Hellton, Bjørn Jamtveit, Martin Jullum, Vinnie Ko, Alexander Koning, Per Mykland, Jonas Moss, Håvard Mokleiv Nygård, Lars Olsen, Catharina Stoltenberg, Gunnar Taraldsen, Ingunn Fride Tvete, Sam-Erik Walker, Lars Walløe, Jonathan Williams, Lan Zhang.

We have also benefited, directly and indirectly, through the collective efforts of 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 Hegre and Hjort) at the Centre for Advanced Study, Academy of Science and Letters, Oslo; the grand *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 grati-

tude a partial support stipend from the Norwegian Non-Fiction Writers and Translators Association (Norsk faglitterær forfatter- og oversetterforening).

(xx Then current time plan, as of 13-Aug-2023, possibly optimistic xx)

Nils Lid Hjort and Emil Aas Stoltenberg
Blindern, some day in 2023

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, Beta, Dirichlet, Poisson, compound Poisson, binomial, multinomial, geometric. 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. Mathematical techniques for deriving crucial properties include those of moment-generating functions. In connection with studying the general exponential family of models, which has various classic models as special cases, we also discuss what it means for a function of a dataset to be sufficient, with related themes returned to in later chapters.

1.A Chapter introduction

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 this 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. Thus a model with probability density $f(y)$ has a cumulative distribution function (c.d.f.) $F(y) = \int_{-\infty}^{y} f(y')\,\mathrm{d}y'$,

its mean is $\mathrm{E}\,Y = \int y f(y)\,\mathrm{d}y$, its median is $F^{-1}(\frac{1}{2})$, its variance $\mathrm{Var}\,Y = \mathrm{E}\,(Y - \mathrm{E}\,Y)^2$, etc. There are also occasions where the double expectation rule $\mathrm{E}\,X = \mathrm{E}\,\mathrm{E}\,(X \,|\, Y)$ of Ex. **??** comes 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. 1.20-1.21), 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 case (see Ex. 1.57-1.59). In that connection we also discuss the notion of a function of a dataset being *sufficient* (see Ex. 1.61-1.63), foreshadowing material in Chs. 5, 6, 8.

(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. 4. xx)

1.B Distributions & densities

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 can be written

Gauß

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

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

(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. 1.32; 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 $f(x) = e^{-\pi x^2}$, 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 $P(X \ge x) \doteq \phi(x)/x$, 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 $P(X \in [x, x+\varepsilon] \,|\, X \ge x) \doteq x\varepsilon$ 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 actually clearly easiest to demonstrate via moment-generating functions, as we come back to in Ex. 1.20, 1.21, 1.22, but here we show this in a more direct fashion.

(a) Let X and Y be independent standard normals. Use the convolution formulae from Ex. A.23 to show that $X + Y \sim \mathrm{N}(0, 2)$. 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 $\sum_{i=1}^m a_i\sigma_i^2$.

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

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

Explain that Y can be expressed as $X_1+\cdots+X_n$, where X_i is a simple 0-1 variable, with $P(X_i=1)=p$, 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 $P(Y=y)$ 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 $\gamma_3=\mathrm{E}\,W^3$, 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

$$\gamma_3=\mathrm{E}\Big[\frac{Y-np}{\{np(1-p)\}^{1/2}}\Big]^3=\frac{1-2p}{\{np(1-p)\}^{1/2}},$$

going to zero with rate $1/\sqrt{n}$. Briefly discuss what this entails regarding the degree of asymmetry for the binomial distribution.

(c) After the skewness comes the so-called kurtosis, defined as $\gamma_4=\mathrm{E}\,W^4-3$, 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 the kurtosis

$$\gamma_4=(1/n)[1/\{p(1-p)\}-6],$$

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

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

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 number 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 $P(X \le x_0, Y \le y_0)$, 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 $P(X \le 15, Y \le 15)$.

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 $j=1,\ldots,n$. 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 $(Y_1,\ldots,Y_k)$.

the multinomial model

(b) Show that the joint probability distribution becomes

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

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.

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

for $y_1 \ge 0, y_2 \ge 0, y_1+y_2 \le n$. Note that formulae from Ex. 1.4 therefore apply to pairs (Y_i,Y_j) here.

(d) Show that $\mathrm{cov}(Y_1,Y_2) = -np_1p_2$, and find the correlation between Y_i and Y_j.

(e) Among the most used acronyms of statistical parlance is *i.i.d.*, for independent 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 density f. Create disjoint cells $C_1, \ldots, C_k$, with $C_j = (a_{j-1}, a_j]$, for $a_0 < \cdots < a_k$. Let then N_j count the number of data points in cell j. The *histogram*, associated with the chosen cells, is then

$$\widehat{f}(x) = \widehat{p}_j/|C_j| \quad \text{for } x \in C_j,$$

where $\widehat{p}_j = N_j/n$ estimates $p_j = P(Y_i \in C_j)$; also, $|C_j| = a_j - a_{j-1}$ is the length of that cell.

(a) Show that $(N_1, \ldots, N_k)$ is multinomial. Find expressions for the mean and variance of $\widehat{f}(x)$.

(b) (xx some easy simulation, from the normal. play with k being small, big, and about right. point to density estimation. xx)

Ex. 1.7 *Hazard rates and survival functions.* 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

$$P(T \in [t, t+\varepsilon] \,|\, T \geq t) = h(t)\varepsilon + O(\varepsilon^2), \tag{1.1}$$

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 $S(t) = P(T \geq t) = \exp\{-H(t)\}$ 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

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

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

Ex. 1.8 *The exponential distribution.* Here and below we shall partly follow the implied tradition of using say T and $f(t)$ and $h(t)$, for random variables and their densities and hazard rate functions, rather han say Y and $f(y)$ and $h(y)$, when these relate to *time*. – A simple but important distribution in probability theory and statistics is the exponential distribution, which with positive parameter θ is the density $f(t, \theta) = \theta \exp(-\theta t)$ for $t > 0$. We write $Y \sim \text{Expo}(\theta)$ 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 $P(T \geq t_0 + t \,|\, T \geq t_0)$ 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.

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

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

the Gamma distribution

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 $m = 1, 2, \ldots$.

(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 $\gamma_3 = 2/a^{1/2}$ and $\gamma_4 = 6/a$. Finally, regarding moments, find that the inverse gamma distributed variable $1/T$ has mean $b/(a-1)$ and finite variance (xx check this xx) $b^2/\{(a-1)^2(a-2)\}$, 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 (1.1), 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 $T_1 + T_2 \sim \text{Gam}(2, \theta)$. With $T_1, \ldots, T_k$ seen as the independent waiting times between events, show that the time to event k is a $\text{Gam}(k, \theta)$.

(f) With $T_1 \sim \text{Gam}(a_1, b)$ and $T_2 \sim \text{Gam}(a_2, b)$ independent, show that $T_1 + T_2 \sim \text{Gam}(a_1 + a_2, b)$. Generalise. This may indeed be accomplished via the convolution formulae from Ex. A.23, but as for other instances it becomes easier to show such statements via moment-generating functions; see Ex. 1.20–1.21.

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 $\text{Expo}(\theta)$ distribution, but the values of θ vary from occasion to occasion.

(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 $\text{E}\,(T \,|\, \theta) = 1/\theta$ and $\text{Var}\,(T \,|\, \theta) = 1/\theta^2$, that $\text{E}\, T = 1/\theta_0$ and $\text{Var}\, T = 1/\theta_0^2 + 2\tau^2$. 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.2 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 $g(t, a, b)$ distribution. This might be used to estimate (a, b) from data. (xx could point to Story iii.2, perhaps with better calibration. xx)

(g) (xx Can use the $n = 799$ nerve impulse data from Hand et al. data collection. 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 $\text{E}\, 1/b = 1/b_0$ and $\text{Var}\, 1/b = \tau^2$. 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)

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

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 $P(Y \le y) = P(X \le h(Y))$, 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

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

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), wth 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 $[0, 2\pi]$. Find also the distribution of $Z = Y/X = \tan A$; see also Ex. 1.13.

(e) 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 $R = X/(X + Y)$. Find the joint density for (R, Z).

Ex. 1.13 *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 $g(v) = \int_0^\infty xf(x)f(vx)\,\mathrm{d}x$. 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 $g(v) = 2\int_0^\infty xf(x)f(vx)\,\mathrm{d}x$. 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. Show that it has no mean. Find its interquartile range.

Ex. 1.14 *The Poisson distribution.* (xx make a pointer here to Poisson processes, which we perhaps have in Ch. 9. xx) Let a count variable Y have the point probabilities

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

We say that Y has the Poisson distribution with parameter θ, and write $Y \sim \text{Pois}(\theta)$.

(a) Show that the probabilities indeed sum to 1, and that the mean and variance are both equal to θ. Letting $W = (Y-\theta)/\sqrt{\theta}$, show that $\gamma_3 = \mathrm{E}\, W^3 = 1/\sqrt{\theta}$, $\gamma_4 = \mathrm{E}\, W^4 - 3 = 1/\theta$. Show also that $\mathrm{Var}\,(Y - \theta)^2 = 2\theta^2 + \theta$.

(b) With $Y \sim \text{Pois}(\theta)$, 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 $\text{Pois}(\theta)$ distribution, in the sense that the point probabilities converge. (xx here a few pointers, to more general poisson limits, and a full process. xx)

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

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

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 histgram.

Ex. 1.15 *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 $\mathrm{E}\,Y = p + (1-p)(1+\mathrm{E}\,Y)$ and solve. Use this representation to also find the variance. Show also that $\mathrm{E}\,(Y-1/p)^3 = (1-p)(2-p)/p^3$.

(d) Find expresssions for $P(Y \geq y)$ and for $P(Y \geq y_0 + y \,|\, Y \geq y_0)$, and comment.

(e) (xx edit and perhaps wash away; we haven't seen the CLT yet; could come back to this after seeing CLT. cc) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the geom(p), with $\bar{Y}$ the sample average. Find the limit distribution of $\sqrt{n}(\bar{Y} - 1/p)$, and then the limit distribution of $\sqrt{n}(\widehat{p} - p)$, where $\widehat{p} = 1/\bar{Y}$.

(f) 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 $P(Y_0 = y) = q^y p$ for $y = 0, 1, \ldots$, writing $q = 1-p$. Show that Y_0 has mean $(1-p)/p$ and variance $(1-p)/p^2$. Show also that $G(s) = \mathrm{E}\,s^{Y_0} = p/(1-qs)$ for $|s| < 1/q$.

(g) (xx something re the two experiments for determining p, the binomial and the geometric. each experiment costs 100 kroner. precision and cost. xx)

Ex. 1.16 *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 $Y \mid \theta \sim \text{Pois}(\theta)$ 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 $\theta \sim \text{Gam}(a, b)$, see Ex. 1.9. Show that $\text{E}\, Y = \theta_0 = a/b$ and that $\text{Var}\, Y = \theta_0(1+1/b)$. 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

$$g(y, a, p) = \frac{\Gamma(a+y)}{\Gamma(a)\, y!}(1-p)^y p^a \quad \text{for } y = 0, 1, \ldots, \tag{1.2}$$

for parameters $a > 0$, $p \in (0, 1)$; see Ex. 1.17 for more details.

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

Ex. 1.17 *The negative binomial.* We met the negative binomial distribution in Ex. 1.16 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 $P(Y = y) = (y+1)q^y p^2$, for $y = 0, 1, \ldots$. For $Y = X_1 + X_2 + X_3$ a sum of three such independent geometric variables, show that $P(Y = y) = \binom{y+2}{2} q^y p^3$ 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

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

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 (1.2) 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 $G_a(s) = \text{E}\, s^Y = \{p/(1-qs)\}^a$ for $|s| < 1/q$, and that $\text{E}\, Y = aq/p$, $\text{Var}\, Y = aq/p^2$.

(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. 1.18 *The uniform.* [xx to be polished. xx] can we simply invert $\{\exp(t)-1\}^n/t^n$, for $n = 2, 3, \ldots$? $\mathrm{E}\, N = e$ for N, the number of uniforms to sum to reach 1. xx] We say 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.

(b) Let $U_1, U_2, \ldots$ be i.i.d. uniforms on the unit interval. Show that the densities for $U_1 + U_2$ and $U_1 + U_2 + U_3$ may be written

$$f_2(x) = \begin{cases} x & \text{when } 0 \le x \le 1, \\ 2 - x & \text{when } 1 \le x \le 2, \end{cases}$$

and

$$f_3(x) = \begin{cases} \frac{1}{2}x^2 & \text{when } 0 \le x \le 1, \\ \frac{1}{2}(-2x^2 + 6x - 3) & \text{when } 1 \le x \le 2, \\ \frac{1}{2}(3-x)^2 & \text{when } 2 \le x \le 3. \end{cases}$$

Show that f_2 is continuous, but that its derivative has a jump at position $x = 1$. Show however that f_3 is smoother, with a continuous derivative.

(c) Generalise the above to the case of $X = U_1 + \cdots + U_n$; show that its density may be written

$$f_n(x) = \frac{1}{(n-1)!} \sum_{j=0}^{\lfloor x \rfloor} (-1)^j \binom{n}{j} (x-j)^{n-1}$$

for $0 \le x \le n$, where $\lfloor x \rfloor$ is the so-called floor function, the largest integer equal to or to the left of x (so $\lfloor 2.99 \rfloor = 2$, $\lfloor 3.00 \rfloor = 3$, $\lfloor 3.01 \rfloor = 3$). In particular, show that $f_n(x) = x^{n-1}/(n-1)!$ for $0 \le x \le 1$. This is sometimes called the Irwin–Hall distribution, from the related and concurrently published papers Irwin (1927), Hall (1927).

(d) Draw the densities f_1, f_2, f_3, f_4, f_5 in a diagram, and comment on their forms.

(e) Let now N be the number of uniforms U_j needed in order for their sum to exceed 1. Show that

$$P(N > n) = P(U_1 + \cdots + U_n < 1) = F_n(1),$$

with F_n the cumulative for f_n. Also show that in fact $F_n(1) = 1/n!$. Use this to deduce that $P(N = n) = 1/\{n\,(n-2)!\}$ and that $\mathrm{E}\,N = e$. As a follow-up computational exercise, simulate a high number of such N to calculate the mathematical constant e to three decimal places.

Ex. 1.19 *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.10 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 X a standard normal, find formulae for $\mathrm{E}\,X^k$ and for $\mathrm{E}\,|X|^k$.

(c) For $r < s$, show that $(\mathrm{E}\,|X|^r)^{1/r} \le (\mathrm{E}\,|X|^s)^{1/s}$, i.e. $h(r) = (\mathrm{E}\,|X|^r)^{1/r}$ is a non-decreasing function in r. (xx have we given the Jensen inequality somewhere? perhaps do it here. xx) 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 $X \sim \mathrm{Expo}(1)$.

(d) For a random variable X with finite fourth moment, we have defined its skewness γ_3 and kurtosis γ_4 in Ex. 1.3. Give expressions $\gamma_3 = h_3(\mu_1, \mu_2, \mu_3)$ and $\gamma_4 = h_4(\mu_1, \mu_2, \mu_3, \mu_4)$ in terms of the moments $\mu_j = \mathrm{E}\,X^j$, and also expressions $\gamma_3 = h_3^*(\mu_2^*, \mu_3^*)$ and $\gamma_4 = h_4^*(\mu_2^*, \mu_3^*, \mu_4^*)$ in terms of the centralised moments $\mu_j^* = \mathrm{E}\,(X - \mu_1)^j$.

Ex. 1.20 *Moment-generating functions: examples.* 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. It 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 $f(y)$, 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.10. [xx the list is a bit too long; may take the chi-squared and the non-central chi-squared to a separate exercise. xx]

(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 $t < b$.

(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. 1.16, 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. 1.21 *Moment-generating functions: properties.* Among the basic properties of moment-generating functions are the following; attempt to demonstrate these.

(a) We have $M(0) = 1$, and when the mean is finite, then $M'(t)$ exists, with $M'(0) = \text{E}\, Y$.

(b) More generally, if $|Y|^r$ has finite mean, then $M^{(r)}(0) = \text{E}\, Y^r$ (the rth derivative of M, at the point zero). So the moment-generating function generates moments!

(c) For $X \sim \text{N}(0, 1)$, show the $M(t)$ for $|X|$ becomes $2\exp(\frac{1}{2}t^2)\Phi(t)$, and use this to find its mean and variance.

(d) 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 $Y \sim \text{Pois}(\theta)$, 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.

(e) 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 $|t|$.

(f) Suppose X and Y are variables with distributions on the unit interval, with identical moment-generating functions, say $\int_0^1 \exp(tx)\, \text{d}F(x) = \int_0^1 \exp(tx)\, \text{d}G(x)$, at least for all t in an interval around zero. Show that X and Y then have identical moment sequences, and that $\int_0^1 p(x)\, \text{d}F(x) = \int_0^1 p(x)\, \text{d}G(x)$ for all polynomials $p(x)$. Use the Weierstraß approximation theorem, see Ex. 4.28, to show that this equality must hold for all continuous functions $p(x)$, and use this to prove $F = G$.

(g) (xx calibrate this, point to full crucial theorem saying $M_X = M_Y$ in interval around zero implies equality of distributions. point to later things. xx) If X and Y are two variables with identical moment-generating functions, then their distributions are identical. [xx There are also 'inversion formulae' in the literature, giving the distribution as a function of M. need to give clear pointer to proofs for this, in Ch. 4. xx]

Ex. 1.22 *Moment-generating functions for sums.* (xx point here to Ex. A.23, 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. 1.23 *The Beta distribution.* An important class of distributions, over the unit interval $(0, 1)$, is the *Beta distribution*, with two positive parameters. We write $p \sim \mathrm{Beta}(a, b)$ if its density is

the Beta distribution

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

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

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

(c) (xx just a bit more. Write $(a, b) = (cp_0, c(1 - p_0))$. Find a formula for $\mathrm{E}\, p^m$, for $m = 1, 2, 3, \ldots$. Find the skewness of p. xx)

(d) (xx some attention to the density $f(p) = (1/\pi)/\sqrt{p(1 - p)}$, which is the Beta$(\frac{1}{2}, \frac{1}{2})$. show that its cumulative is $F(p) = (2/\pi) \arcsin(\sqrt{p})$. Find is quantile $F^{-1}(q)$. point to Jeffreys prior, and also to the 'how much of the time is one of two teams leading' in random walks and Brownian motion. xx)

Ex. 1.24 *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 $X \sim \mathrm{Dir}(a_1, \ldots, a_k)$.

the Dirichlet distribution

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

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

$$\mathrm{E}\, D_i = \frac{a_i}{a}, \quad \mathrm{Var}\, D_i = \frac{1}{a+1} \frac{a_i}{a} \Big(1 - \frac{a_i}{a}\Big).$$

Show also the $\mathrm{cov}(D_i, D_j) = -(a_i/a)(a_j/a)/(a+1)$ 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. Tending to this now, however, show that

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

over the simplex, using Ex. 1.12.

(d) xx

Ex. 1.25 *The Beta-binomial distribution.* [xx intro sentences. sometimes p not quite the same, in a row of experiments. will be used for Story i.3. xx]

(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 rules of Ex. **??**, 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

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

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.

(c) (xx estimating (a, b) from data via sample mean and variance. pointer to Story i.3. xx)

(d) (xx pointer to Bayes things. xx)

Ex. 1.26 *The Dirichlet-multinomial distribution.* Here we deal with the natural extension of the Beta-binomial setup of Ex. 1.25, 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 $p \sim \text{Dir}(cp_0)$, with parameters $cp_0 = (cp_{0,i}, \ldots, cp_{0,k})$. Show that

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

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.

Ex. 1.27 *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. the Laplace distribution

(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 moment-generating function 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 moment-generating function 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 moment-generating function for such a normal scale mixture becomes $M(t)=\mathrm{E}\exp(tX)=M_V(\frac12 t^2)$, where $M_V(s)$ is the moment-generating function 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

$$\begin{aligned}\int_0^\infty \frac{1}{(2\pi)^{1/2}}\exp\Big(-\frac12\frac{x^2}{v}\Big)\frac{\exp(-v)}{v^{1/2}}\,\mathrm{d}v &= 2\int_0^\infty \frac{1}{(2\pi)^{1/2}}\exp\{-(\tfrac12 x^2/w^2+w^2)\}\,\mathrm{d}w\\ &= \tfrac12\sqrt{2}\exp(-\sqrt{2}|x|).\end{aligned}$$

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

(e) (xx point to Ex. 4.61. xx)

Ex. 1.28 *Mixing the normal scale.* (xx at the moment nils thinks this exercise will go away, partly with material in Stpru v.6 and partly elsewhere. point back to and calibrate with Ex. 1.27. 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 $\mathrm{N}(\xi,\sigma^2)$. Assume next that in a population of such X, there is a distribution $\xi\sim\mathrm{N}(\xi_0,\sigma^2_{\rm extra})$ of their means. Show that an X sampled from that population is a $\mathrm{N}(\xi_0,\sigma^2+\sigma^2_{\rm extra})$. 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

$$\bar f(x)=\frac{1}{(2\pi)^{1/2}}\frac{b^a}{\Gamma(a)}\frac{\Gamma(a+\frac12)}{(b+\frac12x^2)^{a+1/2}}.$$

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

$$V=(\tfrac12X^2/b)/(1+\tfrac12X^2/b)\sim\mathrm{Beta}(\tfrac12,a),$$

and express the c.d.f. of X in terms of the c.d.f. of this Beta distribution: $P(|X|\le x)=\mathrm{Be}((\frac12x^2/b)/(1+\frac12x^2/b),\frac12,a)$. 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

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

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

(c)

Ex. 1.29 *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 R^p. We write $Y \sim \mathrm{N}_p(\xi, \Sigma)$ to indicate this distribution.

(a) Show that $\int yf(y)\,\mathrm{d}y$ 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 is

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

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 no zero correlation. This is in general more complicated – construct an example of a joint density $f(x,y)$ where the correlation is zero, even though X and Y are not independent.

Ex. 1.30 *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 is

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

[xx there are actually a couple of different ways of proving the points below, and we ought to think through what's best here. perhaps both, the direct $f(y_1, y_2)/f_2(y_2)$, which must be $\exp(-\frac{1}{2}Q)$ for a quadratic Q, which then needs sorting out. or the other path, via a nice transformation, avoiding the need for patient manipulations. xx]

(a) By carrying out a linear transformation, and using results from Ex. 1.29, 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 $X \,|\, y \sim \mathrm{N}(\rho y, 1-\rho^2)$. [xx some more prose here. if x is seen as a proxy for y, then the stronger the correlation, the more one learns about y from having observed x. xx]

(d) (xx a bit more. pointer to linear regression, and back to Galton c. 1876. if there's multinormality for (x_i, y_i), in dimension $p+1$, then the linear regression model for $y_i \,|\, x_i$ follows perfectly. we could hunt for some of Galton's 1876-ish work, with (height, weight) and so on, xx)

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

Ex. 1.31 *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) If you have not yet seen or bothered to notice this particular aspect of Nils's appearance, 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. 1.32 *The chi-squared.* [xx note to nils; look things carefully through later on, to see that i've landed on a 'natural sequence' for giving these properties, and that i don't foregriper and that i don't repeat things, e.g. regading the convolution properties. these are best handled after the mgf things. xx] This exercise goes through some basic properties of the chi-squared; see also Ex. 1.35 for its eccentric cousin, the noncentral or eccentric chi-squared.

(a) Start with the density

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

for the χ^2_m. Show that its moment-generating function becomes $M(t) = (1-2t)^{-m/2}$, for $t < \frac{1}{2}$. Show also that $\mathrm{E}\, K = m$, $\mathrm{Var}\, K = 2m$, and that the skewness, which is defined as $\mathrm{E}\, W^3$ with $W = (K - \mathrm{E}\, K)/(\mathrm{Var}\, K)^{1/2} = (K-m)/(2m)^{1/2}$, is $(8/m)^{1/2}$.

(b) With this result, show the simple and basic convolution property for the chi-squared, that if $K_1, \ldots, K_n$ are independent and chi-squared distributed with degrees of freedom $m_1, \ldots, m_n$, the the sum $Z = \sum_{i=1}^n K_i$ is chi-squared too, with degrees of freedom $\sum_{i=1}^n m_i$. A generalisation is given in Ex. 1.35.

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

(d) (xx briefly: note and explain the relation between a χ^2_m and a $\mathrm{Gam}(a,b)$, see Ex. 1.9. With $(a,b) = (m/2, 1/2)$, we have the χ^2_m. also, starting with any $Y \sim \mathrm{Gam}(a,b)$, show that $K = 2bY \sim \chi^2_{a/2}$. check this. xx)

Ex. 1.33 *Orthonormal transformations.* [xx check and calibrate with chi-squared things, to get the order right. xx] We have seen in Ex. 1.29 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. 2.4, and will also be utilised (xx in other exercises, like Ex. 3.2 xx). It turns out, however, that the independence of $\bar{Y}$ and S is an instance of a more general phenomenon, to which we turn in Ex. 1.64(h).

(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 orthogonal, 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.

(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.

(f) Show that this implies the following classical and important properties, starting with and 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

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \tag{1.3}$$

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. 1.34 *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. 1.13, 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 $\gamma_4 = 6/(m-4)$ 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 (1.3), show that

$$t = \frac{\bar{Y} - \mu}{\widehat{\sigma}/\sqrt{n}},$$

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

Ex. 1.35 *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 moment-generating function 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) 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.

(c) Also the noncentral chi-squared distributions have convolution properties, generalising those of Ex. 1.32. 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 $\sum_{i=1}^n \lambda_i$.

(d) 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?

(e) 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 moment-generating function and show that it ends to $\exp(\frac{1}{2}t^2)$ for growing λ.

(f) More generally, with $K \sim \chi^2_m(\lambda)$, work out a formula for the moment-generating function $M(t)$ 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. 1.36 *The purely noncentral excentric chi-squared.* The noncentral chi-squared, with its somewhat complicated density expressed as an infinite sum, etc., admits a simple representation, as a pure chi-squared plus a pure noncentral part, as shown in Hjort (1988).

(a) Consider a variable K from the noncentral chi-squared distribution, say $K \sim \chi^2_m(\lambda)$. We found its moment-generating function in Ex. 1.35. Deduce that *if* there is a representation, of the form $K = K_0 + U$, with $K_0 \sim \chi^2_m$ and a certain purely eccentric variable U, then we must have $M(t) = \mathrm{E}\exp(tU) = \exp\{\lambda t/(1 - 2t)\}$. In particular, such a U will then have a distribution depending only on λ, the same for each m.

(b) Let us use the occasion to demonstrate the moment-generating aspect of moment-generating functions. Still under the assumption that there actually is a purely eccentric U, as above, find the four first moments of U, using $\mathrm{E}\, U^j = M^{(j)}(0)$, the derivatives of M at zero. Show indeed that $\mathrm{E}\, U = \lambda$, $\mathrm{Var}\, U = 4\lambda$, and that its skewness is $3/\lambda^{1/2}$.

(c) Indeed, there is such a U. You may check Hjort (1988) to find constructive arguments to *deduce* the distribution of U, as opposed to taking the answer and then verifying that it is correct. Here we are content with the verification: show that with U having the distribution

$$P(U \le u) = \sum_{j=1}^{\infty} p(j)\Gamma_{2j}(u), \quad \text{with} \quad p(j) = \exp(-\tfrac{1}{2}\lambda)(\tfrac{1}{2}\lambda)^j/j!,$$

where $\Gamma_{2j}(u) = P(\chi^2_{2j} \le u)$ is the cumulative for the χ^2_{2j}, its moment-generating function indeed becomes the desired $\exp\{\lambda t/(1 - 2t)\}$. We allow also Γ_0 here, for the chi-squared with zero degrees of freedom, for the variable which is equal to zero with probability one. This is a Poisson mixture of χ^2_{2j} distributions. In particular, U has a positive pointmass at zero, $P(U = 0) = \exp(-\frac{1}{2}\lambda)$.

(d) Use the representation $U \,|\, (J = j) \sim \chi^2_{2j}$, with $J \sim \mathrm{Pois}(\frac{1}{2}\lambda)$, to find the mean, the variance, the skewness, the kurtosis of U. Use these formulae again to find the mean, variance, skewness, kurtosis for $K = K_0 + U$, the noncentral $\chi^2_m(\lambda)$.

(e) What is required, of (m, λ), for $K \sim \chi^2_m(\lambda)$ to have a normal limit distribution (when properly normalised)?

(f) (xx just a bit more. for small λ, an approximation is $(1 - \frac{1}{2}\lambda)\chi^2_m + \frac{1}{2}\lambda\chi^2_{2m+2}$. note its use for power studies of certain tests. there is a full purely noncentral process, say $\{U(\lambda)\colon \lambda \geq 0\}$, perhaps in chapter 8. xx)

Ex. 1.37 *Noncentral chi-squared for empirical variances.* We saw in Ex. 1.33 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 reflecting 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. 1.38 *The F distribution.* As we have seen Ex. 1.33, 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) (xx the mean, the variance, density later. show that when $F \sim F(m_1, m_2)$, then $1/F \sim F(m_2, m_1)$. xx) Show that

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

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, and deduce from this that $F \to_{\mathrm{pr}} 1$. (xx calibrate here; we haven't properly introduced convergence in probability yet. xx)

(c) (xx its approximate distribution, from knowing $\chi^2_m/m = 1+(2/m)^{1/2}N_m$, where N_m tends to the standard normal. so

$$F \sim \frac{1+(2/m_1)^{1/2}N_{m_1}}{1+(2/m_2)^{1/2}M_{m_2}} \approx 1+(2/m_1)^{1/2}N_{m_1} - (2/m_2)N_{m_2}.$$

Deduce that when m_1 and m_2 are not too small, F is approximately a normal, with mean 1 and variance $(2/m_1)+(2/m_2)$. xx)

(d) (xx in this point, tidy up, check, and calibrate notation with g and G for the chi-squared exercise. xx) 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

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

in terms of the cumulative $G(\cdot, m)$ and density $g(\cdot, m)$ 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

$$h(x, m_1, m_2) = \frac{\Gamma(\frac12(m_1+m_2))}{\Gamma(\frac12 m_1)\Gamma(\frac12 m_2)}(m_1/m_2)^{m_1/2}\frac{x^{m_1/2-1}}{\{1+(m_1/m_2)x\}^{(m_1+m_2)/2}}.$$

Ex. 1.39 *The noncentral F distribution.* A ratio of independent chi-squared distributions, modulo a multiplication factor, has the F distribution, as seen in Ex. 1.38. There are certain statistical uses of the extended construction (xx give pointers; power; cc(λ); more) when the nominator is a noncentral chi-squared. Formally, if $X \sim \chi^2_{m_1}(\lambda)$ and $Y \sim \chi^2_{m_2}$ are independent, then

$$F = \frac{X/m_1}{Y/m_2} = \frac{\chi^2_{m_1}(\lambda)/m_1}{\chi^2_{m_2}/m_2} \sim F(m_1, m_2, \lambda),$$

termed a noncentral F distribution with degrees of freedom (m_1, m_2) and excentre parameter λ.

(a) Find the mean and variance of such an F.

(b) Suppose $Y_1, \ldots, Y_n$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$, and consider $F = n\bar{Y}^2/\widehat{\sigma}^2$. Show that $F \sim F(1, n-1, n\mu^2/\sigma^2)$.

Ex. 1.40 *The Weibull distribution.* The Weibull distribution, with positive parameters (a, b), has c.d.f. $F(t) = 1-\exp\{-(t/a)^b\}$ for $t \ge 0$. 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) Work through the details of

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

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 $\mathrm{sd}(T)/\mathrm{E}\, T$ as a function of b. (xx write out. also reparametrisation, with $1 - \exp(-ct^b)$. point to story. xx)

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

Ex. 1.41 *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. 1.42 *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 $P(Y = j) = p_j$, 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 simple Bernoulli variable with $P(Y = 0) = 1 - p$, $P(y = 1) = p$; (ii) the binomial (n, p); (iii) the Poisson with parameter θ; (iv) the geometric with $P(Y = j) = (1 - p)^{j-1}p$.

(b) 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, that a sum of independent Poissons is a Poisson.

Ex. 1.43 *Sums of random lengths.* Let $X_1, X_2, \ldots$ be i.i.d., from a distribution with mean ξ, variance σ^2, and moment-generating function $M_0(t)$. 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 moment-generating function $M(t) = G(M_0(t))$. Show that Z has mean $\lambda\xi$ and variance $\lambda\sigma^2 + \xi^2\tau^2$.

(b) (xx a bit more here; might be used in Ch10 for frailty. edit and polish. xx) Consider the so-called compound Poisson variable $Z = \sum_{i=1}^{N} X_i$, where the X_i are nonnegative with Laplace transform $L_0(s) = \mathrm{E}\, \exp(-sX_i)$ and $N \sim \mathrm{Pois}(\lambda)$. Show that its Laplace transform may be written

compound Pisson

$$\mathrm{E}\, \exp(-sZ) = \mathrm{E}\, L_0(s)^N = \exp[-\lambda\{1 - L_0(s)\}],$$

with mean $\lambda\xi$ and variance $\lambda(\xi^2 + \sigma^2)$, in terms of mean ξ and variance σ^2 for the X_i. For the particular case of $X_i \sim \mathrm{Gam}(a, b)$, show that

$$\mathrm{E}\, \exp(-sZ) = \exp\bigl(-\lambda[1 - \{b/(b+s)\}^a]\bigr).$$

(c) (xx a bit more here. compound Poisson. find an expression for the skewness of $\gamma_3 = \mathrm{E}\,(Z - \lambda\xi)^3/\{\lambda(\xi^2 + \sigma^2)\}^{3/2}$. a special case or two. distribution of a random sum of standard normals. expo. xx)

Ex. 1.44 *The logarithmic distribution.* Consider a variable Y with point probabilities $P(Y = y) = c(p)^{-1}p^y/y$ for $y = 1, 2, \ldots$, 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 moment-generating function $M(t)$, 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 $G(s) = \mathrm{E}\, s^Y = c(ps)/c(p)$.

(b) Consider $Z = \sum_{i=1}^{N} Y_i$, with the Y_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.

Ex. 1.45 *The Tweedie distribution.* As a special case of the construction of Ex. 1.43, study $Z = \sum_{i=1}^{N} X_i$, where $N \sim \mathrm{Pois}(\lambda)$ and the X_i are i.i.d. from a $\mathrm{Gam}(a, b)$.

(a) Show that the c.d.f. can be expressed as

$$H(z) = \sum_{n=0}^{\infty} p(n,\lambda)G(z,na,b) = p(0,\lambda) + \sum_{n=1}^{\infty} p(n,\lambda)G(z,na,b),$$

with $p(n,\lambda)$ the Poisson and $G(z,na,b)$ the c.d.f. of the Gam(na,b). In particular, show that $P(Z=0) = \exp(-\lambda)$, and that $\mathrm{E}\, Z = \lambda a/b$, $\mathrm{Var}\, Z = \lambda\{(a/b)^2 + a/b^2\}$.

(b) Simulate say $n = 500$ outcomes for this distribution, for parameter values (λ, a, b) you decide on. Check if you can estimate these values based on the simulated sample.

(c)

Ex. 1.46 *The Beta Prime distribution.* (xx keep it brief. essentially just $Y = X/(1-X)$ with $X \sim \mathrm{Beta}(a,b)$. xx)

Ex. 1.47 *The Dagum distribution.* [xx to be done. three para. formulae for quantiles etc. xx]

Ex. 1.48 *The Gumbel distribution.* [xx more here. point to later applications. xx]

(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*.

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

(d) Find the median, the mean, and the variance of the Gumbel distribution. (xx and just a bit more. x)

Ex. 1.49 *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. 1.30.

(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 $\mathrm{N}(a + b_1 x_1 + b_2 x_2, \sigma^2)$, as in linear regression models we will study in later chapters; see e.g. Ex. 3.33. 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.

the log-normal distribution

Ex. 1.50 *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 $Y \sim \mathrm{logN}(\xi, \sigma^2)$ 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 $Y \sim \mathrm{logN}(\xi, \sigma^2)$, 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 $Y \,|\, \xi \sim \mathrm{logN}(\xi, \sigma^2)$, and that $\xi \sim \mathrm{N}(\xi_0, \tau^2)$. Show that marginally, $Y \sim \mathrm{logN}(\xi_0, \sigma^2 + \tau^2)$. Make explicit the connection to Ex. 1.49.

(d) Show that a product of independent log-normals is log-normal. Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the $\mathrm{logN}(\xi, \sigma^2)$ 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 $\mathrm{logN}(0, 1)$ 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. 1.51 *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 $Y \,|\, (J = j) \sim \mathrm{N}(\mu_j, \sigma_j^2)$. Show that this Y has the density above; this amounts to a way of representing and interpreting a normal mixture.

(b) From $\mathrm{E}\,(Y \,|\, J) = \mu_J$ and $\mathrm{Var}\,(Y \,|\, J) = \sigma_J^2$, 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. 1.52 *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

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

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. 1.53 *Leftovers.* [xx we push smaller things here, when they belong better in later chapters. in particular, the present chapter should be free of estimators and confidence and inference, and also of lilmits. xx]

(a) If Y_n and Y have moment-generating functions M_n and M, then $M_n(t) \to M(t)$ for all t in a neighbourhood around zero is sufficient for $Y_n \to_d Y$. [xx repair this, point to Ch. 4. since we need to point to what convergence in distribution is. xx]

(b) In particular, if the moment-generating function $M_n(t)$ of some Y_n tends to $\exp(\frac{1}{2}t^2)$ for all t close to zero, then $Y_n \to_d \mathrm{N}(0,1)$. Illustrate this principle for the following situation: Let $Y_1, Y_2, \ldots$ be i.i.d. from the simple symmetric two-point distribution, where Y_i takes on values $-1, 1$ with probabilities $\frac{1}{2}, \frac{1}{2}$. Show that the normalised mean $Z_n = \sqrt{n}\bar{Y}_n = n^{-1/2}\sum_{i=1}^{n} Y_i$ has moment-generating function $\cosh(t/\sqrt{n})^n$, and show that it converges to $\exp(\frac{1}{2}t^2)$. You have now proved the Central Limit Theorem for this special case.

Ex. 1.54 *Moment-generating functions for two or more variables.* (xx needs rounding off and pointers to other matters. xx) The one-dimensional $M(t)$ for a variable Y dealt with in Ex. 1.20, 1.21 generalises neatly to the case of several variables. For a vector $Y = (Y_1, \ldots, Y_p)^{\mathrm{t}}$ the joint moment-generating function is

$$M(t) = M(t_1, \ldots, t_p) = \mathrm{E}\,\exp(t^{\mathrm{t}}Y) = \mathrm{E}\,\exp(t_1Y_1 + \cdots + t_pY_p),$$

involving of course the joint distribution of the random vector.

(a) Assume $Y_1, \ldots, Y_p$ are actually independent, with moment-generating functions $M_1, \ldots, M_p$. Show that $M(t_1, \ldots, t_p) = M_1(t_1)\cdots M_p(t_p)$.

(b) Assume $Y \sim \mathrm{N}_p(\mu, \Sigma)$. Show that $M(t) = \exp(\mu^{\mathrm{t}} t + \frac{1}{2} t^{\mathrm{t}} \Sigma t)$.

(c) Let Y be standard normal. Show that

$$M(s,t) = \mathrm{E}\, \exp(sY + tY^2) = \exp\{\tfrac{1}{2} s^2/(1-2t)\}/(1-2t)^{1/2}$$

for $t < \frac{1}{2}$. This characterises the joint distribution of (Y, Y^2). Comment on the cases $s = 0$ and $t = 0$.

(d) (xx do the trinomial too. xx)

Ex. 1.55 *Product of two normals.* (xx we will see later how this pans out; connect to a cc(ϕ) in Ch7. xx) For independent $X \sim \mathrm{N}(a, 1)$ and $Y \sim \mathrm{N}(b, 1)$, consider their product XY. Can its distribution be close to a normal?

(a) Writing $X = a + U$ and $Y = b + V$, for independent standard normals U and V, show that $Z = XY - ab = aV + bU + UV$, and that XY has variance $\sigma^2 = 1 + a^2 + b^2$. Work also out its skewness and kurtosis; in particular, show that $\mathrm{E}\, Z^3 = 6ab$.

(b) Establish first that $\mathrm{E}\,\{\exp(tZ) \,|\, (U = u)\} = \exp(tbu) \exp\{\frac{1}{2} t^2 (a+u)^2\}$, and use results from Ex. 1.54 to show that Z has moment-generating function

$$M(t) = \frac{1}{(1-t^2)^{1/2}} \exp\{\tfrac{1}{2} a^2 t^2 + \tfrac{1}{2} t^2 (b + at)^2/(1-t^2)^{1/2}\} \quad \text{for } |t| < 1.$$

(c) In order to check if Z/σ might have a distribution not far from the standard normal, plot the function $\log M(t)$, for a few choices of (a, b), and inspects its closeness to $\frac{1}{2} t^2$. (xx round this off. xx)

(d) (xx might drop this, or give a pointer to something in Ch7, re difference of Fieller ratios. xx) Consider the parameters $x_{0,1} = -a_1/b_1$ and $x_{0,2} = -a_2/b_2$, the positions at which two lines $a_1 + b_1 x$ and $a_2 + b_2 x = 0$. It's tricky to test $x_{0,1} = x_{0,2}$, but that hypothesis is equivalent to $a_1 b_2 = a_2 b_1$. May hence construct $Z = \widehat{a}_1 \widehat{b}_2 - \widehat{a}_2 \widehat{b}_1 = Z_1 - Z_2$, say. I find something like

$$\begin{aligned} \mathrm{E}\, \exp(tZ) \;=\; & \frac{1}{(1-t^2)^{1/2}} \exp\{\tfrac{1}{2} a_1^2 t^2 + \tfrac{1}{2} t^2 (b_2 + a_1 t)^2/(1-t^2)^{1/2}\} \\ & \frac{1}{(1-t^2)^{1/2}} \exp\{\tfrac{1}{2} a_2^2 t^2 + \tfrac{1}{2} t^2 (b_1 - a_2 t)^2/(1-t^2)^{1/2}\}. \end{aligned}$$

the point is that the skewness has disappeared, and that $(Z_1 - Z_2)/\tau$ becomes closer to normal, with $\tau^2 = 2 + a_1^2 + b_1^2 + a_2^2 + b_2^2$. still a hard thing to do very well.

Ex. 1.56 *The Gaussian copula.* (xx something here. need to have something beside the multinormal for dependence. xx)

Ex. 1.57 *The exponential family class, I.* 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

$$\begin{aligned} f(y,\theta) &= \exp\{\theta_1 T_1(y) + \cdots \theta_p T_p(y) - k(\theta_1,\ldots,\theta_p)\}h(y) \\ &= \exp\{\theta^{\mathrm{t}} T(y) - k(\theta)\}h(y), \end{aligned} \tag{1.4}$$

for appropriate functions $T_1(y),\ldots,T_p(y)$ and $h(y)$; the $k(\theta)$ function is there to secure integration to one, it is called the normalising function of the cumulant generating function. We then say Y is of the exponential family class, or, more precisely, that is has a density belonging to a *p-parameter exponential family* in *canonical form*, with natural statistics $T(y) = (T_1(y),\ldots,T_p(y))^{\mathrm{t}}$ and natural parameter $\theta = (\theta_1,\ldots,\theta_p)^{\mathrm{t}}$.

The densities in (1.4) are defined with respect to some measures μ, such as the uniform (i.e., Lebesgue) measure on $\mathbb{R}$ for the normal family; or the measure giving weight one to each member of $\{0,1,2,\ldots,\}$ (i.e., the counting measure) for the Poisson family. A change of measure changes the representation of the exponential family. For example, if we define the measure $\nu(A) = \int_A h(y)\,\mathrm{d}\mu(y)$ say, we obtain an exponential family with densities $\tilde{f}_\theta(y,\theta) = \exp\{\theta^{\mathrm{t}} T(y) - k(\theta)\}$ with respect to ν. From the possibility of changing measure it is clear theat the representation in (1.4) is not unique. For example, $\gamma_j = c_j\theta_j$ and $S_j(y) = T_j(y)/c_j$ for $j = 1,\ldots,p$ gives a reparametrised density $f_\gamma(y)$ of the same exponential form.

(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 (1.4). (i) $Y \sim \mathrm{binom}(n,p)$. (i) $Y \sim \mathrm{Pois}(\lambda)$. (iii) $Y \sim \mathrm{Beta}(a,b)$. (iv) $Y \sim \mathrm{Gam}(a,b)$. (v) $Y \sim \mathrm{N}(\xi,\sigma^2)$, first with known σ, then with both parameters unknown. (v) one more.

(b) Show that we must have

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

assumed to be finite for at least some θ. Thus, $k(\theta) = \log(\int \exp\{\theta^{\mathrm{t}} T(y)\}h(y)\,\mathrm{d}y)$ if μ is Lebesgue measure, and $k(\theta) = \log(\sum_y \exp\{\theta^{\mathrm{t}} T(y)\}h(y))$ if μ is counting measure. Let in fact H be the set of θ such that $k(\theta)$ is finite, called the natural parameter space. Show that H is a convex set.

natural parameter space

(c) For the following one-dimensional exponential families having densities with respect to Lebesgue measure, describe the natural parameter space, find expressions for the densities. (i) $h(y) = 1/yI_{(0,1]}$, and $T(y) = \log y$; (ii) $h(y) = I_{(0,\infty)}$, and $T(y) = y$; (iii) $h(y) = y^3 I_{(0,1)}(y)$ and $T(y) = \log(1-y)$; (iv) $h(y) = \exp(-y/2)/yI_{(0,\infty)}(y)$ and $T(y) = \log y$; and (v) $h(y) = \exp(-y^2/2)$ and $T(y) = y$.

(d) Show that the score function corresponding to a density of the form (1.4) becomes $u(y,\theta) = T(y) - \xi(\theta)$, where

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

In particular $\mathrm{E}_\theta\, T(Y) = \xi(\theta)$. Show also that

$$\mathrm{Var}_\theta\, T(Y) = J(\theta) = \partial^2 k(\theta)/\partial\theta\partial\theta^{\mathrm{t}},$$

giving variances and covariances of the $T_j(Y)$ in one matrix formula.

(e) Let $Y_1, \ldots, Y_n$ be i.i.d., each with marginal densities $f_\theta(y) = \exp\{\theta^{\mathrm{t}} T(y) - k(\theta)\}h(y)$. Show that the joint density $f_\theta(y_1, \ldots, y_n)$ is also belongs to an exponential family. Show that the log-likelihood function for a sample $Y_1, \ldots, Y_n$ can be written $\ell_n(\theta) = n\{\theta^{\mathrm{t}}\bar{T} - k(\theta)\}$, with $\bar{T} = (1/n)\sum_{i=1}^n T(Y_i)$ the vector of averages $\bar{T}_j = (1/n)\sum_{i=1}^n T_j(y_i)$, and that this is a concave function. If this is a family with say $p = 3$ parameters, and $n = 10000$, then the full relevant information is captured in the 3 averages $\bar{T}_1, \bar{T}_2, \bar{T}_3$.

(f) In (d) we used a certain smoothness property of exponential families (see, e.g., Schervish (1995, Theorem 2.64, p. 105) or Brown (1986, Theorem 2.2, p. 34)), implying, among other things, that for any function f such that $\int |f(y)| \exp\{\theta^{\mathrm{t}} T(y)\}h(y)\,\mathrm{d}\mu(y)$ is finite for all $\theta \in H$, we can pass the derivative with respect to θ under the integral sign, that is

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int f(y)\exp\{\theta^{\mathrm{t}} T(y)\}h(y)\,\mathrm{d}\mu(y) = \int T(y) f(y)\exp\{\theta^{\mathrm{t}} T(y)\}h(y)\,\mathrm{d}\mu(y). \tag{1.5}$$

Anticipating Ex. 5.5(f), we will, with out loss of generality, verify (1.5) for the function $\exp\{k(\theta)\}$ (i.e., $f(y) = 1$) for a one dimensional parameter θ. Let θ_0 be some point in the interior of the natural parameter space, so that $\exp\{k(\theta)\}$ is finite on some interval around θ_0. For some $\varepsilon > 0$, the derivative at θ_0 is then

$$\lim_{n\to\infty} \frac{e^{k(\theta_0+\varepsilon/n)} - e^{k(\theta_0)}}{\varepsilon/n} = \lim_{n\to\infty} \int \frac{e^{(\theta_0+\varepsilon/n)T(y)} - e^{(\theta_0)T(y)}}{\varepsilon/n} h(y)\,\mathrm{d}\mu(y).$$

Use the inequalities $|e^u - 1| \le |u|e^{|u|}$ and $|u| \le e^{|u|}$ to find an integrable function not depending on n, say $g(y)$, so that the absolute value of the integrand on the right is always smaller than or equal to $g(y)$. We can then appeal to the Dominated convergence theorem, Ex. A.6(d), to conclude that what we did in (d) is legitimate.

Ex. 1.58 *Some moments.* For the time being we stick to one-parameter exponential families. That is, let Y be a random variable with density $f_\theta(y) = \exp\{\theta T(y) - k(\theta)\}h(y)$. We'll take as a fact that as long as $\int |f(y)| \exp\{\theta T(y)\}h(y)\,\mathrm{d}\mu(y)$ is finite, the function $\theta \to \int f(y)\exp\{\theta T(y)\}h(y)\,\mathrm{d}\mu(y)$ has continuous derivatives of all orders (see the references in Ex. 1.57(f)).

(a) Show that $\mathrm{E}_\theta\, T(Y) = k'(\theta)$ and $\mathrm{Var}_\theta T(Y) = k''(\theta)$, where k' and k'' denotes the first and second derivative of the cumulant generating function. Argue that $k(\theta)$ is a convex function, and show that the function $\theta \mapsto \mathrm{E}_\theta\, T(Y)$ is increasing in the natural parameter θ.

(b) Determine the mean and the variance of random variables with the densities you found in Ex. 1.57(c)

(c) Suppose that Y stems from an exponential family with density $f_\theta(y) = \exp\{\theta T(y) - k(\theta)\}h(y)$. Let $M_T(u) = \mathrm{E}_\theta \exp(uT)$ be the moment generating function of $T = T(Y)$ (as introduced in Ex. 1.20). Show that $\log M_T(u) = k(\theta + u) - k(\theta)$, and use this to, once more, derive the result of Ex. 1.21(a), namely that $\mathrm{d}M_T(u)/\mathrm{d}u$ evaluated in $u = 0$ equals $\mathrm{E}_\theta T(Y)$.

Ex. 1.59 *The exponential family class, II.* In Ex. 1.57 we considered exponential families in canonical form. The canonical thing about these densities is that they were parametrised in terms of the natural parameters. We do, however, typically prefer to parametrise our densities in terms of parameters making more intuitive sense, such as the mean and the variance. Consider therefore the moderate jump from (1.4) to

$$f(y,\theta) = \exp\{\eta_1(\theta)T_1(y) + \cdots + \eta_p(\theta)T_p(y) - k(\eta(\theta_1,\ldots,\theta_p))\}h(y), \tag{1.6}$$

where $\eta(\theta) = (\eta_1(\theta),\ldots,\eta_p(\theta))$ is a vector valued function from some parameter space Θ into the natural parameter space H. Densities of the form (1.6) are simply called *p-parameter exponential families*. Thus, from Ex. 1.57(a) again, for the normal distribution with unknown parameter $\theta = (\mu,\sigma^2)$ we have $\eta_1(\theta) = \mu/\sigma^2$ and $\eta_2(\theta) = -1/(2\sigma^2)$, that is, η is a mapping from $\Theta =: \mathbb{R}\times(0,\infty)$ into $H = \mathbb{R}\times(-\infty,0)$.

(a) [xx some more basics xx]

(b) [xx curved exponential families somewhere $\dim(\Theta) < p$ xx]

(c) Important examples of densities such as that in (1.6) arise when working with the regression models falling under the umbrella of generalised linear models. Consider the classical linear regresion model

$$Y_i = x_i^{\mathrm{t}}\beta + \varepsilon_i,$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. $\mathrm{N}(0,\sigma^2)$ for $i = 1,\ldots,n$ with σ^2 known, and the $x_1,\ldots,x_n$ are known constants. Show that the density of Y_i belongs to an exponential family, and find and expression for its natural parameter η_i.

(d) The general idea arising from (c) that if $(x_1,Y_1),\ldots,(x_n,Y_n)$ are independent regresion data, where x_i are p-dimensional covariate vectors, and the Y_i have marginal densities $Y_i \sim f_{\eta_i}(y) = \exp\{\eta_i T(y) - k(\eta_i)\}h(y)$, then a generalised linear model expresses the natural parameter η_i as a function of linear function the known x_i and an unknown regression coefficient $\beta\in\mathbb{R}^p$, that is $\eta_i = x_i^{\mathrm{t}}\beta$. In the case that $Y_i \sim \mathrm{Pois}(\theta_i)$ show that this gives $\theta_i = \log(x_i^{\mathrm{t}}\beta)$. For the case where $Y_i \sim \mathrm{Bernoulli}(\theta_i)$ independent, show that we obtain the logistic regression model, $\theta_i = \exp(x_i^{\mathrm{t}}\beta)/\{1+\exp(x_i^{\mathrm{t}}\beta)\}$.

generalised linear models

(e) With data as described in (d), show that the joint density of $Y = (Y_1,\ldots,Y_n)$ can be expressed as

$$f_\beta(y_1,\ldots,y_n) = \beta^{\mathrm{t}}S_n(y) - k_n(\beta)h_n(y)$$

where $S_n(y) = \sum_{i=1}^n x_i T(y_i)$, $k_n(\beta) = \sum_{i=1}^n k(x_i^{\mathrm{t}}\beta)$, and $h_n(y) = \prod_{i=1}^n h(y_i)$. We now see that the theory developed in Ex. 1.57 applies more or less directly to this regression setting. Show, for example, that the log-likelihood $\ell_n(\beta) = \log f_\beta(y_1,\ldots,y_n)$ is a concave function. Show also that the maximum likelihood estimator $\widehat\beta$ is the solution to $\sum_{i=1}^n x_i\{T(Y_i) - \mathrm{E}_\beta T(Y_i)\} = 0$. Verify that in the normal case of (c), this yields the least squares solution $\widehat\beta = (\sum_{i=1}^n x_i x_i^{\mathrm{t}})^{-1}\sum_{i=1}^n x_i Y_i$.

Ex. 1.60 *The exponential family: marginals and conditionals.* [xx fix details without making it too heavy-handed. xx] Suppose that Y stem from an exponential family with

density of the form $f_{a,b}(y) = \exp\{a^{\mathrm{t}}U(y)+b^{\mathrm{t}}V(y)-k(a,b)\}h(y)$. Then both the density of the marginal distribution of V, as well as the conditional density of U given $V = v$ belong to exponential families. These are densities, however, with respect to some potentially rather weird dominating measures.

(a) We start with some additional facts about exponential families. Let $f_\theta(y) = \exp\{\theta^{\mathrm{t}}T(y)-k(\theta)\}h(y)$ be an exponential family. This means that for all θ, f_θ is a density with respect some dominating measure μ, that is, $P_\theta(A) = \int_A f_\theta(y)\,\mathrm{d}\mu(y)$.

Let θ_0 be some point in the natural parameter space, and explain why $P_{\theta_0}(A) = 0$ implies $P_\theta(A) = 0$ for all θ. In other words, for any θ, the distribution P_θ is absolutely continuous with respect to P_{θ_0}.

(b) Show that the density of P_θ with respect to P_{θ_0} is also of the exponential family form, i.e., $f_\theta(y) = \exp\{(\theta-\theta_0)T(y) - [k(\theta) - k(\theta_0)]\}$, meaning that we can express f_θ as a density with respect to μ as

$$f_\theta(y) = \exp\{(\theta-\theta_0)T(y) - [k(\theta) - k(\theta_0)]\}f_{\theta_0}(y).$$

In other words, any density f_θ can be expressed as an exponential tilt of our chosen density f_{θ_0}.

(c) Now, let us look at an exponential family $f_{a,b}(u,v) = \exp\{au + bv - k(a,b)\}f_0(u,v)$, where $f_0 = f_{a_0,b_0}$ is some (arbitrary) member of the family. For simplicity, we assume that the $f_{a,b}(u,v)$ are densities with respect to Lebesgue measure on $\mathbb{R}^2$, i.e., $F_{a,b}(x,y) = \int_{-\infty}^x \int_{-\infty}^y f_{a,b}(u,v)\,\mathrm{d}v\,\mathrm{d}u$ is the c.d.f. of $(U,V) \sim f_{a,b}$. Show that

$$f^V_{a,b}(v) = \exp\{bv - k_a(b)\}h_a(v),$$

where $h_a(v) = \{\int \exp(au) f_0^{U|V}(u \mid v)\,\mathrm{d}u\}f_0(v)$.

(d) Use the result from (c) to show that the distribution of U given $V = v$ has density

$$f_a^{U|V}(u \,|\, v) = \exp\{au - k(a \,|\, v)\}f_0(u \,|\, v),$$

where $k(a \,|\, v) = \log(\int \exp(au) f_0^{U|V}(u \mid v)\,\mathrm{d}u)$.

(e) We now move on to the repeated sampling situation. Suppose that $Y_1, \ldots, Y_n$ are independent with common density $f_{a,b}(y) = \exp\{aU(y) + bV(y) - k(a,b)\}h(y)$ with respect to μ. Show that the joint density of $(Y_1, \ldots, Y_n)$, say $f^{\text{joint}}_{a,b}(y_1, \ldots, y_n)$ belongs to an exponential family on the same form.

(f) In (e) we saw that $f^{\text{joint}}_{a,b}$ only depends on the sample through $U(y_1, \ldots, y_n)$ and $V(y_1, \ldots, y_n)$. Thus we may define $f^{\text{joint}}_{a,b}(u(y_1, \ldots, y_n), v(y_1, \ldots, y_n)) = f^{\text{joint}}_{a,b}(y_1, \ldots, y_n)$. Define $T = (\sum_{i=1}^n U(Y_i), \sum_{i=1}^n V(Y_i))$, and use the change of variable formula (see Ex. A.7(d) in the appendix) to show that T has density $f^{\text{joint}}_{a,b}(u,v)$ with respect to $\mu^n T^{-1}$, where $\mu^n = \mu \times \cdots \times \mu$ is the product measure on the range of $(Y_1, \ldots, Y_n)$.

(g) In (f), if $\mu^n T^{-1}$ happens to be absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^2$, then (c) and (d) give us the marginal density of $\sum_{i=1}^n V(Y_i)$, and the conditional density of $\sum_{i=1}^n U(Y_i)$ given $\sum_{i=1}^n V(Y_i) = v$, respectively. Generalise (c) and (d) to the situation where (U,V) has some arbitrary two-dimensional distribution.

(h) (xx some concrete examples, and pointers ahead to Ex. 3.26–3.30 xx)

Ex. 1.61 *Sufficiency.* A sufficient statistic is a summary of the data that contains all the information in the data. If you flip a coin ten times, 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 other words, the original sequence, for example $(T, H, T, T, H, H, T, T, H, T)$, can be compressed to a single number, 4 in this case, without any information about θ being lost. When you think about it, if you were told that the number of heads in ten tosses was four, then you would attach the same probability to $(T, H, T, T, H, H, T, T, H, T)$ having occured as to $(H, H, H, H, T, T, T, T, T, T)$ having occured, and so on for all the 210 sequences of ten tosses that contains exactly four heads. And, importantly, this probability would not depend on θ, as the $\theta = \Pr(\text{heads})$ is sort of already swallowed up into the fact that you condition on the number of heads being four.

Well, this leads us to the definition of sufficiency. If X is your data, stemming from a member of the family of distributions $\{P_\theta \colon \theta \in \Theta\}$, then the statistic $T = T(X)$ is sufficient for θ if the distribution of X given $T = t$ is the same for all values of θ. Down the road, in Ex. 1.63, we look more into this definition, and be a bit more formal. sufficient statistic

(a) Here are a few examples. (i) Let $X_1, \ldots, X_n$ be independent Bernoulli(θ) random variables. Show that $T = \sum_{i=1}^n X_i$ is sufficient for θ (ii) Let $Y_1, \ldots, Y_n$ be i.i.d. Pois(θ). Show that $T = \sum_{i=1}^n Y_i$ 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 why T is 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) If you used Bayes theorem in solving (i) and (ii) of Ex. (a) you may already have deduced 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

$$\theta \mapsto \frac{\Pr_\theta\{X_1 = x_1, \ldots, X_n = x_n\}}{P_\theta\{T(x_1, \ldots, x_n) = t\}} \tag{1.7}$$

is constant for every $x_1, \ldots, x_n$. In Ex. 1.63(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 (1.7). 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 $f_\theta(x_1, \ldots, x_n)$, 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. 1.63.

(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. Or, (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.

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. 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. 1.62 *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 $Q_t(B) = \Pr_\theta(X \in B \mid T = t)$. Show that $X' \sim X$ for all θ.

Ex. 1.63 *The factorisation theorem.*[xx move to appendix xx] Above, in Ex. 1.61(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.1(g) for definition of σ-finiteness). First, we must be more formal in our definition of a sufficient statistic. Let $\{P_\theta \colon \theta \in \Theta\}$ be a family of probability distributions on a measurable space $(\mathcal{X}, \mathcal{A})$. The statistic T is sufficient for $\{P_\theta \colon \theta \in \Theta\}$ 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. A.21 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 $\{P_\theta \colon \theta \in \Theta\}$ 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 $\{P_\theta \colon \theta \in \Theta\}$ 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 $\{P_\theta \colon \theta \in \Theta\}$, 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 $\{P_\theta \colon \theta \in \Theta\}$. 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 iff $Q \ll \mu$. 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 $\{P_\theta \colon \theta \in \Theta\} \ll Q \ll \mu$, as described above. Assume that T is sufficient for $\{P_\theta \colon \theta \in \Theta\}$, i.e., there exists $p(A, \cdot)$ that is a version of $P_\theta(A \mid T)$ for every $\theta \in \Theta$. First, show that

$$\int_G Q(A \mid T)(x) \, \mathrm{d}Q(x) = \int_G p(A, x) \, \mathrm{d}Q(x),$$

for all $G \in \sigma(T)$. This shows that T is sufficient for the augmented family $\{P_\theta \colon \theta \in \Theta\} \cup \{Q\}$. 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. A.14). 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. A.14 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.10(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 (1.7) in Ex. 1.61.

(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 moment generating function 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. 1.64 *Minimal sufficiency.* In Ex. 1.61(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 $\{P_\theta : \theta \in \Theta\}$. 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 ways 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) 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 $f_\theta(x) = f^{X|S}(x \mid s) f_\theta^S(s)$, 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. 1.59). 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 $(\theta_1, \ldots, \theta_p)$.

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. **??** as well as Ex. 8.5 for a proper treatments 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. 8.5(g).

(h)

Ex. 1.65 *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. As (d) below clearly demonstrates, 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) Let X and Y be independent $\mathrm{N}(\theta, 1)$, and set $R = X - Y$. Show that R is an ancillary statistic.

(b) In fact, (a) is an instance of a more general result: Let $X_1, \ldots, X_n$ be independent from a family of distribution with density $f(x - \theta)$ with respect to Lebesgue measure. Show that $X_i = Z_i + \theta$, where Z_i has density $f(x)$, and use this to show that the range $\max_{i \le n} X_i - \min_{i \le n} X_i$ is ancillary.

(c) Similarly, let $X_1, \ldots, X_n$ be independent from a family of distribution with density $f(x/\sigma)/\sigma$ with respect to Lebesgue measure. Show that $X_i = \sigma Z_i$, where Z_i has density $f(x)$, and use this to show that any function of the ratios $X_1/X_n, X_2/X_n, \ldots, X_{n-1}/X_n$ is ancillary.

(d) Let $Z \sim \mathrm{Bernoulli}(1/2)$, $X_1 \sim \mathrm{N}(\theta, 1)$, and $X_2 \sim \mathrm{N}(\theta, 2)$, and suppose that all three are independent. Set $X = ZX_1 + (1-Z)X_2$ and suppose that we observe (X, Z). Explain why Z is ancillary, and show that (X, Z) is minimal sufficient for θ, but that X is not sufficient. One says that X is conditionally sufficient given Z.

(e) Let $X_1, \ldots, X_n$ be independent $\mathrm{N}(\theta, 1)$. Show that $S = \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is ancillary.

(f) Let $N \sim \mathrm{Pois}(\lambda)$ for some known λ. Given $N = n$, n independent $\mathrm{Bernoulli}(\theta)$ trials $X_1, \ldots, X_n$ are performed. Show that $(\sum_{i=1}^n X_i, N)$ is minimal sufficient for θ, and that N is ancillary. Show that $N^{-1} \sum_{i=1}^N X_i$ is unbiased for θ, and find its variance.

(g)

Ex. 1.66 *Basu's theorem.* [xx but we have to wait with Basu, because completeness has yet to be introduced xx]

Ex. 1.67 *Nonparametric models.* Above we have seen a wide range of parametric models, each model indexed via a finite and perhaps small number of parameters. Models can also be nonparametric, however, with fewer modelling assumptions placed on their outcomes. (xx just a bit. data with unknown mean and unknown variance, but saying nothing more. symmetry. unimodal. shape constraints. xx)

Ex. 1.68 *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 $P(B_i = 1) = \frac{1}{2}$, 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 $f(x,y) = P(X_n = x, Y_n = y)$, 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 $P(A_i = 1) = p$ and then $P(B_i = 1 \,|\, A_i = 1) = p + a$, $P(B_i = 0 \,|\, A_i = 0) = 1 - p + ap/(1-p)$, 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.)

1.C 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 conceptual framework for hypothesis testing. xx]

I.2

Parameters, estimators, precision

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 also leads to ways of comparing competing estimation methods. Basic concepts include the bias, the variance, the mean squared error of estimators. 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. The more versatile method of maximum likelihood is treated in Ch. 5. To understand the properties of classes of estimators in general models, we learn the basics of normal approximations, involving the Central Limit Theorem and versions of the delta method. This enables one to assess precision and to compare different competing estimators.

2.A Chapter introduction

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.

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

$$\mathrm{E}_\theta \, \widehat{\phi} = \phi + b(\theta). \tag{2.1}$$

unbiased estimator

The footscript signals that the expectation operator is at work at the parameter position θ. The $b(\theta)$ is termed *the bias*, and 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*

mean squared error

$$\text{mse}(\widehat{\phi},\theta) = \text{E}_\theta\,(\widehat{\phi}-\phi)^2 = \text{Var}_\theta\,\widehat{\phi} + b(\theta)^2, \tag{2.2}$$

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 $\text{mse}(\widehat{\phi},\theta)$ of (2.2) 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 Ch. 8.

In various setups one can study distributions, biases, variances 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. We return to such themes in Chs. 4, 5, with more detail and a much wider discussion, but it is fruitful to learn about some of the basic methods and their uses already in this chapter. Thus Ex. 2.8-2.11 provide the basics of convergence in distribution (often to the normal), the Central Limit Theorem (acronym CLT), the delta method, and generally speaking to normal approximations. These methods may be understood, appreciated, seen in action, and used for new situations, without necessarily having been through each δ and ε of their full proofs (but see again Chs. 4, 5 for such detail).

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 some 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 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. 3, along with testing and power and p-values. we do mention a few key concepts here in intro, like unbiasedness, low variance, etc. xx]

(xx meeds a bit of perestroika, as of 13-Aug-2023: we do snikinnfoere confidence intervals, via CLT, but just in a few exercises, with the real thing coming in Ch3. so we need more editorial care to secure that things come in the right order below. xx)

2.B Short and crisp

Ex. 2.1 *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 (2.2).

(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 independent and identically distributed (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. 2.2 *Binomial estimation.* Consider Y being binomial (n, p), as in Ex. 1.3.

(a) To estimate p the canonical choice would be $\widehat{p} = Y/n$. 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 Ur-Bayes estimator $\widehat{p}_B = (Y + 1)/(n + 2)$ (xx pointer to exercise in Ch. 6 xx). 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. 2.3 *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. Here $\bar{Y}$ 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 is 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. 2.4 *Estimating the normal variance.* As in Ex. 2.3, 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. 1.33, $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2\chi^2_m$, 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 $\text{mse}(cZ, \sigma) = \text{E}_\sigma\,(cZ - \sigma^2)^2$. 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

$$\text{mse}(dZ^{1/2}, \sigma) = \text{E}_\sigma\,(dZ^{1/2} - \sigma)^2.$$

Find the best d, according to this criterion.

(d) (xx something more. could briefly investigate $(\log\widehat{\sigma} - \log\sigma)^2$. examine the risk function $\text{mse}(kZ^{1/2}, \sigma) = \text{E}_\sigma\,\{\log(kZ^{1/2}) - \log\sigma\}^2$ and find the best value of k. xx)

(e) A hard-core 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

$$\text{risk}(c_n\widehat{\sigma}, \sigma) = \text{E}_\sigma\,|c_n\widehat{\sigma} - \sigma|$$

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. 2.5 *Normal quantiles.* Consider again the setup of Ex. 2.3, with a sample of Y_i from the normal $\text{N}(\mu, \sigma^2)$ model. In the present exercise we care about quantiles, as opposed to 'only' the mean or the standard deviation.

(a) Writing F for the cumulative distribution function of Y_i, show that

$$F(y) = P(Y_i \le y) = \Phi((y-\mu)/\sigma),$$

with $\Phi(x) = P\{\text{N}(0,1) \le x\}$ the cumulative distribution function 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. 2.4 and $m = n-1$.

(c) (xx a bit more. confidence interval for γ_q. note that this is tighter in the middle than near the edges. xx)

(d) (xx a simple data example. xx)

Ex. 2.6 *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 or spread, 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 (1.3), find formulae for the mean of $1/\widehat{\sigma}^2$ and $\log \widehat{\sigma}$.

(b) Construct unbiased estimators for $1/\sigma^2$ and for $\log \sigma$.

(c) The log-density function is $f(y) = -\log \sigma - \frac{1}{2}(y-\mu)^2/\sigma^2 - \frac{1}{2}\log(2\pi)$. Construct an unbiased estimator for this $\log f(y)$.

(d) (xx more elaborate: constructing unbiased estimator for $f(y)$ on the direct density scale. no, perhaps wait until next exercise. xx)

(e) (xx can ask for estimates with bands of the ratio $f_1(y)/f_2(y)$, perhaps constructed by first estimating the log difference, finding band there, and exp-ing home. could also bake a Type B Story from the birthweights of Oslo boys and Oslo girls, 2001–2008, see (xx Data Story B.2.B xx), with other natural analyses. xx)

Ex. 2.7 *Some probability tail inequalities.* For a random variable X, how can we find useful bounds for tail probabilities, i.e. $P(X \geq a)$? There are several such, as we learn here, with more to come in Ex. 4.33. Their uses include assessing how likely it might be that an estimator is some distance off its target.

(a) Suppose X is nonnegative, with distribution F. Using

$$\mathrm{E}\, X = \int_0^\infty x \,\mathrm{d}F(x) \geq \int_a^\infty x \,\mathrm{d}F(x) \geq \int_a^\infty a \,\mathrm{d}F(x),$$

the Markov inequality

prove the the Markov inequality, that $P(X \geq a) \leq \mathrm{E}\, X/a$. Comment on the following type of consequence: If the average income in your municipality is 1 million kr, then a maximum of 20 percent earns more than 5 million kr.

(b) More generally, if $h(x)$ is a nonincreasing function, show that $P(X \geq a) \leq \mathrm{E}\, h(X)/h(a)$.

the Chebyshov inequality

(c) From this, deduce the Chebyshov inequality: if X has finite variance, then

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

a version of the LLN

With $X_1, \ldots, X_n$ being i.i.d. from such a distribution, with mean ξ and standard deviation σ, show that for the empirical mean $\bar{X}_n$ that $P(|\bar{X}_n - \xi| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2)$, and comment. This is actually a form of the Law of Large Numbers: $P(|\bar{X}_n - \xi| \geq \varepsilon) \to 0$, for any $\varepsilon > 0$; see Ch. 4.

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

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

Ex. 2.8 *Approximate normality and convergence in distribution.* Often the distribution of estimators as well as classes of other statistics are approximately normal. There are several ways of making such a notion precise, and indeed we come back to a more formal apparatus, and to many more details, in Chs. 4, 5. Since approximate normality is so pervasively and powerfully present also when describing behaviour of estimators in simpler settings, we give some preliminary definitions and remarks in this exercise.

(a) With a variable Y_n, with distribution depending on an index n, which often will be or is related to the sample size, we say that Y_n tends to the standard normal in distribution, and write $Y_n \to_d \mathrm{N}(0,1)$, if it is the case that

$$P(a \leq Y_n \leq b) \to P\{a \leq \mathrm{N}(0,1) \leq b\} = \int_a^b \phi(x)\,\mathrm{d}x,$$

for all intervals $[a,b]$, where $\phi(x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}x^2)$ is the standard normal density. Show that $P(|Y_n| \geq 1.96) \to 0.05$. Find the limiting probability of the event that Y_n lands in $[0, 0.50] \cup [1, 1.50] \cup [2, 2.50] \cup [3, 3.50]$.

(b) Since it's so deservedly famous and powerfully useful, from theoretical probability to applied statistics, let's not hesitate to point to the CLT, the Central Limit Theorem. We do come back to the CLT, its proof, ramifications, and some extensions later, in particular in Chs. 4, 5, 9, but we state it here: Suppose $Y_1, Y_2, \ldots$ are i.i.d. with finite mean ξ and standard deviation σ. Then the random sum, normalised to have mean zero and variance one, i.e. the CLT

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

tends to the $\mathrm{N}(0,1)$ in distribution. Note that there are no further assumptions on the distribution of the Y_i, so the exact distribution of Z_n, for a given small or moderate n, might be complicated or strange, but as n increases everything is being smoothened out in the Gaussian fashion. – Use this to show that if $K_n \sim \chi^2_n$, then $\sqrt{n}(K_n/n - 1) \to_d \mathrm{N}(0,2)$.

(c) Let more generally Y_n and Y have cumulative distribution functions F_n and F. We say that Y_n converges to Y in distribution, and write $Y_n \to_d Y$, provided

$$F_n(b) - F_n(a) = P(a < Y_n \leq b) \to P(a < Y \leq b) = F(b) - F(a),$$

for all windows $[a,b]$ where both endpoints a and b are continuity points for F. The point, theoretically, practically, and empirically, is that perhaps complicated Y_n probabilities

might be approximated by perhaps simple Y probabilities. – Suppose $Y_n \to_d \mathrm{N}(0,1)$. Show that $Y_n^2 \to_d \chi_1^2$. [xx later: make sure the $W_{n,q} \to_d \mathrm{N}(0, 1 + \frac{1}{2}z_q^2)$ is seen as a good case in point here. xx]

(d) With $U_1, \ldots, U_n$ being i.i.d. from the uniform distribution on the unit interval, let $M_n = \max_{i \le n} U_i$. Show that $n(1 - M_n)$ tends to the unit exponential in distribution.

(e) Suppose Y_n and Y are integer valued variables, with values in $0, 1, 2, \ldots$, and with probabilities $p_n(j) = P(Y_n = j)$ and $p(j) = P(Y = j)$. Show that $Y_n \to_d Y$ if and only if there is pointwise convergence of these probability functions, i.e. $p_n(j) \to p(j)$ for each j.

(f) With Y_n a binomial (n, p), with fixed p, show that $(Y_n - np)/\{np(1-p)\}^{1/2} \to_d \mathrm{N}(0,1)$, and that $\sqrt{n}(Y_n/n - p) \to_d \mathrm{N}(0, p(1-p))$. With growing n and shrinking p, however, in a manner such that $np \to \lambda$, show that $Y_n \to_d \mathrm{Pois}(\lambda)$.

(g) When $Z_n \sim \mathrm{Pois}(n)$, show that $(Z_n - n)/\sqrt{n} \to_d \mathrm{N}(0,1)$.

(h) Let $X_1, \ldots, X_n$ be independent standard normals. With $M_n(t) = n^{-1} \sum_{i=1}^n \exp(tX_i)$, which we may call the empirical moment-generating function, find the limit distribution for $\sqrt{n}\{M_n(t) - \exp(\frac{1}{2}t^2)\}$.

Ex. 2.9 *Approaching zero.* When dealing with limit distributions it is practical to formalise and build some rules around the notion of variables approaching zero. We say that Z_n converges to zero in probability, and write $Z_n \to_{\mathrm{pr}} 0$, if $P(|Z_n| \ge \varepsilon) \to 0$ for each $\varepsilon > 0$.

(a) When $X_1, X_2, \ldots$ are i.i.d., with mean zero and finite variance, show that the sequence of sample averages tends to zero in probability; see Ex. 2.7.

(b) When $Z_n \to_{\mathrm{pr}} 0$, show that also $h(Z_n) \to_{\mathrm{pr}} 0$, if h is continuous at zero.

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

(d) Show that if $Y_n \to_d Y$ and $\varepsilon_n \to_{\mathrm{pr}} 0$, then $Y_n + \varepsilon_n \to_d Y$ and $Y_n \varepsilon_n \to_{\mathrm{pr}} 0$.

(e) With $X_1, \ldots, X_n$ being i.i.d., with mean ξ, variance σ^2, and finite fourth moment, consider $\widehat{\sigma}_0^2 = (1/n)\sum_{i=1}^n (X_i - \xi)^2$, which uses the ξ, and $\widehat{\sigma}^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)^2$, which uses the sample mean $\bar{X}_n$. Show that these are close; $\sqrt{n}(\widehat{\sigma}^2 - \widehat{\sigma}_0^2) \to_{\mathrm{pr}} 0$.

Ex. 2.10 *Approximate multinormality.* In Ex. 2.8 we described the basics for approximate normality and convergence in distribution, for the one-dimensional case, also pointing to the CLT. We also need to extend this machinery to the multi-dimensional case, also since there are many situations where a one-dimensional estimator or test statistic is a function of several components. Again, more material on large-sample methods, results, techniques is in Chs. 4, 5, 9, but here we work through the basics for the multi-dimensional CLT and see some of its applications.

(a) To approach the notion of 'approximate multinormality' we need an associated notion of convergence in distribution to a multinormal distribution. We will work with more general notions and definitions in Chs. 4, 5, but for the present introduction chapter it is sufficient to say that a random vector $Y_n = (Y_{n,1}, \ldots, Y_{n,p})^{\mathrm{t}}$ converges to a limit distribution variable $Y = (Y_1, \ldots, Y_p)^{\mathrm{t}}$ in distribution if (and actually only if) all linear combinations converge accordingly:

$$Z_n = c^{\mathrm{t}} Y_n = c_1 Y_{n,1} + \cdots + c_p Y_{n,p} \to_d Z = c^{\mathrm{t}} Y = c_1 Y_1 + \cdots + c_p Y_p$$

for each vector $c = (c_1, \ldots, c_p)^{\mathrm{t}}$. – Show that $Y_n \to_d \mathrm{N}_p(0, \Sigma)$ if and only if $c^{\mathrm{t}} Y_n \to_d \mathrm{N}(0, c^{\mathrm{t}} \Sigma c)$ for all c.

(b) Prove that the multi-dimensional CLT then follows from the one-dimensional version: If $X_1, X_2, \ldots$ are i.i.e. from some p-dimensional distribution, with finite mean $\xi = \mathrm{E}\, X_i$ and variance matrix $\Sigma = \mathrm{Var}\, X_i$, then the normalised sum $Z_n = n^{-1} \sum_{i=1}^n (X_i - \xi) = \sqrt{n}(\bar{X}_n - \xi)$ tends to the multinormal $\mathrm{N}_p(0, \Sigma)$.

(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) 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 $\widehat{\sigma}_0^2 = n^{-1} \sum_{i=1}^n (X_i - \xi)^2$.

(e) 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. 2.9. Conclude that the two-dimensional limit distribution result above continues to hold with $\sqrt{n}(\widehat{\sigma}^2 - \sigma^2)$ replacing $\sqrt{n}(\widehat{\sigma}_0^2 - \sigma^2)$.

(f) 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

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - \xi) \\ \sqrt{n}(\widehat{\sigma}^2 - \sigma^2) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2, & 0 \\ 0, & 2\sigma^4 \end{pmatrix} \Big).$$

[xx pointer to later uses, via the delta method, for functions $g(\bar{X}, \widehat{\sigma})$. an example, with moment estimators, is for the $\mathrm{Gam}(a, b)$, see Ex. 5.4. must be others, for moment estimators. xx]

Ex. 2.11 *Approximate variances and the delta method.* 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. More details are discussed in Ch. 4; see in particular Ex. 4.25. Here we are content to work out the basics, and to see its uses in the following few exercises.

(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 $\gamma_4 = \mathrm{E}\,(X_i/\sigma)^4 - 3$ (xx calibrate crosslink later xx). With $\bar{X}$ as usual being the average, show then that

$$\begin{aligned} \mathrm{E}\,\bar{X}^2 &= \sigma^2/n, \\ \mathrm{E}\,\bar{X}^3 &= (\sigma^3/n^2)\gamma_3, \\ \mathrm{E}\,\bar{X}^4 &= (\sigma^4/n^2)(3+\gamma_4/n), \\ \mathrm{Var}\,\bar{X}^2 &= (\sigma^4/n^2)(2+\gamma_4/n). \end{aligned}$$

(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 $\bar{Y}$ 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 $Z_n = g(\bar{Y})$, Since $\bar{Y}$ is close to ξ with high probability, as we learn more formally in Ex. 2.7, 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, if g has three derivatives in a neighbourhood around ξ. Show from the above that

$$\mathrm{Var}\,Z_n = g'(\xi)^2\sigma^2/n + (1/n^2)\{\tfrac{1}{4}g''(\xi)^2\sigma^4(2+\gamma_4/n) + g'(\xi)g''(\xi)\sigma^3\gamma_3\} + o(1/n^2).$$

(d) There is hence a clear leading $O(1/n)$ term, for the variance, with other terms being of size $O(1/n^2)$. To the first order of approximation, show

$$\mathrm{Var}\,Z_n \doteq g'(\xi)^2\sigma^2/n, \quad \sqrt{n}(\bar{Y}-\xi) \approx \mathrm{N}(0,\sigma^2).$$

(e) This is a version of the so-called *delta method*, very useful in probability theory and statistics. Try to prove, with direct methods, using the formal appartus of Ex. 2.8, that

$$\sqrt{n}(A_n - a) \to_d Z \quad \text{implies} \quad \sqrt{n}\{g(A_n) - g(a)\} \to_d g'(a)Z.$$

If the limit Z is a normal $(0,\sigma^2)$, then $g'(a)Z$ is the normal $(0, g'(a)^2\sigma^2)$.

(f) We come forcefully back to the delta method, with more details, generalisations, and uses, in Ch. 4, see e.g. Ex. 4.25. Here we spell out the function of a vector version, since it is so immediately useful; see exercises below. – Suppose $\sqrt{n}(A_n - a) \to_d Z$, as

random vectors of dimension k (xx calibrate with what is said previously in chapter on such convergence; can go to linear combinations xx). With a smooth $g(A_n)$, defined at least in a neighbourhood around a, we have

$$\sqrt{n}\{g(A_n) - g(a)\} \to_d c^{\mathrm{t}} Z = c_1 Z_1 + \cdots + c_k Z_k,$$

where $c = \partial g(a)/\partial a$, the vector of partial derivatives $c_j = \partial g(a)/\partial a_j$, computed at position a. Show that if the limit Z of $\sqrt{n}(A_n - a)$ is multinormal, say $\mathrm{N}_k(\xi, \Sigma)$, then the limit $c^{\mathrm{t}} Z$ is normal $(c^{\mathrm{t}}\xi, c^{\mathrm{t}}\Sigma c)$. The vector version of the delta method is often given precisely in this form. (xx to form this as an exercise, go via linear combinations, which then needs us mentioning the cramer-wold, see Ex. 2.10, with details in Ex. 4.42. we make the readers understand that the vector case follows from the unidim case. xx)

(g) (xx The point is that once a start limit distribution result has been established, perhaps via the CLT, then a string of further limit distribution results come almost for free. simple illustration here; more in Ex. 2.12 and 2.14. xx)

Ex. 2.12 *Applying the delta method.* Here we exercise our delta method muscles, to see how the general recipes may be applied in a few situations.

(a) 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 $\widehat{\gamma} = 2\arcsin(\widehat{p}^{1/2})$.

(b) 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]

(c) Suppose $Y_1, \ldots, Y_n$ are independent from the geometric distribution with $P(Y_i = y) = (1-p)^{y-1}p$ for $y = 1, 2, \ldots$. We learned in Ex. 1.15 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 $\widehat{p} = 1/\bar{Y}$.

(d) 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).

(e) 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.

(f) Suppose one observes $(A_n, B_n) = (4.44, 2.22)$, with $n = 100$. Construct and display a 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. 2.13 *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 $\widehat{\sigma}_n$. Under normality we have precise finite-sample results regarding their distributions, see Ex. 1.33, 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. 2.10, to show that

$$\begin{pmatrix} \sqrt{n}(\bar{Y}_n - \xi) \\ \sqrt{n}(\widehat{\sigma}_n - \sigma) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1, & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3, & \frac{1}{2} + \frac{1}{4}\gamma_4 \end{pmatrix}\Big).$$

(b) Show that $\sqrt{n}(\widehat{\sigma} - \sigma)/\widehat{\kappa} \to_d \mathrm{N}(0,1)$, if κ is a consistent estimator for $\kappa = (\frac{1}{2} + \frac{1}{4}\gamma_4)$. For an application of this, with sample size up to half a million, see Story vii.2.

Ex. 2.14 *Delta method calculus for the normal case.* Let $Y_1, \ldots, Y_n$ be i.i.d. from the normal $\mathrm{N}(\mu, \sigma^2)$. In Exercises [xx fill in xx] we have worked with *exact finite-sample* calculus, for certain basic parameters, like the quantile $\gamma_q = \mu + z_p\sigma$. 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 is 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 [xx above xx] implies that $(\sqrt{n}(\bar{Y} - \mu), \sqrt{n}(\widehat{\sigma} - \sigma))^{\mathrm{t}}$ tends to say $(A, B)^{\mathrm{t}}$, 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 $\sqrt{n}(\bar{Y} - \mu) \sim \mathrm{N}(0, \sigma^2)$ holds exactly, for each finite n.

(b) With $\alpha = g(\mu, \sigma)$, for any smooth function of the two parameters, the natural estimator is $\widehat{\alpha} = g(\bar{Y}, \widehat{\sigma})$. Show that

$$\sqrt{n}(\widehat{\alpha} - \alpha) \to_d cA + dB \sim \mathrm{N}(0, (c^2 + \tfrac{1}{2}d^2)\sigma^2),$$

where c and d are the partial derivatives of g, evaluated at the position (μ, σ). Show how this leads to construction of confidence intervals for the α parameter.

(c) Consider the probability $p = P(Y \geq y_0) = 1 - \Phi((y_0 - \mu)/\sigma)$, 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 $\kappa = \mu/\sigma$, 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) (xx we do one more such. xx)

Ex. 2.15 *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. 2.11. 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. 2.16 *The binomial, the normal approximation, and confidence intervals.* Here is a good occasion to use the CLT, also since some of its immediate *consequences and uses* for this binomial situation must be considered basic knowledge (i.e. even if one does not necessarily know or does not yet care about all the mathematical details under the hood).

(a) So, use the CLT, as formalised e.g. in Ex. 2.8, 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. [xx push to pointers: This theorem is 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] Discuss briefly the skewness result from Ex. 1.3 above in light of the limiting normality.

(b) With $\widehat{p}_B = (Y+1)/(n+2)$, as in Ex. 1.3, 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) Show from the above that

$$P(-1.96 \le W_n \le 1.96) \to 0.95 \quad \text{as sample size increases,}$$

a confidence interval

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.

(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. Then both

asymptotic equivalence

$$W_n = \frac{\widehat{p}-p}{\sigma_n} \quad \text{and} \quad W_n' = \frac{\widehat{p}-p}{\widehat{\sigma}_n}$$

tend to the standard normal in distribution. The first version is that reached and used above (essentially the CLT for Bernoulli variables), whereas the second version requires some additional analysis, returned to in e.g. Ex. 4.14. 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. (independent and identically distributed) 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.

Ex. 2.17 *Limiting distributions via densities.* It is often practical to work with densities, rather than cumulatives, and $f_n \to f$ for densities indeed implies $F_n \to F$ for cumulatives. This is called the Scheffé lemma; see Ex. 4.16 for details.

(a) Show that $t_\nu \to_d \mathrm{N}(0,1)$.

(b) Show that if $X_n \sim \mathrm{Gam}(a_n, b_n)$, and $a_n \to 1, b_n \to b$, then $X_n \to \mathrm{Expo}(b)$.

(c) (xx one or two more. the point is to set it up for use for the quantiles below. xx)

Ex. 2.18 *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

$$h_n(z) \to (2\pi)^{-1/2} 2 f(\theta) \exp\{-\tfrac{1}{2} 4 f(\theta)^2 z^2\}.$$

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.

(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. 1.34. 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. 3.19. 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. 2.19 *Uniform ordering.* Consider $U_1, \ldots, U_n$ i.i.d. from the uniform distribution. Order these, to $U_{(1)} < \cdots < U_{(n)}$.

connection to Beta distributions

(a) Show that $U_{(i)}$ has density

$$g_i(u) = \frac{n!}{(i-1)!\,1!\,(n-i)!} u^{i-1}(1-u)^{n-i} \quad \text{for } u \in (0,1).$$

Explain that $U_{(i)} \sim \mathrm{Beta}(i, n-i+1)$, see Ex. 1.25, show that $\mathrm{E}\, U_{(i)} = p_i = i/(n+1)$, and that $\mathrm{Var}\, U_{(i)} = p_i(1-p_i)/(n+2)$.

(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 $U_{(i)} \in [u, u+\mathrm{d}u]$ and $U_{(j)} \in [v, v+\mathrm{d}v]$, there is a multinomial situation, with five boxes $[0,u], [u,u+\mathrm{d}u], [u+\mathrm{d}u, v], [v, v+\mathrm{d}v], [v+\mathrm{d}v, 1]$, 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 $U_{(1)} \sim \mathrm{Be}(1,n)$. 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 $[u,1]$. 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 $\mathrm{Be}(n-1,2)$ distribution. (xx pointer to exercises in Ch6, Ch7, or perhaps just to Story ii.6, 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 $u_{(1)} < \cdots < u_{(n)}$.

connection to Dirichlet

(e) Use this, in conjunction with Ex. 1.24, to demonstrate that

$$\begin{aligned}(U_{(1)}, U_{(2)}, \ldots, U_{(n)}) &=_d (D_1, D_1+D_2, \ldots, D_1+\cdots+D_n)\\ &=_d (V_1/S, (V_1+V_2)/S, \ldots, (V_1+\cdots+V_n)/S),\end{aligned}$$

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

$$U_{([nq])} = \sum_{i=1}^{[nq]} V_i \Big/ \sum_{i=1}^{n+1} V_i \quad \text{for } 0 \le q \le 1. \tag{2.3}$$

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 $U_{([nq])} \to_{\mathrm{pr}} q$. point to things in Ch9 with full process convergence $\sqrt{n}(U_{[nq]} - q) \to_d W^0(q)$, the Brownian bridge. xx)

Ex. 2.20 *Sample quantiles.* Suppose $Y_1, \ldots, Y_n$ are independent observations coming from the same distribution, with positive density f and cumulative distribution function 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 $q = 0.50$.

sample quantiles

(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. 2.19. 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. 2.18 that $\sqrt{n}(U_{([0.50\,n])} - 0.50) \to_d \mathrm{N}(0, 0.50^2)$, for the uniform median. Then show for a general density f that $\sqrt{n}\{Q_n(0.50) - \mu\} \to_d \mathrm{N}(0, 0.50^2/f(\mu)^2)$, 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 (2.3). 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 $\mathrm{N}(0, 0.50^2)$ limit.

(e) Then generalise to the case of any given quantile level q. Show first that $\sqrt{n}(U_{[nq]} - q) \to_d \mathrm{N}(0, q(1-q))$, and then that the limit distribution is $\mathrm{N}(0, q(1-q)/f(\mu_q)^2)$ 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 $q_1 < q_2$,

$$\begin{pmatrix} \sqrt{n}\{Q_n(q_1) - q_1\} \\ \sqrt{n}\{Q_n(q_2) - q_2\} \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} q_1(1-q_1), & q_1(1-q_2) \\ q_1(1-q_1), & q_2(1-q_2) \end{pmatrix}\Big).$$

Prove this via the explicit density for $(U_{(i)}, U_{(j)})$, given in Ex. 2.19.

(g) It is also instructive to use representation (2.3) 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 $-q_j q_\ell$ 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 $-q_j q_\ell/\{f(\mu_j)f(\mu_\ell)\}$ for $j < \ell$, where $\mu_j = F^{-1}(q_j)$. See Ex. 9.17 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. 2.21 *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 $P(Y_2 - Y_1 \le c) = P(Y_2 - Y_1 > c) = \frac{1}{2}$, for $c = 1 - 1/\sqrt{2} = 0.2929$.

(c) Find also the joint distribution for (R, W) here.

Ex. 2.22 *Ordering exponentials.* (xx check and calibrate the order in which things are told here. xx) Let Y_1, Y_2, Y_3 be independent unit exponentials (with density $\exp(-y)$ for y positive), and order them, to $Y_{(1)} < Y_{(2)} < Y_{(3)}$. Then define the so-called spacings between them, $D_1 = Y_{(1)}$, $D_2 = Y_{(2)} - Y_{(1)}$, $D_3 = Y_{(3)} - Y_{(2)}$.

(a) Find their joint distribution, and show that they are independent. (xx i think this is not true for other start distributions for the data points than the exponential. can we semi-easily prove such a characterisation? xx)

(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) Use this to show that $M_n = \max X_i$ has mean close to $\log n + \gamma$, where $\gamma = 0.5772...$ is the Euler constant, and variance converging to $\pi^2/6$. Finally find the limit or $P(M_n - \log n \le u)$. (xx pointer to things in Ch1 with Gumbel etc. xx)

Ex. 2.23 *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. 2.22.

(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. 2.24 *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 $X_{(i)} = V_iV_{i+1}\cdots V_n$. 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. 2.25 *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 and the Cauchy. 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]

Ex. 2.26 *Which order statistics interval contains the true median?* (xx nilsrant, as of 13-Aug-2023, 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

$$p_i = P\{\mu \in (Y_{(i)}, Y_{(i+1)})\} = P\{\tfrac{1}{2} \in (U_{(i)}, U_{(i+1)})\},$$

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

$$p_i = P\{U_{(i)} < \tfrac{1}{2} < U_{(i+1)}\} = \int_0^{1/2} P\{U_{(i+1)} > \tfrac{1}{2} \,|\, U_{(i)} = u\} g_i(u)\,\mathrm{d}u$$
$$= \int_0^{1/2} \Big(\frac{\frac{1}{2}}{1-u}\Big)^{n-i} \mathrm{be}(u, i, n-i+1)\,\mathrm{d}u,$$

involving a Beta density, as per Ex. 2.20. 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)

Ex. 2.27 *The empirical distribution function.* Assume there is an i.i.d. dataset $Y_1, \ldots, Y_n$ from an unknown distribution, with cumulative distribution function $F(t) = P(Y_i \le t)$. 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 $(-\infty, t]$. Since we know so much about the binomial, we quickly learn a few basic properties of the F_n.

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 2.1, 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) 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 2.1, right panel. [xx nils emil: we might contemplate putting comments such as the following in a 'comments' format, at the end 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

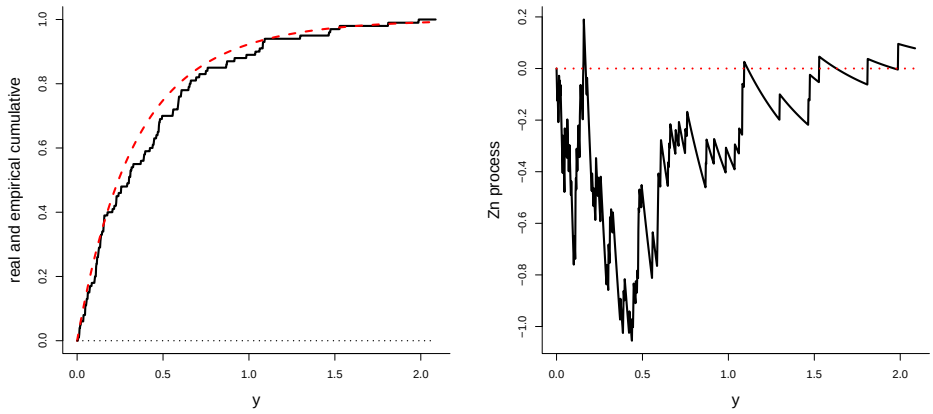

Figure 2.1: *Left panel: The real underlying data-generating $F(t)$ (dashed, red), with the empirical distribution function $F_n(t)$ (full line, black), computed from a sample of $n = 100$ data points from F. Right panel: The process $Z_n(t)$, computed for the data used for the same data. In 95 percent of such cases, the maximum absolute value of the Z_n process will be below 1.358.*

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 $Z_n(t) \to_d \mathrm{N}(0, F(t)\{1 - F(t)\})$. [xx pointer to Ch. 9. the 1.358 limit. kolmogorov-smirnov. and to Glivenko–Cantelli theorem, in Ex. 4.37. xx]

(e) (xx some pointers: 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)

(f) xx can also briefly ask about interesting variations, like

$$F_n^*(t) = \frac{1 + \sum_{i=1}^n I(Y_i \le t)}{n+2} = \frac{1}{n+2} + \frac{n}{n+2} F_n(t),$$

inspired by the Bayes type $(Y + 1)/(n + 2)$ binomial estimator. And the minimax estimator, point to Hjort (1976); Phadia (1973). xx

Ex. 2.28 *Moment fitting 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. 1.15, 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. 1.23, find expressions for the method of moments estimators for a, b.

(c) 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. From your data, use the method of moments to estimate the three parameters.

Ex. 2.29 *Moment fitting estimators for the Gamma distribution.* We now apply the moment matching principle of Ex. 2.28 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. An aspect of what we find here will be used in Story ii.1 (xx check this xx).

(a) With $\bar{Y}$ and S^2 the usual sample mean and sample variance, find explicit formulae for the moment estimators $\widehat{a}, \widehat{b}$.

(b) Use skewness and kurtosis formulae, in combination with Ex. 2.10, to show that

$$\begin{pmatrix} \sqrt{n}(\bar{Y} - a/b) \\ \sqrt{n}(S^2 - a/b^2) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} a/b^2, & 2a/b^3 \\ 2a/b^3, & (2+6/a)a^2/b^4 \end{pmatrix}\Big).$$

Transform this, via the delta method, to find the limit distribution for the moment estimators.

(c) (xx application in Story ii.1. delta method for $g(\widehat{a}, \widehat{b})$, like the median. xx)

(d) xx

Ex. 2.30 *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. 2.28, 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. 2.18, 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. 1.13. 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. 2.31 *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. 1.40.

(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.

Ex. 2.32 *Moment-generating functions close to zero.* (xx need polish; might go to large-sample. xx) Consider a variable Y, with moment-generating function $M(t) = \mathrm{E}\exp(tY)$, assumed to be finite in at least a neighbourhood around zero. We have seen in Ex. 1.21 that $\mathrm{E}\,Y^r = M^{(r)}(0)$. Write ξ and σ^2 for the mean and variance of Y.

(a) 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 moment-generating function $M(t)$ being finite around zero. Show that $Z_n = \sqrt{n}\overline{Y}$ 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. 1.21, with attention to certain further details in Ch 3 xx).

(d) (xx round off, point to CLT, identify remainder term with skewness. xx)

Ex. 2.33 *Estimation via least sum of squares.* (xx to come. the basics. minimising $Q(\theta) = \sum_{i=1}^n \{y_i - \xi_i(\theta)\}^2$. some easy points, finding formulae.

(a) (xx some easy cases. binomial. normal mean. poisson. xx)

(b) Assume $y_1, \ldots, y_n$ 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)

(d) xx

Ex. 2.34 *Linear regression.* Consider the model for observed pairs (x_i, y_i), where $Y_i \sim \mathrm{N}(a_0 + bx_i, \sigma^2)$ for $i = 1, \ldots, n$, with the Y_i being independent. This is classical linear regression, widely used in theoretical and applied statistics. Importantly, the model, along with methods for estimation and inference, will be broadly generalised in Ch. 3, see Ex. 3.33, 3.34, 3.35, to the case of multiple linear regression, allowing several covariates. (xx Least squares things, and also $\widehat{\sigma}$. Estimation also of $a + bx_0$ fixed x_0, and for $P(Y \geq y_0 \,|\, x_0)$. xx)

(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 $Q_0/(n-2)$ is an unbiased estimator of σ^2. (xx more under normality. that $Q_0/\sigma^2 \sim \chi^2_{n-2}$, independent of $(\widehat{a}, \widehat{b})$. xx)

(d) (xx inference for β. and for σ. and for $m(x_0) = \mathrm{E}\,(Y \mid x_0) = a + b(x_0 - \bar{x})$ for a fixed x_0. xx)

Ex. 2.35 *Predicting the next y.* [xx for the next observation, with x_0, where will y_0 land? An application. point to Story iv.1. xx]

(a) (xx The next y_0. xx)

Ex. 2.36 *One or two more on linear regression things.* [xx How to test $\beta = 0$. careful xref with Ch 2 exercises. How to test linear $a + bx$ inside $a + bx + cx^2$. Example. Predicting the next point. Some simple goodness-of-fit things. xx]

Ex. 2.37 *Exponential regression.* Something like $\lambda_i = \lambda_0 \exp(\beta x_i)$. Perhaps we need to write down a log-likelihood, but we don't push it hard, since that comes later. The point is to convey 'yes, we can put up regression models beyond the linear one'.

Ex. 2.38 *Nonparametric estimation.* [xx Something around estimation in iid setup of means, quantiles, functions thereof via delta method. A few points regarding $\widehat{\sigma}^2$, which remains a perfectly good estimator of σ^2 also outside normality, but then with a more complex variance formula, etc. xx]

Ex. 2.39 *Regression outside normality.* [xx We do $\widehat{\beta}$ analysis without assuming that the ε_i are normal. Hint at Lindeberg and such things. An example, with a bit of simulation. xx]

Ex. 2.40 *Extra: empirical covariance matrix.* (xx we'll see later where and how to fit this in. could also be inside likelihood chapter.) Suppose $Y_1, \ldots, Y_n$ are i.i.d. vectors in dimension p, with mean ξ and covariance matrix Σ.

(a) With $A = \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^{\mathrm{t}}$, show that Show that

$$\sum_{i=1}^n (Y_i - \xi)(Y_i - \xi)^{\mathrm{t}} = A + n(\bar{Y} - \xi)(\bar{Y} - \xi)^{\mathrm{t}}.$$

Deduce that $\widehat{\Sigma}_u = A/(n-1)$, the so-called empirical covariance matrix, is unbiased for Σ.

(b) (xx to be pushed to likelihood chapter. xx) Under multinormality, show that the log-likelihood function can be expressed as

$$\ell_n(\xi, \Sigma) = -\tfrac{1}{2} n \log |\Sigma| - \tfrac{1}{2} \sum_{i=1}^n (Y_i - \xi)^{\mathrm{t}} \Sigma^{-1} (Y_i - \xi).$$

Show that this, for any positive definite Σ, is maximised for $\widehat{\xi} = \bar{Y}$, leading to the profiled

$$\ell_{\mathrm{prof}}(\Sigma) = -\tfrac{1}{2} n \{\log |\Sigma| + \mathrm{Tr}(\Sigma^{-1} \Sigma^*)\},$$

in terms of $\Sigma^* = A/n$, i.e. with divisor n not $n-1$ as above. Show in general that if B is some given symmetric positive definite matrix, then the function $h(\Sigma) = \log |\Sigma| + \mathrm{Tr}(\Sigma^{-1} B)$ is minimised for $\widehat{\Sigma} = B$. Deduce that Σ^* is the ML estimator for Σ. (xx this might perhaps most easily be shown using the spectral decomposition $P \Sigma P^{\mathrm{t}} = D$, a diagonal. xx)

Ex. 2.41 *Squares and products of normals.* (xx this might not earn its place, depends on how dennis nils fieller difference work pans out. but it is a good illustration of good clean probability calculus, for a purpose.) For independent $\widehat{a} \sim \mathrm{N}(a,1)$ and $\widehat{b} \sim \mathrm{N}(b,1)$, consider

$$U = 2\widehat{a}\widehat{b} \quad \text{and} \quad V = \widehat{a}^2 + \widehat{b}^2.$$

Aspects of their joint distribution come up in the context of Fieller parameters, of the type a/b, etc.

(a) Show that $|U| \le V$. For

$$T_1 = (\widehat{a}+\widehat{b})/\sqrt{2} \sim \mathrm{N}((a+b)/\sqrt{2},1), \qquad T_2 = (\widehat{a}-\widehat{b})/\sqrt{2} \sim \mathrm{N}((a+b)/\sqrt{2},1),$$

show that these two are independent, and that

$$Z_1 = T_1^2 = \tfrac{1}{2}V + \tfrac{1}{2}U, \quad Z_2 = T_2^2 = \tfrac{1}{2}V - \tfrac{1}{2}U.$$

From $V = Z_1 + Z_2$ and $U = Z_1 - Z_2$, show that (U,V) has joint density

$$h(u,v) = g_1(\tfrac{1}{2}u + \tfrac{1}{2}v, \tfrac{1}{2}(a+b)^2)\, g_1(-\tfrac{1}{2}u + \tfrac{1}{2}v, \tfrac{1}{2}(a-b^2),$$

in terms of the density function $g_1(\cdot,\lambda)$ for the noncentral $\chi^2_1(\lambda)$ (xx mind the determinant; is it a factor 1/8? xx).

(b) Show that the conditional density for U given $V = v$ may be written

$$f_0(u \,|\, v) = \frac{g_1(\frac{1}{2}u + \frac{1}{2}v, \frac{1}{2}(a+b)^2)\, g_1(-\frac{1}{2}u + \frac{1}{2}v, \frac{1}{2}(a-b)^2)}{g_2(v, a^2+b^2)} \quad \text{for } u \in (-v, v).$$

The ratio $R = U/V$ has density $f(r \,|\, v) = v f_0(vr \,|\, v)$ for $r \in (-1,1)$.

(c) (xx can we complete this? xx) For the difference between two Fieller parameters, i think we need something like $U - U'$ and $V + V'$. perhaps

$$\begin{aligned} U^* = U - U' &= Z_1 - Z_2 - (Z_3 - Z_4) = Z_1 - Z_2 - Z_3 + Z_4, \\ V^* = V + V' &= Z_1 + Z_2 + Z_3 + Z_4. \end{aligned}$$

which leads to

$$\tfrac{1}{2}(V^* + U^*) = Z_1 + Z_4, \quad \tfrac{1}{2}(V^* - U^*) = Z_2 + Z_3.$$

So we can write down the joint density for (U^*, V^*):

$$\begin{aligned} h^*(u^*, v^*) = {} & \gamma_2(\tfrac{1}{2}u^* + \tfrac{1}{2}v^*, \tfrac{1}{2}(a_1+b_1)^2 + \tfrac{1}{2}(a_2-b_2)^2) \\ & \gamma_2(-\tfrac{1}{2}u^* + \tfrac{1}{2}v^*, \tfrac{1}{2}(a_1-b_1)^2 + \tfrac{1}{2}(a_2+b_2)^2), \end{aligned}$$

modulo a determinat.

(d)

2.C 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)

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.3

Confidence intervals, testing, and power

In the previous chapters we have learned about classes of distributions, their parameters, and ways of estimating these from data. The present chapter goes on to the fundamental statistical reporting tools of confidence intervals and testing. The former are data-based intervals made to cover the true underlying parameter value with a given degree of confidence, like 95 percent. 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 intervals and 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. (xx more to come in Chapters 5, 7, 8. xx)

3.A Chapter introduction

Fundamental and conspicuous instruments in the statistical toolkits, when summarising and reporting findings of investigations, are *confidence intervals* and *testing of null hypotheses*. This chapter deals with such, their interpretation, construction, properties, performance, and connections.

We have in fact already bumped into confidence intervals in the course of exercises in Ch. 2, but here we give them a more formal treatment, with yet more to come in Chs. 5, 7. Consider data y, perhaps a long vector or a data matrix, from a model with parameter $\theta = (\theta_1, \ldots, \theta_p)$, and suppose $\phi = \phi(\theta_1, \ldots, \theta_p)$ is 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

$$P_\theta\{L(Y) \le \phi \le U(Y)\} = \alpha \quad \text{for all } \theta. \tag{3.1}$$

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 $[L_{\text{obs}}, U_{\text{obs}}] = [L(y_{\text{obs}}), U(y_{\text{obs}})]$, computed based on the actually observed data y_{obs}.

There is a long list of situations where (3.1) does hold precisely, but an even longer list where the statistician only can construct *approximate* confidence intervals, where the coverage probability in question holds approximately rather than with full precision. Often such approximations stem from large-sample calculus, via methods and results of Chs. 4, 5. There are many such versions in exercises below. We note that the notion of confidence intervals can be formulated also in nonparametric situations; as long as ϕ above has a clear interpretation, as a function of the model, we can put up a parallel version of (3.1) without having a θ.

The second major theme of this chapter is that of statistical tests. Consider in general terms, as above, 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. A simple illustration is the test of one component parameter being equal to zero, or testing that two parameters, perhaps for two groups, are equal. A *test* for such an hypothesis, is a rule saying 'we reject H_0 if data y fall in the set R', along with the opposite statement 'we do not reject H_0 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

$$P_\theta(\text{reject } H_0) \le \alpha \quad \text{for all } \theta \in \Theta_0. \tag{3.2}$$

With level 0.01 one is guaranteed such a low chance of falsely claiming that H_0 is wrong, if it is indeed correct. We are also keenly interested in *the power* of a test, which is the detection chance $P_\theta(\text{reject } H_0)$ 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 in the exercises below.

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 $p = P_{H_0}(T \ge t_{\rm obs})$, with $t_{\rm obs}$ 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

$$p = \max\{P_\theta(T \ge t_{\rm obs})\colon \theta \in \Theta_0\}. \tag{3.3}$$

A small p-value, like $p \le 0.01$, 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 $p \le 0.01$.

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. 3.12–3.14. 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.

(xx todo: since we have exponential family in Ch1, we should supplement the exercises with conditional tests with one or two grander things for optimal power for testing inside such exponential families. xx)

(xx then one paragraph with pointers to other chapters and perhaps to a few of the stories. xx)

3.B Be confident, and test it

Ex. 3.1 *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.

(b) 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?

(c) 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. 3.2 *Confidence interval for a normal variance.* (xx easy stuff. to be polished. xx) 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 (xx exercises in chapter 1 xx).

(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 $P_\sigma(\sigma \in \mathrm{ci}) = 0.90$.

(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 $P_\sigma((m/b)\widehat{\sigma}^2 \le \sigma^2 \le (m/a)\widehat{\sigma}^2) = 0.90$; 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. 3.3 *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 statistical world's 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. 2.4. 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. 1.34. 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.

Ex. 3.4 *Confidence intervals via normal approximations.* (xx make connection to Wald tests to come below. 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. 2.9.

(b) Show under these mild and very frequently met assumptions that

$$P(\phi \in \widehat{\phi} \pm 1.96\,\widehat{\tau}/\sqrt{n}) \to 0.95.$$

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. 2.12.

Ex. 3.5 *Confidence for a normal quantile.* (xx need to polish this, calibrating with previous exercises. xx) Consider again a normal sample, observations $Y_1, \ldots, Y_n$ from the normal $\mathrm{N}(\mu, \sigma^2)$. In addition to understanding the behaviour of the natural estimators for the mean μ, the standard deviation σ, the quantile $\mu + z_q\sigma$, and for yet other quantities, one needs of course methods for constructing confidence intervals and tests, for these parameters.

(a) (xx inference for $\gamma_q = \mu + z_q\sigma$. more tricky. for this one. check and edit the following. xx) 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. This leads to the pivot

$$W_{n,q} = \frac{\sqrt{n}(\widehat{\gamma}_q - \gamma_q)}{\widehat{\sigma}} \sim \frac{N + z_q\sqrt{n}\{(K_m/m)^{1/2} - 1\}}{(K_m/m)^{1/2}}.$$

Its distribution is complicated, but independent of parameters, and can be simulated, for any given sample size n and quantile level q. With a_n such that $P(-a_n \le W_{n,q} \le a_n) = 0.95$, the above may be easily converted 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 $m = n - 1$. [xx can also give the large-sample approximation, with limit of $\sqrt{n}\{(K_m/m)^{1/2} - 1\} \to_d \mathrm{N}(0, \frac{1}{2})$. so $W_{n,q} \to_d \mathrm{N}(0, 1 + \frac{1}{2}z_q^2)$. xx]

(b) (xx two-sample data, populations A and B, normal densities $f_A(y)$ and $f_B(y)$. estimators and confidence intervals for $f_A(y)/f_B(y)$, to test for their equality. easiest on the log scale. xx)

(c) (xx simple data illustration. xx)

(d) (xx pointer to delta method to come, whereby we're able to have inference for *any* $g(\mu, \sigma)$, via large-sample approximation to normality. but the cases treated in this exercise admit exact finite-sample analysis, for any finite n. xx)

Ex. 3.6 *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. (xx see Figure 3.1 for 100 simulated confidence intervals, all attempting to capture the true value, here $\sigma = 1$. xx)

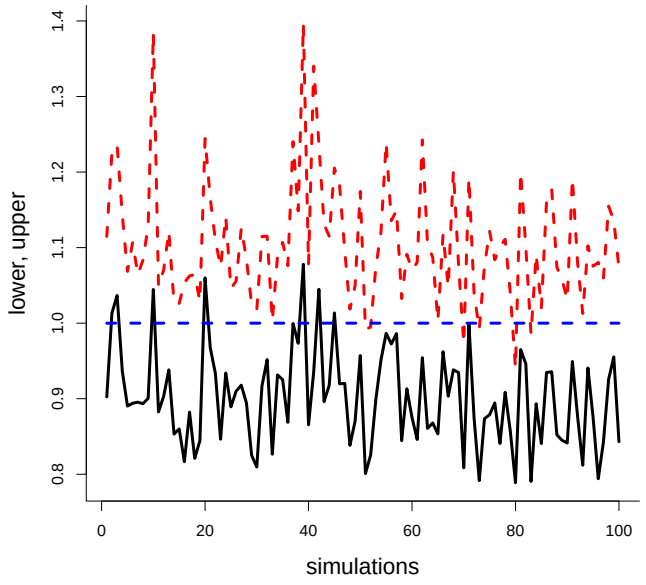

Figure 3.1: *Simulations, with datasets of size* $n = 500$ *from the unit exponential, displaying lower and upper confidence points for 90 percent intervals; the intervals attempt to cover the true value* $\sigma = 1$.

(a) Make sure you understand and can prove that $\widehat{\xi}$ and $\widehat{\sigma}$ are consistent for ξ and σ. (xx calibrate this, with right pointer to LLN with Chebyshov etc.; stronger LLN in next chapter. xx)

(b) Use $S_n^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \xi)^2 - (\bar{Y} - \xi)^2$ to deduce that $\sqrt{n}(S_n^2 - \sigma^2)$ and $\{n^{-1} \sum_{i=1}^{n} (Y_i - \xi)^2 - \sigma^2\}$ must have identical limit distributions, and that this

limit is a $\mathrm{N}(0, \sigma^4(2+\gamma_4))$, in terms of the kurtosis parameter $\gamma_4 = \mathrm{E}\{(Y_i-\xi)/\sigma\}^4 - 3$. The 'subtract 3' is merely a thing of mild convenience, making the kurtosis equal to zero for normal distributions. (xx check, calibrate, avoid defining the kurt several times. xx)

(c) Let us transform the above, from variance to its square root, getting back to the real scale of the measurements: show that $\sqrt{n}(\widehat{\sigma}-\sigma) \to_d \mathrm{N}(0, (\frac{1}{2}+\frac{1}{4}\gamma_4)\sigma^2)$. (xx careful here, tidy up, where is delta method laid out? xx)

(d) 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.

(e) 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 $n = 500$ from this distribution. For each simulated dataset, compute $\widehat{\gamma}_4$, to check how close it is to γ_4, and the approximate 90 percent confidence interval for σ. Construct a version of Figure 3.1. Examine in particular the coverage of your intervals – how often do they contain the correct σ? (xx check, calibrate, with earlier exercises. xx)

(f) Coming back to the general situation, define the skewness as $\gamma_3 = \mathrm{E}\{(Y-\xi)/\sigma\}^3$, which is zero for all symmetric distributions. 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

$$\begin{pmatrix} \sqrt{n}(\widehat{\xi}-\xi) \\ \sqrt{n}(\widehat{\sigma}-\sigma) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3 & \frac{1}{2}+\frac{1}{4}\gamma_4 \end{pmatrix}\Big),$$

(g) Generate a dataset of size $n = 400$ 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.

(h) (xx might also do ci for $\log\sigma$ first, then transforming back to σ. same 1st order asymptotics. xx)

Ex. 3.7 *Testing a null hypothesis.* (xx edit and clean this exercise. we include binomial and trinomial, and point to Story vii.1. "Nullhypotesen er det utsagn hvis feilaktige forkastelse vi fortrinnsvis søker å unngå." a couple of points, qua questions, guiding readers to state hypotheses, and being clear about alternatives. xx)

(a) The probability $p = P(A)$ of a certain event is meant to be 0.33, if the machinery around it works. To test this one carries out the relevant experiment $n = 100$ times, and the event takes place $y = 46$ times. Should you reject the 0.33 hypothesis? Show that with $Y \sim \mathrm{binom}(n, p)$, the statistic $W = (Y-np_0)^2/\{np_0(1-p_0)\}$ 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 $p \neq p_0$.

(b) Suppose (X, Y, Z) is trinomial with sum n and probabilities (p, q, r); see Ex. 1.5. Show that

$$W = (X - np)^2/(np) + (Y - nq)^2/(nq) + (X - nr)^2/(nr)$$

has mean equal to 2. Attempt to show that W is approximately a χ^2_2, using the multidimensional CLT for (X, Y, Z). Explain how this may be used to test the hypothesis H_0 that $(p, q, r) = (p_0, q_0, r_0)$, a given probability vector. This is actually the deservedly famous Pearson chi-squared test for multinomials; see Story vii.1 for details, generalisations, and discussion.

(c) (xx one more case. perhaps $Y \sim \text{geom}(p)$, testing $p = 1/6$ versus $p < 1/6$, since it takes me so long time to roll a 6 with my die. here $P(Y \geq y) = (5/6)^{y-1}$ for $y \geq 1$ under the null. for $y = 15$ not yet really suspicious. xx)

Ex. 3.8 *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 $P_\phi(\phi \in A) = 0.99$, making A a 99 percent confidence region.

(c) (xx a clear example, to see both ways. xx)

Ex. 3.9 *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 $\mu_q = F^{-1}(q)$? (xx there are several approaches here; give pointers to CD chapter and more. but here we give the basic method via F_n. should show later that this first-order equivalent to $\widehat{\mu} \pm 1.645\widehat{\tau}/\sqrt{n}$, where $\widehat{\tau}$ estimates $\frac{1}{2}/f(\mu)$. xx)

(a) Let F_n be the empirical c.d.f. for the data, see Ex. 2.27. If μ_0 is the true median, show that $W_n(\mu_0) = \sqrt{n}\{F_n(\mu_0) - \frac{1}{2}\}/\frac{1}{2}$ is approximately a standard normal, e.g. via Ex. 2.16. 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. 3.8, show that the associated 95 percent confidence interval becomes $\text{ci}_n = \{\mu\colon |F_n(\mu) - \frac{1}{2}| \leq 1.96/(2\sqrt{n})\}$. 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. 2.20.

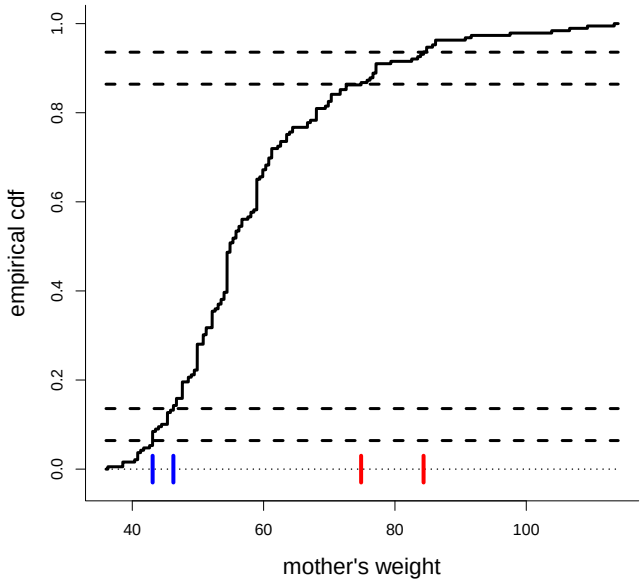

Figure 3.2: *The empirical c.d.f. for the pre-pregnancy weight of 189 mothers, with bands to read off 90 percent confidence intervals for the 0.10 level and 0.90 level quantiles.*

(c) Generalise to the case of any quantile $\mu_q = F^{-1}(q)$. Show that the recipe above leads to the confidence interval $\text{ci}_n = \{\mu_q \colon |F_n(\mu_q) - q| \le z_\alpha \{q(1-q)\}^{1/2}/\sqrt{n}\}$, where $P(|N(0,1)| \le z_\alpha) = \alpha$, the confidence leve.

(d) Discuss ways in which more accurate confidence intervals can be constructed, using the exact binomial distribution; for the median, for example, $nF_n(\mu) \sim \text{binom}(n, \frac{1}{2})$. This leads to slight modification of the horizontal bands when reading off intervals from the empirical c.d.f.

(e) Consider the dataset for $n = 189$ mothers and newborns, along with covariates, given in (xx point here; list these data too in the overview; give reference from Hosmer and Lemeshow (1999) xx). Here we look at the weights of the mothers, pre pregnancy, and ask about 90 percent confidence intervals for the 0.10 and 0.90 deciles. Construct a version of Figure 3.2. Use the method also to construct an interval for the median. (xx answers: 43.10 to 46.20 for 0.10; 74.85 to 84.36 for 0.90; 54.44 to 56.69 for median. point to analogous strategy for Kaplan–Meier quantiles, in Ch10, via band for the t^* at which $A(t^*) = \log 2$. also: point to Story with direct CD and cc methods. xx)

Ex. 3.10 *t testing, one and two samples.* (xx to be well cleaned, and connected to others exercises. xx) 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. 3.8 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. $\mathrm{N}(\xi, \sigma^2)$ and that one wishes to test $H_0 : \xi = \xi_0$ 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 $Z = (\bar{X} - \xi_0)/(\sigma/\sqrt{n}) = \sqrt{n}(\bar{X} - \xi_0)$ is standard normal under H_0, and that $|Z|$ will tend to be bigger than a normal if H_0 is not true. Asssuming to begin with that σ is known, demonstrate 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. 1.34 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. 3.11 *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. The Wald tests can be constructed almost immediately, from the confidence interval construction of Ex. 3.4, via the interval-to-test connection of Ex. 3.8, 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. 3.4 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).

(b) (xx give a pointer perhaps to a story xx). Explain that $p = P(|N(0,1)| \geq W_{n,0,\mathrm{obs}})$ 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 $x_{\mathrm{obs}}=44$.

(d) Suppose 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. 3.12 *The Neyman–Pearson Lemma.* (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 $P_{f_0}(\mathrm{reject}) = \int f_0 T \,\mathrm{d}y$.

(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

$$P_{f_0}(T^* \text{ rejects}) = \int_{y\colon\, f_1(y)/f_0(y)\geq c} f_0(y)\,\mathrm{d}y = \alpha.$$

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. 3.13 *The Neyman–Pearson Lemma: more details.* Here we tend to some further details, related to the Neyman–Pearson Lemma and its proof, in Ex. 3.12.

(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 $p(c) = P_{f_0}(A_c) = \int_{f_1(y)\geq cf_0(y)} f_0(y)\,\mathrm{d}y$ 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 $p(c) = P_{f_0}(A_c)$ 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. 3.14 *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.

(h) xx

(i) xx

Ex. 3.15 *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.

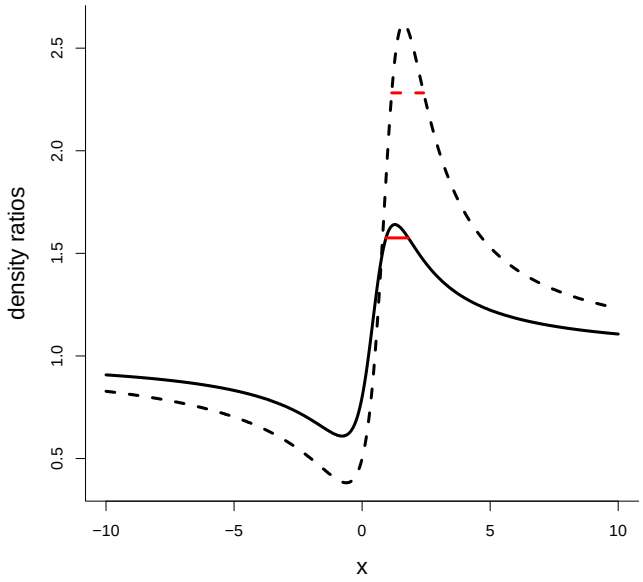

Figure 3.3: *For the Cauchy density model, ratios* $f(y,\theta_1)/f(y,0)$ *(full line) and* $f(y,\theta_2)/f(y,0)$ *(slanted) are shown, for alternative values* $\theta_1 = 0.50$ *and* $\theta_2 = 1.00$ *to the null. Also indicated are the rejection intervals* $[a_1,b_1]$ *and* $[a_2,b_2]$*, for the optimal tests against* θ_1 *and* θ_2*.*

(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 3.3. 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. 3.16 *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. 3.15, 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 of $\bar f(y)/f_0(y) \geq c$, 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. 3.17 *Do the data come from f_0, or rather from f_1?* (xx perhaps a story, not merely an exercise. we use simple tools, CLT and LLN, not much more. perhaps with f_0 the standard normal and f_1 the scaled double expo with variance 1. xx) Suppose i.i.d. data $Y_1, Y_2, \ldots$ are observed, these coming from either f_0 or f_1, two specified densities. We may use Neyman–Pearson to test f_0 against f_1, and also vice versa. We assume that the two variances

$$\tau_{0,1}^2 = \mathrm{Var}_{f_0} \log\{f_0(Y)/f_1(Y)\} = \int f_0\{\log(f_0/f_1) - d_{0,1}\}^2\,\mathrm{d}y,$$
$$\tau_{1,0}^2 = \mathrm{Var}_{f_1} \log\{f_1(Y)/f_0(Y)\} = \int f_1\{\log(f_1/f_0) - d_{1,0}\}^2\,\mathrm{d}y$$

are finite, involving the two Kullback–Leibler distances

$$d_{0,1} = \int f_0 \log(f_0/f_1)\,\mathrm{d}y \quad \text{and} \quad d_{1,0} = \int f_1 \log(f_1/f_0)\,\mathrm{d}y,$$

see Ex. 5.17.

(a) Let $R_n = \sum_{i=1}^n \log f_1(Y_i)/f_0(Y_i)$. Show that

$$R_n/n \to_{\mathrm{pr}} \begin{cases} -d_{01} = -d(f_0,f_1) & \text{if data are from } f_0, \\ d_{10} = d(f_1,f_0) & \text{if data are from } f_1. \end{cases}$$

So a plot of R_n, as a function of growing sample size, will end up positive under f_1 but negative under f_0.

(b) Show that the optimal test for f_0 against f_1, with level say 0.05, is to reject when R_n is big, i.e. bigger than the 0.95 point of the R_n distribution under f_0. One may find this threshold via simulation, for each given n, but show that the first-order approximation is that of rejecting when $Z_{n,0,1} = \sqrt{n}(R_n/n + d_{0,1})/\tau_{0,1} \geq 1.645$.

(c) Show that the detection power for this test, at f_1, with n observations, can be approximated as

$$\begin{aligned}\pi_{n,0,1} = P_{f_1}(\text{reject } f_0) &\doteq P\{(\tau_{1,0}/\tau_{0,1})N + \sqrt{n}d_{1,0}/\tau_{0,1} \geq 1.645\} \\ &= \Phi((\tau_{0,1}/\tau_{1,0})(\sqrt{n}d_{1,0}/\tau_{0,1} - 1.645)),\end{aligned}$$

writing N for a standard normal.

(d) (xx similar for f_1 against f_0. do all of this in two cases, using simulation to arrive at rejection thresholds in the second case. (i) $f_0 \sim \text{Expo}(1.111)$ and $f_1 = \text{Expo}(2.222)$; (ii) f_0 the standard normal and $f_1(x) = \frac{1}{2}\sqrt{2}\exp(-\sqrt{2}|x|)$. make power plots, as a function of n. comment. xx)

Ex. 3.18 *Non-monotone likelihood ratio.* Consider a simple setup with $Y \sim \text{N}(\theta, \theta^2/m)$, for a known m. (xx point to this being approximate situation in several setups. xx) Suppose one needs to test $\theta = \theta_0$, a known value, against the alternative $\theta > \theta_0$.

(a) For some $\theta_1 > \theta_0$, show that the log-likelihood-ratio can be written

$$R(y, \theta_0, \theta_1) = \log f(y, \theta_1) - \log f(y, \theta_0) = \tfrac{1}{2}max^2 - mby - \log(\theta_1/\theta_0),$$

with $a = 1/\theta_0^2 - 1/\theta_1^2$ and $b = 1/\theta_0 - 1/\theta_1$. Simplify this further to

$$R(y, \theta_0, \theta_1) = \tfrac{1}{2}ma\Big(y - \tfrac{1}{1/\theta_0 + 1/\theta_1}\Big)^2 + \text{const.},$$

(b) Set up the Neyman–Pearson optimal test, at level say 0.05, at this alternative point θ_1. Find the associated optimal power, at θ_1. Compare this power with the simpler test which rejects if $m^{1/2}(y - \theta_0)/\theta_0 > 1.645$.

(c) (xx a few details. so there is no UMP, for $Y \sim \text{N}(\theta, \theta^2/m)$, but there *is* a simple UMP for $Y \sim \text{N}(\theta, \theta/m)$. xx)

(d) (xx a bit more, worth doing, in detail. find the ML estimator $\widehat{\theta}$. explore the Wilks test, $W_m = 2\{\ell(\widehat{\theta}) - \ell(\theta_0)\}$, perhaps calibrated to have null distribution closer to χ^2_1. find its power function. compare with NP optimal in a few positions. xx)

(e) xx

Ex. 3.19 *The t test and its power.* (xx repair and polish and simplify here. xx) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the $\text{N}(\xi, \sigma^2)$, with testing of $\xi = 0$ required against $\xi > 0$. This is simple and standard, in the case of σ being known, but requires the t test, as we have seen in Ex. 3.10, in the case of σ being unknown and estimated from the data. Setting up the test requires 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.

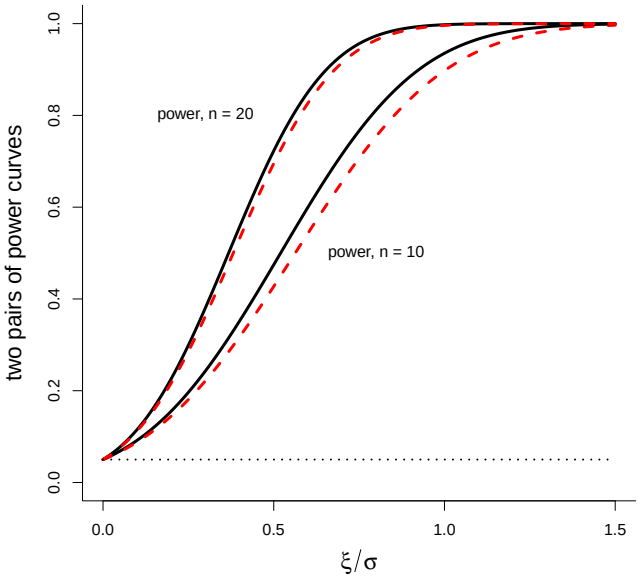

Figure 3.4: *Two pairs of power: Testing $\xi = 0$ against $\xi > 0$, for a normal sample of size 10 (lower pair) and 20 (upper pair), at test level 0.05, as a function of ξ/σ. The lower power is for the t test (slanted, red); the upper power if for the normal based test (full, black), which requires that σ is known.*

(a) 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 3.4.

(c) Then consider the more complex situation where σ is not known, needing the empirical standard deviation $\widehat{\sigma}$ of (1.3). We have seen in Ex. 1.34 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) (xx check things, also where we have said what. xx) For the power of the t test, show that $\pi_{n,t}(\xi/\sigma) = P_{\xi/\sigma}(t \geq t_0)$, 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. Show that the power function can be written

$$\pi_{n,t}(\xi/\sigma) = P_{\xi/\sigma}(t \geq t_{0.m}) = 1 - G_m(t_{0,m}, \sqrt{n}\xi/\sigma),$$

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 3.4, 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 3.4, with the relevant two-sided power functions.

Ex. 3.20 *Power and local power, I.* This exercise is meant to study a 'prototype situation' in some detail; the type of calculations and results will be seen to 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

$$\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i \quad \text{and} \quad M_n = \text{median}(Y_1, \ldots, Y_n).$$

[xx Figure 1: Limiting local power functions for two tests for $\theta \leq \theta_0$ against $\theta > \theta_0$, in the situation with $\mathrm{N}(\theta, \sigma^2)$ data. based on the mean (full line) and on the median (dotted line). can do both exact and local approximation. xx]

(a) For a given value of θ, prove that

$$\sqrt{n}(\bar{Y}_n - \theta) \to_d \mathrm{N}(0, \sigma^2), \quad \sqrt{n}(M_n - \theta) \to_d \mathrm{N}(0, (\pi/2)\sigma^2).$$

Note that the first result is immediate and actually holds with exactness for each n; the second result requires more care, e.g. working with the required density, cf. Exercise xx.

(b) Working under the null hypothesis $\theta = \theta_0$, show that

$$Z_n = \sqrt{n}(\bar{Y}_n - \theta_0)/\widehat{\sigma} \to_d \mathrm{N}(0, 1),$$
$$Z_n^* = \sqrt{n}(M_n - \theta_0)/\{(\pi/2)^{1/2}\widehat{\sigma}\} \to_d \mathrm{N}(0, 1),$$

where $\widehat{\sigma}$ is any consistent estimator of σ.

(c) Conclude from this that the two tests that reject H_0 provided respectively

$$\bar{X}_n > \theta_0 + z_{0.95}\widehat{\sigma}/\sqrt{n} \quad \text{and} \quad M_n > \theta_0 + z_{0.95}(\pi/2)^{1/2}\widehat{\sigma}/\sqrt{n},$$

where $z_{0.95} = \Phi^{-1}(0.95) = 1.645$, have the required asymptotic significance level 0.05;

$$\alpha_n = P\{\text{reject } H_0 \,|\, \theta = \theta_0\} \to 0.05.$$

(There is one such α_n for the first test, and one for the other; both converge however to 0.05.)

(d) Then our object is to study the *local power*, the chance of rejecting the null hypothesis under alternatives of the type $\theta_n = \theta_0 + \delta/\sqrt{n}$. In generalisation of (b), show that

$$\begin{aligned} Z_n &= \sqrt{n}(\bar{Y}_n - \sigma_0)/\widehat{\sigma} \to_d \mathrm{N}(\delta/\sigma, 1), \\ Z_n^* &= \sqrt{n}(M_n - \theta_0)/\{(\pi/2)^{1/2}\widehat{\sigma}\} \to_d \mathrm{N}((\pi/2)^{1/2}\delta/\sigma, 1), \end{aligned}$$

[xx check this xx] where the convergence in question takes place under the indicated $\theta_0 + \delta/\sqrt{n}$ parameter values. (You need to generalise the results of Exercise xx, to the $\delta \neq 0$ case.)

(e) Use these results to show that

$$\begin{aligned} \pi_n(\delta) &= P\{\text{reject} \,|\, \theta_0 + \delta/\sqrt{n}\} \to \Phi(\delta/\sigma - z_{0.95}), \\ \pi_n^*(\delta) &= P\{\text{reject} \,|\, \theta_0 + \delta/\sqrt{n}\} \to \Phi((2/\pi)^{1/2}\delta/\sigma - z_{0.95}), \end{aligned}$$

for the two power functions. Draw these in a diagram, and compare; cf. Figure xx.

(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 $c = \sqrt{2/\pi}$. Compute these sample sizes for the case of $\beta = 0.95$ and $\theta_1 = \theta_0 + \frac{1}{2}\sigma$, when also $\alpha = 0.05$.

(g) One sometimes defines the ARE, the asymptotic relative efficiency of test B with respect to test A, as

$$\mathrm{ARE} = \lim \frac{n_A(\theta_1, \beta)}{n_B(\theta_1, \beta)},$$

the limit in question in the sense of alternatives θ_1 coming closer to the null hypothesis at speed $1/\sqrt{n}$. Show that indeed

$$\mathrm{ARE} = \frac{\sigma^2}{\sigma^2/c^2} = c^2 = 2/\pi = 0.6366$$

in this particular situation – test A needs only ca. 64 percent as many data points to reach the same detection power as B needs.

(h) (xx round off. pointers. xx)

Ex. 3.21 *Power and local power, II.* (xx we do also parallel exercise with two tests for σ, based on the usual χ^2_{n-1}, and on $B_n = (1/n)\sum_{i=1}^n |X_i - M_m|$, with M_n the sample median. xx)

Ex. 3.22 *Two denominators for Wald tests.* (xx this needs work and editing, and might be split in two parts. i need a better prototype example for the 'two denominators' thing than what's here as of 4 april 2021. xx) (xx for $W_n = \widehat{\beta}/D_n$ there are often two choices for the denominator D_n, constructed to be consistent at H_0, or consistent in the wider model. same $\mathrm{N}(0,1)$ limit under H_0, but different powers. need to point to and calibrate wring to monotone likelihood ratio property. we need a good example where this property does not hold. we make a Cauchy exercise too. xx) Consider the simple setup where $\widehat{\theta} \sim \mathrm{N}(\theta, \theta/m)$, with m known, perhaps the sample size behind the estimator $\widehat{\theta}$. We also suppose θ is sufficiently positive so that $\widehat{\theta}$ is also positive with very high probability. One wishes to test the null hypothesis that $\theta = \theta_0$, a known null value, against the alternative that $\theta > \theta_0$.

(a) For a value $\theta_1 > \theta_0$, consider the log-ratio function

$$\begin{aligned} R(y, \theta_1, \theta_0) &= \log\{f(y,\theta_1)/f(y,\theta_0)\} \\ &= -\tfrac{1}{2}m(y-\theta_1)^2/\theta_1 + \tfrac{1}{2}m(y-\theta_0)^2/\theta_0 - \log\theta_1 + \log\theta_0. \end{aligned}$$

Show that R is increasing in x. (xx which implies that $W_{n,0}$ is optimal, on this occasion. need another example to showcase different power behaviour. xx)

(b) Consider the two test ratios

$$W_{m,0} = \frac{\widehat{\theta} - \theta_0}{(\theta_0/m)^{1/2}} \quad \text{and} \quad W_{m,1} = \frac{\widehat{\theta} - \theta_0}{(\widehat{\theta}/m)^{1/2}}.$$

Show that their null distributions both tend to the standard normal, as m grows; in fact $W_{m,0}$ is exactly a $\mathrm{N}(0,1)$ under the null. With a 0.05 level, the tests reject the null if $W_{m,0} \geq z_0$ or $W_{m,1} \geq z_0$, with $z_0 = 1.645$. (xx can ask just a bit more here, a more exact version of the second test, needing then distribution of $(\widehat{\theta}/\theta_0)^{1/2}$. xx)

(c) First consider the local power, at the local alternatives $\theta = \theta_0 + a/m^{1/2}$. Show that both $W_{m,0}$ and $W_{m,1}$ tend to $\mathrm{N}(a/\theta_0, 1)$, and hence that both power functions tend to $\Phi(a/\theta_0 - z_0)$.

(d) Then consider an alternative value, with a non-zero $\delta = \theta - \theta_0$. Show

$$\begin{aligned} W_{m,0} &= \frac{m^{1/2}\{\delta + (\theta/m)^{1/2}Z_m\}}{\theta_0^{1/2}} = (\theta/\theta_0)^{1/2}Z_m + m^{1/2}\delta/\theta_0^{1/2}, \\ W_{m,1} &= \frac{m^{1/2}\{\delta + (\theta/m)^{1/2}Z_m\}}{\widehat{\theta}^{1/2}} = (\theta/\widehat{\theta})^{1/2}Z_m + m^{1/2}\delta/\widehat{\theta}^{1/2}. \end{aligned}$$

where $Z_m \sim \mathrm{N}(0,1)$. Argue that this leads to power functions

$$\begin{aligned} \pi_{m,0}(\theta) &= P\{(\theta/\theta_0)^{1/2}N + m^{1/2}(\theta-\theta_0)/\theta_0^{1/2} \geq z_0\}, \\ \pi_{m,1}(\theta) &= P\{N + m^{1/2}(\theta-\theta_0)/\theta^{1/2} \geq z_0\}, \end{aligned}$$

with N a standard normal; the first power expression is exact, the second an approximation.

(e) xx

(f) xx

Ex. 3.23 *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 $P_\theta\{h(T) = 0\} = 1$ 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. 1.15, 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 M is sufficient and complete.

(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) If T is complete, show that any one-to-one transformation variable $T' = a(T)$ is also complete.

(e) 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. 3.24 *Completeness for the exponential family.* For the large class of exponential family models, see Ex. 1.57 and follow-up exercises, there is a completeness lemma, as follows. Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the model $f(y, \theta) = \exp\{\theta^{\mathrm{t}} T(y) - k(\theta)\} h(y)$, with $T = (T_1, \ldots, T_p)^{\mathrm{t}}$ and $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{t}}$ varying in an open set, then the vector of sample averages $(\bar{T}_1, \ldots, \bar{T}_p)^{\mathrm{t}}$ is not merely sufficient, as seen in Ex. 1.64, 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 $a(\theta) = \mathrm{E}_\theta\, h(\bar{T})$ 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 $\mathrm{N}(\xi, 1)$. 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 $\mathrm{Gam}(a, b)$ 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 $\mathrm{Beta}(a, b)$.

(c) Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from the $\mathrm{N}(\xi, \sigma^2)$. 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 $\mathrm{N}(\theta, \theta^2)$. 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.

Ex. 3.25 *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

$$P_\theta\{U(Y) \in R(v) \,|\, V(Y) = v\} \leq \alpha \quad \text{for all } \theta \in \Theta_0.$$

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. 3.23. 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 $\theta \in \bar{\Theta}_0$. (xx hint: check the function $h(v) = \mathrm{E}_\theta\{T(U, V) \,|\, V = v\} - \alpha$. xx)

Ex. 3.26 *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

$$g_\delta(y \,|\, z) = \frac{\delta \exp(-\delta y)}{\int_0^z \delta \exp(-\delta y')\,\mathrm{d}y'} = \frac{\delta \exp(-\delta y)}{1 - \exp(-\delta z)} \quad \text{for } 0 \leq y \leq z.$$

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. 3.12), 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. 3.26.

(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. 3.27 *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. 1.30, 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. 3.28 *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

$$f = \exp\Big[-\sum_{i=1}^m (\theta_i + \theta_i\gamma) + \sum_{i=1}^m \{(x_i + y_i)\log\theta_i + y_i \log\gamma\}\Big].$$

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. 3.29 *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. 1.52.

(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.9. xx)

Ex. 3.30 *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. 3.3 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

$$\begin{aligned} f &= \frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp\Big\{-\tfrac12\frac{1}{\sigma^2}\sum_{i=1}^n(y_i-\xi)^2\Big\} \\ &= \frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp\Big\{\frac{\xi}{\sigma^2}\sqrt{n}U - \tfrac12\frac{1}{\sigma^2}V - \tfrac12\frac{\xi^2}{\sigma^2}\Big\}. \end{aligned}$$

Note that the testing problem is equivalent to testing $\lambda = 0$ against $\lambda 2 > 0$, with $\lambda = \xi/\sigma^2$, the mathematics indicating that this is a parameter easier to work with than ξ.

(b) Find the distribution of $U \mid (V = v)$, 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 $P_0\{U \geq c(v) \,|\, V = v\} = 0.05$. Then show that this is actually the same as the t test.

(d) (xx rounding off. xx)

Ex. 3.31 *Conditional tests: multiparameter exponential models.* In the rather simple situation of Ex. 3.26, 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. 1.60, 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 $P_{a_0}(U > c(v) \,|\, V = v) = \alpha$, and with consequent power function $\pi(a, b) = P_{a,b}\{U > c(V)\}$.

(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 $P_{a_0}(U > c(v) \,|\, V = v) = \alpha$ 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. $P_{a_0}\{U > c(v) \,|\, V = v\} = 0.05$. 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 $P_{a_0}\{U > c(v) \,|\, V = v\}$ is just below 0.05, and then identify the probability r such that

$$P_{a_0}\{U > c(v) \,|\, V = v\} + r\, P_{a_0}\{U = c(v) \,|\, V = v\} = 0.05,$$

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. 3.32 *Testing correlation.* (xx first 3-parameter model with $(\sigma_1, \sigma_2, \rho)$, then full 5-parameter binormal model. testing $\rho = 0$ against $\rho > 0$. using the theory. xx)

Ex. 3.33 *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 tool, or a bag of related tools, for investigating the extent to which certain 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 $x_i = (x_{i,1}, \ldots, x_{i,p})^{\rm t}$ is of dimension p and y_i of dimension one. The model then postulates that

$$Y_i = x_i^{\rm t}\beta + \varepsilon_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

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 $\beta = (\beta_1, \ldots, \beta_p)^{\rm t}$ 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. 2.34.

(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 $X^{\rm t}X$ 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 $Q(\beta) = \|Y - X\beta\|^2 = \sum_{i=1}^n (Y_i - x_i^{\rm t}\beta)^2$. Show that $\sum_{i=1}^n (Y_i - x_i^{\rm t}\widehat{\beta})x_i = 0$. With Σ_n the $p \times p$ matrix $n^{-1}\sum_{i=1}^n x_i x_i^{\rm t} = n^{-1}X^{\rm t}X$, show that

$$\widehat{\beta} = (X^{\rm t}X)^{-1}X^{\rm t}Y = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i Y_i.$$

(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. 3.34.

Ex. 3.34 *The residuals and their variance.* The setup is an in the previous Ex. 3.33, 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. 1.22. 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. 3.35 *Inference for linear multiple regression.* [xx to be done. perhaps with a real data example, or a pointer to a story. 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 $P(Y \le y_0\,|\,x_0)$. 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 $P(Y \le y_0\,|\,x_0)$. xx)

Ex. 3.36 *How much of the variance is explained?* 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? (xx the below to be polished and illustrated, and with a clearer link to R^2. exact cc(ρ) to come in Ch. 7. xx)

(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. 1.37. 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. 1.39, 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.

(f) (xx round off. data example to see how it works. testing $\beta = 0$ with a clear F test, but rather more; a full cc(ρ), with a point mass at zero, etc. new estimator for ρ is the median confidence estimator $C^{-1}(\frac{1}{2})$. xx)

Ex. 3.37 *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. 1.38 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 $P(a \le F \le b) = 0.95$, 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. 3.6, 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 Stories ii.9–**??**.

(d) xx

Ex. 3.38 *Testing equality of multinormal means.* (xx a brief thing, used in Story ii.8, 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.

3.C 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)

ToDo notes, as of 13-Aug-2023:

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.4

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, already encountered in Ch. 2, 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. 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 famous Lindeberg theorem, giving precise conditions under which a sum of independent components approaches normality. This is then used for handling classes of estimators in regression models. The theory is also used to establish clear limiting normality and related results for classes of minimum criterion function estimators. Methods and results from this chapter are crucial for developing the likelihood theory of Ch. 5, and also for several later chapters.

4.A Chapter introduction

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.

Recall that a random element X is a function defined on a probability space $(\Omega, \mathcal{F}, P)$, and taking its values in some space $\mathcal{X}$, equipped with an appropriate σ-algebra. When $\mathcal{X} \subseteq \mathbb{R}$ we call X a random variable; when $\mathcal{X} \subseteq \mathbb{R}^k$ for some $k \geq 2$, we call it a random vector; and when $\mathcal{X}$ is the space of all continuous functions on the unit interval, for

example, we call X a random or stochastic process. In this chapter we work with convergence of random variables and vectors, with the more involved themes of convergence of stochastic processes studied in Ch. 9.

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. We've seen aspects of this already, in Chs. 2, 3, but in this chapter we dive into more details and learn more. The key convergence concepts, with ensuing approximations and applications, are as follows.

First, if X_n and X are random variables in $\mathbb{R}^k$, defined on the same probability space, we say that X_n *converges to* X *in probability*, written $X_n \to_{\rm pr} X$, if

$$P(\|X_n - X\| \geq \varepsilon) \to 0 \quad \text{for each positive } \varepsilon. \tag{4.1}$$

Here $\|a-b\|$ is simple Euclidean distance, and hence ordinary distance in the one-dimensional case. Typically the limit X is simply a constant. If X_n is an estimator $\widehat{\theta}_n$, 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* X *almost surely*, or with probability one. This means that

$$N = \{\omega \in \Omega\colon \lim X_n(\omega) \neq X(\omega)\} \quad \text{has probability zero}, \tag{4.2}$$

and we write $X_n \to_{\rm a.s.} X$. It is seen to be the same as $P(\limsup \|X_n - X\| \geq \varepsilon) = 0$ for each positive ε. Again, often limits discussed using this concept are constants, and we say that an estimator $\widehat{\theta}_n$ is strongly consistent for θ if $\widehat{\theta}_n \to_{\rm a.s.} \theta_0$, the true value. A strong achievement indeed is the *strong Law of Large Numbers* (LLN), which says that the LLN

$$\text{if } X_1, X_2, \ldots \text{ are i.i.d. with finite mean } \xi, \text{ then } \bar{X}_n = n^{-1}\sum_{i=1}^n X_i \to_{\rm a.s.} \xi, \tag{4.3}$$

with no further assumptions required. One may readily prove weaker versions of the LLN, as in Ex. 4.1, but with the development of sharper tools, and separate valuable results along the way, we reach the strong LLN in Ex. 4.34–4.35. It immediately has many applications and uses, as we shall see.

The third and statistically speaking most important concept is that of *convergence in distribution*, already touched on in Chs. 2-3. If one-dimensional X_n and X 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

$$P(X_n \in (a,b]) \to P(X \in (a,b]) \quad \text{for all continuity points } a \text{ and } b \text{ of } F. \tag{4.4}$$

This is also the same as $F_n(b) - F_n(a) \to F(b) - F(a)$, for all such intervals. We write $X_n \to_d X$ or $F_n \to_d F$ to indicate this, allowing for simplicity also statements like $X_n \to_d {\rm N}(0,1)$. The concept generalises to variables in $\mathbb{R}^k$, where we need $P(X_n \in R) \to P(X \in R)$, for all rectangles R for which $P(X \in \partial R) = 0$, where ∂R is the

boundary of R. The bigger sibling of the LLN is the CLT, which we had occasion to see in action already in Chs. 2–3; the CLT

$$\text{if } X_1, X_2, \dots \text{ are i.i.d. with variance } \sigma^2, \text{ then } \sqrt{n}(\bar{X}_n - \mathrm{E}\, X_1) \to_d \mathrm{N}(0, \sigma^2), \quad (4.5)$$

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 e.g. used to establish approximate normality for estimators in regression models.

We also use the theory to reach limiting normality and related results for classes of minimum criterion function estimators, in the set of exercises Ex. 4.47–4.50. These also pave the way to the more central results for maximum likelihood methods in Ch. 5.

(xx a paragraph with more comments and pointers to later chapters and a few stories. studies of robustness. multivariate CLT and Lindeberg. Markov chains, Bayes and CD approximations, processes, martingales for survival analysis. xx)

4.B Short and crisp

Ex. 4.1 *Convergence in probability.* When working with (4.1) and related quantities it is clear that tail probability bounds of the Markov and Chebyshov type are useful here, as seen already in Ex. 2.7, with more to come in Ex. 4.33.

(a) Let $Y_1, Y_2, \dots$ be i.i.d. random variables with expectation θ and finite variance σ^2. Prove the *Law of of Large Numbers* (in its weak form, we'll get to the to strong version in Ex. 4.34–4.35), namely that $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i \to_{\mathrm{pr}} \theta$ as $n \to \infty$. More generally, 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 $w_i = 1/i^{1.5}$?

(b) Suppose $X_n \to_{\mathrm{pr}} a$, with a being a constant. Show that if g is a function continuous at point $x = a$, then indeed $g(X_n) \to_{\mathrm{pr}} g(a)$.

(c) Suppose more generally that $X_n \to_{\mathrm{pr}} X$, with the limit being a random variable. Show that if g is a function that is continuous in the set in which X falls, then $g(X_n) \to_{\mathrm{pr}} g(X)$.

Ex. 4.2 *Convergence in probability for vectors.* Consider $X_n = (X_{n,1}, X_{n,2})$ and $X = (X_1, X_2)$.

(a) Show that $(X_{n,1}, X_{n,2}) \to_{\mathrm{pr}} (X_1, X_2)$, by the definition of (4.1), is equivalent to $X_{n,1} \to_{\mathrm{pr}} X_1$ and $X_{n,2} \to_{\mathrm{pr}} X_2$, that is, ordinary one-dimensional convergence for each component. Generalise.

(b) Suppose $X_n \to_{\mathrm{pr}} a$, in dimension k, with $a = (a_1, \dots, a_k)$ a constant. If $g(x) = g(x_1, \dots, x_k)$ is a function defined in at least a neighbourhood around a, and continuous at that point, show that $g(X_n) \to_{\mathrm{pr}} g(a)$.

Ex. 4.3 *Quantiles.* Suppose $Y_1, \ldots, Y_n$ are i.i.d. from a distribution with continuous and strictly increasing c.d.f. F.

(a) For a $q \in (0,1)$, let $Q_n = Q_n(q)$ be the empirical q-quantile, defined here to be $Y_{([nq])}$, the $[nq]$ order statistic. 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)

(b) 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. 4.4 *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. 4.5 *Convergence in distribution.* Convergence $X_n \to_d X$, or equivalently $F_n \to_d F$, has been defined in (4.4).

(a) Show that the (4.4) statement is equivalent to convergence $F_n(x) \to F(x)$ for each x at which F is continuous. Show that the set D_F of discontinuity points for F is at most countable.

(b) For dimension two, we say that $X_n \to X$ if $P(X_n \in R) \to P(X \in R)$ for each rectangle $R = (a_1, b_1] \times (a_2, b_2]$ for which the $P(X \in \partial R) = 0$, where ∂R is the boundary of the rectangle. Express $P(X \in R)$ via F. Show that $F_n \to_d F$ if and only if $F_n(x_1, x_2) \to F(x_1, x_2)$ at each point where F is continuous. Generalise to dimension k.

Ex. 4.6 *Convergence in distribution for discrete variables.* Assume X_n and X take values in $\{0, 1, \ldots\}$, with probabilities $p_n(j)$ and $p(j)$, for $j = 0, 1, \ldots$.

(a) Show that $X_n \to_d X$ if and only if there is corresponding convergence for the point probabilities, i.e. $p_n(j) \to p(j)$ for all j.

(b) With $X_n \sim \text{binom}(n, p_n)$, and $np_n \to \lambda$, show that $X_n \to_d \text{Pois}(\lambda)$.

(c) With $X_n \sim \text{binom}(n, p)$, let $Z_n = I\{X_n \geq np\}$. Show that $Z_n \to_d Z \sim \text{binom}(1, \frac{1}{2})$.

Ex. 4.7 *Many small probabilities give a Poisson.* (xx previously in Ch2, now moved here to Ch4. some editing and crossrefing needed. xx) 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. [xx point to von Bortkiewicz (1898), and to Poisson himself. xx]

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

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

(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 $Y_n(t) \to_d \text{Pois}(\frac{1}{2}t^2)$. The limit is actually a full Poisson process in t, with independent increments.

(d) Suppose (X_n, Y_n) has the trinomial distribution, with parameters (n, p, q), see Ex. 1.4. Assume now that p, q become small with n, such that $np \to \lambda_1$, $nq \to \lambda_2$. 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. 4.8 *From discrete to continuous.* Discrete distributions may have continuous limits. To understand the why we in (4.5) only require $F_n(x) \to F(x)$ to hold for continuity points of F, in each of the exercises below, find points x at which $F_n(x) \to F(x)$ fails.

(a) Let X_n have distribution $P(X_n = j/n) = 1/(n+1)$ for $j = 0, 1, \ldots, n$. Show that $X_n \to_d X$, where X has the uniform distribution on the unit interval. With perhaps similar techniques, consider X_n with distribution $P(X_n = j/n) = j/\{n(n+1)/2\}$ for $j = 1, \ldots, n$. Find its limit distribution.

(b) Suppose $X \sim \text{Pois}(\lambda)$ and that λ grows. What is the range of values for $Y_\lambda = (X - \lambda)/\lambda^{1/2}$? Show that $Y_\lambda \to_d \text{N}(0, 1)$.

(c) Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with success probability p. Only using (4.4), show that $\sqrt{n}(\bar{X}_n - p)\} \to_d X$, with X a $\text{N}(0, p(1-p))$ distribution. You may use Stirling's formula $n! \sim \sqrt{2\pi n}(n/e)^n$ here; see Ex. 4.31.

Ex. 4.9 *Maximum of uniforms.* Let $X_1, \ldots, X_n$ be i.i.d. from the uniform distribution on $[0, \theta]$, and let $M_n = \max_{i \le n} X_i$.

(a) Show that $M_n \to_{\text{pr}} \theta$, 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.

Ex. 4.10 *The Portmanteau theorem.* We have taken (4.4) as our definition of convergence in distribution. The theorem proved in the course of this exercise shows that there are several other possible definitions. A stochastic process, for example, (we'll meet such at several occasions in the coming chapters, see e.g., Ch. 9, Ch. 10, and Ch. 15) does not have a c.d.f., and consequently (4.4) can not be taken as defining convergence in distribution. On the other hand, the conditions of the Portmanteau theorem that are equivalent to (4.4) in the case of random variables and vectors, are much more general in the sense that they can be taken as definitions of convergence in distribution, then typically called weak convergence, for much more complicated random objects. Over the coming pages we also prove theorems that could be added to the list of the Portmanteau theorem, see e.g., Ex. 4.21**??** [xx and . . . xx]. Let's turn to the theorem.

Let X_n and X be random variables with distributions P_n and P, that is, $P_n(A) = \Pr\{X_n \in A\}$ and $P(A) = \Pr\{X \in A\}$ (see Ex. A.10 on page 584). In this exercise we will see that the following five statements are equivalent:

(1) $X_n \to_d X$ in the sense of (4.4);
(2) $\liminf P_n(A) \geq P(A)$ for every open set A;
(3) $\limsup P_n(B) \leq P(B)$ for every closed set B;
(4) $P_n(C) \to P(C)$ 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);
(5) $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$ for all bounded and continuous real valued functions g.

We prove this theorem through a string of subexercises.

(a) Show that (1) implies (2) by writing $A = \cup_{j=1}^{\infty} A_j$ for open sets $A_j = (a_j, b_j)$, and, for each j consider interals $(a_j', b_j'] \subset A_j$ where a_j', b_j' are among the continuity points for the disitribution function F of X.

(b) Show that (2) implies (3) using that B is closed if and only if B^c is open.

(c) Show that (3) implies (4), using that the boundary ∂C of any set C is $\partial C = \bar{C} \setminus C^\circ$ where the close $\bar{C}$ is always closed and the interior C° is always open.

(d) Show that (4) implies (5). Since g is bounded, say $a \leq g(x) \leq b$, we can consider the linear transformation $h(x) = (g(x) - a)/(b - a)$ that takes values in $[0, 1]$ and write $\mathrm{E}\, h(X_n) = \int \int_0^1 I\{y \leq h(x)\}\, \mathrm{d}y\, \mathrm{d}P_n(x) = \int_0^1 \Pr\{h(X_n) \geq y\}\, \mathrm{d}y$.

(e) Show that (5) implies (1). You may, for a continuity point x of F, construct continuous and bounded functions that approximate $I\{y \leq x\}$ from above and from below.

(f) Additional things to the Portmanteau: (6) $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$ for all continuous functions that are zero outside a closed and bounded interval. (7) [xx Infinitly smooth functions. See Billingsley 1968 p. 41. xx] (8) All uniformly continuous functions. (9) $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$ for all bounded Lipschitz functions g.

Ex. 4.11 *Tightness, Helly, and Prokhorov.* Let $(X_n)_{n\geq 1}$ be discrete random variables with distribution $\Pr(X_n = x) = \frac{1}{3}$ for $x = 0, \frac{1}{2}, n$, and denote their distribution functions $F_n(x) = \Pr(X_n \leq x)$. The sequence $F_n(x)$ converges pointwise the 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 right-continuous and takes value in $[0, 1]$, but it is not a distribution function: One third of the probability mass of $F_n(x)$ has escaped to inifinity! The condition preventing that probability mass runs away to infinity is called *tightness*. *A sequence $Y_1, Y_2, \ldots$ of random variables is tight if for any $\varepsilon > 0$ there exists a constant K so that $\Pr(|Y_n| > K) < \varepsilon$ for all n.* You may verify that the sequence X_n above is not tight. Tightness

(a) Show that (i) any random variable Y is tight; (ii) the sequence $(Y_n)_{n\geq 1}$ is tight if and only if for any $\varepsilon > 0$ there is a K so that $\limsup_{n\to\infty} \Pr\{|Y_n| > K\} < \varepsilon$; and (iii) 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.

(b) Show that $Y_1, Y_2, \ldots$ is tight if and only if for each $\varepsilon > 0$ there is an interval $(a, b]$ such that $\Pr(Y_n \in (a, b]) < \varepsilon$ for all n, and that, in terms of the distribution functions F_n, this is the same as there being points x and y such that $F_n(y) < \varepsilon$ and $F_n(x) > 1 - \varepsilon$ for all n.

Helly's theorem

(c) Let F_n be a sequence of distribution functions on the real line, or a subset thereof. Consider the infinite array

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

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 $k \to \infty$. 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 $j \to \infty$. Continue like this and argue that the diagonal sequence $n_k = n_{k,k}$ of the array

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

are such that $F_{n_k}(q_j) \to G(q_j)$ for $j = 1, 2 \ldots$ as $k \to \infty$. 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 distribution function, the exception being that it might tend to a limit smaller than one when x tends to infinity, greater than zero when x tends to minus infinity, or both.

Prokhorov's theorem

(d) Let Y_n be a sequence with distribution functions F_n. That tightness ensures that a converging subsequence F_{n_k} of F_n tends to a bona fide distribution function is one part of Prokhorov's theorem. The other part states that if Y_n converges in distribution, then Y_n is tight. Use Helly's theorem to prove the first part, and Ex. 4.10(3) combined with the tightness of each individual X_n to prove the latter.

(e) Suppose that $(Y_n)_{n \geq 1}$ is tight, and that $\varphi_n(t) = \mathrm{E}\exp(itX_n)$ converges to some function $\beta(t)$ as $n \to \infty$. Use Prokhorov's theorem and Ex. 4.18(c) to show that $\beta(t)$ must then be characteristic function of some random variable, say Y, and consequently, $Y_n \to_d Y$ by Ex. 4.21(a).

Continuous mapping

Ex. 4.12 *The continuous mapping theorem.* Let $X_1, X_2, \ldots$ and X be k dimensional random vectors, and $h \colon \mathbb{R}^k \to \mathbb{R}^m$ a function that is continuous on a set $C \subset \mathbb{R}^m$ where X falls with probability one.

(a) Suppose that $X_n \to_d X$ and that $C = \mathbb{R}^m$. Show that $h(X_n) \to_d h(X)$.

(b) Suppose that $X_n \to_d X$ and that C is a proper subset of $\mathbb{R}^m$. Let F be a closed set. Show the inclusion $h^{-1}(F) \subset \overline{h^{-1}(F)} \subset h^{-1}(F) \cup C^c$. Now, combine this inclusion with $\{h(X_n) \in F\} = \{h(X_n) \in h^{-1}(F)\}$ and the Portmanteau theorem to show that $h(X_n) \to_d h(X)$.

(c) Suppose that $X_n \to_{\text{pr}} X$. Show that $h(X_n) \to_{\text{pr}} h(X)$. [xx rewrite this, and compare with Ex. 4.2(b) xx]

Ex. 4.13 *Modes of convergence their implications.* Here we look at some relationships between convergence almost surely, in probability, and in distribution.

(a) Almost sure convergence is defined in (4.2). Show that $X_n \to_{\text{a.s.}} X$ if and only if

$$P(\|X_k - X\| < \varepsilon \text{ for every } k \geq n) \to 1, \quad \text{as } n \to \infty.$$

(b) Show the following implications.

(i) If $X_n \to_{\text{a.s.}} X$ then $X_n \to_{\text{pr}} X$;
(iii) If $\mathrm{E}\,\|X_n - X\| \to 0$ then $X_n \to_{\text{pr}} X$;
(iii) If $X_n \to_{\text{pr}} X$ then $X_n \to_d X$.

(c) Find examples to show that the reverse implications in (i)–(iii) do not hold.

Ex. 4.14 *The Cramér–Slutsky rules.* The utility of the three results in (b), together known as the Cramér–Slutksy rules, will become abundantly clear as we progress.

(a) 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 $(X_n, Y_n) \to_d (X, a)$.

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

(c) Let $Y_1, \ldots, Y_n$ be i.i.d. random variables $\mathrm{E}\,Y_1 = \mu$ and $\mathrm{Var}(Y_1) = \sigma^2$. Let $\widehat{\sigma}_n^2 = 1(n-1)\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ with $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. Show that $\sqrt{n}(\bar{Y}_n - \mu)/\widehat{\sigma}_n \to_d \mathrm{N}(0,1)$

Ex. 4.15 *A few counterexamples.* [xx to what. Perhaps move and expand this exercise. Yes. Must be done xx]

(a) Make an example where $X_n \to_d X$ and $Y_n \to_d Y$, but $X_n + Y_n$ does not converge in distribution to $X + Y$.

(b) Make an example where X_n converges to 0 in probability, but $\mathrm{E}\,X_n$ does not converge to 0.

(c) Let Z be uniform$(0,1)$, and set $X_1 = 1, X_2 = I_{[0,1/2)}(Z), X_3 = I_{[1/2,1)}(Z), X_4 = I_{[0,1/4)}(Z), X_5 = I_{[1/4,1/2)}(Z), \ldots$, and so on. Find the probability limit of X_n. Does X_n converge almost surely to this limit?

Ex. 4.16 *Scheffé's lemma.* Suppose Y_n and Y have densities f_n and f.

(a) If $f_n(y) \to f(y)$ for each y, show that $\int |f_n - f|\,\mathrm{d}y \to 0$, i.e. there is L_1 convergence. Show from this that $Y_n \to_d Y$. This is called Scheffé's lemma. Scheffé's lemma

(b) For the t distribution, show that $t_\nu \to_d \mathrm{N}(0,1)$ as ν increases. Compute actually $\int |f_\nu - \phi|\,\mathrm{d}y$, numerically, for a range of ν values. Detect in this way that degrees of freedom equal to 2.299 is halfway between the Cauchy, i.e. $\nu = 1$, to normal, i.e. $\nu = \infty$. This piece of statistical trivia has a certain statistical relevance in Hjort (1994). – Show also that $t_\nu \to_d \mathrm{N}(0,1)$ in a perhaps simpler way, using the Cramér–Slutsky rules.

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

(d) 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 $F_n(y) = y + (1/n)\sin(n\pi y)$. 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.

Ex. 4.17 *Moment generating functions.* [xx calibrate with Ex. 1.20 and 1.21 in Ch. 2; perhaps we place all the mgf basics there. xx] The moment generating function of a random variable X is defined as $M(t) = \mathrm{E}\, e^{tX}$. In this exercise we will see that if a random variable X has a moment generating function that is finite in some interval around zero, then all its moments are finite, that is, $\mathrm{E}\,|X|^p < \infty$ for $p = 0, 1, 2, \ldots$.

(a) Suppose that there are numbers $s < 0 < t$ such that $M(s)$ and $M(t)$ are both finite. Use the convexity of the exponential function to show that for any $t_0 \in [s,t]$, $M(t_0)$ is also finite.

(b) Let s, t be as in (a), and define $t_0 = \min(-s, t)$. Study the sum $M(-t_0) + M(t_0)$ and conclude that $\mathrm{E}\,|X|^{2k} < \infty$ for $k = 0, 1, 2, \ldots$. Finally, with an application of Jensen's inequality (for concave functions this time, see Ex. 1.19(c)), conclude that $\mathrm{E}\,|X|^{2k+1} < \infty$ for $k = 0, 1, 2, \ldots$ as well.

(c) Consider, for example, the log-normal distribution to conclude that a random variable having finite moments of all orders, does not imply that the moment generating function of that random variable is finite in some interval around zero.

Ex. 4.18 *Characteristic functions.* In some cases it is possible to show convergence in distribution directly from the definition in (4.4), or directly from one of the characterisations given in the Portmanteau theorem. Using the latter is often quite hard, as all the Portmanteau statements involve showing something *for all* functions or sets of a certain type. The utility of characteristic functions derives from the fact that instead of demonstrating $\mathrm{E}\, g(X_n) \to \mathrm{E}\, g(X)$ for all bounded and continuous functions g, we only need to show it for one function g. 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 $t \in \mathbb{R}$.

(a) Show that the characteristic function always exists, and that it 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) Assume that $X_n \to_d X$. Show that $\varphi_n(t) = \mathrm{E}\exp(itX_n) \to \varphi(t) = \mathrm{E}\exp(itX)$. The marvellous thing with characteristic functions is that the converse of this also holds, as we will see in Ex. 4.20 and 4.21.

Ex. 4.19 *Characteristic functions. Moments and derivatives* [xx introtext here xx]

(a) 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

$$|\exp(ix) - \sum_{k=0}^{m}\frac{(ix)^k}{k!}| \le \frac{2|x|^m}{m!} \wedge \frac{|x|^{m+1}}{(m+1)!}$$

for $m = 0, 1, 2, \ldots$.

(b) From the inequality in (a) we see that for all t such that $\lim_{m\to\infty}|t|^m\,\mathrm{E}\,|X|^m/m! = 0$, the characteristic function of X can be expressed as

$$\varphi(t) = \sum_{k=0}^{\infty}\frac{(it)^k}{k!}\,\mathrm{E}\,X^k. \tag{4.6}$$

Show that (4.6) holds if $\mathrm{E}\exp(|tX|) < \infty$, and that this latter inequality holds if the moment generating function $M(t) = \mathrm{E}\exp(tX) < \infty$ for all t.

(c) Provided (4.6) holds, the moments of X can be read off from $\varphi^{(k)}(0) = i^k\mathrm{E}\,X^k$. Show that $\{\varphi(t+h) - \varphi(t)\}/h - \mathrm{E}\{iX\exp(itX)\} \to 0$ as $h \to 0$, provided $\mathrm{E}\,|X| < \infty$. Proceed inductively to show that $\varphi^{(k)}(t) = i^k\mathrm{E}\{X^k\exp(itX)\}$ as long as $\mathrm{E}\,|X|^k < \infty$.

(d) [xx might include some from Nils stk4090 Ex. 59–60 pp. 39–41, or rewrite to be more in line with that exercise xx]

Ex. 4.20 *Uniqueness of characteristic functions.* The characteristic function $\varphi(t)$ of a random variable X uniquely determines its distribution. That is, if two random variables have identical characteristic functions, then their distributions are identical too. In this exercise, we work with real-valued random variables, but the results generalise to higher dimensions.

(a) Let X be any random variable, with cumulative distribution function $F_X(x)$ and characteristic function $\varphi_X(t)$, but nothing assumed about its density. 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

$$F_\sigma(u) = \int \Phi((u-x)/\sigma)\,\mathrm{d}F_X(x), \quad \text{and} \quad f_\sigma(u) = (1/\sigma)\int \phi((u-x)/\sigma)\,\mathrm{d}F_X(x),$$

where $\Phi(z) = \int_{-\infty}^{z} \phi(x)\,\mathrm{d}x$ is the standard normal distribution function, and $\phi(x)$ the standard normal density.

(b) Show that $\mathrm{E}\int \exp\{it(X-u) - \frac{1}{2}t^2\sigma^2\}\,\mathrm{d}t = 2\pi f_\sigma(u)$, which by Fubini's theorem, yields the following expression for the density of U_σ

$$f_\sigma(u) = \frac{1}{2\pi}\int \varphi_X(t)\exp(-itu - \tfrac{1}{2}t^2\sigma^2)\,\mathrm{d}t,$$

and, consequently, that for $a < b$,

$$\begin{aligned} P\{U_\sigma \in [a,b]\} &= F_\sigma(b) - F_\sigma(a) \\ &= \frac{1}{2\pi}\int \frac{\exp(-itb) - \exp(-ita)}{-it}\varphi_X(t)\exp(-\tfrac{1}{2}t^2\sigma^2)\,\mathrm{d}t. \end{aligned}$$

(c) Derive the general inversion formula, $F_X(b) - F_X(a) = \lim_{\sigma\to 0}\{F_\sigma(b) - F_\sigma(a)\}$, valid for all continuity points a, b $(a < b)$.

(d) Assume that X and Y are random variables with identical characteristic functions. Show that X and Y must be equal in distribution.

(e) If X has an integrable characteristic function φ_X, that is, if $\int|\varphi_X(t)|\,\mathrm{d}t < \infty$, show that then X has a continuous density f_X given by $f_X(x) = (1/2\pi)\int \exp(-itx)\varphi_X(t)\,\mathrm{d}t$.

Ex. 4.21 *Lévy's continuity theorem.* We have seen that characteristic functions are deserving of their name in that they uniquely determine the distribution of a random variable (Ex. 4.20(d)). In particular, if $Z_1, \ldots, Z_n$ are independent random variables, then $\varphi_n(t) = \mathrm{E}\exp(it\sum_{j=1}^n Z_j) = \prod_{j=1}^n \mathrm{E}\exp(itZ_j)$, and the product on the right uniquely characterises the distribution of the sum $\sum_{j=1}^n Z_j$. Lévy's continuity theorem says that pointwise convergence of characteristic functions is equivalent to convergence in distribution.

(a) Let $X_1, X_2, \ldots$ and X be random variables with characteristic functions $\varphi_1, \varphi_2, \ldots$ and φ. Assume that $\varphi_n(t) \to \varphi(t)$. Let $Z_1, Z_2, \ldots$ and Z be standard normal random variables independent of $X_1, X_2, \ldots$ and X, respectively. For $\sigma > 0$, and use the densities from Ex. 4.20(b) combined with Ex. 4.10(f) to show that $X_n + \sigma Z_n \to_d X + \sigma Z$ as $n \to \infty$. Conclude that $X_n \to_d X$ as $n \to \infty$ by letting $\sigma \to 0$.

(b) It is crucial for the argument in (a) that the limit φ is indeed a characteristic function. It turns out that if $\varphi_n(t)$ converges pointwise to some function $\beta(t)$ that is not assumed to be a characteristic function, but that is continuous at zero, then $\beta(t)$ is necessarily the characteristic function of some random variable. We prove this through a string of exercises. Start by using Fubini's theorem 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 inequality given above.

(c) For the case of X being a standard normal, check the precision of the tail inequality (the answer appears to be that it's rather unsharp). From $\varphi(t) = 1 - \frac{1}{2}t^2\sigma^2 + o(t^2)$, for a random variable with zero mean and variance σ^2, work out that $\Pr\{|X| \geq 2/\varepsilon\} \leq (1/3)\sigma^2\varepsilon$. Explain why this is blunter, as in less sharp, that with for example the Chebyshov inequality.

(d) [xx this ex must be rewritten, in view of Ex. 4.11(a) and Ex. 4.11(e)]If we know have a collection of random variables, where their characteristic functions have approximately the same level of smoothness around zero, then we should get *tightness*, a guarantee that there is no runaways with mass escaping from the crowd. Assume that $X_1, X_2, \ldots$ have characteristic functions $\varphi_1, \varphi_2, \ldots$, and that $\varphi_n(t) \to \beta(t)$, with $\beta(t)$ a function continuous at zero. For a given $\varepsilon' > 0$, find $\varepsilon > 0$ such that $|1 - \varphi(t)| \leq \varepsilon'$ for $|t| \leq \varepsilon$. Show that

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

We've hence found a broad interval, namely $[-2/\varepsilon, 2\varepsilon]$, inside which each X_n lies, with high enough probability. This is called *tightness* of the X_n sequence.

(e)

Ex. 4.22 *Proving the CLT (under some restrictions).* Let $X_1, X_2, \ldots$ be i.i.d. with distribution F, and assume for simplicity that the mean is zero.

(a) Show that if the moment generating function exists in a neighbourhood around zero, then $M(t) = 1 + \frac{1}{2}\sigma^2t^2 + o(t^2)$ as $t \to 0$, where σ is the standard deviation of X_1.

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

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

and conclude that the CLT holds.

(c) (xx nils pushes an earlier thing from Ch2 to this place; then we edit and prune and clean. xx) Consider a variable Y, with moment-generating function $M(t) = \mathrm{E}\exp(tY)$, assumed to be finite in at least a neighbourhood around zero. We have seen in Ex. 1.21 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 $\log M(t) = \xi t + \frac{1}{2}\sigma^2t^2 + o(t^2)$.

(d) 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

$$M(t) = \exp(t\xi)\,\mathrm{E}\exp(tY_0) = \exp(t\xi)\{1 + \tfrac{1}{2}\sigma^2t^2 + \tfrac{1}{6}\gamma_3t^3 + o(|t|^3)\},$$

where $\gamma_3 = \mathrm{E}\,(Y - \xi)^3$.

(e) Consider $Y_1, \ldots, Y_n$ i.i.d. from a distribution with mean zero and moment-generating function $M(t)$ 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. 1.21, with attention to certain further details in Ch 3 xx).

(f) (xx round off, point to CLT, identify remainder term with skewness. xx)

Ex. 4.23 *Improving on the weak LLN.* Let $X_1, X_2, \ldots$ be i.i.d. with finite mean ξ. We have seen in Ex. 4.1 that the LLN holds, in probability, if the distribution has a finite variance. Here we get rid of the finite variance condition.

(a) Show that the characteristic function for X_i satisfies $\varphi(t) = 1 + i\xi t + o(t)$ as $t \to 0$. Also show that its derivative exists, with $\varphi'(t) = \mathrm{E}\, iX \exp(itX)$; in particular $\varphi'(0) = i\xi$.

(b) Show that $\bar{X}_n \to_d \xi$, and use Ex. 4.13 to argue that we now have the weak LLN, also for distributions with infinite variance, as long as the mean is finite.

Ex. 4.24 *Proving the CLT (again).* As we saw in Ex. 4.17, if one is assuming that a random variable has a moment generating function, then one is effectively assuming that all its moments are finite, a rather restrictive condition. The characteristic function, on the other hand, always exists, so proving the CLT using characteristic functions is gives a more unified and elegant proof than when using moment generating functions, as above.

(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 $\varphi'(0) = i\xi$.

(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. 4.25 *Approximate variances and the delta method, II.* (xx point back to simpler start version, in Ex. 2.11. xx) We often meet variables of the form say $Z = g(Y_1, \ldots, Y_p)$ or similar, where the Y_i have normal or approximately normal distributions. If g is a linear function, say $g(y) = c_0 + c_1 y_1 + \cdots + c_p y_p$, then first of all the mean is transformed in the same fashion, with $\mathrm{E}\, Z = g(\mathrm{E}\, Y) = c_0 + c_1 \mathrm{E}\, Y_1 + \cdots + c_p \mathrm{E}\, Y_p$, and secondly the distribution of Z is normal as long as $(Y_1, \ldots, Y_p)$ is multinormal. This exercise concerns

vertain valid and useful approximations, along with limit theorems. Basically, even though the mean of $\exp(Y)$ is not equal to $\exp(\mathrm{E}\,Y)$, it might be a valid approximation if Y has a small variance. Similarly, as we shall see, if Y is approximately normal, with a small variance, then also $\exp(Y)$ is approximately normal, and so on.

(a) 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.

(b) Suppose $Z_n = \sqrt{n}(A_n - a) \to_d Z \sim \mathrm{N}(0,\tau^2)$, say. Translated to the setting above, we have $A_n = a + Z_n/\sqrt{n} + \varepsilon_n$, with $Z_n \to_d Z$ and $\sqrt{n}\varepsilon_n \to_{\mathrm{pr}} 0$. Let again $g(y)$ be a function smooth in a neighbourhood of a. Applying the previous point, show that $g(A_n) = g(a) + g'(a)Z_n/\sqrt{n} + \varepsilon_n'$, where ε_n' is so small that $\sqrt{n}\varepsilon_n' \to_{\mathrm{pr}} 0$. Show that this implies

the delta method

$$\sqrt{n}\{g(A_n) - g(a)\} \to_d g'(a)Z \sim \mathrm{N}(0, g'(a)^2\tau^2).$$

This is called *the delta method*. We also have the useful approximation $\mathrm{Var}\,g(A_n) \doteq g'(a)^2\,\mathrm{Var}\,A_n$.

(c) If $\sqrt{n}(A_n - a) \to_d \mathrm{N}(0,1)$, find out what happens to $\sqrt{n}\{\exp(A_n) - \exp(a)\}$ and to $\sqrt{n}(A_n^3 - a^3)$.

(d) In extension of the above we have the multidimensional delta method, with ensuing variance approximations. Suppose $\sqrt{n}(A_n - a) \to_d Z$, for variables A_n and Z, and vector a, of dimension p. Assume $g(y) = g(y_1,\ldots,y_p)$ is defined in a neighbourhood around $a = (a_1,\ldots,a_p)^{\mathrm{t}}$, with a continous derivative or gradient there. Show that

$$\sqrt{n}\{g(A_n) - g(a)\} \to_c c^{\mathrm{t}}Z = c_1Z_1 + \cdots + c_pZ_p,$$

where $c = \partial g(a)/\partial a$ is the gradient vector, with components $\partial g(a)/\partial a_j$, evaluated in position a. If in particular Z is multinormal, say $Z \sim \mathrm{N}_p(0,\Sigma)$, then $\sqrt{n}\{g(A_n) - g(a)\} \to_d \mathrm{N}(0,\tau^2)$, with $\tau^2 = c^{\mathrm{t}}\Sigma c$.

Ex. 4.26 *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. 4.27 *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: $Z_n = \sqrt{n}\{\exp(c_nX_n) - 1\}$. xx)

Ex. 4.28 *Bernshteĭn and Weierstraß.* [xx point out the affinity between this exercise and Ex. **??**. 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 $X_n \sim \text{binom}(n,p)$. Note that $B_n(p)$ is a polynomial of degree n.

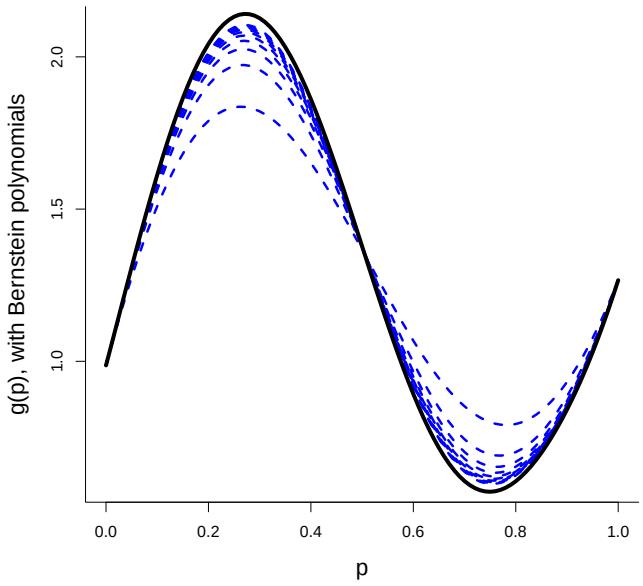

Figure 4.1: The given non-polynomial function $g(p)$ (full black curve), along with approximating Bernshteĭn polynomials, of order $10, 20, \ldots, 90, 100$.

(a) Show that $B_n(p) \to_{\text{pr}} g(p)$, 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

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

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

(b) Consider the marvellous function

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

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 4.1, 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) 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.

(d) xx

Ex. 4.29 *The Borel–Cantelli lemma.* (xx note: decide when cleaning whether this should be in MiniPrimer chapter or here. xx) Let $A_1, A_2, \ldots$ be events, in a relevant probability space, with probabilities $p_i = P(A_i)$. Consider $A_{\text{i.o.}} = \cap_{i \geq 1} \cup_{j \geq i} A_j$, the full-sequence event corresponding to the A_i occurring infinitely often.

(a) 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 $\sum_{i=1}^{\infty} p_i$ is convergent. Show that $P(A_{\text{i.o.}}) = 0$. – So sooner or later, there will be a finite (but random) n, such that none of the A_i will ever occur, for $i > n$.

(c) Assume in addition that the A_i are independent. Show that if $\sum_{i=1}^{\infty} p_i$ is divergent, then $P(A_{\text{i.o.}}) = 1$. – In particular, for the case of independent events, there can't be say a 50 percent chance that there will be infinitely many occurrences.

(d) Consider independent Bernoulli 0-1 variables X_i with $P(X_i = 1) = p_i$. What is the probability for having infinitely many $X_i = 1$, for $p_i = 1/i^{0.99}$, for $p_i = 1/i$, for $p_i = 1/i^{1.01}$?

(e) 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 $X_i \geq 1.01 \log i$?

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

$$P(X_i \geq a) = 1 - \Phi(a) \doteq \phi(a)/a,$$

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 $|X_i| \geq (2 \log i)^{1/2}$.

(g) (xx one or two more. new records, $P(R_n = 1) = 1/n$. xx)

Ex. 4.30 *Proving the CLT (yet again).* We have proven the CLT for i.i.d. random variables two times: Under some moment restrictions in Ex. 4.22, and in a more unified manner in Ex. 4.24. In this exercise we look at a proof that is, in a sense, more direct from the Portmanteau theorem. For this proof of the CLT, the result from Ex. **????** is used actively.

(a) Suppose that $X_1, X_2 \ldots$ are i.i.d. random variables with mean zero and unit variance, and let $Z_1, Z_2, \ldots$ be independent standard normal random variables. Define $X_{n,i} = X_i/\sqrt{n}$ and $Z_{n,i} = Z_i/\sqrt{n}$ for $i \geq 1$. For $n^{1/2}\bar{X}_n \to_d \mathrm{N}(0,1)$ (xx this a bit clearer, emil xx), why is it sufficient to show that

$$\mathrm{E}\, g(\sum_{i=1}^{n} X_{n,i}) \to \mathrm{E}\, g(\sum_{i=1}^{n} Z_{n,i}),$$

for all infinitely differentiable functions g with compact support?

(b) For integers $1 \leq a, b \leq n$, let $X_{a:b} = \sum_{j=a}^{b} X_{n,i}$ if $a < b$, and let $X_{a:b} = 0$ if $a > b$. Convince yourself of the following,

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

(c) Recall that by Taylor's theorem, if the function g is differentiable at y, then there is a function h, such that

$$g(x+y) - g(y) = g'(y)x + \tfrac{1}{2}g''(y)x^2 + h(x+y)x^2,$$

with $\lim_{x\to 0} h(x+y) = 0$. Let now g be an infinitely continuously differentiable function with compact support. Explain why there exists a $K < \infty$, and $\delta > 0$ such that for any $\varepsilon > 0$,

$$|g(x+y) - g(x) - g'(y)x + \tfrac{1}{2}g''(y)x^2| \leq \varepsilon x^2$$

when $|x| \leq \delta$, and

$$|g(x+y) - g(x) - g'(y)x + \tfrac{1}{2}g''(y)x^2| \leq Kx^2$$

otherwise.

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

$$\mathrm{E}\, g(X_{1:k} + Z_{(k+1):n}) - g(X_{1:(k-1)} + Z_{k:n}) = \mathrm{E}\, r_{n,k}(X) + \mathrm{E}\, r_{n,k}(Z), \tag{4.7}$$

where

$$r_{n,k}(X) \leq \frac{\varepsilon}{n} X_k^2 + \frac{K}{n} X_k^2 I_{|X_k| \geq \sqrt{n}\delta}.$$

(e) Please conclude, and you will have shown the CLT for i.i.d. random variables yet again.

the Stirling approximation

Ex. 4.31 *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.

(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 $\sum_{j\ge n}(j-n)p(j,n) = np(n,n)$.

(c) (xx a bit more. xx)

Ex. 4.32 *Characterisations of variables with finite mean.* For certain technical needs we need characterisations of the tails of a distribution with finite mean.

(a) Show that if $X \ge 0$, with distribution function F, then $\mathrm{E}\,X = \int_0^\infty \{1-F(x)\}\,\mathrm{d}x$. Show also that $\mathrm{E}\,X^2 = \int_0^\infty \{1-F(x^{1/2})\}\,\mathrm{d}x$.

(b) Show more generally that for any X,

$$\mathrm{E}\,X = \int_{-\infty}^0 F(x)\,\mathrm{d}x + \int_0^\infty \{1-F(x)\}\,\mathrm{d}x.$$

(c) If X has finite mean, show that $\sum_{i=1}^\infty (1/i^2)\int_{(-i,i)} x^2\,\mathrm{d}F(x) < \infty$.

(d) Upon examining the arguments needed to prove the previous point, one learns that this is an if-and-only-if result. More generally, attempt to prove that

$$\mathrm{E}\,|X|^m < \infty \quad \text{if and only if} \quad \sum_{i=1}^\infty \frac{1}{i^2}\int_{(-i,i)} |x|^{m+1}\,\mathrm{d}F(x) < \infty.$$

Ex. 4.33 *Further tail bound inequalities.* In Ex. 2.7 we learned about the Markov and Chebyshov inequalities; here we work out further tail bounds for our toolboxes.

(a) If X has mean ξ, and a finite fourth moment, show that $P(|X-\xi|\ge\varepsilon) \le \mathrm{E}\,|X-\xi|^4/\varepsilon^4$. With $X_1,\ldots,X_n$ i.i.d. from this distribution, show that $P(|\bar X_n-\xi|\ge\varepsilon) \le K_4/(n^2\varepsilon^2)$, for a suitable positive constant K_4. When is this a sharper result than that of the Chebyshov inequality? Generalise to higher moments. (xx mention von Bahr. his results imply that for any r for which $\mathrm{E}\,|X_i|^r$ is finite, there is a constant K_r such that $P(|\bar X_n-\xi|\ge\varepsilon) \le K_r/(n^r\varepsilon^r)$. xx)

(b) Suppose X has a finite moment-generating function $M(t) = \mathrm{E}\exp(tX)$, as per Ex. 1.20. Show that

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

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 $K'(t_a) = a$.

(c) For $X \sim \mathrm{N}(0,1)$ and a positive, show that $P(x \geq a) \leq \exp(-\frac{1}{2}a^2)$. Show that this is indeed sharper than the tail bound $1/a^2$, from the simpler Chebyshov inequality, for all $a > 0$.

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

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

(xx a bit more here, rounding it off. more inequalities in Ch. 4. xx)

(e) (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 $p_m(a) = P(X \geq m + am^{1/2})$. Show that

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

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

(f) Let $X_1, X_2, \ldots$ be i.i.d. with mean zero and variance one, so that $\sqrt{n}\bar{X}_n \to_d \mathrm{N}(0,1)$. Assume its moment-generating function $M(t) = \exp\{K(t)\}$ is finite. Show that

$$P(\sqrt{n}\bar{X}_n \geq a) \leq \frac{M(t/\sqrt{n})^n}{\exp(ta)} = \exp\{nK(t/\sqrt{n}) - ta\},$$

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)

(g) (xx the Jensen too: $\mathrm{E}\, h(X) \geq h(\mathrm{E}X)$, when h is convex. a few applications. show $(\mathrm{E}\,|X|^r)^{1/r}$ is increasing in r. calibrate with Ex. 1.19, which should perhaps land here. xx)

(h) (xx round off. xx)

Ex. 4.34 *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 (4.3), 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 $P(A_N) = 1$ for all N, then $P(\cap_{N\geq 1}A_N) = 1$ – 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 $P(A) = 1$ if and only if $P(B_{n_0}) \to 0$, 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 $P(B_{n_0}) \leq \sum_{n \geq n_0} P\{|\bar{X}_n| \geq \varepsilon\}$, so it suffices to show, if possible, under appropriate conditions, that $\sum_{n \geq 1} P\{|\bar{X}_n| \geq \varepsilon\}$ is a convergent series. With finite variance σ^2, show that the classic simple Chebyshov bound, see Ex. 2.7, does *not* solve any problem here.

(e) (xx calibrate better with Ex. 2.7. xx) Show, however, that if the fourth moment is finite, then

$$P\{|\bar{X}_n| \geq \varepsilon\} \leq \frac{1}{\varepsilon^4} \mathrm{E}\, |\bar{X}_n|^4 \leq \frac{c}{\varepsilon^4} \frac{1}{n^2},$$

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

$$P\{|\bar{X}_n| \geq \varepsilon\} \leq \frac{1}{(\sqrt{n}\varepsilon)^r} \mathrm{E}\, |\sqrt{n} \bar{X}_n|^r,$$

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. 4.35.

Ex. 4.35 *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

$$P\{\max_{i \leq n} |S_i| \geq \varepsilon\} \leq \frac{\mathrm{Var}\, S_n}{\varepsilon^2} = \frac{1}{\varepsilon^2} \sum_{i=1}^{n} \sigma_i^2.$$

Note that this is a much stronger result than the special case of caring only about $|S_n|$, with $P\{|S_n| \geq \varepsilon\} \leq \mathrm{Var}\, S_n / \varepsilon^2$, which is the Chebyshov inequality. To prove it, work with the disjoint decomposition

$$A_i = \{|S_1| < \varepsilon, \ldots, |S_{i-1}| < \varepsilon, |S_i| \geq \varepsilon\} \quad \text{and} \quad A = \cup_{i=1}^{n} A_i = \{\max_{i \leq n} |S_i| \geq \varepsilon\}.$$

Show that $\mathrm{E}\, S_n^2 \geq \mathrm{E}\, S_n^2 I(A) = \sum_{i=1}^n \mathrm{E}\, S_n^2 I(A_i)$, that

$$\mathrm{E}\, S_n^2 I(A_i) = \mathrm{E}\,(S_i + S_n - S_i)^2 I(A_i) \geq \varepsilon^2 P(A_i),$$

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

$$\lim_{n\to\infty} P\big[\cup_{i,j\geq n}\{|S_i - S_j| \geq \varepsilon\}\big] = 0 \quad \text{for each } \varepsilon > 0.$$

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, ladies and gentlemen. 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 Ex. 4.32, to show that only finitely many Z_i are non-zero, and 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, congratulations. Attempt also to prove the interesting converse that if $\mathrm{E}\,|X_i| = \infty$, then the sequence of sample means is pretty erratic indeed:

$$P\{\limsup_{n\to\infty} \bar{X}_n = \infty\} = 1, \quad P\{\liminf_{n\to\infty} \bar{X}_n = -\infty\} = 1.$$

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. 4.36 *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.

Ex. 4.37 *Glivenko–Cantelli theorem.* For i.i.d. observations $Y_1, \ldots, Y_n$, we form the empirical c.d.f. as in Ex. 2.27, with $F_n(t) = n^{-1}\sum_{i=1}^n I(Y_i \le t)$. Since this is just a binomial ratio, we know from the Law of Large Numbers that $F_n(t) \to_{\mathrm{a.s.}} F(t)$, 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 $P(D_n \to 0) = 1$. 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. 9.10 for more information about 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 $P(\cap_{j=1}^m A_{m,j}) = 1$.

(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

$$D_n(t_j) - \{F(t_{j+1}) - F(t_j)\} \le F_n(t) - F(t) \le D_n(t_{j+1}) + F(t_{j+1} - F(t_j).$$

Deduce that

$$\max_{t_j \le t < t_{j+1}} |D_n(t)| \le B_m + C_m,$$

where $B_m = \max_{1\le j\le m} |D_n(t_j)|$ and $C_m = \max_{1\le j\le m}\{F(t_{j+1}) - F(t_j)\}$.

(c) Show that $P(\limsup D_n \le C_m) = 1$.

(d) For each $\varepsilon > 0$, show that a partition into cells can be arranged, with high m if required, so that $C_m \le \varepsilon$. Conclude that $P(D_n \to 0) = 1$.

(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. 1.51. 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 must be to have $D_n \le 0.01$, say, in a few situations.

Ex. 4.38 *The Liapunov and Lindeberg theorems: main story.* (xx to be edited and polished. xx) 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) As an introductory useful lemma, demonstrate the following. With $a_1, a_2, \ldots$ a sequence of numbers coming closer to zero, we have $\prod_{i=1}^n (1 + a_i) \to \exp(a)$ provided (1) $\sum_{i=1}^n a_i \to a$; (2) $\max_{i \le n} |a_i| \to 0$; and (3) $\sum_{i=1}^n |a_i|$ 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 $K(x)$ is a continuous function such that $|K(x)| \le 1$ for $|x| \le \frac{1}{2}$, and $K(x) \to -\frac{1}{2}$ when $x \to 0$. These statements are valid also when the a_i are the x are complex numbers inside the unit ball, in which case the logarithm is the natural complex extension of the real logarithm.

(b) Show that Z_n has characteristic function

$$\kappa_n(t) = \mathrm{E}\exp(itZ_n) = \varphi_1(t/B_n) \cdots \varphi_n(t/B_n).$$

(c) (xx check this, and calibrate with other exercises. show that

$$|\exp(ix) - 1 - ix - \tfrac{1}{2}(ix)^2 - \ldots - (ix)^m/m!| \le |x|^{m+1}/(m+1)!.$$

xx)

(d) We know that $\varphi_i(s) \doteq 1 - \frac{1}{2}\sigma_i^2 s^2$ 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 not give up until one has found conditions that secure convergence to the desired $\exp(-\frac{1}{2}t^2)$. In view of the lemma of (a), this essentially takes (1) $\sum_{i=1}^n \varepsilon_{n,i}(t) \to 0$;

(2) $\max_{i\le n}\sigma_i^2/B_n^2\to 0$ and $\max_{i\le n}|\varepsilon_{n,i}(t)|\to 0$; and (3) $\sum_{i=1}^n|1-\varphi_i(t/B_n)|$ staying bounded. Show that

$$\begin{aligned}|\varphi_i(s)-(1-\tfrac12\sigma_i^2s^2)| &= \Big|\int\{\exp(isx)-1-isx-\tfrac12(isx)^2\}\,\mathrm{d}F_i(x)\Big|\\ &\le \int|\exp(isx)-1-isx-\tfrac12(isx)^2|\,\mathrm{d}F_i(x)\\ &\le \tfrac16|s|^3\,\mathrm{E}\,|X_i|^3.\end{aligned}$$

(e) 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 $\kappa_n(t)\to\exp(-\frac12t^2)$, which we know is equivalent to the glorious $Z_n\to_d\mathrm{N}(0,1)$. This is (already) a highly significant extension of the CLT. If the X_i are uniformly bounded, for example, with B_n of order $\sqrt{n}$, which would rather often be the case, then the условие Ляпунова 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.

(f) The issue waits however for even milder and actually minimal conditions, and that is, precisely, the Lindeberg condition:

$$\sum_{i=1}^n\mathrm{E}\Big|\frac{X_i}{B_n}\Big|^2 I\Big\{\Big|\frac{X_i}{B_n}\Big|\ge\varepsilon\Big\}\to 0\quad\text{for all }\varepsilon>0.$$

Show that if the Lyapunov condition is in force, then the Lindeberg condition holds (so Lindeberg assumes less than Lyapunov).

(g) (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' moment-generating functions 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.

Ex. 4.39 *The Lindeberg theorems: nitty-gritty details.* (xx to be cleaned and polished. xx) 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. You may also check partly corresponding details in Stoltenberg (2019). 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 creature 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 the Lindeberg condition, that

$$L_n(\varepsilon) = \sum_{i=1}^n \mathrm{E}\Big|\frac{X_i}{B_n}\Big|^2 I\Big\{\Big|\frac{X_i}{B_n}\Big| \ge \varepsilon\Big\} \to 0 \quad \text{for all } \varepsilon > 0.$$

(a) Show that $B_n \to \infty$, and that $\alpha_n = \max_{i \le n}(\sigma_i^2/B_n^2) \to 0$. Show further that this entails

$$\begin{aligned} |\varphi_i(t/B_n) - 1| &\le \int |\exp(itx/B_n) - 1 - itx/B_n| \,\mathrm{d}F_i(x) \\ &\le \tfrac12 t^2 \int (x/B_n)^2 \,\mathrm{d}F_i(x) \le \tfrac12 t^2 \alpha_n, \end{aligned}$$

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 and work with

$$\kappa_n(t) = \log \mathrm{E}\exp(itZ_n) = \sum_{i=1}^n \log \varphi_i(t/B_n)$$

etc.; see the start lemma of the preceding exercise.

(b) In continuation and refinement of arguments above, show that $r_n(t) = \varphi_i(t/B_n) - (1 - \frac12\sigma_i^2 t^2/B_n^2)$ can be bounded, as follows:

$$\begin{aligned} |r_n(t)| &= \Big| \int \{\exp(itx/B_n) - 1 - itx/B_n - \tfrac12 (itx/B_n)^2\} \,\mathrm{d}F_i(x)\Big| \\ &\le \int |\exp(itx/B_n) - 1 - itx/B_n - \tfrac12 (itx/B_n)^2| \,\mathrm{d}F_i(x) \\ &\le \int_{|x|/B_n \le \varepsilon} \tfrac16 \frac{|t|^3|x|^3}{B_n^3} \,\mathrm{d}F_i(x) + \int_{|x|/B_n > \varepsilon} \Big(\tfrac12 \frac{|t|^2|x|^2}{B_n^2} + \tfrac12 \frac{|t|^2|x|^2}{B_n^2}\Big) \,\mathrm{d}F_i(x) \\ &\le \tfrac16 |t|^3 \varepsilon \frac{\sigma_i^2}{B_n^2} + t^2 \,\mathrm{E}\Big|\frac{X_i}{B_n}\Big|^2 I\Big\{\Big|\frac{X_i}{B_n}\Big| \ge \varepsilon\Big\}. \end{aligned}$$

(c) Show that this leads to

$$\sum_{i=1}^n \big|\varphi_i(t/B_n) - (1 - \tfrac12\sigma_i^2 t^2/B_n^2)\big| \le \tfrac16 |t|^3 \varepsilon + t^2 \, L_n(\varepsilon),$$

and via the start lemma of the previous exercise that this secures what we were after, that $\prod_{i=1}^n \varphi_i(t/B_n) \to \exp(-\frac12 t^2)$ and hence triumphantly $Z_n \to_d \mathrm{N}(0,1)$, under the Lindeberg condition only.

Ex. 4.40 *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

$$\sum_{i=1}^n \mathrm{E}\Big|\frac{a_i\varepsilon_i}{B_n}\Big|^2 I\Big(\Big|\frac{a_i\varepsilon_i}{B_n}\Big| \ge \delta\Big) \le \sum_{i=1}^n \frac{a_i^2}{B_n^2}\mathrm{E}\,\varepsilon_i^2\, I(D_n|\varepsilon_i| \ge \delta) \le \int_{|u|\ge \delta/D_n} u^2 \,\mathrm{d}G(u).$$

(b) Conclude that $Z_n \to_d \mathrm{N}(0,1)$ provided $D_n \to 0$.

(c) 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 $a_i = 1/i$.

(d) 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)

(e) xx

Ex. 4.41 *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

$$\varphi(t_1, \ldots, t_k) = \mathrm{E}\,\exp(it^{\mathrm{t}}X) = \mathrm{E}\,\exp\{i(t_1X_1 + \cdots + t_kX_k)\}$$

for $t = (t_1, \ldots, t_k)^{\mathrm{t}}$.

(a) 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.

(b) Show for the multinormal case, where $X \sim \mathrm{N}_k(\xi, \Sigma)$, that $\varphi(t) = \exp(it^{\mathrm{t}}\xi - \frac{1}{2}t^{\mathrm{t}}\Sigma t)$.

(c) (xx spell out basics for $\varphi_n(t) \to \varphi(t)$ as equivalent to $X_n \to_d X$. inversion formula. xx)

Ex. 4.42 *The Cramér–Wold device.* Suppose X_n and X are random variables in $\mathbb{R}^k$.

(a) 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. (xx proof via characteristic functions, so need to set up this properly first. xx)

(b) This very useful device, for proving convergence in distribution for random vectors, can now be used to prove the multivariate CLT: Suppose $X_1, X_2, \ldots$ are i.i.d. with mean ξ and variance matrix Σ. Show that $\sqrt{n}(\bar{X}_n - \xi) \to_d \mathrm{N}_k(0, \Sigma)$.

(c) Let $X_1, X_2, \ldots$ be independent random vectors in dimension k, with finite positive definite variance matrices $\Sigma_1, \Sigma_2, \ldots$. When will

$$Z_n = (\Sigma_1 + \cdots + \Sigma_n)^{-1/2}(X_1 + \cdots + X_n) \to_d \mathrm{N}_k(0, I_k)?$$

(xx give clear conditions. the multidimensional Lindeberg theorem. for each a, need $\sum_{i=1}^n a^{\mathrm{t}} X_i / (\sum_{i=1}^n a^{\mathrm{t}} \Sigma_i a)^{1/2} \to_d \mathrm{N}(0,1)$. to be used for regression models in Ch5 and later. give sets of easier conditions for the general Lindeberg to hold. xx)

Ex. 4.43 *Limiting normality in linear regression.* 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. (xx pointer to other 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 $\widehat{\beta} = \sum_{i=1}^n x_i y_i / M_n^2$, where $M_n^2 = \sum_{i=1}^n x_i^2$. Show further that $\widehat{\beta}$ is unbiased with variance σ^2 / M_n^2. (xx point to least squares, see Ex. 2.33, and more. xx)

(b) Then consider $Z_n = M_n(\widehat{\beta} - \beta)$. Show that it has zero mean and variance σ^2, and that it can be written $\sum_{i=1}^n (x_i / M_n)\varepsilon_i$.

(c) Deduce from Ex. 4.40 that $\widehat{\beta}$ is approximately normal, even if the ε_i are not normal, provided merely that $D_n = \max_{i \le n} |x_i| / M_n \to 0$. If in particular $(1/n)\sum_{i=1}^n x_i^2$ stays bounded, or perhaps has a finite limit, then the natural condition is $(1/\sqrt{n}) \max_{i \le n} |x_i| \to 0$.

(d) Then consider the general linear regression model, $y_i = x_i^{\mathrm{t}}\beta + \varepsilon_i$, with the x_i being p-dimensional covariate vectors and β a p-dimensional vector of regression coefficients. (xx here point to LinReg exercise which we perhaps should move from Ch1 to Ch2. xx) For the least squares estimator we have

$$\widehat{\beta} = \Sigma_n^{-1}(1/n)\sum_{i=1}^n x_i y_i, \quad \text{with} \quad \Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}}.$$

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) (xx spell out, just a bit, that confidence intervals and tests, worked out with the finest precision under exact normality the $y_i \sim \mathrm{N}(x_i^{\mathrm{t}}\beta, \sigma^2)$, work very well, even without the normality. xx)

Ex. 4.44 *Limiting normality for multinomials.* (xx simple basic limits for $\widehat{p}_j = N_j/n$, and delta method. illustration for $\widehat{N} = n_{1,\cdot}n_{\cdot,1}/n_{1,1}$ in the problem with missing $n_{0,0}$. so we refer to this exercise in the Karl Pearson chi-squared story, rather than doing these things there. xx) 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 $p = (p_1, \ldots, p_k)^{\rm t}$.

(a) Show that the relative frequencies $\widehat{p}_j = Y_j/n$ 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 $\sigma_{j,\ell} = p_j\delta_{j,\ell} - p_jp_\ell$. 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 $\widehat{\delta} = 2\arcsin(\widehat{p}^{1/2}) - 2\arcsin(\widehat{q}^{1/2})$.

(d) (xx one more thing. xx)

Ex. 4.45 *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? (xx link to Story iii.9, with further information, using likelihood theory. we make a Srebrenica exercise too, with Brunborg et al. (2003). and to Guatemala Story iii.9, with three sources, and hidden $n_{0,0,0}$. point to Lum et al. (2013), and also to Petersen (1896) for the estimator. also Brunborg et al. (2003), inside Story iii.9. check notation: N^* for Petersen 1896, then $\widehat{N}$ for ML in later exerciss and in story. xx)

(a) Assume in this exercise that factors X and Y are independent, with $P(X = 1) = p = p_{1,\cdot}$ and $P(Y = 1) = q = p_{\cdot,1}$; 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 $n_{1,\cdot}n_{1,\cdot}/N^2$ and $n_{1,1}/N$ are both valid estimates of $p_{1,1}$. Discuss conditions under which $N^* = n_{1,\cdot}n_{1,\cdot}/n_{1,1}$ 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.

(d) 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? (xx pointer again to more on this in Story iii.9, with profiled likelihoods. xx)

(e) 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)

(f) 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. (xx note: in R we may use `intersect`, `union`, `setdiff`. xx)

(g) (xx don't know yet if we should 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)

Ex. 4.46 *Stochastic $O_{\rm pr}$ and $o_{\rm pr}$ symbols.*

Ex. 4.47 *Minimum criterion function estimators, I.* (xx should we bother enough to change from F to G here, for data generating mechanism. but we use $\phi = g(\theta)$ here and there for focus parameters. xx) For observations $Y_1, \ldots, Y_n$ from some distribution F, consider a parameter $\theta_0 = \theta(F)$ defined as the minimiser of the function $H(\theta) = {\rm E}_F\, h(Y, \theta)$, 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 $H_n(\theta) = (1/n)\sum_{i=1}^n h(Y_i, \theta)$, so a natural estimator for θ_0 is $\widehat{\theta} = {\rm argmin}(H_n)$. In fact many important estimators are of this or related types, perhaps minimising somewhat more complicated random functions; cf. the broad maximum likelihood themes of Ch. 5 (xx and more xx). In exercises below we shall develop clear results for how the minimum criterion 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 criterion function estimators

(a) Explain that $H_n(\theta)$ can be written $\int H(y,\theta)\,\mathrm{d}F_n(y)$, with F_n the empirical distribution, having mass $1/n$ at each datapoint; see Ex. 2.27. Explain why $H_n(\theta) \to_{\rm pr} H(\theta)$ for each θ, and find the limit distribution of $\sqrt{n}\{H_n(\theta)-H(\theta)\}$. What we need, tended to in exercises to follow, are conditions under which $\widehat{\theta} = \text{argmin}(H_n)$ tends to $\theta_0 = \text{argmin}(H)$, along with a limit distribution.

(b) A few examples, for the case of one-dimensional Y_i, are as follows. (i) With $h(y,\theta) = (y-\theta)^2$, show that $\theta_0 = \mathrm{E}_F\, Y_i$, with $\widehat{\theta} = \bar{Y}_n$, the sample mean. (ii) Consider $h(y,\xi,\sigma) = (y-\xi)^2 + \{y^2 - (\xi^2+\sigma^2)\}^2$. Show that (ξ_0,σ_0) must be $(\mathrm{E}_F\, Y, \mathrm{sd}_F(Y))$, the true mean and standard deviation for Y. (iii) More generally, if $h(y,\theta) = \{r(y)-\theta\}^{\rm t} V \{r(y)-\theta\}$, for some $r(y) = (r_1(y),\ldots,r_p(y))$ and a symmetric positive definite matrix V, show that $\theta_0 = \mathrm{E}_F\, r(Y)$. (iv) Try $h(y,\theta) = [\exp\{c(y-\theta)\} - 1 - c(y-\theta)]/c^2$. 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) Consider $h_0(x) = x\arctan x - \frac{1}{2}\log(1+x^2)$, and define θ_0 as the minimiser of $\mathrm{E}_F\, h_0(Y-\theta)$. Show that $\widehat{\theta}$, the minimiser of $H_n(\theta) = (1/n)\sum_{i=1}^n h_0(Y_i-\theta)$, 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)

(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 criterion function $(1/n)\sum_{i=1}^n h(Y_i,\xi,\tau)$. Check in particular the case of $p(\tau) = \log\tau$.

(e) (xx briefly, make the connection to moment estimators. $h(y,\theta) = \{y - m(\theta)\}^2$ minimised for $m(\theta_0) = \mathrm{E}\, Y$, or $\theta_0 = m^{-1}(\mathrm{E}\, Y)$. also for vector case. and briefly to quantile fitting estimators too. xx)

(f) The L_2 distance between the data generating density f and a parametrically modelled f_θ is

$$D(f,f_\theta) = \int (f-f_\theta)^2\,\mathrm{d}y = \int f_\theta^2\,\mathrm{d}y - 2\int f f_\theta\,\mathrm{d}y + a(f),$$

where $a(f)$ does not depend on θ. Use this to motivate what we may call the minimum L_2 estimator $\widehat{\theta}$, the minimiser of $D_n(\theta) = \int f_\theta^2\,\mathrm{d}y - 2(1/n)\sum_{i=1}^n f(Y_i,\theta)$. Simulate 100 detapoints from a $\mathrm{Gam}(a,b)$, where you choose (a,b) as you wish, and estimate these using this method. 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. (xx pointer to BHHJ method. xx)

Ex. 4.48 *Minimum criterion function estimators, II.* After having motivated and worked through particular instances of the minimum criterion function estimators, in Ex. 4.47, we now return to the general case, aiming to demonstrate limiting normality, finding recipes for large-sample approximations in the process. So $\theta_0 = \theta_0(F)$ is the minimiser of $H(\theta) = \mathrm{E}_F\, h(y,\theta)$, and $\widehat{\theta}$ is its estimator, the minimiser of $H_n(\theta) = \int h(y,\theta)\,\mathrm{d}F_n(y)$. We shall employ the following regularity conditions: (i) The true

$\theta_0 = \theta_0(F)$ is an inner point in its parameter space inside $\mathbb{R}^p$; thus for each given $s = (s_1, \ldots, s_p)^{\rm t}$, the $A_n(s)$ is well-defined for all large enough n. (ii) The $h(y, \theta)$ is smooth in θ, in a neighbourhood around θ, with three derivatives, say $h'(y, \theta)$ (a vector, with components h'_a), $h''(y, \theta)$ (a matrix, with components $h''_{a,b}$), $h'''(y, \theta)$ (a collection of third derivatives $h'''_{a,b,c}$). (iii) The matrix $J = {\rm E}_F\, h''(Y, \theta_0)$ is finite and positive definite. (iv) The first derivative $h'(Y, \theta_0)$ has finite variance matrix K. (v) The third order derivatives $h'''_{a,b,c}(y, \theta)$ have finite means in a neighbourhood around θ_0.

(a) A key idea is to work with the following random function. Show that

$$\begin{aligned} A_n(s) &= n\{H_n(\theta_0 + s/\sqrt{n}) - H_n(\theta_0)\} \\ &= \sum_{i=1}^n \{h(Y_i, \theta_0 + s/\sqrt{n}) - h(Y_i, \theta_0)\} = U_n^{\rm t} s + \tfrac12 s^{\rm t} J_n s + r_n(s), \end{aligned}$$

with $U_n = (1/\sqrt{n}) \sum_{i=1}^n h'(Y_i, \theta_0)$, $J_n = (1/n) \sum_{i=1}^n h''(Y_i, \theta_0)$. Show that U_n has mean zero and tends to $U \sim {\rm N}_p(0, K)$, and that $J_n \to_{\rm pr} J$. Also, for the remainder term show that $|r_n(s)| \le C_n \|s\|^3/\sqrt{n}$, with C_n bounded in probability.

(b) 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 $B_n(s) = U_n^{\rm t} s + \frac12 s^{\rm t} J_n s$ 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

$$\begin{aligned} \beta_n &= -J_n^{-1} U_n \to_d -J^{-1} U \sim {\rm N}_p(0, J^{-1} K J^{-1}), \\ B_{n,\min} &= -\tfrac12 U_n^{\rm t} J_n^{-1} U_n \to_d -\tfrac12 U^{\rm t} J^{-1} U. \end{aligned}$$

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.

(c) Supposing $|r_n(s)| \le \delta$ for all s in a subset S, show from $A_n = B_n + r_n$ that

$$|\min_{s \in S} A_n(s) - \min_{s \in S} B_n(s)| \le \delta \quad \text{for } s \in S.$$

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 $\alpha_n = O_{\rm pr}(n^{1/8})$, and that $A_{n,\min} - B_{n,\min} \to_{\rm pr} 0$.

(d) 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

$$B_n(\beta_n + v) = B_{n,\min} + \tfrac12 v^{\rm t} J_n v \ge B_{n,\min} + \tfrac12 j_n \varepsilon^2,$$

where j_n is the smallest eigenvalue of J_n. Show then that the event Ω_n, where $A_n(\beta_n + v) \ge B_{n,\min} + \frac14 j_n \varepsilon^2$ for all v with $\|v\| \ge \varepsilon$, must have $P(\Omega_n) \to 1$. Prove from these established statements that $\alpha_n - \beta_n$ must tend to zero in probability. Conclude that

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &\to_d -J^{-1} U \sim {\rm N}_p(0, J^{-1} K J^{-1}), \\ 2n\{H_n(\theta_0) - H_n(\widehat{\theta})\} &\to_d W = U^{\rm t} J^{-1} U. \end{aligned} \tag{4.8}$$

(e) (xx note: i have avoided pointing to continuity of argmin functional in $C([-c,c]^p)$ space and the like, but worked directly with the quadratic approximation. xx)

Ex. 4.49 *Profiling a minimum criterion function, I.* For data $Y_1,\ldots,Y_n$ from some distribution F, we have in Ex. 4.47 considered estimating a parameter $\theta_0 = \text{argmin}(H)$, for $H(\theta) = \mathrm{E}_F\, h(Y,\theta)$, by minimising the criterion function $H_n(\theta) = (1/n)\sum_{i=1}^n h(Y_i,\theta)$. For a focus parameter $\phi = g(\theta)$, a smooth function of $\theta = (\theta_1,\ldots,\theta_p)$, it is useful to work with the associated profile function

profiling a criterion function

$$H_{n,\text{prof}}(\phi) = \min\{H_n(\theta)\colon g(\theta) = \phi\}.$$

(a) As an introductory illustration, consider estimating the parameters (a,b) of a Gamma distribution via minimum L_2, as in Ex. 4.47. Simulate 100 datapoints from a gamma; compute $(\widehat{a},\widehat{b})$; and compute and display also the profile function $H_{n,\text{prof}}(\mu)$ for the mean parameter $\mu = a/b$.

(b) From the full model, with ensuing minimum criterion function estimator $\widehat{\theta}$, show that the consequent $\widehat{\phi} = g(\widehat{\theta})$ becomes normal. With setup and notation as Ex. 4.47, prove indeed that $\sqrt{n}(\widehat{\phi}-\phi_0) \to_d c^{\mathrm{t}}J^{-1}U$, with $c = \partial g(\theta_0)/\partial\theta$, 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. 4.47, show that

$$n\{H_n(\widehat{\theta}+s/\sqrt{n}) - H_n(\widehat{\theta})\} = \tfrac{1}{2}s^{\mathrm{t}}J_n s + r_n(s), \quad \text{with } r_n(s) = O_{\text{pr}}(\|s\|^3/\sqrt{n}).$$

For the profiling, therefore, we must minimise this expression over all s such that $g(\theta) = g(\widehat{\theta}+s/\sqrt{n}) = \phi$. With $g(\theta) = \widehat{\phi} + c_n^{\mathrm{t}}s/\sqrt{n} + O_p(\|s\|^2/n)$, here writing $c_n = \partial g(\widehat{\theta})/\partial\theta$, the essence is to minimise $\frac{1}{2}s^{\mathrm{t}}J_n s$ under $c_n^{\mathrm{t}}s = \sqrt{n}(\phi-\widehat{\phi}) = x$, say. Show, perhaps via Lagrange multiplicator methods, that this minimum becomes $\frac{1}{2}x^2/c_n^{\mathrm{t}}J_n^{-1}c_n = n(\widehat{\phi}-\phi)^2/c_n^{\mathrm{t}}J_n^{-1}c_n$. Fill in more details to prove that with $\phi_0 = g(\theta_0)$ the true parameter in question,

$$\begin{aligned} 2n\{H_{n,\text{prof}}(\phi_0) - H_{n,\text{prof}}(\widehat{\phi})\} &= n(\widehat{\phi}-\phi_0)^2/c_n^{\mathrm{t}}J_n^{-1}c_n + o_{\text{pr}}(1) \\ &\to_d (c^{\mathrm{t}}J^{-1}U)^2/c^{\mathrm{t}}J^{-1}c \sim k\chi_1^2, \end{aligned}$$

with $k = c^{\mathrm{t}}J^{-1}KJ^{-1}c/c^{\mathrm{t}}J^{-1}c$.

(d) (xx to be cleaned. this lemma is for profiling w.r.t. a 1-dimensional $\phi = c^{\mathrm{t}}s$. xx) For a quadratic function $A(s) = \frac{1}{2}s^{\mathrm{t}}Js$, for a positive definite J, we are to examine the minimum of $A(s)$ over all s with $c^{\mathrm{t}}s = x$. Show that the result is $\frac{1}{2}x^2/c^{\mathrm{t}}J^{-1}c$. (xx nils sketching the solution: xx) Lagrange: $\frac{1}{2}s^{\mathrm{t}}Js - \lambda(c^{\mathrm{t}}s-x)$, $Js = \lambda c$, $s_0 = \lambda J^{-1}c$, leading to $c^{\mathrm{t}}s_0 = \lambda c^{\mathrm{t}}J^{-1}c = x$. The minimum becomes $\frac{1}{2}\lambda^2 c^{\mathrm{t}}J^{-1}c = \frac{1}{2}x^2/c^{\mathrm{t}}J^{-1}c$.

(e) (xx to be polished. lifting the above to $\phi = C\theta$. xx) For a $p_0 \times p$ matrix C and an x of dimension p_0, work out the minimum of $\frac{1}{2}s^{\rm t}Js$ over all s with $Cs = x$. Show that this becomes $\frac{1}{2}x^{\rm t}(CJ^{-1}C^{\rm t})^{-1}x$ (xx nils sketching the solution. xx) Lagrange: $\frac{1}{2}s^{\rm t}Js - \lambda^{\rm t}(Cs - x)$, derivative, $Js = C^{\rm t}\lambda$, $s_0 = J^{-1}C^{\rm t}\lambda$, implying $CJ^{-1}C^{\rm t}\lambda = x$. this leads to $\lambda = (CJ^{-1}C^{\rm t})^{-1}x$. nice: minimum becomes

$$\tfrac{1}{2}\lambda^{\rm t}CJ^{-1}C^{\rm t}\lambda = \tfrac{1}{2}x^{\rm t}(CJ^{-1}C^{\rm t})^{-1}CJ^{-1}C^{\rm t}(CJ^{-1}C^{\rm t})^{-1}x = \tfrac{1}{2}x^{\rm t}(CJ^{-1}C^{\rm t})^{-1}x.$$

(xx we should find the following from this: with $\phi = g(\theta) = (\phi_1, \ldots, \phi_{p_0})$,

$$n(H_{n,\min,\rm narr} - H_{n,\min,\rm wide}) \to_d \tfrac{1}{2}U^{\rm t}J^{-1}C(C^{\rm t}J^{-1}C)^{-1}C^{\rm t}J^{-1}U.$$

round off. xx)

(f) (xx 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. 4.50 *Profiling a criterion function, II.* In Ex. 4.47, 4.48, 4.49 we have considered parameters defined as minimisers of functions $H(\alpha) = {\rm E}_F\, h(Y, \alpha)$, and developed the basic theory for the associated minimum criterion 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 $H_n(\theta, \gamma) = \int h(y, \theta, \gamma)\, {\rm d}F_n(y)$ whereas $\widetilde\theta$ for the narrow model minimises $H_n(\theta, \gamma_0) = \int H(y, \theta, \gamma_0)\, {\rm d}F_n(y)$. 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. 4.48 are in force. Efforts of linear and matrix algebra are required in order to handle these models jointly, however. Define therefore

$$J_{\rm wide} = {\rm E}_F \frac{\partial^2 H(Y, \theta_0, \gamma_0)}{\partial\alpha\partial\alpha^{\rm t}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{with inverse} \quad J_{\rm wide}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

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 $\partial H(Y, \theta_0, \gamma_0)/\partial\alpha$. (xx where do we first speak of transpose $x^{\rm t}$, and the cumbersomeness of $x = (x_1^{\rm t}, x_2^{\rm t})^{\rm t}$? xx)

(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 F. 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

$$\begin{aligned} n\{H_{n,\rm wide} - H_n(\theta_0, \gamma_0)\} &\to_d -\tfrac{1}{2}U^{\rm t}J_{\rm wide}^{-1}U, \\ n\{H_{n,\rm narr} - H_n(\theta_0, \gamma_0)\} &\to_d -\tfrac{1}{2}U_0^{\rm t}J_{00}^{-1}U_0, \end{aligned}$$

with $H_{n,\text{wide}} = H_n(\widehat{\theta}, \widehat{\gamma})$ and $H_{n,\text{narr}} = H_n(\widetilde{\theta}, \gamma_0)$. Deduce that

$$W_n = 2n(H_{n,\text{narr}} - H_{n,\text{wide}}) \to_d W = U^{\text{t}} J_{\text{wide}}^{-1} U - U_0^{\text{t}} J_{00}^{-1} U_0. \tag{4.9}$$

Show that this limit variable has mean $\text{Tr}(J_{\text{wide}}^{-1} K_{\text{wide}}) - \text{Tr}(J_{00}^{-1} K_{00})$.

(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 actually later (xx point to Ch11 xx); show that

$$Q = J^{11} = (J_{11} - J_{10} J_{00}^{-1} J_{01})^{-1} \tag{4.10}$$

Similarly, we have $J^{00} = J_{00}^{-1} + J_{00}^{-1} J_{01} Q J_{10} J_{00}^{-1}$. Show also that $J^{00} - J_{00}^{-1} = J^{01} J_{10} J_{00}^{-1}$, $J^{10} = -Q J_{10} J_{00}^{-1}$. (xx point to Claeskens and Hjort (2008b, Section 5.4) and to Ch11. xx)

(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

$$\begin{pmatrix} U_0 \\ V \end{pmatrix} = \begin{pmatrix} U_0 \\ U_1 - K_{10} K_{00}^{-1} U_0 \end{pmatrix} \sim \text{N}_{p+q}(0, \begin{pmatrix} K_{00} & 0 \\ 0 & K_{11} - K_{10} K_{00}^{-1} K_{01} \end{pmatrix}).$$

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):

$$\begin{aligned} W &= U_0^{\text{t}}(J^{00} - J_{00}^{-1}) U_0 + (V + K_{10} K_{00}^{-1} U_0)^{\text{t}} Q (V + K_{10} K_{00}^{-1} U_0) \\ &\quad + 2\, U_0^{\text{t}} J^{01} (V + K_{10} K_{00}^{-1} U_0) \\ &= V^{\text{t}} Q V + U_0^{\text{t}}(J^{00} - J_{00}^{-1} + K_{00}^{-1} K_{01} Q K_{10} K_{00}^{-1} + 2\, J^{01} K_{10} K_{00}^{-1}) U_0 \\ &\quad + U_0^{\text{t}}(J^{01} + K_{00}^{-1} K_{01} Q) V + V^{\text{t}}(J^{10} + Q K_{10} K_{00}^{-1}) U_0. \end{aligned}$$

(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

$$\begin{aligned} Z &= J^{10} U_0 + J^{11} U_1 \\ &= J^{10} U_0 + Q(V + K_{10} K_{00}^{-1} U_0) = Q V + (J^{10} + Q K_{10} K_{00}^{-1}) U_0, \end{aligned}$$

Show that $Z \sim \text{N}_q(0, \Sigma_{11})$, with $\Sigma = J^{-1} K J^{-1}$ the sandwich matrix. The point is now to demonstrate that W above is identical to $W' = Z^{\text{t}} Q^{-1} Z$. Verify first that its mean $\text{Tr}(Q^{-1}\Sigma_{11})$ is identical to the formula found above for $\text{E}\, W$. Work out that

$$\begin{aligned} W' &= [Q V + (J^{10} + Q K_{10} K_{00}^{-1}) U_0]^{\text{t}} Q^{-1} [Q V + (J^{10} + Q K_{10} K_{00}^{-1}) U_0] \\ &= V^{\text{t}} Q V + U_0^{\text{t}}(J^{01} + K_{00}^{-1} K_{01} Q) Q^{-1} (J^{10} + Q K_{10} K_{00}^{-1}) U_0 \\ &\quad + U_0^{\text{t}}(J^{01} + K_{00}^{-1} K_{01} Q) V + V^{\text{t}}(J^{10} + Q K_{10} K_{00}^{-1}) U_0. \end{aligned}$$

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) (xx special case $J = K$: show that $W \sim \chi^2_q$. what about $K = cJ$, do we have $W \sim \chi^q_q$? xx)

(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 $\phi = g(\theta)$. xx)

Ex. 4.51 *Minimisers of convex processes, I.* We have seen useful constructions, methods, and results for minimum criterion function estimators in Ex. 4.47 and 4.49. 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 criterion function $h(t,\theta)$ in $H(\theta) = \mathrm{E}_F\, h(Y,\theta)$ is not smooth. Here we give the basics for how matters simplify, with weaker conditions, if the criterion 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

$$P(\|\alpha_n - \beta_n\| \geq \delta) \leq P\{\Delta_n(\delta) \geq \tfrac{1}{2}h_n(\delta)\},$$

where

$$\Delta_n(\delta) = \sup_{\|s-\beta_n\|\leq\delta} |A_n(s) - B_n(s)| \quad \text{and} \quad h_n(\delta) = \inf_{\|s-\beta_n\|=\delta} B_n(s) - B_n(\beta_n).$$

(c) Basic corollary: Suppose $A_n(s)$ is convex and can be represented as $\frac{1}{2}s^{\mathrm{t}}Js + U_n^{\mathrm{t}}s + C_n + r_n(s)$, where J is symmetric and positive definite, U_n is stochastically bounded, C_n is arbitrary, and $r_n(s) \to_{\mathrm{pr}} 0$ for each s. For the approximation $B_n(s) = \frac{1}{2}s^{\mathrm{t}}Js + U_n^{\mathrm{t}}s + C_n$, 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 $\alpha_n - \beta_n \to_{\mathrm{pr}} 0$; and (ii) that $A_{n,\min} - B_{n,\min} \to_{\mathrm{pr}} 0$.

(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 $-\frac{1}{2}U^{\mathrm{t}}J^{-1}U$. These two statements are what we worked hard for in Ex. 4.48, 4.50 see (4.8)–(4.9), 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. 4.52 *Minimisers of convex processes, II.* The framework worked with now is as in Ex. 4.47 and 4.49, with $Y_1, \ldots, Y_n$ being i.i.d. from some F, a possibly multidimensional parameter θ_0 defined as the minimiser of $H(\theta) = \int h(y,\theta)\,{\rm d}F(y)$, with estimator $\widehat\theta$ the minimiser of the criterion function $H_n(\theta) = \int h(y,\theta)\,{\rm d}F_n(y) = (1/n)\sum_{i=1}^n h(Y_i,\theta)$. 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. 4.51, and allows for nonsmooth criterion functions.

(a) With $h(t,\mu) = |y-\mu|$, show that $\mu_0 = {\rm med}(F)$, 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 $\mu_0 = F^{-1}(q)$, the q quantile.

(b) For an $\alpha \ge 1$, consider the parameter θ_0 being the minimiser of ${\rm E}_F\,|Y-\theta|^\alpha$. Show that the criterion 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 F, 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 $K = {\rm Var}_F\,D(Y)$ 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 F with a density f positive at the median μ. For the median criterion function $|y-\mu|$, show that

$$|y-(\mu+\varepsilon)| - |y-\mu| = D(y)t + R(y,\varepsilon),$$

with $D(y) = I(y \le \mu) - I(y > \mu)$ and

$$R(y,\varepsilon) = \begin{cases} 2\{\varepsilon - (y-\mu)\}\,I(\mu \le y \le \mu+\varepsilon) & \text{if } \varepsilon > 0, \\ 2\{(y-\mu)-\varepsilon\}\,I(\mu+t \le y \le \mu) & \text{if } \varepsilon < 0, \end{cases}$$

with $R(y,0) = 0$. Verify from this that ${\rm E}\,R(Y,\varepsilon) = f(\mu)\varepsilon^2 + o(\varepsilon^2)$ and ${\rm E}\,R(Y,\varepsilon)^2 = (4/3)f(\mu)|\varepsilon|^3 + o(|\varepsilon|^3)$. Deduce that $\sqrt{n}(M_n-\mu) \to_d {\rm N}(0, 1/\{4f(\mu)^2\})$.

(e) Generalise to the case of $\mu_q = F^{-1}(q)$, with $Q_{n,q} = F_n^{-1}(q)$ the empirical q quantile. Show in fact that

$$Z_n(q) = \sqrt{n}(Q_{n,q} - \mu_q) = -f(\mu_q)^{-1}\sqrt{n}\{F_n(\mu_q) - q\} + \varepsilon_n(q),$$

where $\varepsilon_n(q) \to_{\rm pr} 0$ for each q. Derive from this that the limit is a $\mathrm{N}(0, q(1-q)/f(\mu_q)^2)$. (xx point back to sample quantiles things from Ch2. also Ch9. but here easier with different tools, and we get a nice representation. xx)

(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

$$\sqrt{n}(M_{n,\alpha} - \xi_\alpha) \to_d \mathrm{N}(0, \tau_\alpha^2) \quad \text{with } \tau_\alpha^2 = \frac{\mathrm{E}\,|Y-\xi_\alpha|^{2(\alpha-1)}}{\{(\alpha-1)\,\mathrm{E}\,|Y-\xi_\alpha|^{\alpha-2}\}^2}.$$

(xx if distribution is symmetric, the ξ_α is the same for each α. check τ_α for $\alpha \in [1,2]$, the range from median to mean, for the normal, for the Laplace, perhaps the t_ν. check carefully the thing with $\alpha \to 1$. xx)

Ex. 4.53 *Minimisers of processes, III.* (xx to be done. lifting minimum criterion function estimators, with profiling, under smoothness or via convexity, from non-i.i.d., to regression type situations. xx) minimising $\sum_{i=1}^n h(y_i, \theta, x_i)$. having ML in regression in mind.

(a) (xx setup: $\widehat\theta$ minimising $H_n(\theta) = \sum_{i=1}^n h(y_i, x_i, \theta)$, where $Y_i \,|\, x_i$ follows $f(y_i \,|\, x_i, \theta)$. ergodic assumption, that all averages $(1/n)\sum_{i=1}^n p(x_i)$ have limits $\int p(x)\,\mathrm{d}Q(x)$. again we do

$$A_n(s) = n\{H_n(\theta_0 + s/\sqrt{n}) - H_n(\theta_0)\} = U_n^{\rm t} s + \tfrac12 J_n s + r_n(s),$$

with $U_n = (1/\sqrt{n})\sum_{i=1}^n h'(y_i, x_i, \theta_0)$ and $J_n = (1/n)\sum_{i=1}^n h''(y_i, x_i, \theta_0)$. now need Lindeberg for U_n, with $K_n = (1/n)\sum_{i=1}^n \mathrm{Var}\, h'(y_i, x_i, \theta_0) \to K$. with regularity conditions on smoothness, we again have $\sqrt{n}(\widehat\theta - \theta_0) = -J_n^{-1}U_n + o_{\rm pr}(1)$ and $A_{n,\min} \to_d -\frac12 U^{\rm t} J^{-1} U$. subtle things here, when going for least false. there is a least false $\theta_{0,n}$ depending on $x_1, \ldots, x_n$, so we reach results for $\sqrt{n}(\widehat\theta - \theta_{0,n})$. applications to regression estimators later on. again, point to ML in Ch5. xx)

Ex. 4.54 *Minimum criterion function estimators in practice.* (xx to be done suitably. the point is to explain that we need estimating J and K. something for profiling. when we come to Ch5 theory, we point to this exercise for basics. xx) In previous exercises we have learned that $\widehat\theta$, the minimiser of $H_n(\theta) = (1/n)\sum_{i=1}^n h(Y_i, \theta)$, is a natural estimator for θ_0, the minimiser of $\mathrm{E}_F\, h(Y, \theta)$, and that its distribution approaches normality. In order to use such results in practice, for setting confidence intervals, etc., we need to estimate the two crucal matrices $J = \mathrm{E}_F\, h''(Y, \theta_0)$ and $K = \mathrm{Var}_F\, h'(Y, \theta_0)$.

(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 $p_n = (1/n)\sum_{i=1}^n p(Y_i, \theta_0)$ tends to $p_0 = \mathrm{E}_F\, p(Y, \theta_0)$. The best we may do for estimating p_0 in practice is $\widehat p_n = (1/n)\sum_{i=1}^n p(Y_i, \widehat\theta)$. 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 $\widehat p_n \to_{\rm pr} p_0$.

(b) Give conditions under which the natural estimators

$$\widehat{J} = (1/n)\sum_{i=1}^{n} h''(Y_i, \widehat{\theta}) \quad \text{and} \quad \widehat{K} = (1/n)\sum_{i=1}^{n} h'(Y_i, \widehat{\theta})h'(Y_i, \widehat{\theta})^{\mathrm{t}}$$

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 $\Sigma = J^{-1}KJ^1$.

(c) Explain how confidence intervals for the components of θ may be read off from this. More generally, for any focus parameter $\phi = g(\theta)$, with estimator $\widehat{\phi} = g(\widehat{\theta})$, 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 $\widehat{c} = \partial g(\widehat{\theta})/\partial\theta$.

(d) To illustrate how this 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. 4.47. Explain that this means minimising the criterion function

$$H_n(\xi,\sigma) = \int f(y,\xi,\sigma)^2\,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n} f(y_i,\xi,\sigma) = \frac{\frac{1}{2}/\pi^{1/2}}{\sigma} - \frac{2}{n}\sum_{i=1}^{n}\frac{1}{\sigma}\phi\Big(\frac{y_i-\xi}{\sigma}\Big)$$

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 $p(y_0) = P(Y \geq y_0)$, with say $y_0 = 1.00$.

(e) (xx profiling too, to illustrate, for $p(y_0)$. again handled by the general theory. intervals need $k = c^{\mathrm{t}}J^{-1}KJ^{-1}c/c^{\mathrm{t}}J^{-1}c$. xx)

(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.

Ex. 4.55 *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.

(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, and (ii) that $A_k \to_d X$ as $k \to \infty$. In addition, assume that

$$\limsup P(d(X_n, A_{n,k}) \geq \varepsilon) = 0 \quad \text{for each } \varepsilon > 0.$$

Show that $X_n \to_d X$.

(b) (we find a simple application or two here, before we use the two-step method in bigger setups. xx)

Ex. 4.56 *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. 4.55 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. 4.57 *A CLT for m-dependent variables.* In natural generalisation of Ex. 4.56, 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. 4.56 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: $\sqrt{n}\widehat{\rho}(j) \to_d \mathrm{N}(0,1)$, for each j. xx)

Ex. 4.58 *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. 4.59 *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 $\mathrm{E}\,(Y-\theta)^3=\theta$, 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. 4.60 *Limiting normality of the sample variance matrix.* (xx can be better placed, inside Ch 3. results are used for Ex. 5.38. 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. 4.61 *Summing geometrically many terms.* Suppose $Y_1, Y_2, \ldots$ are i.i.d. with mean zero, variance σ^2, and moment-generating function $M_0(t)$. With $P(N = n) = (1-p)^{n-1}p$ 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. 1.42. Show that Z_p has variance σ^2, and that its moment-generating function 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. 1.27.

(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. 1.27.(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 $N \mid \lambda \sim \text{Pois}(\lambda/p)$, and with λ having its own distribution with mean and variance λ_0 and τ_0^2, say. Show that $\text{E}\, pN = \lambda_0$ and that $\text{Var}\, pN \to \tau_0^2$. Also consider the special case for this setup where $\lambda \sim \text{Gam}(a, b)$. From

$$\text{E}\,\{\exp(-spN) \mid \lambda\} = \exp[-(\lambda/p)\{1 - \exp(-sp)\}]$$

deduce

$$\text{E}\,\exp(-spN) = \frac{1}{[1 + (1/b)(1/p)\{1 - \exp(-sp)\}]^a} \to \frac{1}{(1 + s/b)^a},$$

and that $pN \to_d V_{a,b}$, another $\text{Gam}(a, b)$. With the construction above, the normal scale mixture variable X has mgf $1/(1 - \frac{1}{2}t^2/b)^a$, and density

$$\bar f(x) = \int \phi(x/v^{1/2})(1/v^{1/2}) \frac{b^a}{\Gamma(a)} v^{a-1} \exp(-bv)\, \text{d}v.$$

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) $\lambda \sim \text{Gam}(1, b)$, with mean $\lambda_0 = 1/n$. (ii) $N \mid \lambda \sim \text{Pois}(\lambda/p)$. write down $\text{E}\,(pN \mid \lambda)$ and $\text{Var}\,(pN \mid \lambda)$, 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 $\lambda_t \sim \text{Gam}(1, b_t)$. xx)

Ex. 4.62 *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)

the Gumbel distribution

We define the *Gumbel distribution* on the real line by its cumulative distribution function $G_0(u) = \exp\{-\exp(-u)\}$.

(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 $L_0(s) = \text{E}\,\exp(-sU) = \Gamma(1 + s)$, in terms of the gamma function.

the Euler–Mascheroni constant

(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.

(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 $P(-1.097 \le U \le 2.970) = 0.90$.

(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 2.22 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. 4.63 *Weibull and Gamma maxima.* The basic result of Ex. 4.62, 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. 4.64 *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. 1.15.

(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, show that $(Z_n - n \log n)/n \to_d U$, the Gumbel.

(c) Deduce from this that

$$n^s \prod_{i=1}^{n} \frac{(i/n)\exp(-s/n)}{1 - (i/n)\exp(-s/n)} \to \Gamma(1+s).$$

Use the Stirling formula, see Ex. 4.31, to give an approximation to $\sum_{i=1}^{n} \log\{1-(i/n)\exp(-s/n)\}$.

(d) (xx one more thing here. xx)

Ex. 4.65 *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.6 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 $N_i \geq 1$. Show that $W_i \sim \text{geom}(1/n)$, with mean n. These are however

dependent, so the Gumbel limit result of Ex. 4.64 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. 1.15. 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. 4.62. 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) With $U_n' = (V_n - n\log n)/n = \bar T_n - \log n$, show that

$$\begin{aligned} \mathrm{E}\, \exp(-sU_n') &= \exp(s\log n) \prod_{i=1}^n \mathrm{E}\, \exp\{-(s/n)T_i\} \\ &= n^s \prod_{i=1}^n \frac{(i/n)\exp(-s/n)}{1 - (1-i/n)\exp(-s/n)}. \end{aligned}$$

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 $G_n(v) = P(V_n \le v)$ 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

$$\begin{aligned} 1 - G_n(v) &= P(B_1 \cup \cdots \cup B_n) \\ &= \binom{n}{1} P(B_1) - \binom{n}{2} P(B_1 \cap B_2) + \binom{n}{3} P(B_1 \cap B_2 \cap B_3) - \ldots \\ &= \sum_{j=1}^{n} (-1)^{j-1} \binom{n}{j} (1 - j/n)^v, \end{aligned}$$

for $v \ge n$. Use algebra to also derive

$$g_n(v) = P(V_n = v) = \sum_{j=1}^{n} (-1)^{j-1} \binom{n-1}{j-1} (1 - j/n)^{v-1} \quad \text{for } v \ge n.$$

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 $-4 \cdot 1 + 6 \cdot 16 - 4 \cdot 81 + 256 = 24$, 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 $C_r(n) = P_n(W_r < W_{r,\mathrm{obs}}) + \frac{1}{2} P_n(W_r = W_{r,\mathrm{obs}})$. how many Italians in my neighbourhood? xx)

Ex. 4.66 *The 2nd largest, 3rd largest, etc., for exponentials.* Let as in Ex. 4.62 $X_1, \ldots, X_n$ 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

$$g_a(w) = \Gamma(a)^{-1} \exp\{-aw - \exp(-w)\} \tag{4.11}$$

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 (4.11) a generalised Gumbel.

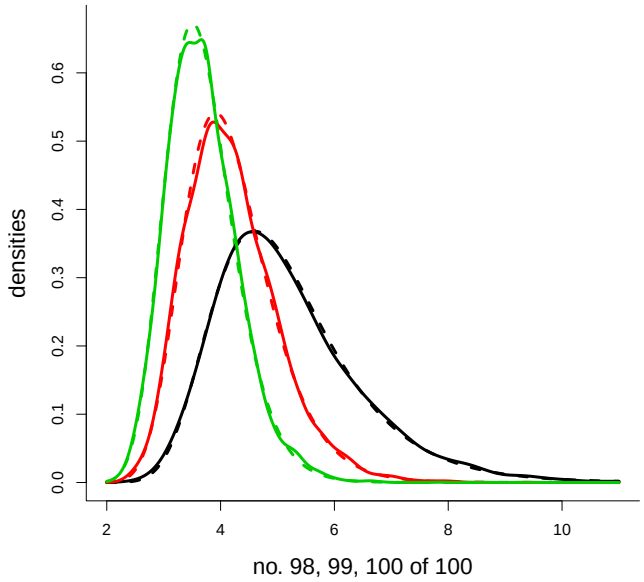

Figure 4.2: *For $n = 100$, the dashed curves are densities for order statistics 98, 99, 100 for i.i.d. exponentials, computed via the limits $\log n + W_i$ for $i = 1, 2, 3$. The full curves are empirical densities based on having simulated 10^4 outcomes for each.*

(b) Deduce that $W_a = \log(1/V_a)$ has mean $-\psi(a)$ and variance $\psi'(a)$. (xx pointer to digamma function. approximations: $\psi(a) \doteq \log a - \frac{1}{2}(1/a)$, $\psi'(a) \doteq 1/a - \frac{1}{2}(1/a^2)$, for growing a. xx)

(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. 4.62. Show that

$$P(X_{(n-i+1)} - \log n \le w) = P(U_{(n-i+1)} \le 1 - (1/n)\exp(-w)),$$

in terms of the order statistics for the uniform.

(d) Use the Beta connection of Ex. 2.20 to deduce that the density of $W_{n,i}$ may be written

$$g_{n,i}(w) = \text{be}(1 - (1/n)\exp(-w), n - i + 1, i)(1/n)\exp(-w)$$

in terms of the Beta density with parameters $(n - i + 1, i)$. Take the limit to prove that $W_{n,i} \to_d W_i$.

(e) (xx something to round this off, an approximation for say the 3rd largest. also point to the representation $X_{(n-i+1)} = D_1 + \cdots + D_{n-i+1}$ with scaled exponentials. Figure 4.2 shows that the theory works well for $n = 100$. the figures features kernel

density estimates, using methods of Chapter 13. can make a Story out of this. insurance company cares for these most extreme outcomes. xx)

Ex. 4.67 *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 moment-generating function $M_0(s) = \mathrm{E}\exp(sU_i)$ 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 $P(J_i = 1) = 1/i$, $P(J_i = 0) = 1 - 1/i$. 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 moment-generating function, 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 $P(U_i = 1) = P(U_i = -1) = \frac{1}{2}$. 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 moment-generating function $M_0(s)$ for the U_i, and then the moment-generating function $M(t)$ 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.

Ex. 4.68 *Characterisations of the normal and the Cauchy.* There are many characterisation theorems in probability theory, results saying that appropriate assumptions or properties fully characterise certain distributions- Here we give a few such, using characteristic functions. (xx check with care. xx)

(a) Suppose F is a distribution, symmetric around zero, such that if X and Y are independent from this distribution, then $(X+Y)/\sqrt{2}$ has the same distribution. Prove that the distribution is normal.

(b) Suppose F is a distribution, symmetric around zero, such that if X and Y are independent from this distribution, then $(X+Y)/2$ has the same distribution. Prove that the distribution is a Cauchy.

(c) (xx one more. xx)

Ex. 4.69 *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 $P(J_i^* = 1) = p = m/n$. 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 $(Z_n - \frac12 n^2 p)/n^{3/2}$ 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 sal-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. 4.70 *Leftovers Ch4.* (xx just a preliminary little station for things that could be put in, perhaps even inside other exercises, in Ch4 xx)

(a) somewhere: suppose $(X_n, Y_n) \to_d (X, Y)$, a binormal zero-mean limit. Show that $X_n \,|\, (|Y_n| \le \varepsilon) \to_d X \,|\, (|Y| \le \varepsilon)$, for each $\varepsilon > 0$, and that $X \,|\, (|Y| \le \varepsilon) \to_d X \,|\, (Y = 0)$ as $\varepsilon \to 0$. Use Ex. 4.55 to put up a condition securing $X_n \,|\, (|Y_n| \le \varepsilon_n) \to_d X \,|\, (Y = 0)$ for ε_n going to zero (at a certain rate?).

(b) 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. 4.20 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. 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 $Z_n^* = \sqrt{n}\bar{X}_n + \delta_n$, with $\delta_n \sim \mathrm{N}(0, \varepsilon_n^2)$. Show that Z_n^* has density

$$f_n^*(z^*) = 1/(2\pi) \int \exp(-itz^*)\varphi(t/\sqrt{n})^n \exp(-\tfrac12 \varepsilon_n^2 t^2)\,\mathrm{d}t,$$

and that this again converges to the normal density $\sigma^{-1}\phi(\sigma^{-1}z^*)$ provided merely that $\varepsilon_n \to 0$.

(c) 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 $Z_n^* = \sqrt{n}\bar{X}_n + \delta_n$, with some small Gaussian added noise $\mathrm{N}_p(0, \varepsilon_n^2\Sigma)$ with $\varepsilon_n \to 0$, has a density $f_n^*(z^*)$ which tends to the same $\mathrm{N}_p(0, \Sigma)$ density.

(d) Suppose $(A_n, B_n)^{\mathrm{t}} \to_d (A, B)^{\mathrm{t}}$, a zero-mean binormal. If there is also density convergence, with (A_n, B_n) having density $f_n(a, b)$ tending to the appropriate binormal density $f(a, b)$, show that $A_n \,|\, (|B_n| \le \varepsilon_n)$ tends to $A \,|\, (B = 0)$, as long as $\varepsilon_n \to 0$. Show that

the same limiting distribution statement holds also when (A_n, B_n) has a discrete distribution, using the 'adding small Gaussian noise to get densities' trick. (xx emil: check the arguments here; does this save our Wilcoxon problem, with stable convergence implied by the regularity assumptions of the CLT and Lindeberg? xx)

4.C Notes and pointers

(xx to come. we point to certain famous things from the past: Kolmogorov (1933), 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 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 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)

ToDo notes, of 13-Aug-2023.

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 $\widehat{\mu}^* = \{1 - c(D_n)\}\widehat{\mu}_{\text{narr}} + c(D_n)\widehat{\mu}_{\text{wide}}$ from model selection. And point to $n^{2/5}$ rates for f estimation.

I.5

Likelihood inference

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. Parts of the theory follow readily from results reached in Ch. 4 for general minimum criterion function estimators, with the criterion function seen to be related to to the Kullback–Leibler divergence. The likelihood methods are practical and versatile, as is demonstrated in several exercises, 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. 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, so results are established outside model conditions. (xx perhaps there is room for empirical likelihood. xx)

5.A Chapter introduction

A fundamental concept in parametric statistical inference is that of *the likelihood function*, most often worked with on logarithmic scale, i.e. via *the log-likelihood function*. Suppose in general terms that a model for observations Y, perhaps a vector or a data matrix, can be expressed as $f_{\text{full}}(y, \theta)$, the full or simultaneous model density, in terms of the model parameter $\theta = (\theta_1, \ldots, \theta_p)^{\text{t}}$, of dimension say p. Then the likelihood function is $L(\theta) = f_{\text{full}}(y_{\text{obs}}, \theta)$, viewed as a function of the model parameters, with y held fixed at the observed value y_{obs}. Similarly the log-likelihood function is

log-likelihood function

$$\ell(\theta) = \log f_{\text{full}}(y_{\text{obs}}, \theta). \tag{5.1}$$

When Y is a vector $Y_1, \ldots, Y_n$ of independent observations, perhaps with separate model densities $f_1(y_1, \theta), \ldots, f_n(y_n, \theta)$, as in a multitude of regression setups, we have log-likelihood $\ell_n(\theta) = \sum_{i=1}^n \log f_i(y_{i,\text{obs}}, \theta)$. This chapter is about the long list of basic

statistical tools, methods, results, applications, associated with the likelihod and log-likelihood functions (with yet more to come in Chs. 6 and 7, with Bayesian methods and with confidence distributions). A fundamental quantity is the *maximum likelihood (ML) estimator*, the $\widehat{\theta} = \widehat{\theta}(y_{\rm obs})$ maximising the likelihood function (equivalent to maximising the log-likelihood function). Via our exercises a full, versatile, very fruitful theory is being developed, built around the maximum likelihood estimator, enabling statistical inference for model parameters, functions thereof, assessing lower-dimensional models, prediction of future outcomes, comparing models, etc. (xx two more paragraphs, also pointing to Stories. highlight the versatility and broad usefulness; the theory works and can be applied also for the not yet invented models. we include basic material for empirical likelihood too. xx)

the ML estimator

[xx calibrate with what's in the abstract. telling the readers about the themes to be worked with. Likelihood, estimators, tests, confidence, approximate multinormality, Wilks theorems and tests, power. lots of uses. Occasionally we can deduce the exact distributions, of ML estimators and test statistics, but more broadly we rely on large-sample approximations to normal and chi-squared distributions. we also include a few on *empirical likelihood*, drawing on Hjort et al. (2009, 2018). 'Learning outcome': you can now construct your own parametric models, fit them via ML, and have clear inference for any focus parameter. So versality! xx]

(xx ToDo, mainly nils: do ML regression outside models, for linear, logistic, Poisson, a bit more; spell out sandwich estimation; Wilks profiling gives $k\chi^2_1$, etc. clean the expofamily exercises. do ML and Wilks from smooth representations, as opposed to deriving it from the regularity conditions things. if we include cc(ϕ), then need to have this on board in one exercise. xx)

5.B Short and crisp

Ex. 5.1 *Log-likelihood for the binomial.* You throw your perhaps not perfectly balanced die $n = 60$ times, and the number of times you have a '1' is $y = 8$.

(a) Draw the log-likelihood function for the probability $\theta = P(\text{having a '1'})$, and show that its maximum occurs at $\widehat{\theta} = 8/60$. In the same diagram, draw the log-likelihood function for the case of having observed $y = 12$, and comment.

(b) With a general y from the binomial (n, θ), set up the log-likelihood function $\ell_n(\theta)$, and show that its maximiser is $\widehat{\theta} = y/n$. Find also an expression for $\widehat{J} = -\ell_n''(\widehat{\theta})$, the sharpness of the peak.

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) Write down the log-likelihood function $\ell_n(\theta)$, and show that it is maximised at $\widehat{\theta} = 1/W_n$, with $W_n = (1/n)\sum_{i=1}^n y_i^{1/2}$. Find also a formula for the Hessian at the maximum position, $\widehat{J} = -\ell_n''(\widehat{\theta})$.

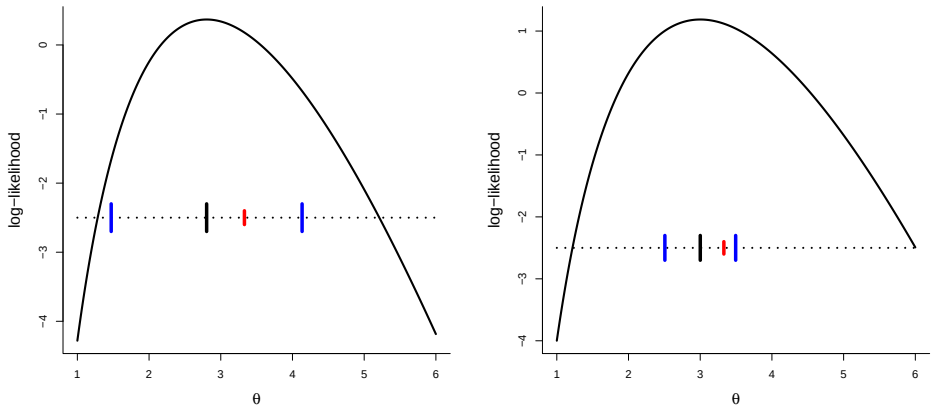

Figure 5.1: *Left panel: The log-likelihood function $\ell_n(\theta)$ for the simple $n = 12$ dataset of Ex. 5.2. The maximum is attained at $\widehat{\theta} = 2.803$. The blue bars indicate the 90 percent confidence interval coming from standard ML estimation theory, see Ex. 5.7, and is $[1.472, 4.134]$. 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 $[2.507, 3.494]$.*

(b) Find the limit distribution of $\sqrt{n}(W_n - 1/\theta)$, using the central limit theorem, and use the delta method to find that the limit distribution of $\sqrt{n}(\widehat{\theta} - \theta)$ is $\mathrm{N}(0, \theta^2)$.

(c) Let θ_0 denote the true parameter underlying the data. Theory to come in later exercises (xx xref xx) says that the distribution of $\widehat{\theta}$, though not normal itself, can be approximated with a $\mathrm{N}(\theta_0, 1/\widehat{J})$ (at least when sample size n is moderate or large). Show that this agrees fully with the limit distribution result reached above. Show further 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 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 distributions (xx pointer xx) 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 *The maximum likelihood estimator.* [xx we work through some examples, setting up the $\ell_n(\theta)$, find the ML, brief comments, no hard theory yet. but we make a little point of being fully able do deal with $\sqrt{n}(\widehat{\theta} - \theta)$ in situations where the ML is a smooth function of sample averages, with more general theory to come below. xx]

(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) 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.

(d) (xx one more. xx)

Ex. 5.4 *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. 1.23, 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 to be turned to in Ex. 5.6, 5.7.

(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. 2.10.

(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 to be worked with in Ex. 5.7; also, we shall again find that ML estimators are more precise than the moment estimators.

(e)

score function

Ex. 5.5 *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

$$u(y, \theta) = \partial \log f(y, \theta)/\partial\theta \quad \text{and} \quad i(y, \theta) = \partial^2 \log f(y, \theta)/\partial\theta\partial\theta^{\mathrm{t}},$$

called the *score function*, with p components, and the *observed 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 that assume 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) Show that the score function has mean zero, that is

$$\mathrm{E}_\theta\, u(Y, \theta) = \int f(y, \theta) u(y, \theta)\, \mathrm{d}\mu(y) = 0.$$

the Fisher information matrix

Let next

$$K(\theta) = \mathrm{Var}_\theta\, u(Y, \theta) \quad \text{and} \quad J(\theta) = -\mathrm{E}_\theta\, i(Y, \theta),$$

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 data set 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. In fact, show that $-i(y,\theta) = K(\theta)$ for all exponential families with the natural parametrisation, see Ex. 1.57.

(c) 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

$$J(\xi,\sigma) = \mathrm{Var}_{\xi,\sigma}\, u(Y,\xi,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

(d) (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)

(e) 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.

(f) Above it was assumed that we could pass the derivative under the integral sign. Here is a lemma: Let $g(y,\theta)$ be a function depending on a one-dimensional parameter θ, and μ a measure on the measurable space in which y lives. If $\partial g(y,\theta)/\partial\theta$ exists and is continuous in θ for all y and all θ in an open interval Θ_0 of the parameter space; there is an integrable function $k(y)$ dominating $|\partial g(y,\theta)/\partial\theta| \le k(y)$ for all $\theta \in \Theta_0$; and if $\int g(y,\theta)\,\mathrm{d}\mu(y)$ exists on Θ_0, then

Derivative under the integral sign

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int g(y,\theta)\,\mathrm{d}\mu(y) = \int \frac{\partial}{\partial\theta}\, g(y,\theta)\,\mathrm{d}\mu(y).$$

You may combine the fundamental theorem of calculus and the dominated converge theorem to prove this lemma.

Ex. 5.6 *Maximum likelihood asymptotics with log-concave likelihood.* (xx edit with care. regularity condition on R. point to expofamily. xx) Suppose $Y_1,\ldots,Y_n$ are i.i.d. from a density $f(y,\theta)$ 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 $\widehat{\theta}$.

log-concave density

(a) Show first that ℓ_n is concave. To start with mild conditions, assume merely that

$$\log f(y,\theta_0+\varepsilon) - \log f(y,\theta_0) = D(y)^{\mathrm{t}}\varepsilon + R(y,\varepsilon), \tag{5.2}$$

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. 4.51 and 4.52 that $\sqrt{n}(\widehat{\theta}-\theta_0) = J^{-1}U_n+o_{\rm pr}(1)$, where $U_n = (1/\sqrt{n})\sum_{i=1}^n u(Y_i,\theta_0)$. Deduce the two fundamental results

$$\begin{aligned} Z_n &= \sqrt{n}(\widehat{\theta}-\theta_0) \to_d {\rm N}_p(0,J^{-1}), \\ W_n &= 2\{\ell_{n,\max}-\ell_n(\theta_0)\} \to_d \chi^2_p. \end{aligned} \tag{5.3}$$

Show that consistency of the ML estimator for θ_0 is a simple consequence of the first statement.

(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 (5.3) 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 $Y_i \sim {\rm Pois}(e_i\theta)$, with different exposures w_i, the point is non-i.i.d. nils does this after mild extra editing for Ex. 4.51 and 4.52, with a bit of Lindeberg too. xx)

(d) (xx to polish. xx) something nontrivial I. 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. 5.7 *Behaviour of the maximum likelihood estimator, under model conditions.* (xx quite a bit of cleaning required here; point back to Ex. 5.6; use Ex. 4.47 and 4.49. make it part of the narrative how this leads to superuseful versatile tools; any model, any focus parameter, we find limiting normality etc. this goes for the next exercise too. xx) Whereas classes of models actually lead to concave log-likelihoods, many others are outside that nice class. Here we deal with general smooth parametric models, establishing versions of (5.3) also outside log-concavity. Let $Y_1,\ldots,Y_n$ be independent from the same density $f(y,\theta)$, where $\theta = (\theta_1,\ldots,\theta_p)^{\rm t}$. Assuming two smooth derivatives, let as in Ex. 5.5 $u(y,\theta)$ and $i(y,\theta)$ be the score function and information function. The log-likelihood is $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i,\theta)$, with first order derivative $U_n(\theta) = \sum_{i=1}^n u(Y_i,\theta)$, and second order derivative $I_n(\theta) = \sum_{i=1}^n i(Y_i,\theta)$, a $p\times p$ matrix. The ML estimator $\widehat{\theta} = \widehat{\theta}_n$ based on the first n observations maximises $\ell_n(\theta)$ and is also a solution to $U_n(\widehat{\theta}) = 0$.

(a) Assume that the model is correct for a certain true parameter point θ_0. Show that $n^{-1}\ell_n(\theta)$ converges with probability 1 to a function $C(\theta)$ which attains its maximum value for $\theta = \theta_0$. This suggests that the maximiser $\widehat{\theta}_n$ of $n^{-1}\ell_n(\theta)$ should tend with probability 1 to the maximiser θ_0 of the limit function. – A rigorous proof requires certain regularity conditions to hold. Try to construct such a proof and see what kind of conditions would suffice. [xx is this all we say concerning the consistency of maximum likelihood estimators? xx]

(b) Maximising $\ell_n(\theta)$ is of course the same as minimising $-\sum_{i=1}^n \log f(Y_i,\theta)$, a criterion function in the language of Ex. 4.47 and 4.49. Establish hence the link from the present likelihood framework to the setup of these exercises, and use results from these to establish the two fundamental results of (5.3), i.e. without log-concavity. Put up a set of sufficient regularity conditions, perhaps utilising the efforts of Ex. 4.48. Note that the Bartlett identity, that the two matrices J and K of 5.5 are identical, under model conditions, plays a role here.

(c) Argue for the delta method consequence of the above: if $\phi = g(\theta)$ is some parameter of interest, a smooth function of the basic model parameter vector, then $\widehat{\phi} = g(\widehat{\theta})$ is the ML estimator, and $\sqrt{n}(\widehat{\phi}-\phi) \to_d c^{\rm t} J(\theta_0)^{-1} U \sim {\rm N}(0,\tau^2)$, with $\tau^2 = c^{\rm t} J(\theta_0)^{-1} c$. Here $c = \partial g(\theta_0)/\partial\theta$. Explain how this may be used to form confidence intervals for ϕ.

(d) How can you test the hypothesis $\theta_1 = \theta_1^0$, where θ_1^0 is a specified value? Also give an approximate 90 percent confidence interval for θ_1.

(e) Recall that if $X \sim {\rm N}_p(\mu,\Sigma)$, then $(X-\mu)^{\rm t}\Sigma^{-1}(X-\mu)$ is χ^2 distributed with p degrees of freedom; see Ex. 1.33. Construct an approximate 90 percent confidence ellipsoïd for the unknown parameter vector. Can you prove that your chosen region has the minimal possible volume, among all asymptotic 90 percent confidence regions for θ?

Ex. 5.8 *Differentiability in quadratic mean.* The key to proving asymptotic normality of the maximum likelihood estimator for the log-concave densities in Ex. 5.6 was the assumption that $\log f(y,\theta_0+\varepsilon) - \log f(y,\theta_0) = h^{\rm t} D(y) + R(y,\varepsilon)$, 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.

Classical maximum likelihood conditions

(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 $\sqrt{n}(\widehat{\theta}_n - \theta_0) = J(\theta_0)^{-1} n^{-1/2} U_n(\theta_0) + o_p(1)$. Do it.

(b) Retain the classical assumptions from (a), except the consistency assumption (as it plays no role here). Show that the expansion in (5.2) of Ex. 5.6(a) holds.

(c)

(d)

Ex. 5.9 *Wilks theorems, I.* (xx more rydding required, for a couple of versions of Wilks, but here we do the first crucial thing, with χ^2_1 for a one-dimensional focus parameter, still under model conditions. we exploit Ex. 4.47 and 4.49 fully and get the essential things for free. xx) In Ex. 5.6 and 5.7 we have seen that ML estimation is a special case of minimum criterion function estimation, as handled in Ex. 4.47 (xx and other Ch4 exercises xx). Methods and results from Ex. 4.48 and 4.51 led readily to the basic results (5.3), under appropriate conditions. There are further fruits within easy reach, however, via the profiling constructions of Ex. 4.49 and 4.50. Assume i.i.d. observations $Y_1,\ldots,Y_n$ follow a parametric model $f(y,\theta)$, of dimension p, with log-likelihood function $\ell_n(\theta)$.

log-likelihood profile, deviance function

(a) For a smooth focus parameter $\phi = g(\theta) = g(\theta_1, \ldots, \theta_p)$, define the profile *log-likelihood profile function* and associated *deviance function*

$$\ell_{n,\text{prof}}(\phi) = \max\{\ell_n(\theta)\colon g(\theta) = \phi\}, \quad D_n(\phi) = 2\{\ell_{n,\max} - \ell_n(\phi)\}.$$

Wilks theorem

Show, under regularity conditions akin to those of the exercises pointed to, that $D_n(\phi_0) \to_d \chi^2_1$ at the true θ_0, where $\phi_0 = g(\theta_0)$. This is called the *Wilks theorem* (or perhaps a Wilks theorem, there being several such, of varying generality, see below). Note that the Bartlett identity plays a role here, in that the two matrices in the sandwich matrix $J^{-1}KJ^{-1}$, in the notation of Ex. 4.49 and 4.50, are equal, under model conditions.

(b) An important consequence of the Wilks theorem is that confidence intervals can be constructed from the profile and deviance. Show that

$$I_n = \{\phi\colon D_n(\phi) \le \gamma_{1,\alpha}\} \quad \text{has } P_\theta(I_n) \to \alpha,$$

with $\gamma_{1,\alpha} = \Gamma_1^{-1}(\alpha)$ the appropriate χ^2_1 quantile. The Wilks theorem hence allows construction of confidence intervals, at any given level, for any focus parameters. (xx about the large-sample equivalence to $\widehat{\phi} \pm 1.645\,\widehat{\kappa}/\sqrt{n}$. xx)

(c) More generally, looking and testing for lower-dimensional structure, consider the setup of Ex. 4.50, with a parameter vector $\alpha = (\theta, \gamma)$ of dimension $p + q$. We think in terms of a wide model, with these $p + q$ parameters being free, and a narrow model of dimension p, where $\gamma = \gamma_0$, a prespecified value. With $(\widehat{\theta}, \widehat{\gamma})$ the ML estimator in the wide model, and $(\widetilde{\theta}, \gamma_0)$ the ML estimator in the narrow model, define the log-likelihood maxima $\ell_{\max,\text{wide}} = \ell_n(\widehat{\theta}, \widehat{\gamma})$ and $\ell_{\max,\text{narr}} = \ell_n(\widetilde{\theta}, \gamma_0)$. Use results from the exercises pointed to to show that

$$W_n = 2(\ell_{\max,\text{wide}} - \ell_{\max,\text{narr}}) \to_d W = \chi^2_q.$$

This could start from the representation $W = Z^{\mathrm{t}}Q^{-1}Z$ found there, and then using $Z \sim \mathrm{N}_q(0, Q)$.

(d) (xx more. examples. round off. testing H_0 that $\gamma = \gamma_0$. freedom of parametrisation. point to outside model too. xx)

Ex. 5.10 *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. 5.7 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. 5.5, 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) (xx perhaps something moderately unusual here. xx)

Ex. 5.11 *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. xx)

Ex. 5.12 *ML machinery in practice, I.* (xx put in: nice if explicit formulae may be found, but we do manage quite well without. xx) The aim of the present exercise is to showcase how the general maximum likelihood theory can be applied also in new situations, perhaps with models outside the usual repertoire. Even with a freshly invented model one may fit parameters, read off approximate standard deviations, construct confidence intervals, test hypotheses, as long as the log-likelihood can be programmed. The machinery also applies to any interest functions of the model parameters, via the delta method. It is to be noted that general-purpose numerical optimisation methods and algorithms, along with routines for computing gradients and Hessians, i.e. first and second order derivatives, are wondrously helpful here, essentially with only modest extra efforts needed beyond having programmed the log-likelihood.

Of course methods also apply for standard models, where there might be packages or established routines accomplishing the fitting and testing, but the spirit for the modern statistician should be that of building and trying out also new models for new purposes. This might be even more important for regression type models, where similar programming and implementation schemes work; see Ex. 5.29.

We build our present illustrations around the following dataset, with $n = 201$ time-to-failure measurements for certain machine parts, taken from the SAS User's Guide Ch. 44. We shall fit the data to the gamma and Weibull models, and also to a three-parameter extension of these, which we call the gamma-Weibull model.

```
  [1]  620  470  260   89  388  242  103  100   39  460  284 1285  218  393  106
 [16]  158  152  477  403  103   69  158  818  947  399 1274   32   12  134  660
 [31]  548  381  203  871  193  531  317   85 1410  250   41 1101   32  421   32
 [46]  343  376 1512 1792   47   95   76  515   72 1585  253    6  860   89 1055
 [61]  537  101  385  176   11  565  164   16 1267  352  160  195 1279  356  751
 [76]  500  803  560  151   24  689 1119 1733 2194  763  555   14   45  776    1
 [91] 1747  945   12 1453   14  150   20   41   35   69  195   89 1090 1868  294
```

```
[106]   96  618   44  142  892 1307  310  230   30  403  860   23  406 1054 1935
[121]  561  348  130   13  230  250  317  304   79 1793  536   12    9  256  201
[136]  733  510  660  122   27  273 1231  182  289  667  761 1096   43   44   87
[151]  405  998 1409   61  278  407  113   25  940   28  848   41  646  575  219
[166]  303  304   38  195 1061  174  377  388   10  246  323  198  234   39  308
[181]   55  729  813 1216 1618  539    6 1566  459  946  764  794   35  181  147
[196]  116  141   19  380  609  546
```

(a) Consider the three-parameter density function

$$f(y,a,b,c) = k(a,b,c)y^{a-1}\exp(-by^c) \quad \text{for } y>0.$$

Prove that the normalisation constant must be $k(a,b,c) = cb^{a/c}/\Gamma(a/c)$. Show that $c=1$ gives the $\text{Gam}(a,b)$ distribution, and that the special case $a=c$ corresponds to c.d.f. $F(y) = 1-\exp(-by^c)$, which is a Weibull (though with a different parametrisation than with Ex. 1.40). Due to these special cases we may call this three-parameter family the gamma-Weibull distribution. Prove also the mean formula

$$\text{E}\,Y = \frac{\Gamma(a/c+1/c)}{\Gamma(a/c)}\frac{1}{b^{1/c}},$$

and verify that mean formulae for the gamma and the Weibull indeed are special cases. (xx nils notes: a/b for $c=1$ and $\Gamma(1+1/c)/b^{1/c}$ for $a=c$, the weibull case. xx)

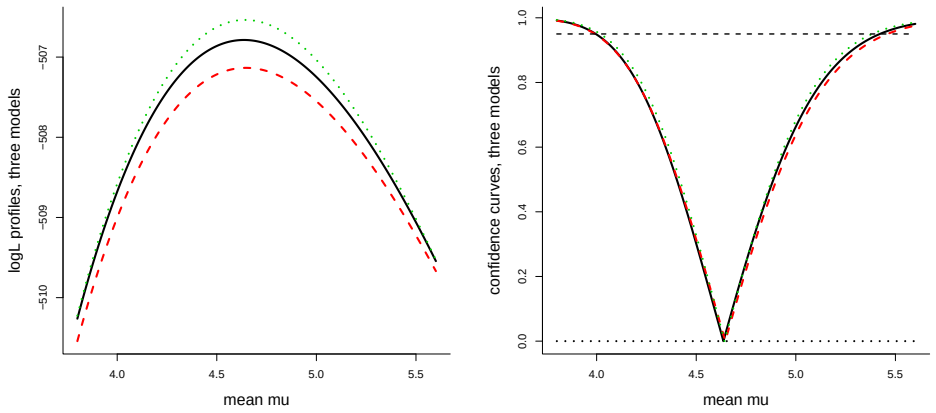

Figure 5.2: *For the time-to-failure data, left panel shows log-likelihood profile functions for the mean μ, via the* $\text{Gam}(a,b)$ *model, the Weibull* (b,c) *model, and the three-parametric* (a,b,c) *model. In the right panel, these are transformed to confidence curves via the Wilks theorem (xx pointer back xx), where e.g. 95 percent intervals can be read off. These are in essential agreement here.*

(b) First read the data into your computer suitably. It helps accurate numerics to scale them with a factor of e.g. 1/100; in the scripts below this is what we call `yy`. Show that the following little script succeeds in computing the log-likelihood for the $\text{Gam}(a,b)$ model:

```
logL1 = function(para)
{
a = para[1]
b = para[2]
aux = (a-1)*log(yy)-b*yy + a*log(b) - lgamma(a)
sum(aux)
}
```

To maximise the log-likelihood, along with the Fisher information matrix $\widehat{J} = -\partial^2 \ell_n(\widehat{\theta})/\partial\theta\,\partial\theta^{\mathrm{t}}$, it is practical to use the general-purpose non-linear minimisation algorithm `nlm` in R. Define the function `minuslogL1` as `-logL1`, and then use

```
starthere1 = c(1,1)
fit1 = nlm(minuslogL1,starthere1,hessian=T)
ML1 = fit1$estimate
Jhat1 = fit1$hessian
se1 = sqrt(diag(solve(Jhat1)))
show1 = cbind(ML1,se1)
```

Carry out this scheme, and explain what the different steps involve and achieve; sometimes a bit of fiddling might be required with the start point `starthere1` to secure convergence of the numerical iterative minimisation procedure. Read off both ML estimates $(\widehat{a}, \widehat{b})$ and approximate 95 percent confidence intervals for them. Test the hypothesis that the data are actually from the simpler exponential model.

(c) For any focus parameter $\mu = \mu(\theta)$ of the model parameters, the delta method says that the approximate variance of $\widehat{\mu}_{\mathrm{ml}} = \mu(\widehat{\theta}_{\mathrm{ml}})$ is $\widehat{\kappa}^2 = \widehat{d}^{\mathrm{t}} \widehat{J}^{-1} \widehat{d}$, with $\widehat{d} = \partial\mu(\widehat{\theta})/\partial\theta$. To illustrate the general machinery, consider the mean $\mu = \mathrm{E}\,Y$, which for the gamma model is the simple a/b. Here partial derivatives etc. are easily found, but to explain the general practical principle start by defining the function `mu1 = function(para)` as `a/b`, and then carry out

```
mu1hat = mu1(ML1)
der1 = grad(mu1,ML1)
kappa1 = sqrt( t(der1) %*% solve(Jhat1) %*% der1 )
```

where `grad` is available via the library `numDeriv`. Construct a 95 percent interval for the mean using this. Modify your code to similarly find estimate and interval for the median.

(d) Having accomplished the above for the gamma model, modify your code to handle also the Weibull model, with $F(y) = 1 - \exp(-by^c)$. Find ML estimates, their estimated standard deviations, test exponentiality; then find estimates and intervals for the mean and the median. Part of the intended experience here is that passing from one model to another often does not take many extra efforts, as results flow from having programmed the log-likelihood.

(e) There are packages and routines available handling the gamma and Weibull models, but perhaps not our three-parameter extension. Programme the appropriate $\ell_n(a, b, c)$, then find ML estimates $(\widehat{a}, \widehat{b}, \widehat{c})$, along with estimates and confidence intervals for the mean and median. For these tasks it is indeed helpful to have an explicit formula for

the normalisation constant $k(a,b,c)$, but it is useful for other models and situations to learn that one may manage without, via numerical integration routines, typically in the format of `integrate(g,0,Inf)$value`. For the learning experience, redo the fitting of the (a,b,c) model without using the $k(a,b,c)$ formula.

(f) (xx flex your ML machinery programming muscles by attempting one or two more models. may take $F(y,\theta,a,b) = \text{Be}(1-\exp(-\theta y),a,b)$. xx)

(g) The methods and programmes above lead to confidence intervals of the first-order large-sample approximation type, say $\widehat{\mu} \pm 1.96\,\widehat{\kappa}$. Supplement your efforts by programming also the log-profile-likelihood functions, $\ell_{n,\text{prof}}(\mu) = \max\{\ell_n(\theta)\colon \mu(\theta)=\mu\}$, for the three models. Here you are helped by having explicit formulae for the μ. Construct a version of Figure 5.2, left panel. The profiles are in good agreement hee, and the three-parameter model does not lead to any significant increase over the two two-parameter models. Then construct a version of the *confidence curves* in the right panel, as follows. With $D(\mu) = 2\{\ell_{n,\max} - \ell_{n,\text{prof}}(\mu)\}$ the deviance, explain that $D(\mu) \approx_d \chi^2_1$, at the true position in the parameter space, via the Wilks theorem. Deduce that $\text{cc}(\mu) = \Gamma_1(D(\mu))$ has the uniform distribution, and that the random set $\{\mu\colon \text{cc}(\mu) \le 0.95\}$ has probability 0.95 of covering the true value. We return at length to such confidence curves and distributions in Ch. 7.

(h) (xx also to be put in, but briefly: the data fit well to each of the models here, and the above at-the-model variance calculations work fine. but do the outside-the-model sandwich things too. also: do wilks. xx)

(i) (xx round off. a figure. point to AIC things. and to Ex. 5.29, where the lifting from i.i.d. to general parametric regression models turns out to be relatively modest. rather similar confidence intervals for the mean here. simple first-order to the left; Wilks based to the right. xx)

```
                    low   up    low   up
gamma(a,b)          3.931 5.342 3.996 5.421
weibull (b,c)       3.923 5.368 3.992 5.459
three-para (a,b,c) 3.964 5.321 4.019 5.382
```

Ex. 5.13 *The empirical correlation coefficient, I.* (xx R_n studied under binormality. xx) For i.i.d. data pairs (X_i,Y_i), the famous empirical correlation coefficient is

$$R_n = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\{\sum_{i=1}^n (X_i-\bar{X})^2\}^{1/2}\{\sum_{i=1}^n (Y_i-\bar{Y})^2\}^{1/2}}. \tag{5.4}$$

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. 1.29. Use results from Ex. 1.30, including $Y_i \,|\, X_i \sim \text{N}(\rho X_i, 1-\rho^2)$, to derive expressions for $\text{E}\,X_i^2Y_i^2$, $\text{E}\,X_i^3Y_i$, $\text{E}\,X_iY_i^3$. Use these to show that (xx check this xx)

$$\Sigma = \text{Var}\begin{pmatrix} X_i^2 \\ Y_i^2 \\ X_iY_i \end{pmatrix} = \begin{pmatrix} 2, & 2\rho^2, & 2\rho \\ 2\rho^2, & 2, & 2\rho \\ 2\rho, & 2\rho, & 1+\rho^2 \end{pmatrix}.$$

Use the CLT to argue that

$$\begin{pmatrix} A_n \\ B_n \\ C_n \end{pmatrix} = \sqrt{n} \begin{pmatrix} n^{-1}\sum_{i=1}^n X_i^2 - 1 \\ n^{-1}\sum_{i=1}^n Y_i^2 - 1 \\ n^{-1}\sum_{i=1}^n X_i Y_i - \rho \end{pmatrix} \to_d \begin{pmatrix} A \\ B \\ C \end{pmatrix} \sim \mathrm{N}_3(0, \Sigma).$$

With $R_{n,0} = C_n/(A_n B_n)^{1/2}$, 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) Here we have found the the limit distribution for $\sqrt{n}(R_{n,0}-\rho)$ directly, by representing $R_{n,0}$ as a smooth function of three averages, then using the CLT and the delta method. Another route is a follows. Show first that $R_{n,0}$ is the ML estimator for ρ, in this three-parameter zero-mean binormal setup. Find the Fisher information matrix $J = J(\sigma_1, \sigma_2, \rho)$ and its inverse.

(d) 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 (5.4) but using ξ_1, ξ_2 instead of $(\bar{X}, \bar{Y})$. Then, finally, show the Real Thing, that $\sqrt{n}(R_n - \rho)$ must have the same limit distribution.

(e) Argue 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 clever 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.

variance stabilising transformation

Fisher's zeta

(f) For the full five-parameter binormal model, find ML estimators for $\xi_1, \xi_2, \sigma_1, \sigma_2, \rho$, and show in particular that ML estimator for ρ is in fact R_n. Argue from the Cramér-Rao lower bounds of Ex. 5.16 that we for large n cannot do better than the standard deviation $(1-\rho^2)/\sqrt{n}$ achieved by R_n. Assume however that it is known that σ_1, σ_2 are equal to a common σ. What are now the ML estimators for σ and ρ, say σ^* and ρ^*? How much is won, when estimating ρ, by knowing that $\sigma_1 = \sigma_2$?

Ex. 5.14 *The empirical correlation coeffcient, II.* (xx R_n studied outside binormalilty. xx) Here we use some of the arguments of Ex. 5.13 to find the limit distribution of the empirical correlation R_n of (5.4) 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 $\sqrt{n}(R_n - R_{n,0}) \to_{\rm pr} 0$, 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 $(A_n, B_n, C_n)^{\rm t} \to_d (A, B, C)^{\rm t} \sim {\rm N}_3(0, \Sigma)$, for the variance matrix Σ of $(U_i^2, V_i^2, U_i V_i)^{\rm t}$. Spell out the elements of this matrix, using the $a_{j,k}$. Check that this agrees with the Σ of Ex. 5.13 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) (xx perhaps a dataset, doing ρ both under binormality and without that assumption. xx)

(d) (xx perhaps something with working out τ^2 for a case of dependent but not binormal (X, Y), requiring $a_{4,0}, a_{0,4}, a_{3,1}, a_{1,3}, a_{2,2}$. can take $X = \Phi(X_0), Y = \Phi(Y_0)$, with (X_0, Y_0) binormal. point to copulae. xx)

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.

(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, $J(\theta) = \int f(y, \theta) u(y, \theta)^2 \, {\rm d}y$. Let $T = T(Y)$ be any estimator unbiased for θ, and assume that $f(y, \theta)$ satisfies the conditions of Ex. 5.5(f). From ${\rm E}_\theta T = \int T(y) f(y, \theta) \, {\rm d}y = \theta$, use the conditions on $f(y, \theta)$ to deduce that $({\rm d}/{\rm d}\theta) \int T(y) f(y, \theta) u(y, \theta) \, {\rm d}y = 1$. Show that ${\rm cov}_\theta\,(T, u(Y, \theta)) = 1$, and, consequently $1 \le {\rm Var}_\theta\, T \, {\rm Var}\, u(Y, \theta)$, from which the we get one-dimensional classic Cramér–Rao inequality

$${\rm Var}_\theta\, T \ge 1/J(\theta).$$

(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

$$u(y_1, \ldots, y_n, \theta) = (\partial/\partial\theta) \sum_{i=1}^n \log f(y_i, \theta) = \sum_{i=1}^n u_0(y_i, \theta)$$

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.

(c) Show from this that if $T = T(Y_1,\ldots,Y_n)$ is an unbiased estimator for θ, then

Cramér–Rao lower bound

$$\mathrm{Var}_\theta\, T \geq \frac{1}{nJ_0(\theta)} = \frac{1/J_0(\theta)}{n}.$$

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

$$\mathrm{Var}_\theta\, T \geq \frac{1}{n}\frac{\{1+b'(\theta)\}^2}{J_0(\theta)}.$$

In particular, show that there's a lower bound on the mean squared error for *any* estimator (i.e. not merely the unbiased ones):

$$\mathrm{mse}(\theta) = \mathrm{E}\,\{T(Y_1,\ldots,Y_n)-\theta\}^2 \geq (1/n)\{1+b'(\theta)\}^2/J_0(\theta) + b(\theta)^2.$$

(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.* (xx put in here, suitably, that we're learning that ML achieves the best possible accuracy, for large n. xx) 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) 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. Deduce from $\mathrm{E}_\theta\, T = \int T(y)f(y,\theta)\,\mathrm{d}y$ that

$$(\partial/\partial\theta)\,\mathrm{E}_\theta\, T = \int T(y)f(y,\theta)u_0(y,\theta)^{\mathrm{t}}\,\mathrm{d}y = I,$$

the $p\times p$ identity matrix.

(b) Then work out that

$$\mathrm{Var}_\theta\,\{T - J_0(\theta)^{-1}u_0(Y,\theta)\} = \mathrm{E}_\theta\,\{T - J_0(\theta)^{-1}u_0(Y,\theta)\}\{T - J_0(\theta)^{-1}u_0(Y,\theta)\}^{\mathrm{t}}$$

van be expressed as $\mathrm{Var}_\theta\, T - J_0(\theta)^{-1}$. We have then shown a multidimensional version of the Cramér–Rao inequality, that

$$\mathrm{Var}_\theta\, T \geq J_0(\theta)^{-1}.$$

(c) 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

$$J_n(\theta) = \mathrm{Var}_\theta\, \frac{\partial \log\{f(Y_1,\theta)\cdots f(Y_n,\theta)\}}{\partial\theta} = nJ_0(\theta),$$

Cramér–Rao lower bound, matrix case

and that for *any* unbiased estimator $T = T(Y_1, \ldots, Y_n)$ of θ, we must have

$$\mathrm{Var}_\theta\, T \geq \{nJ_0(\theta)\}^{-1} = (1/n)J_0(\theta)^{-1}.$$

Here $A \geq B$, or $A - B \geq 0$, means that $A - B$ is nonnegative definite, which is equivalent to $c^{\mathrm{t}}Ac \geq c^{\mathrm{t}}Bc$ for all c. In particular, $a_{i,i} \geq b_{i,i}$ for all diagonal elements, but we do not necessarily have $a_{i,j} \geq b_{i,j}$ outside the diagonal.

(d) 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 $\mathrm{Var}_\theta\,[T - \{I + b'(\theta)\}J_0(\theta)^{-1}u_0(Y,\theta)]$ to demonstrate that

$$\mathrm{Var}_\theta\, T \geq \{I + b'(\theta)\}J_0(\theta)^{-1}\{I + b'(\theta)\}^{\mathrm{t}}.$$

(e) Generalise to the case of n i.i.d. observations, to reach

$$\mathrm{Var}_\theta\, T \geq (1/n)\{I + b'(\theta)\}J_0(\theta)^{-1}\{I + b'(\theta)\}^{\mathrm{t}}.$$

(f) 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 *The Kullback–Leibler distance, from one density to another.* 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

the Kullback–Leibler distance

$$d(g,f) = \int g \log \frac{g}{f}\,\mathrm{d}y.$$

It is an important concept and tool for communication and information theory, and also for probability theory and statistics. In particular, it turns out that the KL distance is intimately connected to maximum likelihood, to the most well-used model selection method AIC (the Akaike Information Criterion, see Ch. 11), etc.

(a) The $\log(g/f)$ term will be both positive and negative, in different parts of the domain. Show nevertheless that indeed $d(g,f) \geq 0$, and that $d(g,f) = 0$ only when the two densities are equal a.e. The 'a.e.' is a measure theoretic little standard miniphrase, meaning 'almost everywhere', i.e. the set where $g(y) \neq f(y)$ is so small that it has Lebesgue measure zero (the integral does not change its value if the integrand function changes its value in a finite number of points, or, for that matter, if $g(y)$ somewhat artificially should change its value in every rational number). Try to prove nonnegativity via Jensen's inequality.

(b) A useful way of proving nonnegativity, since it opens a little door to certain generalisations, is as follows. Write first

$$d(g,f) = \int \Big\{ g \log \frac{g}{f} - (g-f) \Big\}\,\mathrm{d}y,$$

and then show that the function which for fixed g is equal to $A(u) = g\log(g/u) - (g-u)$, has its minimum position at $u = g$, where $A_{\min} = A(g) = 0$.

(c) 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 $\frac{1}{2}(\xi_2-\xi_1)^2/\sigma^2$.

(d) Find the KL distance from one Poisson to another.

(e) 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

(f) 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.

(g) Consider a parametric density $f(y,\theta)$, with score function $u(y,\theta)$ and information matrix $J(\theta) = \mathrm{Var}_\theta\, u(y,\theta)$. Show that

$$d(f(\cdot,\theta), f(\cdot,\theta+\varepsilon)) = \tfrac{1}{2}\varepsilon^{\mathrm{t}} J(\theta)\varepsilon + O(\varepsilon^3).$$

(h) 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

$$d(g,f) \approx \tfrac{1}{2}\int g(f/g-1)^2\,\mathrm{d}y = \tfrac{1}{2}\Big(\int f^2/g\,\mathrm{d}y - 1\Big).$$

(xx some words indicating that the root-KL might have an easier interpretation. xx)

(i) (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. 5.18 *What is the maximum likelihood aiming for?* (xx the following to be reworked, in view of generalities above, from Ch4 exercises on minimum criterion function estimators. xx) Assume independent observations $Y_1, Y_2, \ldots$ become available, from a certain data generating mechanism g, the envisaged true but typically unknown data density. With a parametric model f_θ, with $f_\theta(y) = f(y,\theta)$, what it the ML method aiming for? We learn here that there is a clear answer, intimately connected to the Kullback–Leibler distance from truth to approximation. We learn also elsewhere that ML is interrelated with KL, as with the AIC of Ch. 11.

(a) Consider the usual log-likelihood function $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i,\theta)$. The framework of Ex. 5.7 (xx and one more xx) involved the assumption that the model was actually correct, and then we saw that the ML estimator $\widehat\theta$ is consistent for the true parameter θ_0. Now there is no 'true parameter', however. But show that

$$C_n(\theta) = n^{-1}\ell_n(\theta) \to_{\rm pr} C(\theta) = \mathrm{E}_g \log f(Y,\theta) = \int g \log f_\theta\,\mathrm{d}y \quad \text{for each } \theta.$$

Note that this involves the Kullback–Leibler distance, since $d(g,f_\theta) = \int g\log g\,\mathrm{d}y - C(\theta)$. Under reasonable regularity conditions [xx which we'll be coming back to, where? xx], it will then be the case that the maximiser of C_n, which is the ML estimator $\widehat\theta$, will tend to the maximiser θ_0 of C, which is also the minimiser of the KL distance $d(g,f_\theta)$ – we do assume that there is precisely one such minimiser. Attempt to formalise such regularity conditions, going from (i) $C_n(\theta) \to_{\rm pr} C(\theta)$ for each θ to (ii) $\mathrm{argmax}(C_n) \to_{\rm pr} \mathrm{argmax}(C)$.

(b) So we've uncovered what goes on in the mindset of the ML operator – it aims for the *least false parameter*, the θ_0 minimising the Kullback–Leibler distance $d(g,f_\theta)$. The principle itself does not say or claim to say how well this might be working, as the size of the minimal distance

least false parameter value

$$d_{\rm min} = \min_{\text{all }\theta} d(g,f_\theta) = d(g,f(\cdot,\theta_0))$$

will depend on both g and the parametric family being used.

(c) Suppose data $Y_1, Y_2, \ldots$ are recorded on the positive halfline, from some underlying density g. Suppose that the exponential model $\theta \exp(-\theta y)$ is being used. What is the ML estimator $\widehat{\theta}$ aiming for?

(d) Assume independent data $Y_1, Y_2, \ldots$ stem from some density g on the line, with finite mean ξ_0 and standard deviation σ_0. Using the normal model $\mathrm{N}(\xi, \sigma^2)$, show that

$$d(g, f(\cdot, \xi, \sigma)) = \int g \log g \, \mathrm{d}y + \log \sigma + \tfrac{1}{2} \frac{\sigma_0^2 + (\xi - \xi_0)^2}{\sigma^2},$$

and that this is being minimised, over all (ξ, σ) pairs, for precisely $\xi = \xi_0 = \mathrm{E}\, Y$ and $\sigma = \sigma_0 = (\mathrm{Var}\, Y)^{1/2}$.

Ex. 5.19 *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 5 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.

(e) (xx one more. round off. xx)

Ex. 5.20 *Behaviour of the maximum likelihood estimator, under agnostic conditions.* (xx to be restructured and simplified, in view of Ch4 things. we get results readily. xx) Luckily, it might be fair to say, ML estimation still manages to make sense, even when the parametric model employed is not 100 percent correct. Statistics would have been a somewhat different discipline, with lower ambition level and bragging rights, if all its methods had a Red Warning Flag on top of all papers and algorithms and applications, saying 'can only be used if the model is perfect'. The aim here is to uncover and understand more of what happens with the ML estimator, in the case that the true density g is outside the $\{f_\theta \colon \theta \in \Omega\}$ in question.

(a) Let $Y_1, \ldots, Y_n$ be independent realisations from an underlying g, with $\widehat{\theta}$ the ML estimator. We have seen that $\widehat{\theta} \to_{\rm pr} \theta_0$, the least false parameter value, as judged by the Kullback–Leibler distance $d(g, f_\theta)$. With terms and notation from Ex. 5.7, establish that the score function has mean zero, at this true parameter value, $\mathrm{E}_g\, u(Y, \theta_0) = 0$. Explain in detail why this generalises a corresponding result for the 'under the model' case.

(b) Under model conditions, certain essential things could be told using only one matrix, namely Fisher's information matrix $J = J(\theta)$. Now we are in need of as many as two matrices, it turns out. Define

$$\begin{aligned} J &= -\mathrm{E}_g\, i(Y, \theta_0) = -\int g(y) \frac{\partial^2 \log f(Y, \theta_0)}{\partial\theta\partial\theta^{\rm t}}, \\ K &= \mathrm{Var}_g\, u(Y, \theta_0) = \int g(y) u(y, \theta_0) u(y, \theta_0)^{\rm t}\, {\rm d}y, \end{aligned}$$

assumed to be finite and positive definite. These are identical under model conditions. Now use the machinery of minimum criterion function estimators (xx pointer Ch4 xx) to establish that $U_n = (1/\sqrt{n})\partial\ell_n(\theta_0/\partial\theta$ tends to a $U \sim \mathrm{N}_p(0, K)$, along with the two fundamental results

$$\begin{aligned} Z_n &= \sqrt{n}(\widehat{\theta} - \theta_0) \to_d J^{-1}U \sim \mathrm{N}_p(0, J^{-1}KJ^{-1}), \\ W_n &= 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d U^{\rm t} J^{-1} U. \end{aligned} \tag{5.5}$$

These are the appropriate generalisations of (5.3) to situations outside model conditions. In particular, the limit distribution now has the sandwich matrix $J^{-1}KJ^{-1}$ instead of the simple $J(\theta_0)^{-1}$. Also, the quadratic form $W = U^{\rm t}J^{-1}U$ does not have a χ^2_p distribution, outside model conditions; show that its mean is $p^* = \mathrm{Tr}(J^{-1}K)$.

(c) (xx also put in here: using efforts of Ch4 to read off limits for $D_n(\phi) = 2\{\ell_{n,\max} - \ell_n(\phi)\}$, extending the χ^2_1 results. xx)

(d) (xx well, we put this in Ch4 already, see Ex. 4.54, and point to this here. xx) Consider a function $h(y, \theta)$, with finite mean $h_0 = \mathrm{E}_{\theta_0}\, h(Y, \theta_0)$ at position θ_0. It is clear that $\widehat{h} = n^{-1}\sum_{i=1}^n h(Y_i, \widehat{\theta}) \to_{\rm pr} h_0$ under natural conditions. Show that if $|h(y, \theta_0 + \varepsilon) - h(y, \theta_0)| \le M(y)|\varepsilon|$, for all small $|\varepsilon|$, for some function $M(y)$ with finite mean, then indeed $\widehat{h} \to_{\rm pr} h_0$.

(e) Natural estimators for J and K, needed for estimating the sandwich from data, are

$$\widehat{J} = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log f(Y_i, \widehat{\theta})}{\partial\theta\partial\theta^{\rm t}} \quad \text{and} \quad \widehat{K} = \frac{1}{n}\sum_{i=1}^n u(Y_i, \widehat{\theta}) u(Y_i, \widehat{\theta})^{\rm t}.$$

Give conditions under which $\widehat{J}$ and $\widehat{K}$ are consistent for J and K. This is what we need to have a consistent estimator for the sandwich matrix $J^{-1}KJ^{-1}$.

Ex. 5.21 *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 $Z = (Y - \mathrm{E}\, Y)/\mathrm{sd}(Y) = (Y - \xi_0)/\sigma_0$; see Ex. 2.10. 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

$$\begin{pmatrix} \sqrt{n}(\widehat{\xi} - \xi) \\ \sqrt{n}(\widehat{\sigma} - \sigma) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1, & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3, & \frac{1}{2} + \frac{1}{4}\gamma_4 \end{pmatrix} \Big).$$

Note that this is a 'rediscovery' of what we found in Ex. 2.10 and 2.13, 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. 5.22 *A log-likelihood function process.* Consider i.i.d. observations $Y_1, \ldots, Y_n$ from some density g, with a model $f(y,\theta)$ fitted via ML. Thus $\widehat{\theta}$ maximises the log-likelihood function θ. It is fruitful to work the random function

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0).$$

(a) Simulate a sample of $n = 25$ points from the exponential model with $\theta_0 = 3.33$. Compute and display the $A_n(s)$ function. Then do this with say ten different samples, from the same model and the same n, and display the ten A_n curves in a diagram.

(b) (xx then point back to earlier general efforts, to get main points across. $A_n(s) \to_d U^{\mathrm{t}} s - \frac{1}{2} s^{\mathrm{t}} J s$. argmax to argmax, max to max. xx)

(c) (xx to be altered. xx) Go back to your ten simulated versions of $A_n(s)$ for the exponential case, where the true $\theta_0 = 3.33$. Use the above results to test the hypothesis that $\theta = 4.44$.

Ex. 5.23 *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 $h_0 = \int h(x)\,{\rm d}Q(x)$ for $n^{-1}\sum_{i=1}^n h(x_i)$, for an appropriate distribution Q on the covariate space. (xx can discuss this a little bit more. most often we do not see or need Q, beyond its postulated existence. xx)

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. 5.7, we need to understand $n^{-1/2}U_n(\theta_0) = n^{-1/2}\sum_{i=1}^n u(Y_i \,|\, x_i, \theta_0)$. Show that it has mean zero and variance matrix $J_n = n^{-1}\sum_{i=1}^n J(x_i)$, where $J(x_i) = {\rm Var}_{\theta_0} u(Y_i \,|\, x_i, \theta_0)$. We have $J_n \to J$, under ergodic conditions. Give Lindebeberg type conditions under which $U_n(\theta_0) \to_d U \sim {\rm N}_p(0, J)$.

(b) Extend techniques from Ex. 5.7 to deduce that $\sqrt{n}(\widehat\theta - \theta_0) \to_d {\rm N}_p(0, J^{-1})$, in this general regression setting, under these Lindeberg type conditions. Show also that the observed Fisher information matrix

observed Fisher information matrix

$$\widehat J_{n,{\rm full}} = -\partial^2 \ell_n(\widehat\theta)/\partial\theta\partial\theta^{\rm t}, \tag{5.6}$$

i.e. the Hessian mastrix associated with the maximisation of the log-likelihood, satisfies $\widehat J_n = (1/n)\widehat J_{\rm full} \to_{\rm pr} J$. Deduce from this that $\widehat\theta \approx_d {\rm N}_p(\theta_0, \widehat J_{\rm full}^{-1})$.

(c) Next make clear that the log-likelihood function process $A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - A_n(\theta_0)$, a natural extension of the process studied in Ex. 5.22, has the same limiting behaviour as in that earlier i.i.d. setup, with a random limit function $A(s) = U^{\rm t}s - \frac12 s^{\rm t}Js$. Explain how the basic $\sqrt{n}(\widehat\theta - \theta_0) \to_d {\rm N}_p(0, J^{-1})$ follows from this.

(d) (xx rounding this off. the point is that i.i.d. results all extend to the broad regression cases. xx)

Ex. 5.24 *Linear regression revisited.* (xx edit and clean. xx) Consider the linear regression model of Ex. 3.33, with $Y_i \,|\, x_i \sim {\rm N}(x_i^{\rm t}\beta, \sigma^2)$, 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 $\ell_n(\beta, \sigma) = -n\log\sigma - \frac12 Q(\beta)/\sigma^2 - \frac12 n\log(2\pi)$, with $Q(\beta) = \sum_{i=1}^n (y_i - x_i^{\rm t}\beta)^2$ Show that the ML estimator for β is the least squares estimator $\widehat\beta$, given in 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

$$u(y_i \,|\, x_i, \beta, \sigma) = \begin{pmatrix} (1/\sigma^2)(y_i - x_i^{\rm t}\beta)x_i \\ -1/\sigma + (1/\sigma^3)(y_i - x_i^{\rm t}\beta)^2 \end{pmatrix} = \begin{pmatrix} (1/\sigma)\varepsilon_i x_i \\ (1/\sigma)(\varepsilon_i^2 - 1) \end{pmatrix}$$

in terms of $\varepsilon_i = (y_i - x_i^{\rm t}\beta)/\sigma$, which are independent standard normals under the model. With $(\beta_0, \sigma_0$ the true parameters, deduce that the $(p+1) \times (p+1)$ Fisher information matrix becomes

$$J_n = (1/n)\sum_{i=1}^n {\rm Var}_{\beta_0,\sigma_0} u(Y_i \,|\, x_i, \beta_0, \sigma_0) = (1/\sigma_0^2){\rm diag}(\Sigma_n, 2),$$

with $\Sigma_n = (1/n)\sum_{i=1}^n x_i x_i^{\rm t} = (1/n)X^{\rm t}X$. Show also that the observed Fisher information matrix becomes $\widehat{J}_{n,{\rm full}} = (n/\widehat{\sigma}^2){\rm diag}(\Sigma_n, 2)$.

(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. 5.25 *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

$$p_i = P(Y_i = 1 \,|\, x_i) = H(x_i^{\rm t}\beta) = \frac{\exp(x_i^{\rm t}\beta)}{1 + \exp(x_i^{\rm t}\beta)} \quad \text{for } i = 1, \ldots, n, \tag{5.7}$$

with $H(u) = \exp(u)/\{1 + \exp(u)\}$ the so-called logistic transform.

(a) Show that $H(u) = p$ means $u = H^{-1}(p) = \log\{p/(1-p)\}$, so that the model can be represented as $\log\{p_i/(1-p_i)\} = x_i^{\rm t}\beta$.

(b) Show that $P(Y_i = y \,|\, x_i) = p_i^y(1-p_i)^{1-y}$, for the two outcomes, and deduce that the log-likelihood function can be written

$$\ell_n(\beta) = \sum_{i=1}^n \{y_i \log p_i + (1-y_i)\log(1-p_i)\} = \sum_{i=1}^n \big[y_i x_i^{\rm t}\beta - \log\{1 + \exp(x_i^{\rm t}\beta)\}\big].$$

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

$$J_{n,{\rm full}}(\beta) = -\frac{\partial^2 \ell_n(\beta)}{\partial\beta\partial\beta^{\rm t}} = \sum_{i=1}^n p_i(1-p_i)x_i x_i^{\rm t} = \sum_{i=1}^n H(x_i^{\rm t}\beta)\{1 - H(x_i^{\rm t}\beta)\}x_i x_i^{\rm t}.$$

(c) Show that, under model conditions, $\widehat{\beta} \approx_d {\rm N}_p(\beta, \widehat{J}_{n,{\rm full}}^{-1}/n)$, where $\widehat{J}_{n,{\rm full}} = J_{n,{\rm full}}(\widehat{\beta})$ is the observed Fisher information matrix (the Hessian matrix of minus the normalised log-likelihood function, at the ML position). It is assumed here to be positive definite, which in particular requires $n \geq p$.

(d) Consider an individual, perhaps outside the dataset, with covariate vector x_0. Show that $x_0^{\rm t}\widehat{\beta}$ is approximately a normal $(x_0^{\rm t}\beta, x_0^{\rm t}\widehat{J}_{n,{\rm full}}^{-1}x_0)$, and use this to construct a confidence interval for $p(x_0) = P(Y_0 = 1 \,|\, x_0)$.

Ex. 5.26 *Poisson regression.* (xx to be spelled out. $\mu_i = \exp(x_i^{\rm t}\beta)$, approximate normality, delta method thing. xx) 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 $Y_i \sim \text{Pois}(\mu_i)$, with $\mu_i = \exp(x_i^{\rm t}\beta)$.

(a) Show that the log-likelihood function becomes

$$\ell_n(\beta) = \sum_{i=1}^n \{-\mu_i + y_i \log(\mu_i)\} = \sum_{i=1}^n \{y_i x_i^{\rm t}\beta - \exp(x_i^{\rm t}\beta)\},$$

and that the equations $\sum_{i=1}^n \{y_i - \mu_i(\beta)\}x_i = 0$ define the ML estimators.

(b) Show that $-\partial^2 \ell_n(\beta)/\partial\beta\partial\beta^2 = \sum_{i=1}^n \mu_i x_i x_i^{\rm t}$, with the observed Fisher information matrix $\widehat{J}_{n,\rm full} = \sum_{i=1}^n \widehat{\mu}_i x_i x_i^{\rm t}$, where $\widehat{\mu}_i = \exp(x_i^{\rm t}\widehat{\beta})$.

(c) For a case with covariate vector x_0, estimate the associated expected count $\mu_0 = \exp(x_0^{\rm t}\beta)$, and construct a confidence interval.

Ex. 5.27 *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 $Y_i \,|\, (x_i, w_i) \sim \text{N}(x_i^{\rm t}\beta, \sigma_i^2)$, for $i = 1, \ldots, n$, with covariate vectors x_i of length p and variance related covariates w_i of length q, influencing $\sigma_i = \sigma \exp(\gamma^{\rm t} w_i)$. These w_i could be a subset of the x_i or functions thereof. It is convenient to normalise these such that $\bar{w} = (1/n)\sum_{i=1}^n w_i = 0$, 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

$$\ell_n(\beta, \gamma) = -n \log \sigma - \tfrac{1}{2}(1/\sigma^2) Q(\beta, \gamma), \quad \text{with } Q(\beta, \gamma) = \sum_{i=1}^n \frac{(y_i - x_i^{\rm t}\beta)^2}{\exp(2w_i^{\rm t}\gamma)}.$$

Show that minimising $Q(\beta, \gamma)$ over β, for fixed γ, takes place for

$$\widehat{\beta}(\gamma) = \Big\{\sum_{i=1}^n \frac{x_i x_i^{\rm t}}{\exp(2\gamma^{\rm t} w_i)}\Big\}^{-1} \sum_{i=1}^n \frac{x_i y_i}{\exp(2\gamma^{\rm t} w_i)}.$$

Demonstrate that this leads to the profiled log-likelihood $\ell_{n,\rm prof}(\gamma) = -n \log \widehat{\sigma}(\gamma) - \frac{1}{2}n$, 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 $\Sigma_n(\gamma) = (1/n)\sum_{i=1}^n x_i x_i^{\mathrm{t}}/\exp(2\gamma^{\mathrm{t}}w_i)$ and $M_n = (1/n)\sum_{i=1}^n w_i w_i^{\mathrm{t}}$.

(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 $\mu(x_0, w_0)$, $x_0^{\mathrm{t}}\widehat{\beta} \pm 1.96\,\widehat{\sigma}\exp(\widehat{\gamma}^{\mathrm{t}}w_0)$. pointer to Story iv.4. xx)

Ex. 5.28 *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 $\mathrm{Pois}(\lambda)$, with $k(\lambda, 1) = \exp(\lambda)$. This two-parameter generalised Poisson model is from Schweder and Hjort (2016, Examples 4.18, 8.16).

(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. 5.29 *ML machinery in practice, II.* (xx nils starts ranting. point back to Ex. 5.12. nils needs to calibrate with Ex. 5.28. make the point that some famous models are part of standard packages, as with `glm` in R, but that we can attach also fresh new models. xx) We have seen the ML machinery in practice in Ex. 5.12, for i.i.d. models, where the central message is that as long as one can programme the log-likelihood function, one may often apply generic optimisation algorithms to find ML estimates, their standard errors, find confidence intervals for focus parameters, test hypotheses, etc. The aim of the present exercise is to showcase how essentially the same machinery works also for regression models, whether these are part of the standard statistical repertoire or are freshly invented with new twists and ingredients. The main reason for this is that the central parts of ML theory extend from i.i.d. to regression models, as we have seen in (xx point to exercises xx).

Our illustration will be in terms of the following relatively simple and small dataset, pertaining to y, the number of different bird species living on páramos on fourteen islands outside Ecuador. The task is to attempt to understand how y is influenced by x_1, the distance from Ecuador, in km; and x_2, the area, in thousands of square km (and perhaps on yet other covariates not taken on board here). The grander purposes relate to understanding biological variation and to prediction of species abundance on other islands.

```
     x1   x2  y         x1   x2  y
1 0.036 0.33 36    8 0.958 0.14 13
2 0.234 0.50 30    9 0.995 0.05 17
3 0.543 2.03 37   10 1.065 0.07 13
4 0.551 0.99 35   11 1.167 1.80 29
5 0.773 0.03 11   12 1.182 0.17  4
6 0.801 2.17 21   13 1.238 0.61 18
7 0.950 0.22 11   14 1.380 0.07 15
```

(a) For the birds-on-islands dataset, first carry out ordinary Poisson regression, taking $y_i \sim \text{Pois}(\mu_i)$ with $\mu_i = \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})$. Show indeed that β_1 is significantly negative, that β_2 is significantly positive, and give interpretations of these initial findings; cf. columns 1:3 in the table.

(b) We then pass to the extended Poisson regression model introduced in Ex. 5.28, taking distribution

$$f(y_i, \mu_i, \gamma) = k(\mu_i, \gamma)^{-1} \mu_i^{y_i} / (y_i!)^\gamma \quad \text{for } y_i = 0, 1, 2, \ldots,$$

with normalisation constant $k(\mu_i, \gamma) = \sum_{y=0}^{\infty} \mu_i^y / (y!)^\gamma$. Write up a script for the log-likelihood function $\ell_2(\beta_0, \beta_1, \beta_2, \gamma)$, in the style of what is carried out in Ex. 5.12. The following works – here we have used `X = cbind(one,x1,x2)`, with `one` the vector of 1, and `pp = ncol(X)`; put an `aux = 0*(1:n)` in preparation; and made an initial script for $k(\mu, \gamma)$:

```
logL2 <- function(para)
{
beta = para[1:pp]
gam = para[(pp+1)]
mu = exp(X %*% beta)
for (i in 1:nn)
{aux[i] = -mu[i]+yy[i]*log(mu[i])-gam*lgamma(yy[i]+1) - log(k(c(mu[i],gam)))}
sum(aux)
}
```

Maximise the log-likelihood, with steps similar to those in Ex. 5.12, and reproduce columns 4:6 in the table. Deduce that an approximate 90 percent interval for the γ parameter is $[0.187, 0.949]$, indicating overdispersion compared to Poisson. Also carry out log-likelihood-profiling, computing $\ell_{2,\text{prof}}(\gamma)$, for a somewhat more accurate 90 percent interval, using (xx point to wilks theorem exercise xx).

	model M1			model M2				model M3		
	estim	se	ratio	estim	se	ratio		estim	se	ratio
beta0	3.429	0.139	24.597	1.941	0.806	2.410		1.927	0.805	2.393
beta1	-0.814	0.151	-5.400	-0.472	0.216	-2.181		-0.506	0.236	-2.144
beta2	0.312	0.072	4.347	0.181	0.089	2.031		1.567	0.705	2.224
				gam 0.568	0.232	2.450	alpha0	-0.232	0.388	-0.597
							alpha1	0.320	0.078	4.098

(c) The dispersion parameter γ is perhaps not quite constant, across the different islands. A finer model worth working through takes $\gamma_i = \exp(\alpha_0 + \alpha_1 w_i)$, with $w_i = (x_{i,2} - \bar{x}_2)/\text{sd}(x_2)$, i.e. the normalised x_2. This helps stable numerics and eases the interpretaion of α_0 and α_1. Now programme the appropriate log-likelihood function, say $\ell_3(\beta_0, \beta_1, \beta_2, \alpha_0, \alpha_1)$. Find ML estimates and their estimated standard deviations, and produce a version of columns 7:9 of the table. Give an interpretation of these results.

(d) For the three models, record the attained log-likelihood maxima, say $\ell_{1,\max}$, $\ell_{2,\max}$, $\ell_{3,\max}$; these are found as easy byproducts of the maximisation algorithms in the first place. Compare Models 1, 2, 3, via Wilks testing (xx point xx). Compute also Pearson type chi-squared statistics, of the type $W = \sum_{j=1}^n (y_j - \widehat{y}_j)^2/\widehat{y}_j$ over the $n = 14$ islands, where $\widehat{y}_j = n\widehat{f}_j(y_j)$ is the estimated y_j for the model considered. It will be seen, also via AIC and other model selection criteria in Ch. 11, that the five-parameter Model 3 is the best one here. To check for differences, produce a version of Figure 5.3, with the estimated probabilities $\widehat{f}_1, \widehat{f}_2, \widehat{f}_3$, for a few positions $(x_{1,0}, x_{2,0})$ in the covariate space. This particular figure has 'big island, far from Ecuador' to the left and 'small island, close to Ecuador' to the right.

(e) (xx just a few more things, then round off. main point is again to showcase the versatility and relative ease of the ML theory, via the log-likelihood function. compute logL maxima, carry out Wilks testing; model M3 is best here. it also wins AIC. Show a figure with $\widehat{f}_1, \widehat{f}_2, \widehat{f}_3$, plotted for say $y = 0, 1, \ldots, 40$. use this to explain which features are not well captured by the simple poisson. can also take poisson-gamma overdisperson model, from Ex. 5.26. from $y_i \,|\, \mu_i \sim \text{Pois}(\mu_i)$ and then $\mu_i \sim \text{Gam}(\mu_i/\tau, 1/\tau)$; τ small

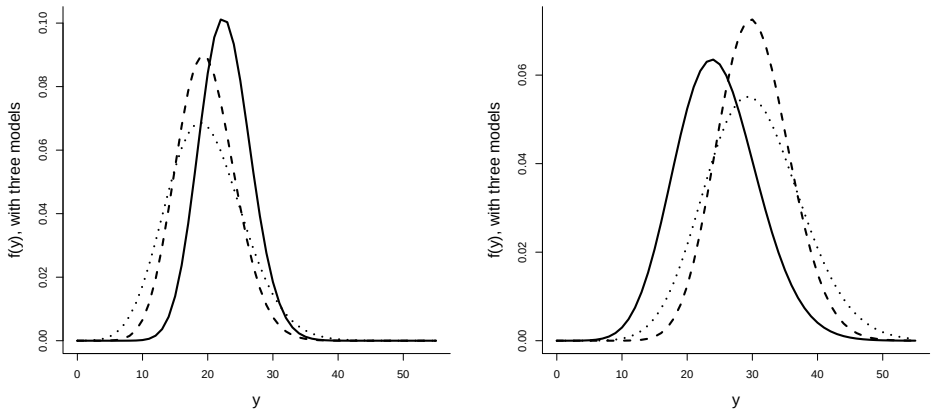

Figure 5.3: *Estimated probability distributions* $\widehat{f}_1, \widehat{f}_2, \widehat{f}_3$*, for the number of bird species on two imagined islands, based on models 1 (simple Poisson), 2 (extended Poisson, with one* γ*), 3 (extended Poisson, with* γ_i*). Left panel:* (x_1, x_2) *taken to be their max values, i.e. big island, far from Ecuador; right panel:* (x_1, x_2) *at the min values, i.e. small island, close to Ecuador. The full black curves are for the five-parameter* f_3*, the best model.*

corresponds to ordinary poisson. but the dispersion model worked with above has the abiity to reflect both under- and overdispersion. we have Ex. 1.16, and ought to have more here in Ch5. xx)

Ex. 5.30 *We can do things.* (xx very tentative, but nils wishes to include a couple of things in the spirit of 'here is a natural but not off-the-shelf model, and we can estimate parameters etc. xx)

(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 $(1-p)\,\text{Pois}(\theta) + p\,\text{Pois}(c\theta)$, 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) (xx one or two more points in this spirit. putting up log-likelihood and then maximising etc. xx)

Ex. 5.31 *Extending theory and methods to regression setups, II.* (xx then outside model conditions. xx)

Ex. 5.32 *Logistic regression, II.* (xx nils ranting on a bit; to be cleaned later. building on Ex. 5.25 but now analysis outside model conditions. xx)

(a) When the parametric form of the model cannot be trusted, characterise the underlying least false model parameter $\beta_{0,n}$ for which the ML estimator is aiming. Then show that $\widehat{\beta} \approx_d \mathrm{N}(\beta_{0,n}, \Sigma_n/n)$, with sandwich matrix $\Sigma_n = \widehat{J}_n^{-1}\widehat{K}_n\widehat{J}_n^{-1}$, with

$$\widehat{J}_n = (1/n)\sum_{i=1}^n \widehat{p}_i(1-\widehat{p}_i)x_i x_i^{\mathrm{t}}, \quad \widehat{K}_n = (1/n)\sum_{i=1}^n (y_i-\widehat{p}_i)^2 x_i x_i^{\mathrm{t}}.$$

(b) For a new individual, or object, with covariate vector x_0, explain that $x_0^{\mathrm{t}}\widehat{\beta} \approx_d \mathrm{N}(x_0^{\mathrm{t}}\theta, \widehat{\tau}_0^2/n)$, where $\widehat{\tau}_0^2 = x_0^{\mathrm{t}}\widehat{J}^{-1}x_0$, and use this to form a 90 percent confidence interval for the associated $P(Y_0 = 1 \,|\, x_0)$.

(c) 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

$$\sum_{i=1}^n (y_i - p_i) = 0, \quad \sum_{i=1}^n (y_i - p_i)x_i = 0,$$

and that

$$J_n(a,b) = \sum_{i=1}^n \frac{\exp(a+bx_i)}{\{1+\exp(a+bx_i)\}^2}\begin{pmatrix} 1, & x_i \\ x_i, & x_i^2 \end{pmatrix}$$

(d) (xx we point to an illustration. xx)

Ex. 5.33 *The exponential family class and ML.* [xx exercises on the exponential family class and maximum likelihood estimation here. Pointers to Ex. 1.57–**??** . Noen subopp-gaver som passer best here. Skriv dem inn. xx]

(a) Show that the ML estimator $\widehat{\theta}$ is the unique solution to the equation $\bar{T} = \xi(\theta)$, or $\bar{T}_j = \xi_j(\theta_1, \ldots, \theta_p)$ for $j = 1, \ldots, p$. (xx brief comment on optimisers and equation solvers; each sensible algorithm will find the ML, and its Hessian matrix $\widehat{J}_n = -(1/n)\partial^2 \ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}} = J(\widehat{\theta})$.

(b) (xx about $\sqrt{n}\{\bar{T}_n - \xi(\theta)\} \to_d \mathrm{N}(0, J(\theta))$, and about $\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathrm{N}(0, J(\theta)^{-1})$. xx)

(c) (xx we'll see later where to put what. this is the start about understanding the c/n bias of ML for exponential class models. xx) Suppose $Y_1, \ldots, Y_n$ are i.i.d. from an exponential family model $f(y,\theta) = \exp\{\theta^{\mathrm{t}}T(y) - k(\theta)\}h(y)$, leading to the ML estimator $\widehat{\theta}$, solving $\bar{T} = \xi(\theta)$, with $\xi(\theta) = \mathrm{E}_\theta\, T = \partial k(\theta)/\partial\theta$. Here $\xi(\theta) = (\xi_1(\theta), \ldots, \xi_p(\theta))^{\mathrm{t}}$, with length p, the dimension of $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{t}}$.

Write $\widehat{\theta} = A(\bar{T})$, with A the inverse function of ξ; $\xi(\theta) = t$ is equivalent to $\theta = A(t)$. With θ_0 the true parameter value, write $\xi_0 = \xi(\theta_0)$. Use Taylor expansion to derive

$$\widehat{\theta}_j = A_j(\xi_0) + A_j'(\xi_0)^{\mathrm{t}}(\bar{T} - \xi_0) + \tfrac{1}{2}(\bar{T} - \xi_0)^{\mathrm{t}}A_j''(\xi_0)(\bar{T} - \xi_0) + O_{\mathrm{pr}}(1/n^{3/2}),$$

for components $j = 1, \ldots, p$. Show first from this that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}_p(0, \Sigma), \quad \text{with } \Sigma = A'(\xi_0)^{\mathrm{t}}J(\theta_0)A'(\xi_0),$$

and that $\Sigma = J(\theta_0)^{-1}$. (xx check details and formalisation here. we use $A(\xi(\theta)) = \theta$ and find matrix identity. xx)

(d) Then show that $\mathrm{E}\,\widehat{\theta}_j = c_j/n + o(1/n)$ with

$$c_j = \tfrac{1}{2}\mathrm{Tr}\{A_j''(\xi_0)J(\theta_0)\}.$$

(e)

Ex. 5.34 *Influence functions.* [xx this exercise to be used in Ch. 9. xx] 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, a quantile, a threshold probability. The *influence function* for $\theta = T(F)$ is a very useful quantity, as we shall see. It is defined as

$$\mathrm{IF}(F, y) = \lim_{\varepsilon \to 0}(1/\varepsilon)\{T((1-\varepsilon)F + \varepsilon\delta(y)) - T(F)\}. \tag{5.8}$$

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 $\theta(F) = \mathrm{E}_F\, h(Y) = \int h(y)\,\mathrm{d}F(y)$, 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 $\theta = \mathrm{E}_F\, Y$.

(b) Then consider the class of smooth functions of means. For mean type parameters $\gamma_1 = \mathrm{E}_F\, h_1(Y), \ldots, \gamma_k = \mathrm{E}_F\, h_k(Y)$, 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 $\mu_F = \mathrm{E}_F\, Y$ for the mean, show for the variance parameter $\sigma_F^2 = \mathrm{E}_F\, Y^2 - \mu_F^2$ that its influence function becomes

$$\mathrm{IF}(F, y) = -2\,\mu_F(y - \mu_F) + y^2 - \mathrm{E}_F\, Y^2 = (y - \mu_F)^2 - \sigma_F^2.$$

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. 2.27, Show that $T(F_n)$ is the ML estimator, and that its influence function becomes $\mathrm{IF}(F, y) = J^{-1}u(y, \theta_0)$.

(e) (xx the median and the quantile. smooth function of quantiles and means. and something more involved too. interquartile range. MAD statistic. xx)

Ex. 5.35 *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 $\gamma(F) = \mathrm{E}_F\, h(Y) = \int h \,\mathrm{d}F$. 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. 5.34, 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

$$\widehat{\theta} - \theta = T(F_n) - T(F) = n^{-1}\sum_{i=1}^n \mathrm{IF}(F, Y_i) + b/n + \varepsilon_n, \tag{5.9}$$

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 (5.9) 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. [xx give reference, with easy conditions, or we might write out the basics. xx]

(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. 2.12.

(e) xx

Ex. 5.36 *Leave-one-out statistics.* (xx to come here, after influence functions. make sure ML is board too, with $\mathrm{IF}(y) = J^{-1}u(y, \theta_0)$. delta method on top of this. jackknife and cross-validation approx for AIC. xx) Consider a parameter functional $T(F)$ with influence function $\mathrm{IF}(F, y)$, so that the representation (5.9) holds for $\widehat{\theta} = T(F_n)$, based on observations $y_1, \ldots, y_n$.

(a) Let $\widehat{\theta}_{(i)}$ be the estimator in question computed for the dataset where y_i is pushed out, with $\widehat{\theta}_{\mathrm{jack}} = (1/n)\sum_{i=1}^n \widehat{\theta}_{(i)}$ their average again. From the above representation, deduce $\mathrm{E}\,\widehat{\theta} = \theta + b/n + o(1/n)$ and $\mathrm{E}\,\widehat{\theta}_{(i)} = \theta + b/(n-1) + o(1/n)$. (xx work from this: to $\widehat{b}_{\mathrm{jack}}/n = (n-1)(\widehat{\theta}_{\mathrm{jack}} - \widehat{\theta})$. and from this to $n\widehat{\theta} - (n-1)\widehat{\theta}_{\mathrm{jack}}$ as bias corrected estimator. xx)

(b) For numbers $a_1, \ldots, a_n$ with average $\bar{a}$, show that the average of those remaining after having pushed out a_i can be presented as

$$\bar{a}_{(i)} = \frac{1}{n-1}\sum_{j\neq i} a_j = \frac{(n-1)\bar{a} - (a_i - \bar{a})}{n-1} = \bar{a} - \frac{a_i - \bar{a}}{n-1}.$$

(c) Let $\widehat{\theta}_{(i)}$ be the estimator in question computed for the dataset where y_i is pushed out. Show that (xx care here xx)

$$\widehat{\theta}_{(i)} \doteq \widehat{\theta} - (n-1)^{-1}\mathrm{IF}(F_n, y_i)$$

(xx point is to give what is practical for cross-validation things and then the AIC connection. xx)

(d) Show that

$$\Sigma_n = \sum_{i=1}^n (\widehat{\theta}_{(i)} - \widehat{\theta})(\widehat{\theta}_{(i)} - \widehat{\theta})^{\mathrm{t}} \doteq (1/n^2)\sum_{i=1}^n \mathrm{IF}(F_n, y_i)\,\mathrm{IF}(F_n, y_i)^{\mathrm{t}},$$

and argue that this is a natural estimator of the variance matrix for $\widehat{\theta}$. (xx make connections to other ways of estimating this; all first order the same. so we can handle both bias and variance and approximate normality simply by leave-one-out things. xx)

Ex. 5.37 *0-1 regression from modelling in covariate space, I.* In classical logistic regression, see Ex. 5.25, one models $P(Y_i = 1 \,|\, x_i)$ directly, as a function of the covariate vector x_i, without considering how these x_i actually behave, in the two implicit groups $Y_i = 0$ and $Y_i = 1$; in statistical language, all the x_i have been conditioned upon. Such models may however also be *derived* from working with the distributions of x of the two types.

(a) Suppose there are two groups of covariates, group 0, where x follow density f_0, and group 1, where the x follow density f_1. Assume also that these two groups have prior group probabilities π_0 and π_1. Show that

$$P(Y_i = 1 \,|\, x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} = \frac{(\pi_1/\pi_0) f_1(x)/f_0(x)}{1 + (\pi_1/\pi_0) f_1(x)/f_0(x)} = \frac{\exp\{R(x)\}}{1+\exp\{R(x)\}}.$$

Deduce that if $\log\{f_1(x)/f_0(x)\}$ is linear in x, we have *derived* the logistic regression equation from the models of f_0 and f_1.

(b) Consider the one-dimensional case, with $f_0 \sim \mathrm{N}(\xi_0, \sigma^2)$ and $f_1 \sim \mathrm{N}(\xi_1, \sigma^2)$. Show that

$$\begin{aligned} R(x) &= \log(\pi_1/\pi_0) + \log\{f_1(x)/f_0(x)\} \\ &= \{(\xi_1 - \xi_0)/\sigma^2\}(x - \tfrac{1}{2}(\xi_0 + \xi_1)) + \log(\pi_1/\pi_0). \end{aligned}$$

Deduce that this is logistic regression $P(Y = 1 \,|\, x) = H(a + bx)$, with

$$b = \frac{\xi_1 - \xi_0}{\sigma^2} \quad \text{and} \quad a = \log\frac{\pi_1}{\pi_0} - \tfrac{1}{2}(\xi_1 + \xi_0)b = \log\frac{\pi_1}{\pi_0} - \tfrac{1}{2}\frac{\xi_1^2 - \xi_0^2}{\sigma^2}.$$

(c) Suppose there are group-wise data, $x_{0,1},\ldots,x_{0,n_0}$ from group 0 and $x_{1,1},\ldots,x_{1,n_1}$ from group 1. Assume for simplicity here that π_0 and π_1 are known, and that data have been generated with group sizes reflecting these, so that $\pi_0 = n_0/n$ and $\pi_1 = n_1/n$, with $n = n_0+n_1$ the total sample size. Take now $\widehat{b} = (\widehat{\xi}_1 - \widehat{\xi}_0)/\widehat{\sigma}^2$, where $\widehat{\xi}_0 = \bar{x}_0$ and $\widehat{\xi}_1 = \bar{x}_1$ are data averages, and $\widehat{\sigma}^2 = (n_0/n)\widehat{\sigma}_0^2 + (n_1/n)\widehat{\sigma}_1^2$, combining the two sample variances. Show that

$$\begin{aligned}\sqrt{n}(\widehat{\xi}_0 - \xi_0) &\to_d (\sigma/\pi_0^{1/2})N_0, \quad \sqrt{n}(\widehat{\xi}_1 - \xi_1) \to_d (\sigma/\pi_1^{1/2})N_1,\\ \sqrt{n}(\widehat{\sigma}^2 - \sigma^2) &\to_d \pi_0^{1/2}\sqrt{2}\sigma^2 M_0 + \pi_1^{1/2}\sqrt{2}\sigma^2 M_1 = \sqrt{2}\sigma^2 M,\end{aligned}$$

with N_0, N_1, M_0, M_1 independent standard normals. Deduce that

$$\sqrt{n}(\widehat{b} - b) \to_d \mathrm{N}(0,\tau^2) \quad \text{with } \tau^2 = \frac{1}{\pi_0\pi_1\sigma^2} + \frac{2(\xi_1-\xi_0)^2}{\sigma^4}.$$

(d) With $\widehat{a} = \log(\pi_1/\pi_0) - \frac{1}{2}(\widehat{\xi}_0 + \widehat{\xi}_1)\widehat{b}$, complement the above result by finding the joint limit distribution for $(\sqrt{n}(\widehat{a}-a), \sqrt{n}(\widehat{b}-b))$. Then find the limit distribution of $\widehat{a}+\widehat{b}x_0$, for some given x_0. (xx the point is to compare the variance with that of $\widetilde{a}+\widetilde{b}x_0$, from logistic regression. how much do we earn? some rough notes, to be checked and simplified, and compared to the vector case later. the main point is to compare with plain logistic regression for $a+bx_0$. xx) First, show that

$$\begin{aligned}\sqrt{n}(\widehat{d} - d) &\to_d \sigma(N_1/\pi_1^{1/2} - N_0/\pi_0^{1/2}),\\ \sqrt{n}(\tfrac{1}{2}\widehat{\xi}_0 + \tfrac{1}{2}\widehat{\xi}_1 - \bar{\xi}) &\to_d \tfrac{1}{2}\sigma(N_0/\pi_0^{1/2} + N_1/\pi_1^{1/2}),\\ \sqrt{n}(\widehat{\sigma}^2 - \sigma^2) &\to_d \sqrt{2}\sigma^2 M.\end{aligned}$$

From this, for $\widehat{b} = \widehat{d}/\widehat{\sigma}^2$ estimating $b = d/\sigma^2$, show that

$$\sqrt{n}(\widehat{b} - b) \to_d (1/\sigma)(N_1/\pi_1^{1/2} - N_0/\pi_0^{1/2}) - \sqrt{2}(d/\sigma^2)M$$

and show that its limit variance can be expressed as

$$\tau_0^2 = (1/\sigma^2)\{1/(\pi_0\pi_1) + 2\delta^2\},$$

with $\delta^2 = d^2/\sigma^2$ the squared Mahalanobis distance. Next, with $\widehat{a} = \log(\pi_1/\pi_0) - \frac{1}{2}(\widehat{\xi}_0 + \widehat{\xi}_1)\widehat{b}$, show that

$$\begin{aligned}\sqrt{n}(\widehat{a} - a) \to_d &-\tfrac{1}{2}b\sigma(N_0/\pi_0^{1/2} + N_1/\pi_1^{1/2})\\ &-\bar{\xi}\{(1/\sigma)(N_1/\pi_1^{1/2} - N_0/\pi_0^{1/2}) - \sqrt{2}(d/\sigma^2)M\}.\end{aligned}$$

Deduce that for any given x_0,

$$\begin{aligned}\sqrt{n}(\widehat{a} + \widehat{b}x_0 - a - bx_0) \to_d &(x_0 - \bar{\xi})/\sigma\,(N_1/\pi_1^{1/2} - N_0/\pi_0^{1/2})\\ &-(x_0 - \bar{\xi})\,\sqrt{2}(d/\sigma^2)M - \tfrac{1}{2}(d/\sigma)(N_0/\pi_0^{1/2} + N_1/\pi_1^{1/2}).\end{aligned}$$

Show that its limit variance becomes

$$\tau^2 = \frac{(x_0-\bar{\xi})^2}{\sigma^2}\Big(\frac{1}{\pi_0\pi_1} + 2\delta^2\Big) + \frac{\delta^2}{4\pi_0\pi_1} - \frac{x_0-\bar{\xi}}{\sigma}\frac{d}{\sigma}(1/\pi_1 - 1/\pi_0).$$

(xx again, the crux is to show that this is smaller than the corresponding variance from plain logistic regression. xx)

(e) (xx check these things. also simulations. and compare with the variance of $\widehat{b}_{\rm ml}$ from logistic regression. the covariate modelling approach wins, but not by so much. more. show that the simpler logistic regression estimators $(\widetilde{a},\widetilde{b})$ have large-sample variance matrix J^{-1}/n, with

$$J=\int H(a+bx)\{1-H(a+bx)\}\{\pi_0f_0(x)+\pi_1f_1(x)\}\begin{pmatrix}1,x\\x,x^2\end{pmatrix}{\rm d}x.$$

round off. xx)

(f) (xx include something here, but not too long, about $f_0\sim{\rm N}(\xi_0,\sigma_0^2)$ and $f_1\sim{\rm N}(\xi_1,\sigma_1^2)$, different σ parameters. Show that this leads to logistic regression in x,x^2. a bit on comparing the estimation schemes. xx)

Ex. 5.38 *0-1 regression from modelling in covariate space, II.* Here we extend the setting and generalise results reached in Ex. 5.37, from the one-dimensional to the multidimensional case. Let again the group probabilities be π_0 and π_1, and suppose the covariate vector x has group distributions ${\rm N}_p(\xi_0,\Sigma)$ and ${\rm N}_p(\xi_1,\Sigma)$,

(a) Show that this again leads to logistic regression $P(Y_i=1\,|\,x)=H(a+b^{\rm t}x)$, with

$$b=\Sigma^{-1}(\xi_1-\xi_0)\quad\text{and}\quad a=\log(\pi_1/\pi_0)-\tfrac12(\xi_1^{\rm t}\Sigma^{-1}\xi_1-\xi_0^{\rm t}\Sigma^{-1}\xi_0)$$

and H the logistic transform (5.7). Show that we also have

$$a=\log(\pi_1/\pi_0)-\bar\xi^{\rm t}b=\log(\pi_1/\pi_0)-\bar\xi^{\rm t}\Sigma^{-1}d,$$

where $d=\xi_1-\xi_0$ and $\bar\xi=\frac12(\xi_0+\xi_1)$. The logistic regression method, with $(\widetilde{a},\widetilde{b})$, has large-sample precision J^{-1}/n, with

$$J=\int H(a+b^{\rm t}x)\{1-H(a+b^{\rm t}x)\}\{\pi_0f_0(x)+\pi_1f_1(x)\}\begin{pmatrix}1,x^{\rm t}\\x,xx^{\rm t}\end{pmatrix}{\rm d}x.$$

(b) (xx rough notes just now. xx) We next should learn how $\widehat{b}=\widehat{\Sigma}^{-1}(\widehat{\xi}_1-\widehat{\xi}_0)$ fares compared to the $\widetilde{b}$ from logistic regression. First, with $d=\xi_1-\xi_0$, show that

$$\sqrt{n}(\widehat{d}-d)\to_d(1/\pi_1^{1/2})\Sigma^{1/2}N_1-(1/\pi_0^{1/2})\Sigma^{1/2}N_0\sim\{1/(\pi_0\pi_1)^{1/2}\}\Sigma^{1/2}N,$$

with N_0,N_1 denoting independent draws from ${\rm N}_p(0,I_p)$, and N also having that distribution; the limit variance here is hence $1/(\pi_0\pi_1)\,\Sigma$. Second, with $\widehat{\Sigma}=(n_0/n)\widehat{\Sigma}_0+(n_1/n)\widehat{\sigma}_1$, show via results of Ex. 4.60 that

$$\sqrt{n}(\widehat{\Sigma}-\Sigma)\to_d\pi_0^{1/2}M_0+\pi_1^{1/2}M_1=M,$$

where M is a zero-mean normal matrix with a certain covariance structure for ${\rm cov}(M_{i,j},M_{k,l})$, given there, and independent of N. From that same exercise, $\sqrt{n}(\widehat{\Sigma}^{-1}-\Sigma^{-1})\to_d M^*=-\Sigma^{-1}M\Sigma^{-1}$, with covariance structure ${\rm cov}(M^*_{i,j},M^*_{k,l})=\sigma^{i,k}\sigma^{j,l}+\sigma^{i,l}\sigma^{j,k}$ for indexes i,j,k,l. Finally, from the joint convergence of $\widehat{\Sigma}^{-1}$ and $\widehat{d}$, derive

$$\sqrt{n}(\widehat{b}-b)\to_d M^*d+(\pi_0\pi_1)^{-1/2}\Sigma^{-1/2}N.$$

(c) (xx then round this off with something informative regarding how much we earn by using active modelling of the two group distributions, compared to jumping to logistic regression. we should compare variances of $x_0^{\rm t}\widehat{b}$ and $x_0^{\rm t}\widetilde{b}$, under model conditions. i don't know yet if this gain is larger for $p \geq 2$ than for $p = 1$. we have

$$\sqrt{n}(x_0^{\rm t}\widehat{b} - x_0^{\rm t}b) \to_d x_0^{\rm t}M^*d + (\pi_0\pi_1)^{-1/2}x_0^{\rm t}\Sigma^{-1/2}N.$$

Show that this is a $\mathrm{N}(0,\tau^2)$, with variance

$$\tau^2 = x_0^{\rm t}\Sigma^{-1}x_0\, d^{\rm t}\Sigma^{-1}d + (x_0^{\rm t}\Sigma^{-1}d)^2 + (\pi_0\pi_1)^{-1}x_0^{\rm t}\Sigma^{-1}x_0.$$

a bit more. the point is to show this is smaller than that of $x_0^{\rm t}(J^{-1})_{11}x_0$. xx)

(d) (xx if we manage, with not too complicated answers: joint convergence $\sqrt{n}(\widehat{a}-a)$ and $\sqrt{n}(\widehat{b}-b)$, with $\widehat{a} = \log(\pi_1/\pi_0) - \frac{1}{2}(\widehat{\xi}_0+\widehat{\xi}_1)^{\rm t}\widehat{\Sigma}^{-1}\widehat{d}$. xx) let's see. The crux of the matter is $\bar\xi^{\rm t}b$. Show that

$$\begin{pmatrix}\sqrt{n}(\widehat{d}-d)\\ \sqrt{n}(\frac{1}{2}\widehat{\xi}_0+\frac{1}{2}\widehat{\xi}_1-\bar\xi)\end{pmatrix} \to_d \begin{pmatrix}(1/\pi_1^{1/2})\Sigma^{1/2}N_1 - (1/\pi_0^{1/2})\Sigma^{1/2}N_0\\ \frac{1}{2}(1/\pi_0^{1/2})\Sigma^{1/2}N_0 + \frac{1}{2}(1/\pi_1^{1/2})\Sigma^{1/2}N_1\end{pmatrix} = \begin{pmatrix}U\\V\end{pmatrix},$$

where U, V are jointly multinormal with zero means, $\mathrm{Var}\,U = 1/(\pi_0\pi_1)\,\Sigma$, $\mathrm{Var}\,V = \frac{1}{4}/(\pi_0\pi_1)\,\Sigma$, $\mathrm{cov}(U,V) = \frac{1}{2}(1/\pi_1 - 1/\pi_0)\Sigma$. Also, U is the same variable as was expressed as $(\pi_0\pi_1)^{-1/2}\Sigma^{1/2}N$ above. Deduce that

$$\begin{aligned}\sqrt{n}((\tfrac{1}{2}\widehat{\xi}_0+\tfrac{1}{2}\widehat{\xi}_1)^{\rm t}\widehat{b} - \bar\xi^{\rm t}b) &\doteq \bar\xi^{\rm t}\sqrt{n}(\widehat{b}-b) + \sqrt{n}(\tfrac{1}{2}\widehat{\xi}_0+\tfrac{1}{2}\widehat{\xi}_1-\bar\xi)^{\rm t}b\\ &\to_d \bar\xi^{\rm t}(M^*d + \Sigma^{-1}U) + V^{\rm t}b,\end{aligned}$$

and from this that

$$\begin{pmatrix}\sqrt{n}(\widehat{a}-a)\\ \sqrt{n}(\widehat{b}-b)\end{pmatrix} \to_d \begin{pmatrix}-\bar\xi^{\rm t}(M^*d+\Sigma^{-1}U) + V^{\rm t}b\\ M^*d + \Sigma^{-1}U.\end{pmatrix}$$

(e) For any given x_0, show that

$$\sqrt{n}(\widehat{a} + x_0^{\rm t}\widehat{b} - a - x_0^{\rm t}b) \to_d (x_0-\bar\xi)^{\rm t}(M^*d - \Sigma^{-1}U) + V^{\rm t}b.$$

(xx this is a $\mathrm{N}(0,\tau^2)$, and the point is to compare this variance with that for $\widetilde{a} + x_0^{\rm t}\widetilde{b}$, from the Fisher information matrix limit J. this needs tidying but i think

$$\begin{aligned}\tau^2 &= (x_0-\bar\xi)^{\rm t}\Sigma^{-1}(x_0-\bar\xi)\{(\pi_0\pi_1)^{-1} + d^{\rm t}\Sigma^{-1}d\} + \{(x_0-\bar\xi)^{\rm t}\Sigma^{-1}d\}^2\\ &\quad + \tfrac{1}{4}(\pi_0\pi_1)^{-1}b^{\rm t}\Sigma b - (1/\pi_1 - 1/\pi_0)(x_0-\bar\xi)^{\rm t}\Sigma^{-1}\Sigma b\\ &= (x_0-\bar\xi)^{\rm t}\Sigma^{-1}(x_0-\bar\xi)\{(\pi_0\pi_1)^{-1} + \delta^2\} + \{(x_0-\bar\xi)^{\rm t}\Sigma^{-1}d\}^2\\ &\quad + \tfrac{1}{4}(\pi_0\pi_1)^{-1}\delta^2 - (1/\pi_1 - 1/\pi_0)(x_0-\bar\xi)^{\rm t}\Sigma^{-1}d,\end{aligned}$$

with $\delta^2 = (\xi_1-\xi_0)^{\rm t}\Sigma^{-1}(\xi_1-\xi_0)$ the Mahalanobis distance between the two group distributions.

Ex. 5.39 *Logistic regression with two Beta groups.* In the setting of Ex. 5.37, where group densities $f_0(x)$ and $f_1(x)$ for the covariate vector leads to

$$P(Y_i = 1 \,|\, x) = \frac{\exp\{R(x)\}}{1+\exp\{R(x)\}} \quad \text{with} \quad R(x) = \log \frac{\pi_1 f_1(x)}{\pi_0 f_0(x)},$$

suppose for a different type of situation that the x is one-dimensional and inside the unit interal, and that $f_0 \sim \text{Beta}(a_0, b_0)$ and $f_1 \sim \text{Beta}(a_1, b_1)$.

(a) Show that this leads to

$$P(Y_i = 1 \,|\, x) = \frac{\exp\{\alpha + \beta_1 \log x + \beta_2 \log(1-x)\}}{1+\exp\{\alpha + \beta_1 \log x + \beta_2 \log(1-x)\}},$$

with $\beta_1 = a_1 - a_0$ and $\beta_2 = b_1 - b_0$; also, find a formula for α. This is logistic regression in $\log x$ and $\log(1-x)$.

(b) There are now two (and even more) estimation schemes, for getting at α, β_1, β_2. The first is via direct modelling estimation scheme, using ML for the $x_{0,i}$ data from group 0 and the $x_{1,i}$ data from group 1, to reach $\widehat{\beta}_1 = \widehat{a}_1 - \widehat{a}_0$ and $\widehat{\beta}_2 = \widehat{b}_1 - \widehat{b}_0$. The second is via logistic regression, disregarding the modelling of f_0 and f_1. Set up a simulation experiment, with group probabilities π_0, π_1, sample sizes $n_0 = n\pi_0, n_1 = n\pi_1$, and Beta distributions with $(a_0, b_0) = (c_0\xi_0, c_0(1-\xi_0))$ and $(a_1, b_1) = (c_1\xi_1, c_1(1-\xi_1))$, centred at ξ_0, ξ_1. Compare the two estimation methods, for some choices of these parameters.

(c) Your simulations should show that one earns a lot from separate modelling of the group densities, in this case, when it comes to precision of α, β_1, β_2, under model conditions. Comment further on the pluses and minuses here.

(d) Extend the situation above to the case of two covariates x_1, x_2, taken independent and Beta distributed, with parameters say (a_0, b_0) and (c_0, d_0) for group 0, and (a_1, b_1) and (c_1, d_1) for group 1. Show that this leads to a logistic regression formula in terms of $\log x_1, \log(1-x_1), \log x_2, \log(1-x_2)$. Use again simulation to show that there is much to gain by exploiting these group densities, compared to the usual logistic regression.

(e) xx

Ex. 5.40 *0-1 regression from modelling in covariate space, III.* (xx nils is ranting on a bit, we'll see what can be included and also whether we could write a shortish paper on this. xx) Consider again 0-1 type outcome data, where also the available covariates $x_1, \ldots, x_k$ are 0-1. The theme is to see modelling in the x space might improve on the dafault logistic regression.

(a) For a binomial $Y \sim \text{binom}(n, p)$ we know $\sqrt{n}(\widehat{p}-p) \to_d \text{N}(0, p(1-p))$, where $\widehat{p} = Y/n$. Now transform to $p = H(\theta) = \exp(\theta)/\{1+\exp(\theta)\}$, or $\theta = H^{-1}(p) = \log\{p/(1-p)\}$. Show that

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d \text{N}(0, \tau^2), \quad \text{where } \tau^2 = \frac{1}{p(1-p)} = \frac{\{1+\exp(\theta)\}^2}{\exp(\theta)}. \tag{5.10}$$

(b) Assume $x_1, \ldots, x_k$ are independent 0-1 variables, with parameters $p_0 = (p_{0,1}, \ldots, p_{0,k})$ for group 0 and $p_1 = (p_{1,1}, \ldots, p_{1,k})$ for group 1. We transform these logistically to $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,k})$ and $\theta_1 = (\theta_{1,1}, \ldots, \theta_{1,k})$. Show that

$$f_1(x) = \prod_{j=1}^k \{(1-p_{1,j})^{1-x_j} p_{1,j}^{x_j}\} = \prod_{j=1}^k \frac{\exp(\theta_{1,j} x_j)}{1+\exp(\theta_{1,j})},$$

with a corresponding formula for $f_0(x)$. Show from this that

$$P(Y_i = 1 \,|\, x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} = \frac{\exp\{R(x)\}}{1+\exp\{R(x)\}},$$

with

$$\exp\{R(x)\} = \frac{\pi_1}{\pi_0} \prod_{j=1}^k \frac{1+\exp(\theta_{0,j})}{1+\exp(\theta_{1,j})} \prod_{j=1}^k \exp\{(\theta_{1,j}-\theta_{0,j})x_j\}.$$

Argue that this hence is logistic regression in $x_1, \ldots, x_k$, but with more knowledge behind the coefficients in $H(a + b_1 x_1 + \cdots + b_k x_k)$.

(c) For $b_j = \theta_{1,j} - \theta_{0,j}$, estimated bu $\widehat{b}_j = \widehat{\theta}_{1,j} - \widehat{\theta}_{0,j}$, show that

$$\sqrt{n}(\widehat{b}_j - b_j) \to_d N_{1,j}/\pi_1^{1/2} - N_{0,j}/\pi_0^{1/2},$$

where $N_{0,j}$ and $N_{1,j}$ are independent zero-mean normals with variances $\tau_{0,j}^2$ and $\tau_{1,j}^2$, as per (5.10) above.

(d) (xx then comes comparing these variances, and also for the crucial $\widehat{a} + \widehat{b}^{\mathrm{t}} x_0$, with those from logistic regression. xx) The direct logistic regression method leads to estimating $p(x) = H(a + b_1 x_1 + \cdots + b_k x_k)$ via the direct binary log-likelihood function, i.e. $\sum_{i=1}^n \{y_i \log p_i + (1-y_i)\log(1-p_i)\}$. Here $(\sqrt{n}(\widetilde{a}-a), \sqrt{n}(\widetilde{b}-b)) \to_d \mathrm{N}_{k+1}(0, J^{-1})$, with

$$\begin{aligned} J &= \sum_{\text{all } x} p(x)\{1-p(x)\}\{\pi_0 f_0(x) + \pi_1 f_1(x)\} \begin{pmatrix} 1, & x^{\mathrm{t}} \\ x, & xx^{\mathrm{t}} \end{pmatrix} \\ &= \sum_{\text{all } x} \pi_0 f_0(x) \frac{\exp(a+b^{\mathrm{t}}x)}{1+\exp(a+b^{\mathrm{t}}x)} \begin{pmatrix} 1, & x^{\mathrm{t}} \\ x, & xx^{\mathrm{t}} \end{pmatrix}. \end{aligned}$$

The sum extends over all 2^k possible outcomes for $x = (x_1, \ldots, x_k)^{\mathrm{t}}$. This may be computed numerically, for given θ_0, θ_1. (xx round off. xx)

(e) (xx point also to the very conservative approach of modelling $2^k + 2^k$ probabilities separately. xx)

Ex. 5.41 *Two-group classification.* (xx a little rant, perhaps to be placed in Ch. 16, or in Ch. 6, but it has relevance to binary regression. xx) Two groups, probabilities π_0 and π_1, with feature vectors having densities f_0 and f_1 from 0 and 1. In this exercise we assume that these two class densities are known; in practice they need to be estimated from training data.

(a) Suppose an object comes from group g, and our decision after seeing its x is $\widehat{g}$. Let the loss function be of symmetric 0-1 type, $L(g,\widehat{g}) = I(\widehat{g} \neq g)$. Show that the optimal Bayes solution is to claim '1' if $p(x) \geq \frac{1}{2}$ and '0' if $p(x) < \frac{1}{2}$, where

$$p(x) = P(g = 1 \,|\, x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}.$$

Show that this is equivalent to claiming '1' if $R(x) = \pi_1 f_1(x)/\{\pi_0 f_0(x)\} \geq 1$.

(b) Show that the two group-conditional error rates, using this optimal classifier, is

$$\begin{aligned} \mathrm{err}_0 &= P(\text{saying } 1 \,|\, g = 0) = \int_A f_0 \,\mathrm{d}x, \\ \mathrm{err}_1 &= P(\text{saying } 0 \,|\, g = 1) = \int_{A^c} f_1 \,\mathrm{d}x, \end{aligned}$$

with A the set $\{x\colon R(x) \geq 1\}$. Deduce that the total optimal error rate, seen if this classifier is used for a high number of cases, is

$$\mathrm{err} = \pi_0 \,\mathrm{err}_0 + \pi_1 \,\mathrm{err}_1 = \int_A \pi_0 f_0 \,\mathrm{d}x + \int_{A^c} \pi_1 f_1 \,\mathrm{d}x.$$

(c) Then work out that

$$I = \int |\pi_1 f_1 - \pi_0 f_0| \,\mathrm{d}x = 2 \int_A (\pi_1 f_1 - \pi_0 f_0) \,\mathrm{d}x - (\pi_1 - \pi_0),$$

and from this that

$$\mathrm{err} = \pi_1 - \int_A (\pi_1 f_1 - \pi_0 f_0) \,\mathrm{d}x = \tfrac{1}{2} - \tfrac{1}{2} I.$$

In other words, $\frac{1}{2} + \frac{1}{2} \int |\pi_1 f_1 - \pi_0 f_0| \,\mathrm{d}x$ is the success rate, $P(\widehat{g} = g)$. In particular, for the balanced case of $\pi_0 = \pi_1 = \frac{1}{2}$, show that $\mathrm{err} = \frac{1}{2} - \frac{1}{4} \int |f_1 - f_0| \,\mathrm{d}x$, with success rate $\frac{1}{2} + \frac{1}{4} \int |f_1 - f_0| \,\mathrm{d}x$.

(d) (xx more. Mahalanobis distance. mild but important extension to losses $L_{0,1}$ and $L_{1,0}$. $\mathrm{E}\,\{L(g,0) \,|\, x\} = L_{1,0}(1 - p(x))$ and $\mathrm{E}\,\{L(g,1) \,|\, x\} = L_{0,1} p(x)$. tidy up here. wish to see methods like 'say 1 if $p(x) \geq 0.90$'. xx)

(e) xx

Ex. 5.42 *Estimating in a Weibull model.* Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the one-parameter model with $P(Y_i \leq y) = 1 - \exp(-y^\theta)$ for $y \geq 0$.

(a) Set up a formula for the log-likelihood function. Find the limit distribution for $\sqrt{n}(\widehat{\theta} - \theta)$. (xx this needs: $f(y,\theta) = \exp(-y^\theta)\theta y^{\theta-1}$, with $\log -y^\theta + \log\theta + (\theta - 1)\log y$, then score function

$$u(y,\theta) = 1/\theta + \log y - y^\theta \log y = (1/\theta)(1 + \log v - v \log v),$$

in terms of $v = y^\theta$, a unit exponential. xx)

(b) (xx Analyse the log-likelihood-ratio function. no simple sufficient statistic. check wilks approximation. confidence interval, tests. xx)

(c) xx

Ex. 5.43 *Estimating in the Cauchy location model.* Consider the Cauchy density model $f(y,\theta) = (1/\pi)\{1+(y-\theta)^2\}^{-1}$; see Ex. 1.13.

(a) Show that the score function becomes

$$u(y,\theta) = \partial \log f(y,\theta)/\partial\theta = 2(y-\theta)/\{1+(y-\theta)^2\},$$

and hence that the Fisher information can be expressed as $J(\theta) = \text{Var}\,\{2X/(1+X^2)\}$, with X a unit Cauchy, with density $(1/\pi)(1+x^2)^{-1}$.

(b) We have seen in Ex. 1.13 that X can be represented as U/V, with U and V independent standard normals. Write $U = R\cos\theta$ and $V = R\sin\theta$, to get to

$$2X/(1+X^2) = 2\cos\theta\sin\theta,$$

with θ uniform on $[0,2\pi]$. Then show that $J(\theta) = \frac{1}{2}$. With i.i.d. data $Y_1,\ldots,Y_n$ from the Cauchy model, show for the ML estimator that $\sqrt{n}(\widehat{\theta}_{\rm ml}-\theta) \to_d \text{N}(0,2)$.

(c) xx

Ex. 5.44 *Three tests for the Cauchy location parameter.* Continuing the study of the Cauchy model, from Ex. 5.43, let $X_1,\ldots,X_n$ be i.i.d. from thid Cauchy location model, with density $(1/\pi)/\{1+(x-\theta)^2\}$, and suppose one needs to test θ_0 vs. $\theta \neq 0$.

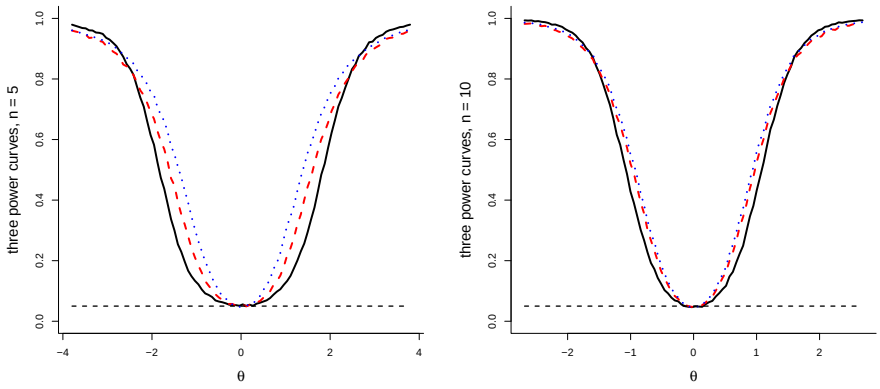

Figure 5.4: *Power functions for three tests for the Cauchy location model, based on A_n (black, full), B_n (red, slanted), D_n (blue, dotted) of Ex. 5.44. The sample sizes are $n = 5$ (left panel) and $n = 10$ (right panel). Note that the θ scale is different in the two panels; from $n = 5$ to $n = 10$, the three powers have increased.*

(a) Show that the log-likelihood function is

$$\ell_n(\theta) = \sum_{i=1}^n \big[-\log\{1 + (x_i - \theta)^2\}\big],$$

modulo a constant. Show, via concrete data examples of size e.g. $n = 5$, that the log-likelihood function can have more than one maximum, i.e. several bumps. Hence one needs to compute the ML estimator $\widehat\theta$ with numerical care. (xx calibrate here, where is what. xx) Show that $\sqrt{n}(\widehat\theta - \theta) \to_d \mathrm{N}(0, \tau^2)$, with $\tau = \sqrt{2}$. (xx compare with median, from Ch2 exercise. xx)

(b) For testing the null, at level say 0.05, three choices are as follows: (i) Compute $A_n = \sqrt{n}\widehat\theta/\tau$, and reject when $|A_n| \geq a_n$, the 0.95 point in the null distribution of $|A_n|$. (ii) Compute $B_n = \sqrt{n}\widehat\theta/\widehat\tau$, and reject when $|B_n| \geq a_n$, the 0.95 point in the null distribution of $|B_n|$; here $\widehat\tau^2 = -(1/n)\partial^2\ell_n(\widehat\theta)/\partial\theta$ is the normalised Hessian associated with the maximisation procedure for finding $\widehat\theta$. (iii) Compute $D_n = 2\{\ell_{n,\max} - \ell_n(0)\}$, and reject when $D_n \geq d_n$, the 0.95 point in the null distribution of D_n. Show that $a_n \to 1.96$, $b_n \to 1.96$, $d_n \to \Gamma_1^{-1}(0.95) = 1.96^2$, as n increases. For our power analysis below we have used exact non-asymptotic values of a_n, b_n, d_n, however, computed via simulations. (xx a bit more: need perhaps $n \geq 30$ for asymptotics to have kicked in properly. xx)

(c) We may now compute the associated power functions

$$\pi_{A,n}(\theta) = P_\theta(|A_n| \geq a_n),\ \pi_{B,n}(\theta) = P_\theta(|B_n| \geq b_n),\ \pi_{D,n}(\theta) = P_\theta(D_n \geq d_n),$$

via simulation. Produce a version of Figure 5.4, where we have used $n = 5$ and $n = 10$. With these two n, and with 10^5 simulations for the null distributions, we arrived at $2.959, 2.387$ for a_n, $2.981, 2.291$ for b_n, $2.144^2, 2.012^2$ for d_n.

(d) (xx round off, some comments. moral: direct $\sqrt{n}\widehat\theta/\tau$ loses to the less direct $\sqrt{n}\widehat\theta/\widehat\tau$, even though τ is a known quantity, for the main area around the null value. but it does win, over the B_n and D_n, at a certain distance away from the null, where the power is high for all competitors. deviance test a bit better than the others, for the main neighbourhood. also: more equal powers with $n = 10$ than with $n = 5$, etc. xx)

(e) xx

Ex. 5.45 *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. 3.16, 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

$$R_n = \frac{\bar{f}(y_1,\ldots,y_n)}{f_0(y_1,\ldots,y_n)} = \int \exp\{\ell_n(\theta)-\ell_n(\theta_0)\}\,\mathrm{d}w(\theta)$$
$$= \exp\{\ell_n(\widehat{\theta})-\ell_n(\theta_0)\}\int \exp\{\ell_n(\theta)-\ell_n(\widehat{\theta})\}\,\mathrm{d}w(\theta).$$

(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 $J_n = -(1/n)\partial^2\ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ 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

$$R_n \doteq \exp(\tfrac{1}{2}D_n)(2\pi)^{p/2}|nJ_n|^{1/2}w(\widehat{\theta}),$$
$$2\log R_n \doteq D_n - p\log n + \log|J_n| + p\log(2\pi) + 2\log w(\widehat{\theta}).$$

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, alsmost regardless of the weighting measure.

(e) (xx round this off. example. xx)

Ex. 5.46 *Finding the ML estimate via simulation.* We have seen in Ex. 1.57 that if Y comes from a model of the form $f(y,\theta)=\exp\{\theta^{\mathrm{t}}T(y)-k(\theta)\}h(y)$, then the ML estimate is the solution to $\xi(\theta)=\mathrm{E}_\theta\,T(Y)=t_{\mathrm{obs}}$, where $t_{\mathrm{obs}}=T(y_{\mathrm{obs}})$. To find $\widehat{\theta}$, for a given dataset, there are various options. If the equation is easy to work with, one solves it; in many other situations, one may programme the log-likelihood function, and throw it to an optimiser. There is a third option, however, described now.

(a) Suppose that there is no easy way of solving $\xi(\theta)=t_{\mathrm{obs}}$, and that the log-likelihood function cannot be worked with, perhaps because the $k(\theta)$ is too complicated or impossible to compute. Suppose however that one may *simulate outcomes* for each θ. With a high number B of outcomes Y^*, for a given θ, explain why $\widehat{\xi}(\theta)=(1/B)\sum_{j=1}^B T(Y_j^*)$ is a good (but simulation based) estimate of $\xi(\theta)$. Explain also how one can assess its variance. Argue furthermore that this may be used to solve $\widehat{\xi}(\theta)=t_{\mathrm{obs}}$, as a simulation based estimate of the real ML estimate.

(b) Work through the following simple illustration of this principle. Let $Y_1,\ldots,Y_n$ be i.i.d. $\mathrm{N}(0,\sigma^2)$. Show that the log of the joint likelihood is $-n\log\sigma-\frac{1}{2}nT_n/\sigma^2$, with $T_n=(1/n)\sum_{i=1}^n y_i^2$, and that the ML estimator admits the exact formula $\widehat{\sigma}=T_n^{1/2}$. Here we pretend that we are not clever enough to do this bit of mathematics, and that

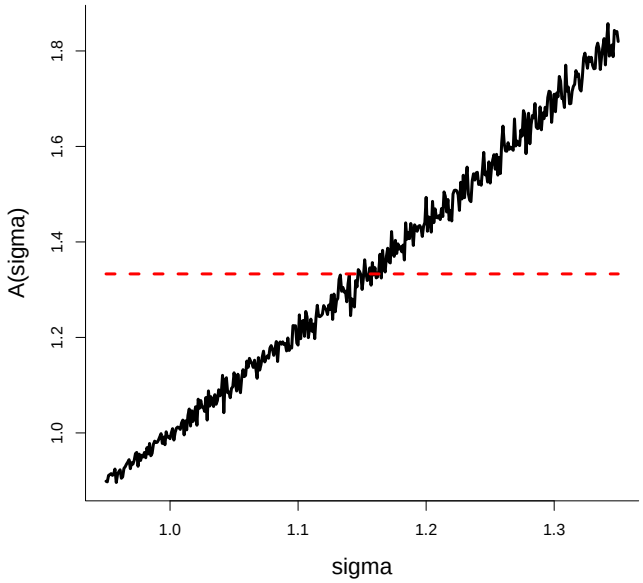

Figure 5.5: *The estimated* $A(\sigma) = \mathrm{E}_\theta\, T_n$*, from* 10^3 *simulated outcomes of* T_n *at each* σ*, to see when it is equal to* $t_{n,\mathrm{obs}} = 1.333$.

we neither are clever enough to programme the log-likelihood function – so we tend to simulations instead. Assume $n = 10$ and that we observe $t_{n,\mathrm{obs}} = 1.333$. Simulate say 10^3 outcomes of T_n for each σ, to compute $A(\sigma) = \mathrm{E}_\sigma\, T_n$. Construct a version of Figure 5.5. Read off the consequent simulation based estimate σ^*, and compare it to the ecact ML estimate $\widehat{\sigma}_{\mathrm{obs}}$.

(c) (xx one more example, to make it interesting. point to magis squares Story vi.3, where we have a subquestion on estimating λ from an MCMC run. xx)

(d)

Ex. 5.47 *The density power divergence estimation method.* (xx to be edited and shortened, but with an example somewhere. we get main results from Ch4 exercises. xx) Here we work through the basics of a robustification method, which can be applied for most models to produce estimation and inference less influenced by extreme and perhaps erroneous data values, when compared with the plain ML. The method 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 et al. (1998), Jones et al. (2001), or the density power divergence 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

$$d_a(g, f_\theta) = \int \Big\{ f_\theta^{1+a} - \Big(1 + \frac{1}{a}\Big) g f_\theta^a + \frac{1}{a} g^{1+a} \Big\} \, \mathrm{d}y, \tag{5.11}$$

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.

(b) Use Taylor expansion for f_θ^a and g^a for small a, to demonstrate that the integrand in (5.11) may be written

$$f_\theta - g + g \log(g/f_\theta) - a(g - f_\theta) \log f_\theta + \tfrac{1}{2} a g \{ (\log g)^2 - (\log f_\theta)^2 \} + O(a^2).$$

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

$$H_n(\theta) = \int f(y,\theta)^{1+a} \, \mathrm{d}y - (1 + 1/a) n^{-1} \sum_{i=1}^n f(y_i, \theta)^a. \tag{5.12}$$

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.

(d) (xx basic theory. sandwich matrix. xx)

(e) (xx examples. xx)

(f) (xx influence function, $\mathrm{IF}(G, y) = J_a^{-1}\{f(y,\theta_0)^a - \xi_a\}$. this is typically bounded in y. xx)

Ex. 5.48 *Gamma regression.* (xx a few versions of gamma regression. also 'doubly linear', in two parameters. this could mean $Y_i \sim \mathrm{Gam}(a_i, b_i)$ with $a_i = \exp(x_i^{\mathrm{t}}\beta)$ and $b_i = \exp(z_i^{\mathrm{t}}\gamma)$. xx)

Ex. 5.49 *Log-linear mixing of densities.* (xx don't know yet how this pans out in the end, but i write down things, before they converge. xx) Consider a density model $f(y,\theta)$, where a certain parameter value θ_0 has some prior credit. There may then be a choice between using $f(y,\theta_0)$ and $f(y,\widehat{\theta})$, the latter using the ML estimate, based on i.i.d. data $Y_1, \ldots, Y_n$. Rather than merely choosing one of them, perhaps via a test procedure or a model selection rule, it might be fruitful to consider the intermediate model

$$\widehat{f}(y,\lambda) = f(y,\theta_0)^{1-\lambda} f(y,\widehat{\theta})^\lambda / R_n(\lambda), \quad \text{where } 0 \leq \lambda \leq 1,$$

and $R_n(\lambda)$ is the required normalisation value $\int f(y,\theta_0)^{1-\lambda} f(y,\widehat{\theta})^\lambda \, \mathrm{d}y$. The endpoints $\lambda = 0$ and $\lambda = 1$ correspond to the null model and the ML estimated model, respectively.

(a) Show that the derivative of $R_n(\lambda)$, at the endpoints 0 and 1, can be written

$$R_n'(0) = -d(f(\cdot,\theta_0), f(\cdot,\widehat{\theta})), \quad R_n'(1) = d(f(\cdot,\widehat{\theta}), f(\cdot,\theta_0)),$$

in terms of the Kullback–Leibler distance $d(g,f)$ discussed in Ex. 5.17 and later on. In particular, the $R_n(\lambda)$ has negative derivative at the start and positive derivative at the end.

(b) One idea for choosing the λ is to use the pseudo-log-likelihood function, constructed as if $f(y,\widehat{\theta})$ were a known density, with only λ unknown. Show that this leads to $\ell_n^*(\lambda) = \lambda S_n - n\log R_n(\lambda)$, with $S_n = \sum_{i=1}^n \{\log f(y_i,\widehat{\theta}) - \log f(y_i,\theta_0)\}$. Show in particular that $\ell_n^*(0) = 0$, that $\ell_n^*(1) = S_n$, and that the derivative at zero is positive. Show also that ℓ_n^* is concave.

(c) To more easily see the behaviour of the $\ell_n^*(\lambda)$, consider the exponential family case described in Ex. 1.57, with $f(y,\theta) = \exp\{\theta^{\rm t}T(y) - k(\theta)\}h(y)$. Show that

$$S_n = n\big[(\widehat{\theta}-\theta_0)^{\rm t}\bar{T} - \{k(\widehat{\theta}) - k(\theta_0)\}\big]$$

and that this actually is identical to $nR_n'(1)$. Hence deduce that the scheme above does not really lead to a new estimator, but back to the ML density $f(y,\widehat{\theta})$.

(d) (xx a bit more, with Bayes; perhaps the full exercise is to land in Ch. 6. xx) Consider a setup with two known densities f_0 and f_1, and then the log-linear mixing models $f(y,\lambda) = f_0(y)^{1-\lambda}f_1(y)^\lambda/R(\lambda)$. With a prior $\pi_0(\lambda)$ for the mixing parameter, over $[0,1]$, show that the posterior density becomes proportional to $\pi_0(\lambda)\exp(\lambda S_n)$, with $S_n = \sum_{i=1}^n \log\{f_1(y_i)/f_0(y_i)\}$. In particular, with a uniform prior for λ, show that

$$\pi(\lambda\,|\,\text{data}) = \frac{\exp(\lambda S_n)}{\{\exp(S_n)-1\}/S_n} \quad \text{for } \lambda\in[0,1].$$

(e) If data really come from f_0, show that $P(\lambda\in[0,0.01]\,|\,\text{data})\to 1$, with probability 1; if data instead really come from f_1, show that $P(\lambda\in[0.99,1]\,|\,\text{data})\to 1$.

(f) (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)

Ex. 5.50 *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 $-n\log\sigma - \frac12 Q_n(\beta)/\sigma^2$, 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 $\sqrt{n}(\widehat{\beta}-\beta) \to_d \mathrm{N}(0,\sigma^2\Sigma^{-1})$, 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. 5.51 *Estimating cycle length.* Consider the model with mean structure $m_i = a\cos(2\pi i/\omega+\phi)$, common in time series models, with ω related to the length of a cycle. Assume first that a,ϕ are known. Work out the asymptotics for the ML estimator $\widehat{\omega}$. (xx answer: it is surprisingly sharp. xx) Then assume all three parameters need to be estimated. (xx again: ω can be estimated with high precision i think. xx)

(a) For the case of known a,ϕ, and cycle length ω estimated by minimising $Q_n(\omega) = \sum_{i=1}^n\{y_i - a\cos(2\pi i/\omega\phi)\}^2$, show that

$$n^{3/2}(\widehat{\omega}-\omega) \to_d \frac{\sqrt{6}}{2\pi}\omega^2\frac{\sigma}{a}\mathrm{N}(0,1).$$

Ex. 5.52 *ML behaviour under local alternatives, I.* Consider a parametric density $f(y,\theta)$ for observations $Y_1,\ldots,Y_n$. Assume however that these do not precisely follow the model, but rather a density in its vicinity. What can we deduce for the behaviour of the ML estimator $\widehat{\theta}$? (xx results are useful for local power etc., and also for FIC in Ch 11, see Ex. 5.53. xx)

(a) Assume the real state of affairs is a density $f_n(y) = f(y,\theta_0)+\delta h(y)/\sqrt{n}$. Use notation and the same line of arguments as in Ex. 5.7 to show that

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \{-n^{-1}I_n(\theta_0)\}^{-1}n^{-1/2}U_n(\theta_0) + o_{\mathrm{pr}}(1).$$

Show next that $n^{-1/2}U_n(\theta_0)$ has mean $b\delta$, with $b=\int h(y)u(y,\theta_0)\,\mathrm{d}y$, and variance matrix $J+O(1/n)$. Use the Lindeberg theorem to infer that $n^{-1/2}U_n(\theta_0)\to_d \mathrm{N}_p(b\delta,J)$.

(b) Show also that $-n^{-1}I_n(\theta_0)\to_{\mathrm{pr}} J$, i.e. even when the true density f_n is $O(1/\sqrt{n})$ away from $f(y,\theta_0)$. Conclude that $\sqrt{n}(\widehat{\theta}-\theta_0)\to_d \mathrm{N}_p(J^{-1}b\delta,J^{-1})$.

(c) (xx invent a simple illustration. point to local power things already in Ch 3. xx)

Ex. 5.53 *ML behaviour under local alternatives, II.* (xx point here to FIC of Ch. 11. xx) As a useful special case of the above, assume that a certain narrow model with density $f_0(y,\theta)$ is used, with dimension p, but that the reality is that data come from a wider model, needing an extra parameter γ of dimension q. Suppose indeed that the real density is $f_{\mathrm{true}}(y) = f(y,\theta_0,\gamma_0+\delta/\sqrt{n})$, where setting $\gamma=\gamma_0$ in the $p+q$-dimensional model $f(y,\theta,\gamma)$ gives us back the $f_0(y,\theta)$ model. For this wider model, introduce score vector component $u_0(y) = \partial\log f(y,\theta_0,\gamma_0)/\partial\theta$ and $u_1(y) = \partial\log f(y,\theta_0,\gamma_0)/\partial\gamma$. We also need to introduce

$$J=\begin{pmatrix}J_{00}&J_{01}\\J_{10}&J_{11}\end{pmatrix} \quad\text{with inverse}\quad J^{-1}=\begin{pmatrix}J^{00}&J^{01}\\J^{10}&J^{11}\end{pmatrix},$$

the $(p+q)\times(p+q)$ Fisher information matrix for the wider model, with inverse, computed at the position (θ_0,γ_0).

(a) The $q \times q$ submatrix $Q = J^{11}$ will serve a crucal role, here and in the building of Focused Information Criteria in Ch. 11, so we give it its own name, the Q matrix. By multiplying out $JJ^{-1} = I_{p+q}$, show the following.

$$Q = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}, \quad J^{01} = -J_{00}^{-1}J_{01}Q.$$

(b) Show now that

$$\begin{pmatrix} U_{n,0} \\ U_{n,1} \end{pmatrix} = \begin{pmatrix} n^{-1/2}\sum_{i=1}^n u_0(Y_i) \\ n^{-1/2}\sum_{i=1}^n u_1(Y_i) \end{pmatrix} \to_d \begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim \mathrm{N}_{p+q}(\begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix}, \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}).$$

We also write $U_{n,0} = \sqrt{n}\bar{U}_0$, $U_{n,1} = \sqrt{n}\bar{U}_1$. With $\widehat{\theta}_{\rm narr}$ the ML estimator in the narrow model, show that

$$\sqrt{n}(\widehat{\theta}_{\rm narr} - \theta_0) = J_{00}^{-1}\sqrt{n}\bar{U}_{n,0} + o_{\rm pr}(1) \to_d J_{00}^{-1}U_0 \sim \mathrm{N}_p(J_{00}^{-1}J_{01}\delta, J_{00}^{-1}).$$

(c) Next consider ML estimation in this wide model, with $(\widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide})$. Show that

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{\rm wide} - \theta_0) \\ \sqrt{n}(\widehat{\gamma}_{\rm wide} - \gamma_0) \end{pmatrix} \to_d J^{-1}\begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim \mathrm{N}_{p+q}(\begin{pmatrix} 0 \\ \delta \end{pmatrix}, J^{-1}).$$

In particular, note that

$$D_n = \sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d D \sim \mathrm{N}_q(\delta, Q). \tag{5.13}$$

(d) So how do narrow model estimation and wide model estimation pan out, when it comes to some given focus parameter? Consider such a $\phi = \phi(\theta, \gamma)$, with true value $\phi_{\rm true} = \phi(\theta_0, \gamma_0 + \delta/\sqrt{n})$. The two estimators under consideration are $\widehat{\phi}_{\rm narr} = \phi(\widehat{\theta}_{\rm narr}, \gamma_0)$ and $\widehat{\phi}_{\rm wide} = \phi(\widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide})$. Show that

$$\begin{aligned} \sqrt{n}(\widehat{\phi}_{\rm narr} - \phi_{\rm true}) &\to_d \mathrm{N}(\omega^{\rm t}\delta, \tau_0^2), \\ \sqrt{n}(\widehat{\phi}_{\rm wide} - \phi_{\rm true}) &\to_d \mathrm{N}(0, \tau_0^2 + \omega^{\rm t}Q\omega), \end{aligned} \tag{5.14}$$

featuring

$$\tau_0^2 = (\tfrac{\partial\mu}{\partial\theta})^{\rm t}J_{00}^{-1}\tfrac{\partial\mu}{\partial\theta} \quad \text{and} \quad \omega = J_{10}J_{00}^{-1}\tfrac{\partial\mu}{\partial\theta} - \tfrac{\partial\mu}{\partial\gamma},$$

with these partial derivatives computed at the null position (θ_0, γ_0).

(e) So when is narrow model estimation still best, even if that model might be somewhat wrong? Narrow model means bias but smaller variance; wide model means lower bias but higher variance. This bias-variance trade-off is nicely captured in the (5.14) limits. Show that the limit risks, mean squared errors for the limit distributions, are

$$\mathrm{mse}_{\rm narr} = \tau_0^2 + (\omega^{\rm t}\delta)^2 \quad \text{and} \quad \mathrm{mse}_{\rm wide} = \tau_0^2 + \omega^{\rm t}Q\omega.$$

Conclude that narrow is best, provided $|\omega^{\rm t}\delta| \le (\omega^{\rm t}Q\omega)^{1/2}$.

(f) (xx more here, rounding it off, with a bit of commentary, an exercise finding the tolerance radius, $\sqrt{n}|\gamma - \gamma_0| \le \kappa$ for one-extra-para case, it is sample-size dependent as it should, make sure regression models mentioned properly too, pointing to FIC things for Ch. 11. xx)

Ex. 5.54 *ML behaviour under local alternatives, III.* (xx to come here, nils hoping to make it not long and not complicated. to be used in Ch7 and some stories. starting point: $y \sim \mathrm{N}(\theta, 1)$, where we know $\theta \geq 0$. then ML is $\widehat{\theta} = y\, I(y \geq 0)$, and with a positive probability of being zero. then $\sqrt{n}(\bar{y}_n - \delta/\sqrt{n})$ in the case of $\theta = \delta/\sqrt{n}$. then general ML and profiled, aiming for $D_n = \sqrt{n}(\widehat{\gamma} - \gamma_0)$ when $\gamma = \gamma_0 + \delta/\sqrt{n}$ and we have $\gamma \geq \gamma_0$ a priori. consequences also for log-likelihood-ratio. xx)

Ex. 5.55 *Proving the Wilks theorem.* (xx as of 13-Aug-2023, we need an opprydning and editing and cleaning and outpushing in the various Wilks things in this chapter. here nils klipper inn points for proving the Wilks, to be calibrated with other material, proofs, profiling, uses. xx) Suppose $Y_1, \ldots, Y_n$ are i.i.d., where two models are considered: a narrow one, namely $f_0(y, \theta)$, with θ of dimension p; and a wide one, namely $f(y, \theta, \gamma)$, needing a further parameter vector γ of dimension q. We need the narrow model to be inside the wide one, so we assume that there is a γ_0 for which $f_0(y, \theta) = f(y, \theta, \gamma_0)$. We assume that γ_0 is an inner point in its parameter domain. We wish to construct a test for the hypothesis H_0 that the narrow model holds, and this is equivalent to testing $\gamma = \gamma_0$.

Let $\ell_n(\theta, \gamma)$ be the log-likelihood function for the wide model, which also means $\ell_n(\theta, \gamma_0)$ is the log-likelihood function in the narrow model. Let $(\widehat{\theta}, \widehat{\gamma})$ be ML estimates in the wide model and $(\widetilde{\theta}, \gamma_0)$ ML estimates in the narrow model. Assuming that H_0 is in force, with density $f(y, \theta_0, \gamma_0)$ for the appropriate θ_0, we already know the principal answers regarding limit distributions for $\sqrt{n}(\widehat{\theta} - \theta_0, \widehat{\gamma} - \gamma_0)$ and $\sqrt{n}(\widetilde{\theta} - \theta_0)$ separately, but now we need to study them jointly, which calls for accurate representations and for linear matrix algebra to sort things out. Let the $(p+q) \times (p+q)$ information matrix $J = J(\theta_0, \gamma_0)$ and its inverse be partitioned into blocks:

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{and} \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}.$$

(a) Here and (xx point to later settings xx) the matrix $Q = J^{11}$ serves a special role. Show via matrix manipulations of $JJ^{-1} = I = J^{-1}J$ that $Q = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, similarly that $J^{00} = (J_{00} - J_{01}J_{11}^{-1}J_{10})^{-1}$, and that $J^{01} = -J_{00}^{-1}J_{01}J^{11}$.

(b) Show that there is simultaneous convergence in distribution

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_0) \\ \sqrt{n}(\widehat{\gamma} - \gamma_0) \end{pmatrix} \to_d \begin{pmatrix} A \\ B \end{pmatrix} = J^{-1} \begin{pmatrix} U \\ V \end{pmatrix} \quad \text{and} \quad \sqrt{n}(\widetilde{\theta} - \theta_0) \to_d C = J_{00}^{-1}U,$$

where

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathrm{N}_{p+q}(0, J) \quad \text{and hence} \quad \begin{pmatrix} A \\ B \end{pmatrix} = J^{-1} \begin{pmatrix} U \\ V \end{pmatrix} \sim \mathrm{N}_{p+q}(0, J^{-1}).$$

Show in particular that $B_n = \sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d B \sim \mathrm{N}_q(0, Q)$ under the narrow model.

(c) Do Taylor expansion around $(\widehat{\theta}, \widehat{\gamma})$ to show that

$$\sum_{i=1}^{n} \{\log f(Y_i, \widehat{\theta}, \widehat{\gamma}) - \log f(Y_i, \widetilde{\theta}, \gamma_0)\} = \tfrac{1}{2} n \begin{pmatrix} \widehat{\theta} - \widetilde{\theta} \\ \widehat{\gamma} - \gamma_0 \end{pmatrix}^{\mathrm{t}} J_n^* \begin{pmatrix} \widehat{\theta} - \widetilde{\theta} \\ \widehat{\gamma} - \gamma_0 \end{pmatrix} + \varepsilon_n,$$

where the J_n^* matrix tends to J in probability and $\varepsilon_n \to_{\rm pr} 0$. Hence conclude that

$$\Delta_n = 2(\ell_{\max,\text{wide}} - \ell_{\max,\text{narr}}) \to_d \Delta = \begin{pmatrix} A-C \\ B \end{pmatrix}^{\rm t} \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \begin{pmatrix} A-C \\ B \end{pmatrix},$$

provided the narrow model holds, i.e. under H_0.

(d) It remains to establish that the limiting variable Δ has the advertised nice chi squared distribution. This is not obvious from its expression above – but do it by first discovering $A - C = -J_{00}^{-1} J_{01} B$ and then plugging in to simplify the expression for Δ. The result is $\Delta = B^{\rm t} Q^{-1} B$, which is a χ^2_q. – A rephrasing of this important result is as follows: If $\mathcal{M}_0$ is a model contained in a bigger $\mathcal{M}_1$ model, then twice the difference of maximised log-likelihoods, which is also by definition the *deviance distance* from the narrow model to the wider model, goes under the narrow model conditions to $\chi^2_{\rm df}$, with $\text{df} = \dim(\mathcal{M}_1) - \dim(\mathcal{M}_0)$.

(e) xx should do local power too, under $f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$. i think that $B_n = \sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d B \sim {\rm N}_q(\delta, Q)$, and that

$$\Delta_n \to_d \Delta = B^{\rm t} Q^{-1} B \sim \chi^2_q(\delta^{\rm t} Q^{-1} \delta).$$

also, of separate interest: the log-LR Δ_n test is asymptotically equivalent to $\Delta_n' = B_n^{\rm t} \widehat{Q}^{-1} B_n = n(\widehat{\gamma} - \gamma_0)^{\rm t} \widehat{Q}^{-1} (\widehat{\gamma} - \gamma_0) \to_d B^{\rm t} Q^{-1} B$. write this out. xx

(f) (xx extension to regression models. xx)

Ex. 5.56 *Wilks Theorem for k-dim subsets of p-dim parameter space.* (xx perhaps a Leftover; check with that is part of intro exercises about Wilks, via efforts of Ch4. xx) Material on Wilks Theorems is not 'naturally completed' before we also come to and include the lifting from dimension 1 to dimension k, so to speak. The basic story is simple to summarise, though not necessarily easy to prove with all the required steps, also since there are different versions and setups. The main story, at any rate, is as follows. Suppose we have n observations from a model $f(y, \theta)$, perhaps with regression parameters etc. Here θ is 'the full parameter vector', belonging to a parameter region Ω, in say p-dimensional space. Then there's a well defined log-likelihood function, say

$$\ell_n(\theta) = \sum_{i=1}^n \log f_i(y_i, \theta).$$

Suppose one is interested in testing whether $\theta \in \Omega_0$ a subset of lower dimension $k < p$; perhaps this corresponds to having $\theta_j = 0$ for $p - k$ of the components. Then we may define and compute

$$\ell_{\max,\text{all}} = \max\{\ell_n(\theta) \colon \theta \in \Omega\} \quad \text{and} \quad \ell_{\max,H_0} = \max\{\ell_n(\theta) \colon \theta \in \Omega_0\},$$

the maximised log-likelihood values under the full model and under the hypothesis H_0 that θ lies in this smaller space. Maxing over a bigger space yields a bigger number than maxing the same function over a small space. Then consider

$$\Delta_n = 2(\ell_{\max,\text{all}} - \ell_{\max,H_0}).$$

Then the splendidly useful Wilks theorem, going all the way back to his 1938 paper, says that under H_0 conditions,

$$\Delta_n \to_d \chi^2_{\mathrm{df}}, \quad \text{with df} = p - k.$$

This is often presented, and made easier to remember and to use, by 'counting the degrees of freedom' as the dimension a priori minus the dimension under the hypothesis.

(a) Assume the H_0 in question is the simple one of $\theta = \theta_0$, so Ω_0 is a single point, of dimension zero. Verify that the Wilks theorem then is the same as what we've seen earlier, e.g. from Ex. 5.22 [xx and one more? xx].

(b) Assume next that H_0 corresponds to $\phi = h(\theta) = \phi_0$, with $h(\theta)$ a smooth one-dimensional function. Note that saying $h(\theta) = \phi_0$ amounts to characterising a $(p-1)$-dimensional subspace of Ω. Verify that the general Wilks theorem above then corresponds to what we've worked with in the previous few exercises, with the deviance function, its limiting χ^2_1 distribution at the hypothesised value, etc.

(c) (xx a bit more. xx)

Ex. 5.57 *Minimum divergence estimators.* (xx leftover: perhaps fully covered by things in Ch4. as of 13-Aug-2023, a little nils rant, which might be formalised and polished into something useful and generic, with M estimation and other schemes as special cases, also BHHJ. xx) Consider in general terms some discrepancy measure $d(g, f_\theta)$, interpreted as the distance from density g to some parametric approximation f_θ. Suppose further that with observed data $Y_1, \ldots, Y_n$ from g, there is a connection from such a discrepancy to an empirical $Q_n(\theta)$. We take this to mean that $Q_n(\theta) \to_{\mathrm{pr}} Q(\theta)$, with $Q(\theta)$ and $d(g, f_\theta)$ having the same minimiser, the least false parameter value θ_0. We call the minimiser $\widehat{\theta} = \widehat{\theta}_n$ of θ a minimum divergence estimator. (xx the above not fully written out yet. xx)

(a) To motivate further work below, establish that maximum likelihood estimation is actually a special case: with $Q_n(\theta) = -\ell_n(\theta)/n$, show that the limit becomes $Q(\theta) = -\int g \log f_\theta \, \mathrm{d}y$, which is a constant away from the KL distance. A more general theory for minimum divergence estimators hence subsumes ML estimation as a special case.

s

(b) Consider the first and second order derivatives $Q'_n(\theta) = \partial Q_n(\theta)/\partial\theta$ and $Q''_n(\theta) = \partial^2 Q_n(\theta)/\partial\theta\partial\theta^{\mathrm{t}}$. Assume (i) that $U_n = \sqrt{n} Q'_n(\theta_0) \to_d U$ and (ii) that $Q''_n(\theta_0) \to_{\mathrm{pr}} J$. Use Taylor expansion

$$0 = Q'_n(\widehat{\theta}) = Q'_n(\theta_0) + Q''_n(\theta_0)(\widehat{\theta} - \theta_0) + \varepsilon_n,$$

where ε_n goes sufficiently quickly to zero in probability, to show that under regularity conditions,

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -Q''_n(\theta_0) U_n + \varepsilon'_n \to_d -J^{-1} U.$$

It might as easy for nils-emil to work with the process

$$\begin{aligned} H_n(s) &= n\{Q_n(\theta + s/\sqrt{n}) - Q_n(\theta_0)\} \\ &= U_n^{\mathrm{t}} s + \tfrac{1}{2} s^{\mathrm{t}} Q_n''(\theta_0) s + \varepsilon_n'' \to_d H(s) = U^{\mathrm{t}} s + \tfrac{1}{2} s^{\mathrm{t}} J s. \end{aligned}$$

From this, show the two basic results, (i) that $\sqrt{n}(\widehat{\theta} - \theta_0) \to_d -J^{-1}U$, (ii) that $n\{Q_n(\theta_0) - Q_{n,\min}\} \to_d \frac{1}{2} U^{\mathrm{t}} J^{-1} U$.

(c) Recreate the basic ML estimation method results from this.

(d) (xx then to M-estimation, a couple of examples, including arcus tangens estimator. xx)

(e) (xx then to BHHJ. xx)

5.C Notes and pointers

[xx CR bound: In its simplest form, the inequality goes back to Cramér (1946) and Rao (1945). 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).

[xx check and calibrate what's here and what's in Ch. 7, regarding CD things. 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. [xx pointers to Ch. 7, 15. xx]

6.A Chapter introduction

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 $P_{\theta_0}(\theta_0 \in I_n) = 0.95$; 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 $P(\theta \in J_n \,|\, y) = 0.95$. 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).

(xx two more paragraphs to come. action space, loss function, Master Recipe is to minimise posterior expected loss, and pointers to Bayesian nonparametrics in Ex. 15. xx)

6.B Short and crisp

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.

(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 Pois(θ) 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. $\sim$ Pois(θ) 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 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 curves $p(\theta \,|\, \text{data}_j)$, where data_j is $y_1, \ldots, y_j$, along with the prior density. 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 $j = 1, \ldots, 10$.

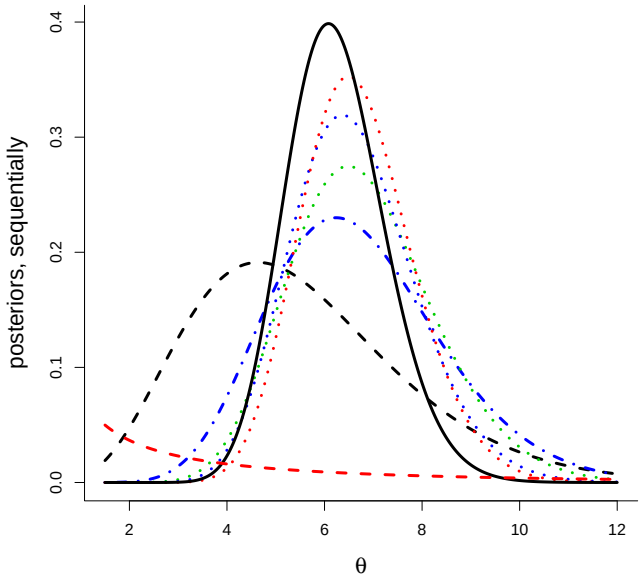

Figure 6.1: *Seven curves are displayed, corresponding to the* Gam$(0.1, 0.1)$ *intial prior density for the Poisson parameter* θ*, along with the six first updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3.*

(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.

Ex. 6.2 *The Master Recipe for finding the Bayes solution.* Due to the importance of Bayes solutions we start by showing how to derive them. (xx more rydding here, re intro here and intro text above, also to be calibrated with Ch 7. need to define $L(\theta, a)$, an action space, etc. xx)

(a) Show that the posterior density of θ, that is, the distribution of the parameter given the data, takes the form

$$\pi(\theta \,|\, y) = m(y)^{-1} f_\theta(y)\pi(\theta),$$

where $m(y)$ is the required integration constant $\int_\Theta f_\theta(y)\pi(\theta)\,\mathrm{d}\theta$. 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'.

(b) Show also that the *marginal distribution* of the data y is $m(y)$.

(c) Show that the Bayes risk may be expressed as

$$\mathrm{BR}(a,\pi) = \mathrm{E}_\theta\, \mathrm{E}_\pi\, [L(\theta, a(Y)) \,|\, Y] = \int_{\mathcal{Y}} \left\{ \int_\Theta L(a(y),\theta)\pi(\theta \,|\, y)\, \mathrm{d}\theta \right\} m(y)\, \mathrm{d}y.$$

The inner integral, or 'inner expectation', is $\mathrm{E}_\pi\, \{L(\theta, a(y)) \,|\, Y = y\}$, that is, the expected loss given data.

(d) 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 $g = g(a)$ is the expected posterior loss,

$$g(a) = \mathrm{E}_\theta\, \{L(\theta, a) \,|\, y\}.$$

The function g is evaluated and minimised over all a, for the given data y. This is the Bayes recipe.

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

$$\widehat{\theta}_B = \frac{\mathrm{E}\,\{w(\theta)\theta \,|\, \text{data}\}}{\mathrm{E}\,\{w(\theta) \,|\, \text{data}\}}.$$

In particular, when estimating a positive θ using loss $(a-\theta)^2/\theta$, show that the Bayes estimator is $1/\mathrm{E}\,\{(1/\theta) \,|\, \text{data}\}$.

(c) Consider the natural absolute loss function, $L(\theta, a) = |a-\theta|$. Show that the Bayes solution becomes the posterior median, i.e. $\widehat{\theta}_B = G^{-1}(\frac{1}{2} \,|\, \text{data})$, where $G(\theta \,|\, \text{data})$ 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 $L(\theta, a) = (a-\theta)^{\mathrm{t}} M (a-\theta)$, 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?* (xx to be done. unknown number: N. likelihood $(1/N)^n\, I(y_1 \le N, \ldots, y_n \le N)$, for n numbers observed. first 203, then also 157, 222. different priors. create your own. find $\widehat{N}$ and 90 percent interval. xx)

Ex. 6.5 *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 (xx point to Ch 2 stuff xx), but we now consider this problem from a Bayesian viewpoint.

(a) The decision space is $\{\text{accept}, \text{reject}\}$. For the loss function, take $L(\theta, \text{accept})$ equal to 0 or L_0, if θ is inside or outside Ω_0, and $L(\theta, \text{reject})$ equal to 0 or L_1, if θ is outside or inside Ω_0. Show that

$$\mathrm{E}\,\{L(\theta, \text{accept}) \,|\, \text{data}\} = L_0\, p(\text{data}), \quad \mathrm{E}\,\{L(\theta, \text{reject}) \,|\, \text{data}\} = L_1\,\{1 - p(\text{data})\},$$

where $p(\text{data}) = P(\theta \in \Omega_0^c \,|\, \text{data})$, 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 $p(\text{data})$ for its falseness is sufficiently overwhelming, namely when $p(\text{data}) \ge L_1/(L_0 + L_1)$. – 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 $\{\text{accept}, \text{reject}, \text{doubt}\}$, 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)

(d) xx

Ex. 6.6 *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 $p_j(\text{data}) = P(\theta \in \Omega_j \,|\, \text{data})$. 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.

Ex. 6.7 *The binomial-beta setup.* Let Y given θ be a binomial (n, θ), and for θ take a Beta(a, b) prior, see Ex. 1.25. 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 $r(\theta) = \text{E}_\theta\,(\widehat{\theta} - \theta)^2$, 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. 6.8 *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 1.24.

(a) Show the important and useful result that $(p_1, \ldots, p_k) \,|\, (y_1, \ldots, y_k) \sim \text{Dir}(a_1 + y_1, \ldots, a_k + y_k)$.

(b) Show that the Bayes estimator under squared error loss becomes

$$\widehat{p}_{i,B} = \text{E}\,(p_i \,|\, \text{data}) = \frac{a_i + y_i}{a + n} = (1 - w_n)p_{0,i} + w_n(y_i/n)$$

for $i = 1, \ldots, k$, with prior means $p_{0,i} = a_i/a$, with $a = a_1 + \cdots + a_k$, and weight $w_n = n/(a + n)$.

(c) Find also the posterior variance and posterior correlation between p_i and p_j-

(d) (xx just a bit more. xx)

Ex. 6.9 *Gott würfelt nicht.* For the multinomial-Dirichlet setup of Ex. 6.8, we reached the posterior characterisation $p \,|\, \text{data} \sim \text{Dir}(a_1 + y_1, \ldots, a_k + y_k)$. 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. 1.24. Concretely, show that one may use $p_1 = G_1/G, \ldots, p_k = G_k/G$, with independent $G_1 \sim \text{Gam}(a_1 + y_1), \ldots, G_k \sim \text{Gam}(a_k + y_k)$, 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', $\text{Dir}(1, 1, 1, 1, 1, 1)$; (ii) 'symmetric and more confident', $\text{Dir}(3, 3, 3, 3, 3, 3)$; (iii) 'unwilling to guess', $\text{Dir}(0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$, for the probabilities $(p_1, \ldots, p_6)$, to assess the posterior distribution of each of the following quantities (xx how do we do aligning here xx):

$$\alpha = p_6/p_1, \qquad \beta = (1/6)\sum_{j=1}^{6}(p_j - 1/6)^2,$$
$$\gamma = (1/6)\sum_{j=1}^{6}|p_j - 1/6|, \qquad \delta = (p_4p_5p_6)^{1/3}/(p_1p_2p_3)^{1/3}.$$

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) (xx polish and xref to mixtures. xx) 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 ok, then, perhaps $\rho = 1$, $\alpha = 0$, $\beta = 0$, $\gamma = 1$, after all. This motivates using a mixture prior which allows a positive chance for $p = p_0$. Please therefore redo the Bayesian analysis above, with the same $(2, 5, 3, 7, 5, 8)$ data, for the prior $\frac{1}{2}\,\delta(p_0) + \frac{1}{2}\,\text{Dir}(1, 1, 1, 1, 1, 1)$. 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 $\rho, \alpha, \beta, \gamma$.

(e) xx

Ex. 6.10 *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. 6.11, 6.12.

(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. 1.30 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 $1/\tau^2=v/\sigma^2$, show that

$$\bar f(y_1,\ldots,y_n)=\frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp(-\tfrac12Q_0/\sigma^2)\Big(\frac{v}{v+n}\Big)^{1/2}\exp\{-\tfrac12d_n(\bar y-\xi_0)^2/\sigma^2\},$$

where $d_n=(1/n+1/v)^{-1}$.

(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. 6.11 *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. 6.10 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. 6.10(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) (xx an illustration here. we ask readers to fix the GN prior according to wishes for σ and for $\xi \,|\, \sigma$. perhaps cigarette consumption per state data, data in 2.B. we ask for more than merely ξ and σ inference, could e.g. ask for $P(Y \geq y_0)$, and for prediction. xx)

(g) (xx same data, but with the uninformative version of the GN prior. discuss differences in results and interpretation. xx)

Ex. 6.12 *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. 6.11, 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. 6.11. Show first that the likelihood times the prior, $L_n(\lambda,\xi)p(\lambda,\xi)$, can be expressed as

$$\frac{\lambda^{n/2}}{(2\pi)^{n/2}}\exp[-\tfrac12\lambda\{Q_0+n(\xi-\bar y)^2\}]\frac{b^a}{\Gamma(a)}\lambda^{a-1}(v\lambda)^{1/2}\exp[-\lambda\{b+\tfrac12 v(\xi-\xi_0)^2\}].$$

Then integrate out the ξ to get

$$\frac{1}{(2\pi)^{n/2}}\frac{b^a}{\Gamma(a)}\lambda^{a+n/2-1}\Big(\frac{v}{v+n}\Big)^{1/2}\exp\{-\lambda(b+\tfrac12 Q_0+\tfrac12 d_n(\bar y-\xi_0)^2\},$$

with $d_n=(1/v+1/n)^{-1}$ as per Ex. 6.11. Show then that this leads to the marginal density being

$$\bar f(y_1,\dots,y_n)=(2\pi)^{-n/2}\Big(\frac{v}{v+n}\Big)^{1/2}\frac{b^a}{\Gamma(a)}\frac{\Gamma(a+n/2)}{\{b+\frac12 Q_0+\frac12 d_n(\bar y-\xi_0)^2\}^{a+n/2}}.$$

(d) xx

Ex. 6.13 *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 (xx xref here xx) where

$$y_i=x_{i,1}\beta_1+\cdots+x_{i,k}\beta_k+\varepsilon_i=x_i^{\mathrm t}\beta+\varepsilon_i\quad\text{for } i=1,\dots,n,$$

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 $x_i=(x_{i,1},\dots,x_{i,k})$ 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

$$\begin{aligned}L_n(\beta,\sigma)&=\sigma^{-n}\exp\Big\{-\tfrac12\frac{1}{\sigma^2}\sum_{i=1}^n(y_i-x_i^{\mathrm t}\beta)^2\Big\}\\&=\sigma^{-n}\exp\Big[-\tfrac12\frac{1}{\sigma^2}\{Q_0+n(\beta-\widehat\beta)^{\mathrm t}M_n(\beta-\widehat\beta)\}\Big],\end{aligned}$$

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

$$\lambda = 1/\sigma^2 \sim \mathrm{Gam}(a, b) \quad \text{and} \quad \beta \,|\, \sigma \sim \mathrm{N}_k(\beta_0, \sigma^2 M_0^{-1}).$$

Show that this prior may be expressed as

$$p(\lambda, \beta) \propto \lambda^{a-1} \lambda^{k/2} \exp\big[-\lambda\{b + \tfrac{1}{2}(\beta - \beta_0)^{\mathrm{t}} M_0 (\beta - \beta_0)\}\big].$$

(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,

$$(\lambda, \beta) \,|\, \text{data} \sim \text{GN}_k(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_0, \widehat{\beta}, nM_n).$$

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) Go again to the dataset 2.B, for illustration and for flexing your operational muscles. For y use the lung 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

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{for } i = 1, \ldots, 44,$$

with ε_i taken i.i.d. $\text{N}(0, \sigma^2)$. Specifically, we wish point estimates along with 95% credibility intervals for (i) each of the three parameters α, β, σ; (ii) the probability that $y \geq 25.0$, for a country with cigarette consumption $x = 35.0$; (iii) the lung cancer death rates y_{45} and y_{46}, per 100,000 inhabitants, for countries 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:

. 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 α and β as independent normals $(15.0, (2.0\,\sigma)^2)$ and $(0.0, (2.0\,\sigma)^2)$, given σ.

. Then, the simpler and partly non-informative one that takes a flat prior for (α, β) 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. 6.14 *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. 6.9. 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 $\pi(\theta) = \int \pi_\alpha(\theta)\,\text{d}G(\alpha)$ is a mixture of $\pi_\alpha(\theta)$ priors, with $\text{d}G(\alpha)$ 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 $\pi(\theta \,|\, y) = \int \pi_\alpha(\theta \,|\, y)\,\text{d}G(\alpha \,|\, \text{data})$, where $\text{d}G(\alpha \,|\, \text{data})$ 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

Ex. 6.15 *(xx a simple intro one-para illustration.)* (xx about here: a simple but real illustration that for almost any one-parameter model, we may carry out Bayes inference, with modest numerical efforts; bigger models need bigger tools, to be pointed to. xx)

Ex. 6.16 *The Jeffreys prior.* (xx to come. with examples. $\pi_0(\theta) \propto |J(\theta)|^{1/2}$. $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ for the binomial. but a different one for the Geometric, even though the likelihoods are the same. also do the multinomial, where it is $\text{Dir}(\frac{1}{2}, \ldots, \frac{1}{2})$. xx)

Ex. 6.17 *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

$$f(y, \theta) = 1 + \theta(y - \tfrac{1}{2}) \quad \text{for } y \in [0, 1].$$

(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

$$J(\theta) = \int_{-1/2}^{1/2} \frac{x^2}{1 + \theta x}\,\text{d}x.$$

(xx perhaps more, a formula. xx)

(b) Compute and display the Jeffreys prior.

(c) Take $\theta_{\text{true}} = 0.333$, 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.

(d) Then, with a uniform prior on $[-2, 2]$, compute and display the posterior distribution for θ.

(e) 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)

Ex. 6.18 *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

$$\begin{aligned}\mathrm{E}\,\{L_c(\theta,t)\,|\,\mathrm{data}\} &= \mathrm{E}\,[\exp\{c(t-\theta)\} - 1 - c(t-\theta)\,|\,\mathrm{data}] \\ &= \exp(ct)M(-c) - 1 - c(t-\widehat{\xi}),\end{aligned}$$

where $\widehat{\xi}$ is the posterior mean and $M(-c) = \mathrm{E}\,\{\exp(-c\theta)\,|\,\mathrm{data}\}$, 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 approximation results from Ex. 2.32, 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. 6.10(b).

(f) (xx rounding off for now; point to Ex. 6.24, 6.25. xx)

Ex. 6.19 *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. 6.20 for a very useful approximation method for more complex cases.

(a) Let $y_1,\ldots,y_n$ be independent Bernoulli variables with $P(y_i = 1\,|\,\theta) = \theta$, 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. 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. 6.20 *Approximating the marginal distribution.* In the setup of Ex. 6.19, 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

$$\begin{aligned}\bar f &\doteq \exp(\ell_{n,\max})\int \exp\{-\tfrac12 n(\theta-\widehat\theta)^{\mathrm t}J_n(\theta-\widehat\theta)\}\pi(\theta)\,\mathrm{d}\theta\\ &= \exp(\ell_{n,\max})\int \exp(-\tfrac12 s^{\mathrm t}J_n s)\pi(\widehat\theta+s/\sqrt n)\,\mathrm{d}s/n^{p/2}\\ &\doteq \exp(\ell_{n,\max})\pi(\widehat\theta)(2\pi)^{p/2}|J_n|^{-1/2}/n^{p/2}.\end{aligned}$$

(b) (xx a couple of things here. check how successful the approximation is in two setups. the formula

$$\log\bar f \doteq \ell_{n,\max} - \tfrac12 p\log n + \log\pi(\widehat\theta) - \tfrac12\log|J_n| + \tfrac12 p\log(2\pi),$$

with its two leading terms, lead to the BIC in Ch. 11. xx)

Ex. 6.21 *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 maximum likelihood estimator, and $J_n = -(1/n)\partial^2 \ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ the normalised observed information, as per (xx pointer to Ch 4 xx). 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 $\widehat{\theta} \to_{\text{pr}} \theta_{\text{true}}$ and $J_n \to_{\text{pr}} J = J(\theta_0)$, 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

$$D_n = \int |g_n(z) - \phi_p(z, 0, J^{-1})| \, \mathrm{d}z \to_{\text{pr}} 0.$$

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:

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_{\text{true}}) &\to_d \mathrm{N}(0, J^{-1}), \\ \sqrt{n}(\theta - \widehat{\theta}) \,|\, \text{data} &\to_d \mathrm{N}(0, J^{-1}), \text{ in probability.} \end{aligned}$$

(d) Check two clear situations in detail, comparing the exact posterior density $\pi(\theta \,|\, \text{data})$ with the normal approximation: (i) where $Y \,|\, \theta \sim \text{binom}(n, \theta)$, and $\theta \sim \text{Beta}(a_0, b_0)$; (ii) where $Y_1, \ldots, Y_n \,|\, \theta$ are i.i.d. Pois(θ), and $\theta \sim \text{Gam}(a_0, b_0)$. Choose n and (a_0, b_0), and also the true θ_0, 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. 6.22 *MCMC, I: simulating from a given distribution.* (xx the MCMC basics. Metropolis algorithm. simulating from a couple of distributions. xx)

Ex. 6.23 *MCMC, II: simulating from the posterior.* (xx applying the above to the generic Bayesian posterior distribution. a chain $\theta_1, \theta_2, \ldots$ in your computer, with $\pi(\theta \,|\, \text{data})$ as its equilibrium distribution. some easy start examples. 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. 6.18. 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

$$r_c(\theta)=\mathrm{E}_\theta\,[\exp\{c(X+d-\theta)\}-1-c(X+d-\theta)]=\exp(cd+\tfrac12 c^2)-1-cd,$$

constant in θ, and that the best estimator of this sort is $\theta^*=X-\frac12 c$. Show also that the risk achieved, by this estimator, is $\frac12 c^2$.

(b) Now consider Bayes estimation, with the prior $\theta\sim\mathrm{N}(0,\tau^2)$. Show via Ex. 6.10 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

$$\mathrm{E}\,\{L_c(\theta,t)\,|\,x\}=\exp\{c(t-wx)+\tfrac12 wc^2\}-1-c(t-wx).$$

Deduce that the Bayes estimator is $\widehat\theta_B=wx-\frac12 wc$, and that the posterior expected loss is $\mathrm{E}\,\{L_c(\theta,\widehat\theta_B)\,|\,x\}=\frac12 wc^2$, independent of x.

(c) Show that $\theta^*=X-\frac12 c$ 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 $\theta^*=\bar X-\frac12 c/n$ 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.* (xx clean this, and calibrate with earlier. xx) For the linex loss, studied initially in Ex. 6.18 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 $c\to0$, 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. 6.21, an approximation to the posterior distribution is $\theta\,|\,\text{data}\sim\mathrm{N}(\widehat\theta_{\rm ml},\widehat\sigma^2/n)$, in terms of the maximum likelihood estimate and estimated inverse Fisher information. Deduce that $M_n(-c)\doteq\exp(-c\widehat\theta_{\rm ml}+\frac12 c^2\widehat\sigma^2/n)$, in the notation above, and that this leads to the approximation $\widehat\theta_B=\widehat\theta_{\rm ml}-\frac12 c\widehat\sigma^2/n$ for the Bayes estimator under linex loss. Show also that the posterior expected loss is approximately $\frac12 c^2\widehat\sigma^2/n$.

(d) (xx an example where we can check the approximation with the exact Bayes estimator, e.g. with Poisson and gamma. xx)

6.C Notes and pointers

A few remarks.

(xx mention Varian (1975); Zellner (1986); Claeskens and Hjort (2008a) for the linex loss; but this should perhaps be in Ch 7. xx)

Stories, so far: nils puts in a brief version of Abel.

ToDo: nils needs to get this started. pointers to other chapters.

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.

7.A Chapter introduction

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,

$$P_{\theta_0}\{a \le C(\phi_0, Y) \le b\} = b - a \quad \text{for each } a, b \in [0, 1].$$

Assuming this random c.d.f. has a unique inverse, then, we have

$$P_\theta\{C^{-1}(0.05, Y) \le \phi \le C^{-1}(0.95, Y)\} = 0.90, \tag{7.1}$$

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. 3. The *confidence curve* is a related summary graph, most often computed from the CD via $\text{cc}(\phi, y) = |1 - 2\,C(\phi, y)|$. It has the practical property that $\{\phi\colon \text{cc}(\phi, y) \le 0.90\}$ give the 90 percent interval directly; similarly, all intervals at any desired confidence level can be read off from the confidence curve.

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. 3. 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.

[xx as of 13-Aug-2023, the chapter needs a mild perestroika, more detail, and a few exercises on the combining information things. nils attempts roughly following the structure (i) lots of CDs and cc(ϕ) things, with some three basic recipes, for which we also point to Ch3 Ch5 with Wilks; (ii) seeing CDs as natural for boundary parameters, and pointmasses at zero are fine; (iii) some optimality, with applications; (iv) then combination things, brief meta-analysis, some II-CC-FF. and pointers to stories. xx]

7.B Short and crisp

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. $F(x) = P(X \le x)$. 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 $y_{\rm obs}$. 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 $K_\theta(x) = P_\theta(\widehat\theta \le x)$; its distribution is here required to depend only on θ, not on other aspects of the underlying

model employed. Consider Recipe One, the construction

$$C(\theta, y_{\rm obs}) = P_\theta(\widehat\theta \geq \widehat\theta_{\rm obs}) = 1 - K_\theta(\widehat\theta_{\rm obs}),$$

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 $f(y, \theta) = \frac12 \exp(-|y - \theta)$, 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

$$C(\sigma, y_{\rm obs}) = P_\sigma(\widehat\sigma \geq \widehat\sigma_{\rm obs}) = 1 - \Gamma_m(m\widehat\sigma^2_{\rm obs}/\sigma^2).$$

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 $\text{cc}(\rho, \text{data})$.

Ex. 7.4 *Computing a CD with simulation and isotonic repair.* (xx to be polished. we use this is Story iii.6 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 $\bar{y}_{\rm obs} = 1.313$.

(a) Though we do not actually need this in the CD computations here, find an estimate of b based on $\text{E}\, Y_i = \Gamma(1 + 1/b)$; see Ex. 1.40. Show that $C(b) = P_b(\bar{Y} \le \bar{y}_{\rm obs})$ 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 $\bar{y}_{\rm obs}$. Compute also the confidence curve $\text{cc}^*(b) = |1 - 2\, C^*(b)|$. For this simple example it is easy to accomplish this with a high B, say 10^5, to make $C^*(b)$ and $\text{cc}^*(b)$ smooth and very close to the real $C(b)$ and $\text{cc}(b)$; for this illustration, however, make the simulation size as relatively small as $B = 100$, and plot the curves, as in Figure 7.1.

(c) We learn that with a low or moderate simulation size B, the $C^*(b)$ and $\text{cc}^*(b)$ will be wiggly. We can do better, using the prior knowledge that $C(b)$ is increasing. There 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.

isotonic regression

(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 $C(\theta, y) = P_\theta(\widehat{\theta} \ge \widehat{\theta}_{\rm obs})$ 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 $C^*(\theta) = P_\theta(\widehat{\alpha} \ge \widehat{\alpha}_{\rm obs})$ gives precisely the same CD as without the g transformation.

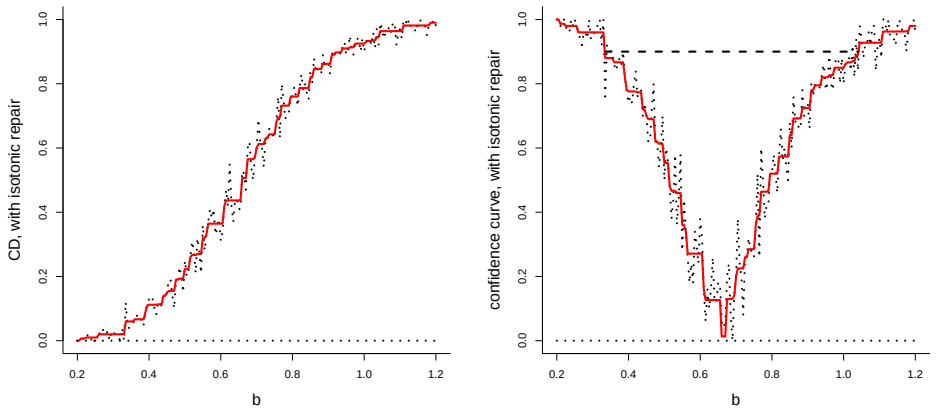

Figure 7.1: *Simulation based confidence distribution* $C^*(b)$ *and confidence curve* $\text{cc}^*(b)$ *for the Weibull parameter* b*, based on the observed sample mean* $\bar{y}_{\text{obs}} = 1.313$ *for* $n = 25$ *data points, along with isotonic repairs. The simulation size here is the low* $B = 100$*.*

(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 $C(\theta, y) = P_\theta(Z \geq z_{\text{obs}})$ 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. 4 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. 4 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. 2.11, 4.25), 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 $C(\theta,y)=P_\theta(\widehat{\theta}\geq\widehat{\theta}_{\rm obs})$ 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 $P_{\mu,\sigma}(Z\geq z_{\rm obs})$ schemes work, in that the Z in question has a distribution depending on σ, but not μ. Attempt to work with $C^*(\mu,y)=P_{\mu,\sigma}(\bar{Y}\geq\bar{y}_{\rm obs})$ – 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. 1.34, 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. 5.9.

(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 $P_{\theta_0}(\mathrm{cc}(\phi_0)\leq\alpha)\to\alpha$ 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 estimator $\widehat{\theta}$. Use also theory of Ch. 5 to write down a normal approximation to the distribution of $\widehat{\theta}$.

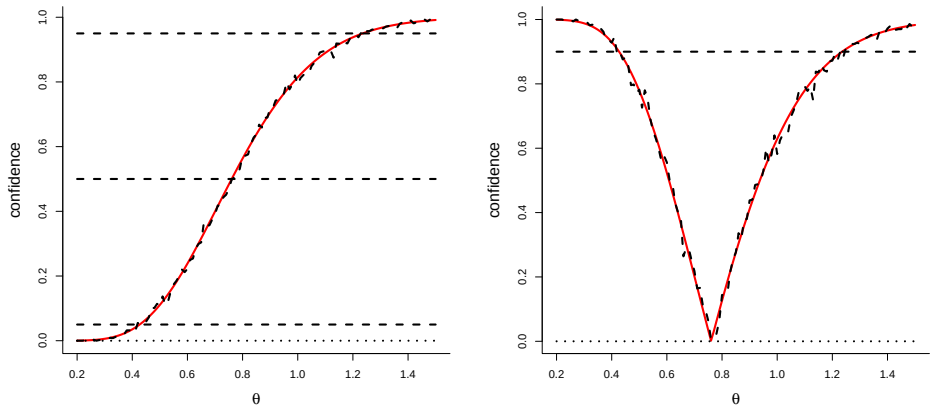

Figure 7.2: *For the simple data example of Ex. 7.9: Left panel: confidence distribution $C(\theta)$, via simulations (black and wiggly curve) and via exact calculations (red and smooth curve); right panel: the two versions of the associated confidence curve* $\mathrm{cc}(\theta)$. *From these we read off the median confidence estimate* $\widehat{\theta}_{0.50} = 0.76$, *and the 90 percent confidence interval* $[0.43, 1.24]$.

(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
```

Estimate θ and compute the CD $C(\theta) = P_\theta(\widehat{\theta} \geq \widehat{\theta}_{\rm obs})$, 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 *Something.* (xx nils invents one more cool enough one-dim example, showcasing that $D_n(\theta)$ works, even when we can't do it explicitly. could take $1 - \exp(-t^b)$ weibull, clear enough cc for b, and for any good function of b. xx)

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

$$C(\phi) = P_\theta(\widehat{\phi} \geq \widehat{\phi}_{\rm obs}) = 1 - G_0(\widehat{\phi}_{\rm obs} - \phi).$$

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

$$C(\phi) = P_\theta(\phi + \kappa N \geq \widehat{\phi}_{\mathrm{obs}}) = P(N \geq (\widehat{\phi} - \phi)/\kappa) = \Phi((\phi - \widehat{\phi})/\kappa),$$

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

$$C(\phi) = P_\theta((\widehat{\phi} - \phi)/\widehat{\kappa} \geq (\widehat{\phi}_{\mathrm{obs}} - \phi)/\widehat{\kappa}_{\mathrm{obs}}) = 1 - G((\widehat{\phi}_{\mathrm{obs}} - \phi)/\widehat{\kappa}_{\mathrm{obs}})$$

is a CD. If G is not known, or too difficult to derive, use simulations, of $t^* = (\widehat{\phi}^* - \widehat{\phi}_{\mathrm{obs}})/\widehat{\kappa}^*$, via datasets simulated at position $\widehat{\theta}_{\mathrm{obs}}$. 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 $P_\theta(t \leq u) = G(u, \phi, \gamma)$. Show that

$$H(\phi, \gamma) = P_{\phi,\gamma}((\widehat{\phi} - \phi)/\widehat{\kappa} \geq (\widehat{\phi}_{\mathrm{obs}} - \phi)/\widehat{\kappa}_{\mathrm{obs}}) = 1 - G((\widehat{\phi}_{\mathrm{obs}} - \phi)/\widehat{\kappa}_{\mathrm{obs}}, \phi, \gamma).$$

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

$$C^*(\phi) = 1 - \widehat{G}((\widehat{\phi}_{\mathrm{obs}} - \phi)/\widehat{\kappa}_{\mathrm{obs}}, \phi) \quad \text{where } \widehat{G}(u, \phi) = G(u, \phi, \widehat{\gamma}).$$

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 $t^* = (\widehat{\phi}^* - \widehat{\phi}_{\mathrm{obs}})/\widehat{\kappa}^*$.

Ex. 7.12 *Recipe Six: CDs in exponential families.* We have worked with the general exponential family in previous chapters, see Ex. 1.57. In particular we learned in Ex. 3.31

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

$$C^*(a) = P_a\{U(Y) \geq u_{\rm obs} \,|\, V(Y) = v_{\rm obs}\}.$$

This construction is actually optimal, in a power function sense we come back to (xx pointer xx), but we can already start working with this definition and see how it applies in various situations.

(a) Verify from arguments in Ex. 3.31 that $C^*(a)$ indeed depends only on a, not on b.

(b) For an illustration, consider the pair of exponentials of Ex. 3.26. 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

$$C^*(\delta) = P_\delta\,(Y \leq y_{\rm obs} \,|\, Z = z_{\rm obs}) = \frac{1-\exp(-\delta y_{\rm obs})}{1-\exp(-\delta z_{\rm obs})}.$$

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

$$C^*(\delta) = P_\delta(U \leq u_{\rm obs} \,|\, Z_1 = z_{1,\rm obs}, \ldots, z_{m,\rm obs}),$$

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. 3.26 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 *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 $y_{1,0} \sim \text{binom}(m_{i,0}, p_{1,0})$.

m_1	m_0	y_1	y_0
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. 7.14 *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.

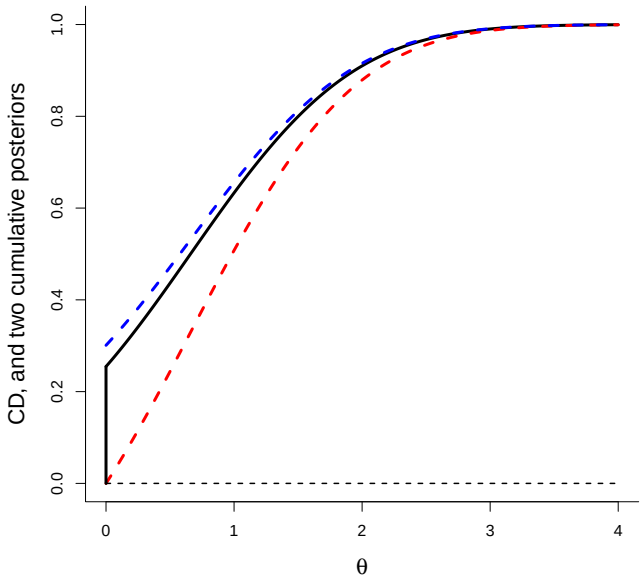

Figure 7.3: With $y_{\mathrm{obs}} = 0.66$ for the $\mathrm{N}(\theta, 1)$ model, the black curve is the natural CD, with positive point mass 0.255 at zero. The red and the blue curves are Bayesian posterior distributions, for the flat prior on the halfline, and for the mixture prior with $\frac{1}{2}$ at zero and $\frac{1}{2}$ flat on the halfline, respectively.

(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 $y_{\rm obs}$ 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) (xx repair this. 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. In his FocuStat Blog Post, Schweder (2017) dares to very much disagree with Nobel Prize Winner (that is to say, the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel Winner) Professor Christoper Sims. It would have been better for Sims, in his chosen example featuring Bayesian methodology, to not use flat priors on positive halflines, but to allow pointmasses at zero too.

(e) In general terms, for the case of $y \,|\, \theta \sim \mathrm{N}(\theta, 1)$, 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 $f_0(y) = \int \phi(y-\theta)\pi_0(\theta)\,\mathrm{d}\theta$ and $f_1(y) = \int \phi(y-\theta)\pi_1(\theta)\,\mathrm{d}\theta$ 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.)

(f) 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 7.3.

(g) 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 $y_{\rm obs}$.

(h) Generalise the formulae above to the case of $y_1, \ldots, y_n$ i.i.d. $\mathrm{N}(\theta, \sigma^2)$, with known σ.

Ex. 7.15 *CDs for regression parameters with boundary constraints.* (xx to come here: more on boundary parameters, now in simple regression models. pointer to Story iii.6. 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. 7.16 *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 maximum likelihood 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 maximum likelihood 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. 7.17 *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

$$\theta^N \exp\Big[-\theta\Big\{\sum_{t_i \le t_0} t_i + (n-N)t_0\Big\}\Big].$$

(b) Show that the maximum likelihood estimator is

$$\widehat{\theta} = N/R = N/\Big\{\sum_{t_i \le t_0} t_i + (n-N)t_0\Big\}.$$

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 $C_n(\theta) = P_\theta(\widehat{\theta} \ge \widehat{\theta}_{\rm obs})$ 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 maximum likelihood estimator; and a t-bootstrap based version. Comment on your findings.

Ex. 7.18 *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.

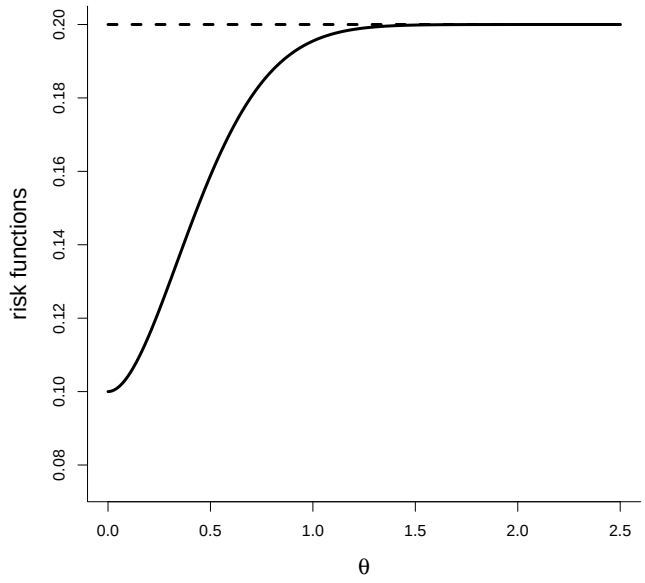

Figure 7.4: *Risk function for the CD of Ex. 7.18(d), in the setup with $Y_1, \ldots, Y_n$ being i.i.d. from $\mathrm{N}(\theta, 1)$, with $n = 10$, but with the restriction $\theta \geq 0$. It starts out at $1/n$ and then grows to the $2/n$ risk of the unrestricted case as θ grows.*

(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 7.4.

(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 maximum likelihood 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 7.4. 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 $C_B(\theta) = P_\theta(\widehat{\theta}_B \geq \widehat{\theta}_{B,\rm obs})$, and comment.

Ex. 7.19 *Risk functions for three CDs in a variance components model.* Consider the simple variance component model with independent observations $y_i \sim {\rm N}(0, \sigma^2 + \tau^2)$ 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

$$\text{risk}(C, \tau) = {\rm E}_\tau\, |\tau_{\rm cd} - \tau| = {\rm E}_\tau \int |\tau_{\rm cd} - \tau|\,{\rm d}C(\tau_{\rm cd}, Y),$$

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.

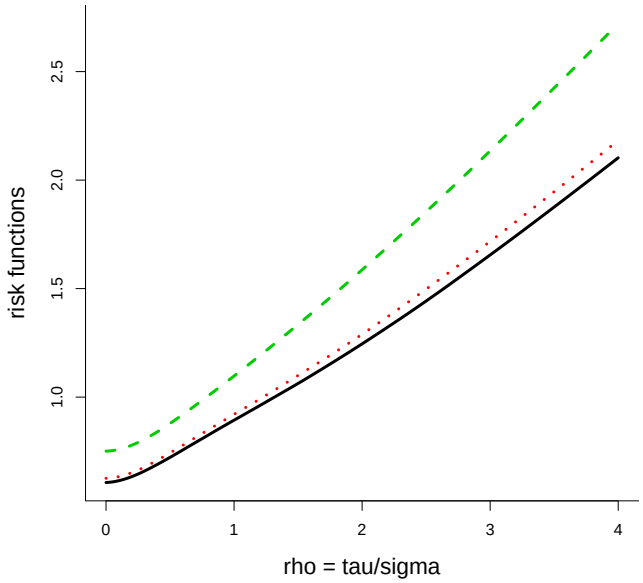

Figure 7.5: For the variance component model, with $p = 4$ and $\sigma = 1$, risk functions $r(C, \tau)$ for three CDs for τ. The one based on $Z = \sum_{i=1}^{p} Y_i^2$ is best, closely followed by the one using $A = \sum_{i=1}^{p} |Y_i|$, whereas the one using the range $R = \max Y_i - \min Y_i$ does worse.

(b) Show that a random draw τ_{cd} from the first of these, i.e. C_Z, for a given dataset, can be represented as $\tau_{\text{cd}} = (Z_{\text{obs}}/K - \sigma^2)_+^{1/2}$, 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

$$\tau_{\text{cd}} - \tau = \{(\sigma^2 + \tau^2)K_0/K - \sigma^2\}_+^{1/2} - \tau = \sigma\big[\{(1 + \rho^2)K_0/K - 1\}_+^{1/2} - \rho\big],$$

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 7.5.

(c) Then consider the C_A option. Show that a random draw from an observed $C_A(\tau, \text{data})$ can be written $\tau_{\text{cd}} = \{(A_{\text{obs}}/A)^2 - \sigma^2\}_+^{1/2}$. Deduce that for random data behind the CD, we have the representation

$$\tau_{\text{cd}} - \tau = \{(\sigma^2 + \tau^2)(A_0/A)^2 - \sigma^2\}_+^{1/2} - \tau = \sigma\big[\{(1 + \rho^2)(A_0/A)^2 - 1\}_+^{1/2} - \rho\big],$$

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 7.5.

(e) Use your programme to explore the three risk functions for other values of p.

Ex. 7.20 *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)

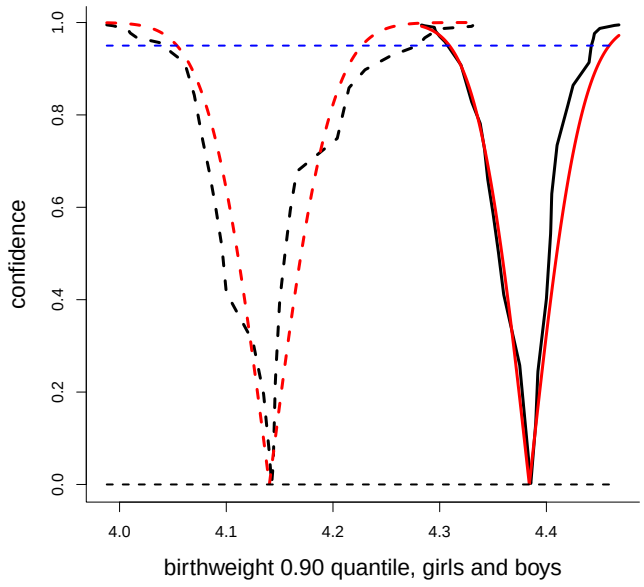

Figure 7.6: Confidence curves $\mathrm{cc}(\mu_{0.90})$, for the 0.90 quantiles of the birthweight distributions for girls (to the left) and boys (to the right). The black curves use the Beta method $\mathrm{cc}_n^*(\mu, \mathrm{data})$, with linear interpolation, whereas the slanted curves use the large-sample 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.

(a) It is not difficult to construct a first-order correct CD via large-sample results reached in Chapter 2, see in particular Ex. 2.20. 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

$$P_f(\mu \le Y_{(i)}) = P(\tfrac{1}{2} \le U_{(i)}) = 1 - \mathrm{Be}(\tfrac{1}{2}, i, n-i+1) \quad \text{for } i = 1, \ldots, n.$$

Here $U_{(i)} = F(Y_{(i)})$; these form an ordered sample from the standard uniform, and we saw in Ex. 2.20 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 $P_f(\mu_q \le Y_{(i)}) = 1 - \mathrm{Be}(q, i, n-i+1)$, and use linear interpolation:

$$C_n^*(\mu_q, \mathrm{data}) = \text{interpolation with } 1 - \mathrm{Be}(q, i, n-i+1) \text{ at } y_{(i)}, \tag{7.2}$$

the Beta method CD for quantiles

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.

(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 7.6. point to Story **??**. xx)

Ex. 7.21 *CDs for quantiles: how well do they work?* In Ex. 7.20 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 $\text{sim} = 10^5$ values of $V_n = C_n(\mu_{q,\text{true}}, Y_1, \ldots, Y_n)$ and $V_n^* = C_n^*(\mu_{q,\text{true}}, Y_1, \ldots, Y_n)$. 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.

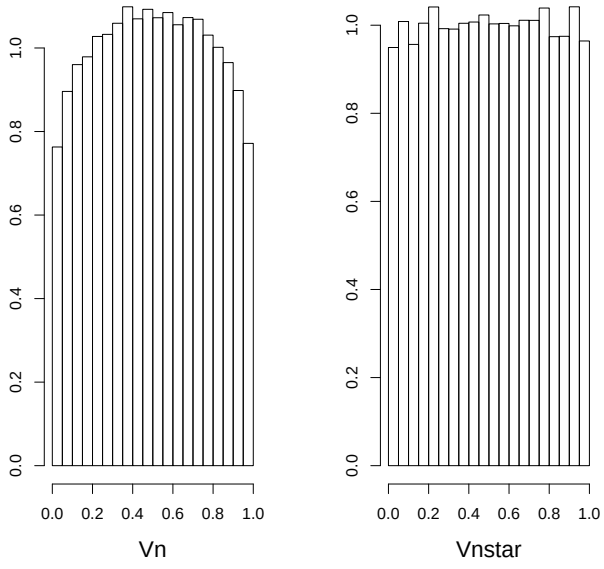

Figure 7.7: For the case of f being standard normal, $n = 100$, $q = 0.50$: histograms of V_n and V_n^* based on $\text{sim} = 10^5$ simulations. Also in other cases the V_n^* is much closer to uniformity than is V_n.

(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. 2.27. 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. 7.22 *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 13-Aug-2023. xx)

Ex. 7.23 *Optimal CDs and confidence curves.* (xx the main recipe here, with a a few applications. point back to optimal testing in Ch. 3, and to theory in Ch. 8. exponential family etc. going back to exercise in Ch 2, with testing for δ with expo pairs, we here give full confidence curves, with two methods. xx)

Ex. 7.24 *Comparing Poisson parameters.* (xx ranting on a bit, to be edited. point to Story i.2, 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

$$\begin{aligned} C(\gamma) &= P_\gamma(Y_1 > y_{1,\text{obs}} \,|\, z) + \tfrac12 P_\gamma(Y_1 = y_{1,\text{obs}} \,|\, z) \\ &= 1 - B_z(y_{1,\text{obs}}, m_1\gamma/(m_0+m_1\gamma)) + \tfrac12 b_z(y_{1,\text{obs}}, m_1\gamma/(m_0+m_1\gamma)). \end{aligned}$$

(xx point to Story i.2, 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. 7.25 *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

$$C(\theta,y) = P_\theta(y > y_{\text{obs}}) + \tfrac12 P_\theta(y = y_{\text{obs}}).$$

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. 7.26 *Aboriginals and invaders in Watership Down.* Suppose a population of rabbits has been living for a long time on an island, in Hardy–Weibnerg equilibrium $(p_0, q_0) = (0.25, 0.75)$, which means that pairs of alleles aa, Aa, AA occur with frequencies $(p_0^2, 2p_0q_0, q_0^2)$. Suppose next that there's an invading populations 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 $1 - \lambda$).

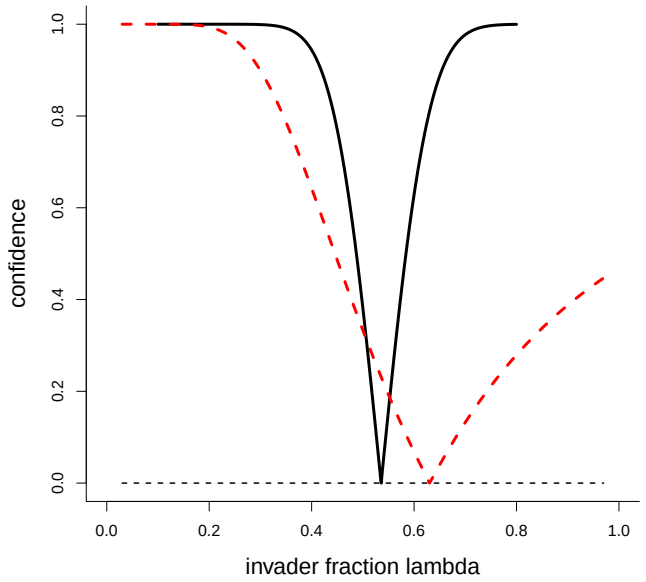

Figure 7.8: *Confidence curves for the unknown fraction λ of newcomers, after having counted $(X, Y, Z) = (118, 438, 444)$ of allele pairs aa, Aa, AA. The start population has HW parameters $(p_0, q_0) = (0.25, 0.75)$. (i) The black smooth $\mathrm{cc}_1(\lambda)$ is computed using the knowledge that 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.*

(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

$$\mathrm{pr}_1 = (1-\lambda)p_0^2 + \lambda p^2, \quad \mathrm{pr}_2 = (1-\lambda)2p_{=}q_0 + \lambda 2pq, \quad \mathrm{pr}_3 = (1-\lambda)q_0^2 + \lambda q^2.$$

Note that $\mathrm{pr}_1 + \mathrm{pr}_2 + \mathrm{pr}_3 = 1$.

(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 7.8. 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 7.8. Comment on your findings. For your computer script, 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. 7.27 *CD and cc for comparing binomials.* In Ex. 7.25 we learn how to construct CDs for separate binomial parameters. Consider now the 2×2 table setup with two binomials, as with Ex. 3.29, 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

$$C(\gamma) = P_\gamma(Y_1 > y_{1,\text{obs}} \,|\, Z = z_{\text{obs}}) + \tfrac{1}{2} P_\gamma(Y_1 = y_{1,\text{obs}} \,|\, Z = z_{\text{obs}}),$$

with $Z = Y_0 + Y_1$. The conditional distribution in question is the eccentric hypergeometric, found in Ex. 3.29. (xx do simple example here. this CD is used in both Stories i.1 and i.9. we use Ex. 3.29. xx)

Ex. 7.28 *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 $P_\theta\{C^{-1}(0.05, Y) \le \phi \le C^{-1}(0.95, Y)\} = 0.90$, 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 Ch. 6. 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 $\text{Pois}(\theta)$, and with a prior $\theta \sim \text{Gam}(a, b)$; see Ex. (xx suitable exercise Ch 5 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 $\text{Pois}(\theta_0)$, for a certain θ_0. Show that

$$\begin{aligned} B_n(\theta_0 \,|\, Y_1, \ldots, Y_n) &= P(\theta \le \theta_0 \,|\, Y_1, \ldots, Y_n) = G(\theta_0, a + n\bar{y}, b + n) \\ &\doteq \Phi((\theta_0 - \widehat{\theta}_B)/\widehat{\tau}_B) \to_d \Phi(\text{N}(0,1)) \sim \text{unif}, \end{aligned}$$

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

$$\begin{aligned} B_n(\theta_0\,|\,Y_n) &= P(\theta \le \theta_0\,|\,Y_n) = \text{Be}(\theta_0, a+Y_b, b+n-Y_n) \\ &\doteq \Phi((\theta_0-\widehat{\theta}_B)/\widehat{\tau}_B) \to_d \Phi(\text{N}(0,1)) \sim \text{unif}, \end{aligned}$$

with probability 1.

(c) (xx to land somewhere in Ch. 6, sorted under Bernshteĭn–von Mises. xx) consider the posterior distribution $\pi_n(\theta\,|\,\text{data}) \propto \pi_0(\theta)L_n(\theta)$, and then normalise to $Z_n = \sqrt{n}(\theta-\widehat{\theta})$, with $\widehat{\theta}$ the maximum likelihood estimate. Show that Z_n, given the data, has density

$$\begin{aligned} g_n(z\,|\,\text{data}) &\propto \pi_0(\widehat{\theta}+z/\sqrt{n})L_n(\widehat{\theta}+z/\sqrt{n})(1/\sqrt{n})^p \\ &\propto \exp\{\ell_n(\widehat{\theta}+z/\sqrt{n})-\ell_n(\widehat{\theta})\}\pi_0(\widehat{\theta}+z/\sqrt{n}) \\ &\doteq \exp\{-\tfrac{1}{2}(\widehat{\theta}-\widehat{\theta})^{\rm t}J_n(\theta-\widehat{\theta})\}\pi_0(\widehat{\theta}), \end{aligned}$$

with J_n the normalised Hesse matrix $-(1/n)\partial^2\ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\rm t}$. Show from this, under mild conditions, that

$$\sqrt{n}(\theta-\widehat{\theta})\,|\,(Y_1,\ldots,Y_n) \to_d \text{N}_p(0,J^{-1}) \quad \text{with probability 1.}$$

(d) xx

Ex. 7.29 *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 $\widehat{a} \sim \text{N}(a,1)$ and $\widehat{b} \sim \text{N}(b,1)$ 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 $\ell_{\rm prof}(x_0) = -\frac{1}{2}Q_{\rm prof}(x_0)$, where $Q_{\rm prof}(x_0) = \min\{Q(a,b)\colon x_0 = -a/b\}$.

(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

$$\text{cc}(x_0) = \Gamma_1((\widehat{a}+\widehat{b}x_0)/(1+x_0^2))$$

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. 7.30 *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 $\text{N}_p(\theta,\Sigma)$ 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.

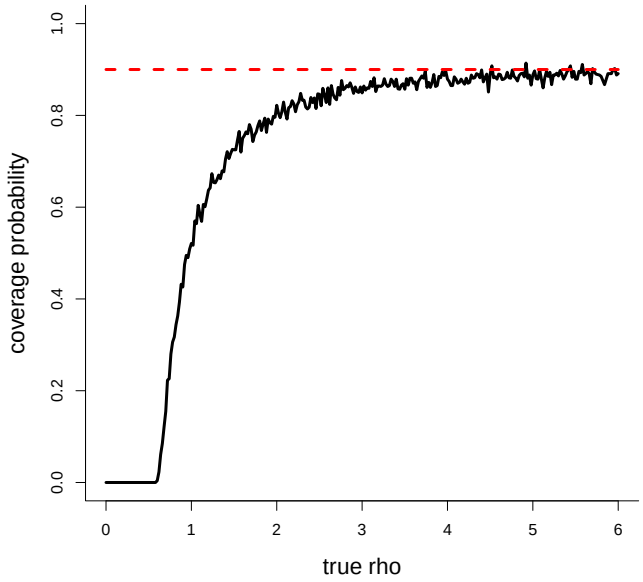

Figure 7.9: As a function of the true ρ, for dimension $p = 3$, the figure shows the actual coverage probability of the Bayesian 90 percent credibility interval, based on the posterior stemming from a flat prior for the θ.

(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 maximum likelihood 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 7.9 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 maximum likelihood 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. 7.31 *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. 2.22 to show that the joint distribution of these f first spacings may be written

$$\begin{aligned} g_r(d_1, \ldots, d_r) &= n(n-1)\cdots(n-r+1) \\ &\quad \exp\big[-\{n(d_1 + \cdots + d_r) - d_2 - 2d_3 - \cdots (r-1)d_r\}\big], \end{aligned}$$

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

$$C_r(n) = P_n(Y_{(r)} \le Y_{(r),\mathrm{obs}}) = \mathrm{Be}(F(Y_{(r),\mathrm{obs}}), r, n-r+1).$$

(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. 7.32 *(xx another discrete model thing. xx)* (xx to come. xx)

Ex. 7.33 *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 $y_j \sim \mathrm{N}(\phi, \sigma_j^2)$, 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 $\widehat{\phi} \sim \mathrm{N}(\phi, \kappa^2)$, 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 $y_j \,|\, \phi_j \sim \mathrm{N}(\phi_j, \sigma_j^2)$, with these individual mean parameters not being equal, but having their own distribution, say $\phi_j \sim \mathrm{N}(\phi_0, \tau^2)$. The task is then to reach inference for both the overall mean ϕ_0 and the spread τ among the ϕ_j. Show that $y_j \sim \mathrm{N}(\phi_0, \sigma_j^2 + \tau^2)$, 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

$$\ell_{\mathrm{prof}}(\tau) = -\tfrac{1}{2} \sum_{j=1}^k \Big[\log(\sigma_j^2 + \tau^2) + \frac{\{y_j - \widehat{\phi}_0(\tau)\}^2}{\sigma_j^2 + \tau^2} \Big], \quad \text{for } \tau \geq 0,$$

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 $D \le 0$, then $\widehat{\tau}_{\rm ml} = 0$, and if that $D > 0$, then $\widehat{\tau}_{\rm ml}$ 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 $\ell_{\rm prof}(\tau)$. partly from CLP. derivative at zero. xx) By maximising over ϕ_0, for each given τ, show that

$$\ell_{\rm prof}(\tau) = -\tfrac{1}{2} \sum_{j=1}^{k} \Big[\log(\sigma_j^2 + \tau^2) + \frac{\{y_j - \widehat{\phi}(\tau)\}^2}{\sigma_j^2 + \tau^2} \Big].$$

Ex. 7.34 *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 $N_j = \phi^{-1}(C_j(\phi_{\rm true}))$ 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. 7.33. xx)

(c)

Ex. 7.35 *From CD to likelihood.* (xx to come. with illustrations. normal conversion. $\ell(\phi) = -\frac{1}{2}\Gamma_1^{-1}({\rm cc}_j(\phi))$. xx)

Ex. 7.36 *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.2. xx)

Ex. 7.37 *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*.

7.C 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), ...

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.

8.A Chapter introduction

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. 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

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

8.B Short and crisp

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. 8.8 we learn that $\delta_{a(n)}$ is indeed minimax.

Ex. 8.2 *Uniformly minimum variance unbiased estimators.* As the sketch you made Ex. 8.1 illustrates, comparing risk functions is not always straight forward, 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. 5.5 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. 1.57–**??**.

(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. 1.61 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 extremely 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. 8.15 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. 2.19 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. $\mathrm{Pois}(\lambda)$. 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 $P_\theta\{h(T) = 0\} = 1$, for all θ. We recall from Ex. 3.23 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. 1.64 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 *Tools for minimaxity.* In Exercise. 8.1 we compared three different estimators. That's fine, but we ultimately want to say something about the performance of our estimators 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. 8.9 *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 $\theta_0 = 1/2$ (see Exercise 8.1(d)). Conclude.

Ex. 8.10 *Minimaxity and sequences of priors.* In Exercise 8.8(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 Poisson(θ). 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.

Ex. 8.11 *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. 8.11(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 $X \sim \mathrm{unif}(0, \theta)$ 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. 8.12 *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 $\mathrm{N}(\theta, \sigma^2)$. Suppose θ is given the improper prior $\pi(\theta) = 1$. Show that $\pi(\theta \mid x_1, \ldots, x_n) = \mathrm{N}(\bar{X}_n, \sigma^2/n)$. A generalised Bayes estimator is the estimator δ minimising the posterior expected loss $\mathrm{E}\{L(\delta, \theta) \mid \text{data}\}$. 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. 8.13 *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.

(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 $L(\delta, \theta) = (\delta - \theta)^2$. Show that $\widehat{\theta} = \bar{X}_n$ is admissible.

Ex. 8.14 *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 $\delta_{\rm ml}$ takes $\delta_{{\rm ml},i}(Y) = Y_i$ for $i = 1, \ldots, p$. Clevenson and Zidek (1975) showed that $\delta_{\rm ml}$ is inadmissible by constructing an estimator, say $\delta_{\rm CZ}$, such that $R(\theta, \delta_{\rm CZ}) < R(\theta, \delta_{\rm ml})$ 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 $\mathrm{E}\,(Y_i \mid Z) = Z\pi_i$ and $\mathrm{Var}(Y_i \mid Z) = Z\pi_i(1-\pi)$ 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

$$\mathrm{E}\, g(X)/\theta = \mathrm{E}\, g(X+1)/(X+1).$$

(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. [xx hmm, this is too hard. make it into a separate exercise xx].

Ex. 8.15 *Rao–Blackwellisation.* Recall that a function is convex if g is a convex if for all x, y in its domain, and all $\lambda \in [0, 1]$,

$$g(\lambda x + (1-\lambda)y) \le \lambda g(x) + (1-\lambda) g(y).$$

Jensen's inequality states that if g is convex, then

$$g(\mathrm{E}\, X) \le \mathrm{E}\, g(X),$$

with equality only if $g(x) = a+bx$. Jensen's inequality also holds conditional expectations [xx point to appendix here xx]. 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 hmm, 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.

Ex. 8.16 *Admissibility of X.* [xx sketch of exercise. Point to an exercise in Chapter 4, Cramér–Rao stuff xx]. Assume that $X \sim \mathrm{N}(\theta, 1)$. We wish to estimate θ under the squared error loss function $L(\delta, \theta) = (\delta - \theta)^2$.

(a) Find the Cramér–Rao lower bound.

(b)

Ex. 8.17 *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. 8.18 Suppose $X_1, \ldots, X_n$ are i.i.d. from $\mathrm{N}(0, \sigma^2)$. We are to estimate σ^2 under the loss function

$$L(\delta, \sigma^2) = \frac{(\delta - \sigma^2)^2}{\sigma^2}. \tag{8.2}$$

(a) Find the maximum likelihood estimator and its risk function.

(b) Consider the prior distribution given by density

$$\sigma^2 \sim \frac{b^a}{\Gamma(a)}(1/\sigma^2)^{a+1} \exp(-b/\sigma^2), \quad \sigma^2 > 0,$$

with $a > 1$ and $b > 0$. This is the density of an inverse gamma distribution. Find the prior expectation of σ^2. Find also the prior expectation of $1/\sigma^2$.

(c) Find the posterior distribution $\sigma^2 \mid x_1, \ldots, x_n$, 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 $\delta_\alpha = \alpha\widehat{\sigma}^2_{\mathrm{ml}}$.

(e) Is $\delta^\star$ admissible? *Hint*: Use Blyth's method.

Ex. 8.19 [xx change loss in above exercise xx] 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

$$L(\delta, \sigma) = \frac{(\delta - \sigma)^2}{\sigma}. \tag{8.3}$$

(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 (8.3).

(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.

Ex. 8.20 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

$$\delta_i(Y) = Y_i - b_i(Y). \tag{8.4}$$

Show that the difference in risk between Y and estimators of the form (8.4) can be expressed as

$$R(\delta, \theta) - R(Y, \theta) = \mathrm{E}_\theta\, D(Y),$$

where

$$D(y) = \sum_{i=1}^{p} \{b_i(y)^2 - 2 b_{i,i}(y)\}.$$

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 $\theta_1,\dots,\theta_p$. 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

$$\delta_{\mathrm{JS}}(Y) = \big(1 - \frac{p-2}{\|Y\|^2}\big)Y. \tag{8.5}$$

This estimator is known as the James–Stein estimator. Show that the risk function of this estimator can be expressed as

$$R(\delta_{\mathrm{JS}},\theta) = p - (p-2)^2 \mathrm{E}_\theta\, \frac{1}{\|Y\|^2}.$$

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 $R(\delta_{\mathrm{JS}},0)$ in this point.

(f) We'll now make a connection to empirical Bayes procedures. Start with a prior that takes $\theta_1,\dots,\theta_p$ independent from $\mathrm{N}(0,\tau^2)$. Show that the Bayes solution is $\delta^B = (\delta_1^B,\dots,\delta_p^B)$, with

$$\delta_i^B(Y) = \alpha\, Y_i, \quad i = 1,\dots,p, \quad \text{where} \quad \alpha = \frac{\tau^2}{\tau^2+1}. \tag{8.6}$$

(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 $y_1,\dots,y_p$ is a product of $\mathrm{N}(0,1+\tau^2)$ 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 $\delta_{\mathrm{EB}}(Y) = \tilde\alpha Y$. What's noticeable about this estimator?

Ex. 8.21 *Resolving the paradox.* [xx make an exercise based on insights from Stigler (1990), perhaps?]

Ex. 8.22 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.

Ex. 8.23 *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

$$p(\theta) = P(Y \in [a, b]) = \int_a^b f(y, \theta) \, \mathrm{d}y$$

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

$$p^* = \mathrm{E}\,(\widehat{p} \,|\, T) = P(Y_1 \in [a, b] \,|\, T).$$

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)

8.C Notes and pointers

[xx some notes and pointers here xx]

I.14

Bootstrapping

Part II

Stories

Part III

Solutions

Part IV

Appendix

IV.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]

1.A 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.

1.B Essentials of measure, integration, and probability

Ex. A.1 *Quick introduction to measure and integration, I: Measures, measurable spaces, measurable functions.* The purpose of our first three exercises is to point to, explain, helping the readers go through a list of key tools and properties from the general theory of measure and integration, without tending to all details and the wider stories.

(a) We start with a *measurable space*, say $(\Omega, \mathcal{A})$, with Ω any non-empty set and $\mathcal{A}$ a collection of subsets; these subsets are later to be given values, perhaps probabilities, in terms of a measure. For $\mathcal{A}$, we demand (i) that the full set Ω is in $\mathcal{A}$; (ii) that complements are in $\mathcal{A}$ (if A is there, then $A^c = \Omega \setminus A$ needs to be there); (iii) that countable unions of

a σ-algebra of sets

sets in $\mathcal{A}$ are in $\mathcal{A}$. An $\mathcal{A}$ with these properties is called a σ-algebra. – Show that 2^Ω, the collection of absolutely all subsets of Ω, often called the power set, is a 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 $A_1, A_2, \ldots$ in that class, 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 $A_1 \cap (A_2 \cup A_3 \cup A_4)^c \cap A_5$ must be 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 $\mathcal{B}_0$. That is, $\mathcal{B} = \bigcap\{\mathcal{A} \text{ is } \sigma\text{-algebra}\colon \mathcal{B}_0 \subset \mathcal{A}\}$. The σ-algebra $\mathcal{B}$ is said to be *generated* by $\mathcal{B}_0$, and since $\mathcal{B}_0$ is contained in the collection of absolutely all subsets of Ω, namely 2^Ω, there is always such a σ-algebra. 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. A famous and important example is the *Borel sets on the real line*, say $\mathcal{B}(\mathbb{R})$, defined as the σ-algebra generated by all intervals (a, b). Show that sets like $[a, b]$, $(-\infty, b)$, finite unions of intervals, etc., are then also in $\mathcal{B}(\mathbb{R})$. Similarly we define $\mathcal{B}^k = \sigma(\mathcal{B}_0^k)$, the Borel sets of $\mathbb{R}^k$, as the smallest σ-algebra containing all open rectangles $(a_1, b_1) \times \cdots \times (a_k, b_k)$. Show that $\mathcal{B}^k$ then also must contain all closed rectangles $[a_1, b_1] \times \cdots \times [a_k, b_k]$, and also all open sets of $\mathbb{R}^k$.

the Borel sets

(d) Consider a real function $f\colon \Omega \to \mathbb{R}$ on a measurable space $(\Omega, \mathcal{A})$. We say that f is *measurable* provided $f^{-1}(a, b) = \{\omega\colon a < f(\omega) < b\}$ is a measurable set, that is, $f^{-1}(a, b)$ is in $\mathcal{A}$ for each interval (a, b). Show that if f is measurable, then $\mathcal{B}^\star = \{B \in \mathcal{B}\colon f^{-1}(B) \in \mathcal{A}\}$ is a σ-algebra. Show also that if f is measurable, then also the much more general inverse sets $f^{-1}(B) = \{\omega\colon f(\omega) \in B\}$ are in $\mathcal{A}$, for any Borel set B.

(e) Show that $f_1, \ldots, f_n$ are measurable functions, then $f_{\max} = \max(f_1, \ldots, f_n)$ and $f_{\min} = \min(f_1, \ldots, f_n)$ are also measurable.

(f) Show further that if $0 \le f_1 \le f_2 \le \cdots$, a sequence of functions where $f_n(x)$ is nondecreasing for each x, then the limit function f, with $f(x) = \lim_n f_n(x)$, is also measurable. A simple function is one taking on only finitely many values, so if g is a simple function it can be written $g = \sum_{j=1}^k a_j I_{A_j}$ for a measurable decomposition $A_1, \ldots, A_k$ of Ω, that is, $A_i \cap A_j = \emptyset$ for all $i \ne j$ and $\cup_{j=1}^k A_j = \Omega$. With f any nonnegative measurable function $\Omega \to \mathbb{R}$, show that the sequence of simple functions

Simple functions

Approximation by simple functions

$$f_n(x) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{A_{n,k}}(x) + n I_{A_n}(x),$$

with $A_{n,k} = \{x\colon (k-1)/2^n \le f(x) < m/2^n\}$ and $A_n = \{x\colon f(x) \ge n\}$, is measurable, and is such that $0 \le f_1 \le f_2 \le \cdots$ and $f(x) = \lim_n f_n(x)$ for each x.

(g) A *measure* ν on a measurable space $(\Omega, \mathcal{A})$ is a function $\mathcal{A} \to [0, \infty]$, giving values to all sets A in $\mathcal{A}$, with the following properties: (i) $\nu(\emptyset) = 0$, for the empty set; (ii) with

Measures and measurable spaces

$A_1, A_2, \ldots$ disjoint sets in $\mathcal{A}$, $\nu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \nu(A_i)$. The resulting triple $(\Omega, \mathcal{A}, \nu)$ is called a *measure space*. We say that ν is a *finite measure* if $\nu(\Omega)$ is finite, and prime examples are those where $\nu(\Omega) = 1$; such measures are *probability measures*, to be returned to below. If Ω can be represented as a countable union $\cup_{i=1}^{\infty} A_i$, with each $\nu(A_i)$ finite, we say that ν is a σ-finite measure. Show that if A and B are sets in $\mathcal{A}$, then (1) $B \subset A$ entails that $\nu(B) \leq \nu(A)$; (2) provided $\nu(A) < \infty$, $\nu(A \setminus B) = \nu(A) - \nu(A \cap B)$; and (3) $\nu(A \cup B) = \nu(A) + \nu(B) - \nu(A \cap B)$ provided both A and B have finite measure. Show also that $\nu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \nu(A_i)$, for any sequence of A_i in $\mathcal{A}$.

σ-finite measure

(h) Suppose that $A_1 \subset A_2 \subset \cdots$ are sets in $\mathcal{A}$. We say that $(A_i)_{i \geq 1}$ is a nondecreasing sequence of sets. Show that $\nu(\cup_{i=1}^{\infty} A_i) = \lim_{n \to \infty} \nu(A_n)$. Suppose now that $A_1 \supset A_2 \supset \cdots$ is a decreasing sequence of sets in $\mathcal{A}$. Show that $\nu(\cap_{i=1}^{\infty} A_i) = \lim_{n \to \infty} \nu(A_n)$, provided $\nu(A_k) < \infty$ for some k.

Continuity of measure

(i) Suppose that ν is a *finitely additive* measure. That is, ν is a function $\mathcal{A} \to [0, \infty]$, such that (i) $\nu(\emptyset) = 0$, and $\nu(A \cup B) = \nu(A) + \nu(B)$ for all disjoint sets A and B. Suppose that for any nondecreasing sequence $A_1 \subset A_2 \subset \cdots$ in $\mathcal{A}$ it is the case that $\nu(\cup_{i=1}^{\infty} A_i) = \lim_{n \to \infty} \nu(A_n)$. Show that ν is a measure, in other words, that countably additivity is a continuity property in disguise.

(j) 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

(k) 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.

Ex. A.2 *Limits of sets and Fatou's lemma.* Recall that for sequences $(a_n)_{n \geq 1}$, we define $\liminf_{n \to \infty} a_n = \lim_{n \to \infty} \inf_{n \geq m} a_m$ and $\limsup_{n \to \infty} a_n = \lim_{n \to \infty} \sup_{n \geq m} a_m$. 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.

(a) Show that $\liminf_{n \to \infty} I_{A_n}(x) = 1$ if and only if $x \in \cup_{n \geq 1} \cap_{m \geq n} A_m$. Show also that $\limsup_{n \to \infty} I_{A_n}(x) = 1$ if and only if $x \in \cap_{n \geq 1} \cup_{m \geq n} A_m$. These two equivalences motivate the definitions

$$\liminf_{n \to \infty} A_n = \cup_{n \geq 1} \cap_{m \geq n} A_m, \quad \text{and} \quad \limsup_{n \to \infty} A_n = \cap_{n \geq 1} \cup_{m \geq n} A_m.$$

In probability and statistics one often also encounters events that occur *infinitely often*, or i.o., this notion is defined by A_n i.o. $= \limsup_{n \to \infty} A_n$.

(b) Prove that for any sequence of sets $A_1, A_2, \ldots$ in a measure space $(\Omega, \mathcal{A}, \nu)$,

$$\nu(\liminf_{n \to \infty} A_n) \leq \liminf_{n \to \infty} \nu(A_n).$$

Fatou's lemma

This inequality is known as Fatou's lemma. Here you have proven it for sequences of indicator functions. As we will se in Ex. A.6(b), it holds for any sequence of nonnegative measurable functions.

Ex. A.3 *Quick introduction to measure and integration, II: Lifting good candidates to bona fide measures.* An important and useful general result is *Carathéodory's Extension Theorem*, which we now briefly summarise; we will not ask you to prove it here (xx give reference xx), but there will be occasions where we need to verify its conditions. – Consider a nonempty space Ω, and an *algebra* $\mathcal{A}_0$ of subsets; this requires the system of sets to contain the emptyset and to be closed under complements and finite unions. The point is often to start with such an $\mathcal{A}_0$, with simpler subsets, before we go to $\mathcal{A} = \sigma(\mathcal{A}_0)$, the fuller σ-algebra generated by $\mathcal{A}_0$ subsets. Suppose ν is a function working on A_0 sets in $\mathcal{A}_0$, with the properties (i) $\nu(\emptyset) = 0$; (ii) ν is finitely additive on $\mathcal{A}_0$, which means $\nu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \nu(A_i)$ if these are disjoint and in $\mathcal{A}_0$; and (iii) that it has the *continuity property*, that if $A_1 \supset A_2 \supset A_3 \cdots$ is a full sequence of sets in $\mathcal{A}_0$, with $\cap_{i=1}^\infty A_i = \emptyset$, then $P(A_n) \to 0$. Then ν on $\mathcal{A}_0$ can be lifted to a full measure μ^* on $\mathcal{A} = \sigma(\mathcal{A}_0)$, with $\nu^*(A) = \nu(A)$ for all $A \in \mathcal{A}_0$; also, this extension is unique if ν is sigma-finite.

Carathéodory's Extension Theorem

the continuity property

(a) Suppose $\mathcal{A}_1$ and $\mathcal{A}_2$ are two different algebras, both generating the same σ-algebra, i.e. $\mathcal{A} = \sigma(\mathcal{A}_1) = \sigma(\mathcal{A}_2)$. Assume that ν_1 and ν_2 are finitely additive on $\mathcal{A}_1$ and $\mathcal{A}_2$, both satisfying the continuity condition, and that they are equal for at least all sets $A_3 \in \mathcal{A}_3$, say, another algebra generating $\mathcal{A}$. Then show that ν_1 and ν_2 can both be lifted to $\mathcal{A}$, and that they must be equal. – Go through the details for the following application, with subsets of $\mathbb{R}$, taking $\mathcal{A}_1$ as all (a, b), with finite unions and complements, and $\mathcal{A}_2$ as all $[a, b]$, with finite unions and complements.

(b) Consider in particular a $\nu\colon \mathcal{B}_0 \to [0, \infty]$, defined on the system $\mathcal{A}_0$ of subsets of $\mathbb{R}$, which are open intervals, finite unions of such, and the complements of all these again; also, the emptyset is included. Show that $\sigma(\mathcal{A}_0)$ is the same as $\mathcal{B} = \sigma(\mathcal{B}_0)$, the Borel sets generated by all open sets. Assume that $\nu(\emptyset) = 0$, that $\nu(A \cup B) = \nu(A) + \nu(B)$ for disjoint A, B in $\mathcal{A}_0$, and that it has the continuity property. Show that ν can be lifted to a bona fide measure on $(\mathbb{R}, \mathcal{B})$; this extension is also unique if ν is sigma-finite.

(c) (xx spell out, for $(\mathbb{R}, \mathcal{B})$ and $(\mathbb{R}^k, \mathcal{B}^k)$, that measures and indeed probability measures can be constructed by starting simple, with open intervals and open rectangles. link to cumulative distribution functions. xx)

(d) A fundamental measure is the so-called Lebesgue measure, say λ, on the real line with its Borel sets; again, these are $\mathcal{B} = \sigma(\mathcal{B}_0)$, the smallest σ-algebra containing all open sets. Its basic property is that $\lambda(a, b) = b - a$, the length of the interval in question. For $A = \cup_{j \in J} A_i$, a finite union of disjoint open intervals, we define $\lambda(A) = \sum_{j \in J} \lambda(A_j)$, the total length of A. Show that this is unambiguous, giving the same $\lambda(A)$ for possibly different representations of this type of sets A. Show also that λ must be sigma-finite. – What is not yet fully clear is that λ, defined on open intervals and unions of such, can be lifted to a bona fide measure on the full measurable space $(\mathbb{R}, \mathcal{B})$. This is the business of having a clear definition of $\lambda(A)$ also for very complicated sets A, and still keeping the simple $\lambda(a, b) = b - a$ for intervals. Show via Carathéodory's Extension Theorem that this is indeed possible.

Lebesgue measure

(e) 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.

(f) 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.9 and A.12 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.4 *Quick introduction to measure and integration, III: Integrals, probabilities, extensions.* After having defined measures and measurable functions, the next goal is to form a well-defined *integral*, say $\int f \,\mathrm{d}\nu = \int f(x)\,\mathrm{d}\nu(x)$, with $(\Omega, \mathcal{A}, \nu)$ a measure space, and with $f\colon \Omega \to \mathbb{R}$ a measurable function.

(a) We start with $f = I_A$, an indicator function, with $f(x) = 1$ if $x \in A$ and $f(x) = 0$ if $x \notin A$. Here define $\int I_A(x)\,\mathrm{d}\nu(x) = \int I_A \,\mathrm{d}\nu = \nu(A)$. Next, for f a nonnegative simple function, taking on only finitely many values, say $f(x) = \sum_{j=1}^k c_j I_{A_j}(x)$, define

$$\int f \,\mathrm{d}\nu = \sum_{j=1}^k c_j \nu(A_j).$$

Show that this is unambiguous, giving the same value for different representations of the same simple function.

(b) Let now $f \geq 0$ be any measurable function. By earlier efforts, see Ex. A.1(f), there is a monotone sequence $0 \leq f_1 \leq f_2 \leq \cdots$ of simple functions, each taking on only finitely many values, with $f_n(x) \to f(x)$ for each x. We define the integral $\int f \,\mathrm{d}\nu = \int f(x)\,\mathrm{d}\nu(x)$ as the limit of these $\int f_n \,\mathrm{d}\mu$, that is

Lebesgue integral

$$\int f \,\mathrm{d}\nu = \lim_n \int f_n \,\mathrm{d}\nu.$$

Show that the limit $\int f \,\mathrm{d}\nu$ is the same for all nondecreasing sequences of functions converging pointwise to f. For a fully general measurable f, show that we may represent it as $f = f_+ - f_-$, with these two being nonnegative. We then define

$$\int f \,\mathrm{d}\nu = \int f_+ \,\mathrm{d}\nu - \int f_- \,\mathrm{d}\nu,$$

provided one or both of these two terms are finite, so we avoid coming into ∞ minus ∞ type trouble. If A is measurable, and f a measurable function, show that fI_A is measurable too; we can hence define $\int_A f \,\mathrm{d}\nu = \int f I_A \mathrm{d}\nu$.

Lebesgue integral properties

(c) With f and g nonnegative simple functions, show from the above definitions that (i) if $f \leq g$, in the sense of $f(x) \leq g(x)$ for all x, then $\int f \,\mathrm{d}\nu \leq \int g \,\mathrm{d}\nu$; (ii) if $A \subset B$ are measurable sets, then $\int_A f \,\mathrm{d}\nu \leq \int_B f \,\mathrm{d}\nu$; (iii) if $\nu(A) = 0$, then $\int_A f \,\mathrm{d}\nu = 0$; (iv) if $f(x) = 0$ for all $x \in A$, then $\int_A f \,\mathrm{d}\nu = 0$; and, (v) $\int af \,\mathrm{d}\nu = a \int f \,\mathrm{d}\nu$ for any constant a.

(d) (xx showing that classical Riemann integration is a special case. xx) Suppose $f\colon [a,b] \to \mathbb{R}$ is a continuous function, on some interval $[a,b]$. Show that f is 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 in Ex. A.1(d)

Ex. A.5 *Almost surely, Borel–Cantelli, and convergence in measure.* Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. The set (or event) A in $\mathcal{A}$ is said to be true *almost surely* or *almost everywhere* if $\mu(A) = 1$. If μ is a probability measure, we say that A is true, or A occurs or happens, *with probability* 1.

(a) Let $A_1, A_2, \ldots$ be a sequence of events in $\mathcal{A}$, and suppose that $\sum_{n=1}^{\infty} \mu(A_n) < \infty$. Show that $\mu(\limsup_{n\to\infty} A_n) = 0$. When μ is a probability measure, this lemma has a converse, but that requires the notion of independence, to which we turn in Ex. A.11.

(b) A sequence of measurable functions f_n converges almost surely if the set

$$A = \{\omega\colon \limsup_{n\to\infty} f_n(\omega) - \liminf_{n\to\infty} f_n(\omega) = 0\},$$

has measure one, i.e. $\mu(A) = 1$. If $\mu(A) = 1$ then $\lim_{n\to\infty} f_n$ exists, and to represent this limit we pick a function f that is equal to $\lim_{n\to\infty} f_n$ almost everywhere. This means that if $\tilde{f}$ is another function and $\mu(\{x\colon \tilde{f}(x) \neq f(x)\}) = 0$, then it is also true that $\lim_{n\to\infty} f_n = \tilde{f}$. For all this to make sense, we helps to know that $\inf_n f_n$, $\sup_n f_n$, $\liminf_n f_n$, $\limsup_n f_n$, and $\lim_n f_n$ are measurable functions. Please show that they are.

(c) A sequence of measurable functions f_n converges *in measure* to a measurable function f if for any $\varepsilon > 0$, $\mu(\{x\colon |f_n(x) - f(x)|\}) \to 0$ as $n \to \infty$. When μ is a probability measure, this type of convergence is called convergence in probability, a notion meet we will innumerable times in this book. Assume that the functions $f_1, f_2, \ldots, f$ are defined on a measure space with finite measure, say $\mu(\Omega) = K$. Show that if $f_n \to f$ almost surely, then $f_n \to f$ in measure.

(d) Suppose that $f_n \to f$ in measure. Show that there exists a subsequence $f_{n_1}, f_{n_2}, \ldots$ that converges almost surely to f. [xx perhaps elaborate a bit more here xx]

Ex. A.6 *Convergence theorems.* One of the main objectives of the integration theory developed in the preceding exercises is to find general criteria for when $\lim_{n\to\infty} \int f_n\,\mathrm{d}\nu = \int \lim_{n\to\infty} f_n\,\mathrm{d}\nu$. The theorems that give various sets of conditions for when we can pass the limit under the integral sign like this, are the convergence theorems of measure theory. Remember that the measure ν 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 the Riemann integrals $\int f_n(x)\,\mathrm{d}x$. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and $f_1, f_2, \ldots$ a sequence of measurable functions.

(a) Let A be a set with finite measure, and suppose that $f_n(x) = 0$ on the set A^c for all n. Show that if $|f_n| \leq M$ for all n and $f_n \to f$ in measure, then $\lim_{n\to\infty} \int f_n\,\mathrm{d}\nu = \int f\,\mathrm{d}\nu$. Bounded convergence

(b) Suppose that $f_n \geq 0$ for all n. Show that Fatou's lemma

$$\int \liminf_{n\to\infty} f_n \, d\nu \leq \liminf_{n\to\infty} \int f_n \, d\nu.$$

Monotone convergence

(c) Suppose that $0 \leq f_1 \leq f_2 \leq \cdots$ almost surely, and that $f_n \to f$ almost surely. Show that $\lim_{n\to\infty} \int f_n \, d\nu \to \int f \, d\nu$.

Dominated convergence

(d) Let $f_1, f_2, \ldots$ be a sequence of measurable functions such that $f_n \to f$ almost surely. Suppose there is a nonnegative integrable function g so that $|f_n| \leq g$ for all n. Show that $\lim_{n\to\infty} \int f_n \, d\nu \to \int f \, d\nu$ and that $\lim_{n\to\infty} \int |f_n - f| \, d\nu = 0$.

(e) Show that the Dominated convergence theorem also holds under the weaker assumption $f_n \to f$ in measure.

Ex. A.7 *Bootstrapping the integral.* [xx some exercises where we establish various properties of the integral, using what we learned above xx]

(a) We can use the monotone convergence theorem to prove properties of the Lebesgue integral, for example (and crucially!) linearity,

$$\int (\alpha f + \beta g) \, d\nu = \alpha \int f \, d\nu + \beta \int g \, d\nu,$$

for all measurable functions f and g and constants α and β. Start by showing that linearity holds for simple functions $f = \sum_{i=1}^{k} a_i I_{A_i}$ and $g = \sum_{j=1}^{n} b_j I_{B_j}$. Next, appeal Ex. A.1(f) and the monotone convergence theorem to show that linearity holds for measurable nonnegative functions, not merely simple ones. Finally, extend it to a general measurable functions, that is, functions that might attain both positive and negative values.

(b) Show that $\int_A \sum_{j=1}^{\infty} f_j \, d\nu = \sum_{j=1}^{\infty} \int_A f_j \, d\nu$, from which it follows (prove it) that for any pairwise disjoint sequence $A_1, A_2, \ldots$ of measurable sets $\int_{\cup_{j=1}^{\infty} A_j} f \, d\nu = \sum_{j=1}^{\infty} \int_{A_j} f \, d\nu$. Deduce from this that if a nonnegative f has finite integral, then $\kappa(A) = \int_A f \, d\nu$ defines a finite measure. Here we also say that f is the density of κ with respect to ν, and write $f = d\kappa/d\nu$; see also Ex. A.12.

bootstrap arguments

(c) In (a) we touched on a proof strategy that appears again and again, so often that some call it a bootstrapping argument (not to be confused with bootstrapping in statistics). Here is an example: Let $\kappa(A) = \int_A f \, d\nu$ be the measure introduced in (b). We want to show that that for any measurable function g,

$$\int g \, d\kappa = \int gf \, d\nu.$$

First, prove this for indicator functions $g = I_A$. Second, by linearity go to simple functions $g = \sum_{j=1}^{k} a_j I_{A_j}$. Third, use Ex. A.1(f) and the monotone convergence theorem, and deduce that it extends to nonnegative measurable functions. Finally, using linearity again, show that it holds for all measurable functions g, provided g is integrable with respect to κ.

change of variable

(d) Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces, and let ν be a measure on $\mathcal{A}$, and let $T\colon \mathcal{X} \to \mathcal{Y}$ be a measurable function. Define the set function $\nu T^{-1}(B) = \mu(T^{-1}(A))$. – Show that νT^{-1} is measure in $\mathcal{B}$. Let $f\colon \mathcal{Y} \to \mathbb{R}$ be an integrable function. Show that

$$\int_{T^{-1}(B)} f(T(x))\mathrm{d}\nu(x) = \int_B f(y)\,\mathrm{d}\nu T^{-1}(y).$$

Here, you may once again use a bootstrapping argument. [xx fT is integrable with respect to ν iff f integrable w.r.t. νT^{-1} xx]

Ex. A.8 *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 fixed set; $\mathcal{A}$ a σ-algebra of subsets of this Ω (see Ex. A.1); and $P\colon \mathcal{A} \to [0,1]$ a *probability measure*, defined simply to be a measure, in the sense of Ex. A.1, 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.1, and for convenience stated again here, for the present case of $P(\Omega) = 1$. We demand

axioms for a probability space

(i) that $P(A) \geq 0$ for all $A \in \mathcal{A}$;

(ii) that $P(\Omega) = 1$;

(iii) that if $A_1, A_2, \ldots$ are disjoint sets in $\mathcal{A}$, then we have *countable additivity*, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

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) Deduce that $P(\emptyset) = 0$. For all events A and B deduce that $P(A) \leq 1$; that $P(A) = 1 - P(A^c)$; and that $P(A \setminus B) = P(A) - P(A \cap B)$. Show also that countable additivity implies finite additivity, namely $P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$, for each finite collection $A_1, \ldots, A_n$ of disjoint events.

(b) Deduce the following continuity properties. First, if $A_1 \subset A_2 \subset \cdots$, then $P(\cup_{i=1}^{\infty} A_i) = \lim P(A_i)$; secondly, if $A_1 \supset A_2 \supset \cdots$, then $P(\cap_{i=1}^{\infty} A_i) = \lim P(A_i)$. Show furthermore that either of these two statements could replace (iii) in the axiom list above. These are hence highly related to the continuity condition for measures, see Ex. A.3.

(c) From the axioms, show that if $A \subset B \subset C$, then $P(A) \leq P(B) \leq P(C)$. Show that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, for any A, B. Deduce also that $P(A \cup B) \leq P(A) + P(B)$, and, by induction, that $P(A_1 \cup \cdots \cup A_n) \leq P(A_1) + \cdots + P(A_n)$, for all events $A_1, \ldots, A_n$. Show that this rule also holds with infintely many events, $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$.

(d) Show that

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(A \cap B) - P(A \cap C) - P(A \cap B \cap C) + P(A \cup B \cup C),$$

and try to generalise to the union of four or more events.

(e) If A and B have probabilities 0.95, or more, show that $P(A \cap B) \geq 0.90$. Generalise. This simple lower-bounding of certain types of probabilities is sometimes called the Bonferroni method, or Bonferroni correction.

(f) For a σ-algebra $\mathcal{A}$, show that intersections, finite and countable, must be in – hence also sets like $\cup_{i=1}^{\infty} \cap_{j=i}^{\infty} A_j$, etc.

(g) 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

$$P(A) = \sum_{j\colon \omega_j \in A} p_j, \quad \text{for any subset } A,$$

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.1(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.9 *Distribution functions.* Consider the case where the probability space is $(\mathbb{R}, \mathcal{B}, P)$, with $\mathcal{B}$ the Borel sets on the real line (the smallest σ-algebra containing all intervals), and P is some probability 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

$$F(t) = P(A_t), \quad \text{with } A_t = (-\infty, t],$$

where we also allow the simpler notation $P(\infty, t]$ for $P((-\infty, t])$.

(a) Show that F is nonincreasing, right continuous, with $F(t) \to 1$ and $F(t) \to 0$ as $t \to \infty$ and $t \to -\infty$, respectively. Show also that $F(t - 1/n) \to P(-\infty, t)$, and that

$$P(a + 1/n, b - 1/n] = F(b - 1/n) - F(a + 1/n) \to F(b-) - F(a) = P[a, b),$$

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 $P(\{t\})$, 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. $F_1 = F_2$. From the mathematical analysis fact that every open set on the real line can be expressed as a finite or countably infinite union of disjoint open intervals, deduce that $P_1(A) = P_2(A)$ for all open A, and for all closed sets, too. – An important fact is that if $F_1 = F_2$, then indeed $P_1(A) = P_2(A)$ for *all* Borel sets; it is impossible to construct a perhaps very complicated set A for which the probabilities would disagree, if their cumulatives are equal. Show this, from Carathéodory's Extension Theorem of Ex. A.3, or, alternatively from Ex. A.17. 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) (xx briefly here about $\mathbb{R}^2$ and $\mathbb{R}^k$ too. sufficient to define probabilities on all boxes. xx) Suppose P is a probability measure on $(\mathbb{R}^2,\mathcal{B}^2)$, where $\mathcal{B}^2 = \sigma(\mathcal{B}_0^2)$ is the collection of Borel sets on the plane, the smallest σ-algebra containing all boxes, or rectangles, $(a_1,b_1)\times(a_2,b_2)$. Define the cumulative distribution function for the pair as

$$F(t_1,t_2) = P(A_{t_1,t_2}) = P((-\infty,t_1]\times(-\infty,t_2]).$$

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 Carathéodory Extension Theorem of Ex. A.3 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. – Attempt to generalise this to dimension k, i.e., to $(\mathbb{R}^k,\mathcal{B}^k)$; 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 $k=2$.

Ex. A.10 *Random variables.* Speaking mathematically, and more to the point in the language of measure theory and integration, a *random variable* is a measurable function on a probability space. In detail, with $(\Omega,\mathcal{A},P_0)$ a 'background' probability space, we may construct random variables as measurable functions $X\colon \Omega\to\mathcal{X}$, where $(\mathcal{X},\mathcal{B})$ is a measurable space where $X(\omega)$ lands. Measurability means that the inverse images $A = X^{-1}(B) = \{\omega\colon X(\omega)\in B\}$ are measurable, that is, $X^{-1}(B)$ belongs to $\mathcal{A}$ for any measurable B in the image space $\mathcal{X}$.

(a) Show that the probability distribution for X, say P, inherited from its ingredients, or, P is the probability distribution induced by X on the range of X, is

$$P(B) = P_0(X\in B) = P_0(\{\omega\colon X(\omega)\in B\}) = P_0(X^{-1}(B)), \quad \text{for } B\in\mathcal{B}.$$

We often write the probabilities like this, $P_0(X \in B)$, though we sometimes write $P_0\{X \in B\}$ or $P_0(\{X \in B\})$, for emphasis of the set, in the background probability space, behind the event in the image probability space. Show that $P = P_0X^{-1}$ indeed is a probability measure on $(\mathcal{X}, \mathcal{B})$; with pedantic care, we define P via $(P_0X^{-1})(A) = P_0(X^{-1}(A))$.

(b) Often what matters is the distribution of X, rather than particularities of the background space. Indeed there may be different spaces $(\Omega_j, \mathcal{A}_j, P_{0,j})$ and random variables $X_j\colon \Omega_j \to \mathcal{X}$ inducing precisely the same distribution, i.e., the different $P_j = P_{0,j}X_j^{-1}$ 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.9 that what matters is the c.d.f. $\Pr(X_j \le t) = P_j(\{\omega_j\colon X_j(\omega_j) \le t\}) = F_j(t)$; as long as these are equal, the distributions $P_j = P_{0,j}X_j^{-1}$ are identical. Give three separate such constructions of the standard normal distribution. – The general type construction becomes more useful when working with several random variables at the same time, say $X_1, \dots, X_n$, in which case it becomes practical to have them defined on the same background probability space.

the expectation

(c) If $X\colon \Omega \to \mathbb{R}$ is a real random variable, defined on a background probability space $(\Omega, \mathcal{A}, P_0)$, its *mean*, or *expected value*, is defined as

$$\mathrm{E}\, X = \int X \,\mathrm{d}P_0 = \int X(\omega)\,\mathrm{d}P_0(\omega),$$

as long as this integral is finite. This general definition requires measure and integration theory, but it may be sufficient to think of the integral $\int X\,\mathrm{d}P_0$ as the limit of simpler Riemann-like approximations, say via X_n taking on values c_j on sets A_j, with $X_n(\omega) \to X(\omega)$ (such a sequence can always be found, see Ex. A.1(f)); we then have $\int X_n\,\mathrm{d}P_0 = \sum c_j P_0(A_j)$, and under mild conditions, see Ex. A.6, $\int X_n\,\mathrm{d}P_0 \to \int X\,\mathrm{d}P_0$. – With $g(x)$ any nonnegative measurable function, use a bootstrap argument (see Ex. A.6(c)) to show that

$$\mathrm{E}\, g(X) = \int g(X(\omega))\,\mathrm{d}P_0(\omega) = \int g(x)\,\mathrm{d}P(x), \quad \text{with } P = P_0X^{-1}.$$

In particular, only the distribution of X matters, not the details associated with the background probability space.

the variance

(d) With the mean of a real random variable well defined, we may of course go on to other and higher moments. For a real random variable $X\colon \Omega \to \mathbb{R}$, as above, with $\xi = \mathrm{E}\, X$, show that

$$\mathrm{E}\,(X-\xi)^2 = \int (X-\xi)^2\,\mathrm{d}P_0 = \int_{-\infty}^{\infty} (x-\xi)^2\,\mathrm{d}P(x) = \int_0^{\infty} y\,\mathrm{d}Q(y),$$

with Q the distribution of $Y = (X-\xi)^2$; so there's no ambiguity. This quantity is of course *the variance of* X, denoted $\mathrm{Var}\, X$, the square of *the standard deviation of* X.

(e) Consider real random variables X, Y defined on the same background probability space $(\Omega, \mathcal{A}, P_0)$. Show that $\mathrm{E}\,(aX + bY) = a\,\mathrm{E}\, X + b\,\mathrm{E}\, Y$, and generalise. In particular,

for 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. For variables with finite second moments, show that $\mathrm{Var}\,X = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2$.

Ex. A.11 *Independence.* Here we define and work through basic properties of *independence*, for events and for random variables.

(a) For a probability space $(\Omega, \mathcal{A}, P)$, we start out saying that two events A and B are independent if $P(A \cap B) = P(A)P(B)$. Show that then also 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 $P(A \cap B \cap C) \neq P(A)P(B)P(C)$. Hence care is needed when defining independence for more than two events. We say that $A_1, \ldots, A_n$ are independent if $P(A_{i_1} \cap \cdots A_{i,k}) = P(A_{i,1}) \cdots P(A_{i,k})$ holds, for any finite subset $\{i_1, \ldots, i_k\}$ of $\{1, \ldots, n\}$. Show that then also $A_1^c, \ldots, A_n^c$ are independent.

(c) Consider X, Y defined on the same probability space $(\Omega, \mathcal{A}, P_0)$, with distributions $P_1 = P_0 X^{-1}$ and $P_2 = P_0 Y^{-1}$. We say that X and Y are independent if $\{X \in A\}$ and $\{Y \in B\}$ are independent events, i.e.

$$\Pr(X \in A, Y \in B) = P_0(\{\omega\colon X(\omega) \in A, Y(\omega) \in B\}) = P_1(A)P_2(B),$$

for any A, B. Show the existence of a product measure $Q = P_1 \times P_2$, on the σ-algebra $\mathcal{A}^2 = \mathcal{A} \times \mathcal{A}$, generated by all $A \times B$ sets with $A, B \in \mathcal{A}$, such that $Q(A \times B) = P_1(A)P_2(B)$ for all such sets.

(d) So $Q(C) = \Pr((X, Y) \in C)$ 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.

(e) Show that $\mathrm{E}\,g(X)h(Y) = \mathrm{E}\,g(X)\,\mathrm{E}\,h(Y)$ for each $g(x)$ and $h(y)$ for which the means exist. (xx a bit more; round it off. covariance, correlation. xx)

(f) We must of course extend the above to the case of more than two independent random variables. With $X_1, \ldots, X_n$ defined on the same underlying probability space $(\Omega, \mathcal{A}, P_0)$, their distributions are $P_1 = P_0 X_1^{-1}, \ldots, P_n = P_0 X_n^{-1}$. We say that $X_1, \ldots, X_n$ are independent if

$$\begin{aligned} \Pr(X_1 \in A_1, \ldots, X_n \in A_n) &= P_0(\{\omega\colon X_1(\omega) \in A_1, \ldots, X_n(\omega) \in A_n\}) \\ &= P_1(A_1) \cdots P_n(A_n) \end{aligned}$$

for all Borel sets $A_1, \ldots, A_n$. Show that this is equivalent to having $\{X_1 \in A_1\}, \ldots, \{X_n \in A_n\}$ independent for every $A_1, \ldots, A_n$. Show also that this gives rise to a well-defined product probability measure $Q = P_1 \times \cdots \times P_n$ on the σ-algebra $\mathcal{A}^n = \mathcal{A} \times \cdots \times A$, generated by all $A_1 \times \cdots \times A_n$.

(g) Show that if $X_1, \dots, X_n$ independent, and $g_1, \dots, g_k$ are measurable functions, then also $g_1(X_1), \dots, g_k(X_k)$ 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.

(h) (xx with care: $\mathrm{Var}\,(X_1 + \cdots + X_k) = \mathrm{Var}\, X_1 + \cdots + \mathrm{Var}\, X_k$, for independent variables. point briefly to stigler and seven pillars. xx)

(i) xx

Ex. A.12 *Probability densities.* We have seen in Ex. A.9 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 $P(A)$ for all A, is the most convenient (and traditional) way in which to define a probability distribution.

(b) Suppose in general terms that ν is a σ-finite measure on a measurable space $(\mathcal{X}, \mathcal{A})$. That ν is σ-finite means that there is a countable division $\mathcal{X} = A_1 \cup A_2 \cup A_3 \cdots$, where each $\nu(A_i)$ is finite. Suppose next that the probability measure P is dominated by ν, meaning that $\nu(A) = 0$ implies $P(A) = 0$; one also says that P is absolutely continuous with respect to ν. Under these conditions, the Radon–Nikodym theorem says that there is *a density*, say $f(x)$, such that

absolute continuity

the Radon–Nikodym theorem

$$P(A) = \int_A f(x)\,\mathrm{d}\nu(x) \quad \text{for all } A \in \mathcal{A}. \tag{A.1}$$

The density f is often denoted $\mathrm{d}P/\mathrm{d}\nu$, to remind us that this is a density of P with respect to ν. Explain that what we have seen above, where F has a derivative and is the integral of this derivative, matches this more general setup, where ν is the so-called 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) Absolute continuity is related to the ε-and-δ definition of continuity. Suppose that ν and μ are finite measures. Show that if for any $\varepsilon > 0$ there exists $\delta > 0$, such that $\nu(A) < \varepsilon$ whenever $\mu(A) < \delta$, then $\nu \ll \mu$. Prove the converse.

(d) The strength of the general $f = \mathrm{d}P/\mathrm{d}\nu$ 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 $\mathcal{X}$, $\nu(A)$ equal to the number of numbers $j \in A$, that is, ν is the counting measure on the integers, which, from Ex. A.1(k), we know is a σ-finite measure. Show that with P having a Poisson distribution P, with mean θ, that there is a density $f = \mathrm{d}P/\mathrm{d}\nu$, given by $f(x) = \exp(-\theta)\theta^x/x!$ for $x = 0, 1, 2, \ldots$, in the sense given above.

(e) 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 ν, which has pointmasses 1 and 1 at the points 0 and 1, and is uniform inside $(0, 1)$. Find the probability density $f(x) = \mathrm{d}P(x)/\mathrm{d}\nu$.

(f) Suppose P is dominated by a sigma-finite ν, with $f(x) = \mathrm{d}P(x)/\mathrm{d}\nu$ the probability density, as per (A.1). With X having distribution P, if you have not already done so in Ex. A.6(c), show that for any $g(x)$ for which the mean is finite (with respect to P), that

$$\mathrm{E}\, g(X) = \int g(x)\,\mathrm{d}P(x) = \int g(x)\,\frac{\mathrm{d}P(x)}{\mathrm{d}\nu}\,\mathrm{d}\nu(x) = \int g(x) f(x)\,\mathrm{d}\nu(x).$$

(g) (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 13-Aug-2023, 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.13 *Proving the Radon–Nikodym theorem.* To prove the Radon–Nikodym theorem

Ex. A.14 *Radon–Nikodym derivatives* Let p, q, μ be σ-finite measures on the measurable space $(\Omega, \mathcal{A})$.

(a) Show that if $p \ll q$ and $q \ll \mu$, then $p \ll \mu$, then

$$\frac{\mathrm{d}p}{\mathrm{d}\mu} = \frac{\mathrm{d}p}{\mathrm{d}q}\frac{\mathrm{d}q}{\mathrm{d}\mu}, \quad \mu \text{ almost surely.}$$

(b)

(c)

Ex. A.15 *Modes of convergence*[xx check if such an exercise already exists in the large-sample chapter xx]

(a) Suppose that X_n converges almost surely to X. Show that X_n converges in probability to X.

(b) Suppose that $X_n \to_p X$ as $n \to \infty$. Show that there exists a subsequence $(X_{n_k})_{k\geq 1}$ so that $X_{n_k} \to X$ almost surely.

Ex. A.16 *Uniform integrability.* Let $(\Omega, \mathcal{A}, P)$ be a probability space, and write $\{X > M\} = \{\omega \in \Omega\colon X(\omega) > M\}$.

(a) Suppose that X is a random variable. Show that X is integrable if and only if there is a real number M such that $\mathrm{E}\,|X|I_{|X|>M} < \infty$.

(b) Suppose that X is integrable. Show that $\lim_{M\to\infty} \mathrm{E}\,|X|I_{|X|>M} = 0$.

(c) Let $X_1, X_2, \ldots$ be a sequence of integrable random variables, converging almost surely to X. Show that $\liminf_{n\to\infty} |X_n|I_{|X_n|>M} \geq |X|I_{|X|>M}$, almost surely.

(d) Let $X_1, X_2, \ldots$ be a sequence of integrable random variables, converging almost surely to X. Show that if $\mathrm{E}\,|X_n|I_{|X_n|>M} \leq \delta$ for all n, then $\mathrm{E}\,|X|I_{|X|>M} \leq \delta$.

Uniform integrability

(e) A sequence of random variables $X_1, X_2, \ldots$ is called *uniformly integrable* if

$$\lim_{M\to\infty} \sup_{n\geq 1} \mathrm{E}\,|X_n|I_{|X_n|>M} = 0.$$

Assume that $X_1, X_2, \ldots$ is a uniformly integrable sequence of random variables converging almost surely to X. Show that X is integrable.

(f) Assume that $X_1, X_2, \ldots$ is a uniformly integrable sequence of random variables converging to X almost surely. Show that $\lim_{n\to\infty} \mathrm{E}\,X_n = \mathrm{E}\,X$.

(g) Show that in Ex. (f), it is sufficient that $X_n \to_p X$.

(h) Show that if $(X_n)_{n\geq 1}$ is uniformly integrable and $X_n \to_p X$, then $\mathrm{E}\,|X_n - X| \to 0$.

(i) Let $X_1, X_2, \ldots$ be random variables such for some $\delta > 0$ it is the case that $\mathrm{E}\,|X_n|^{1+\delta} < \infty$ for all n. Show that $(X_n)_{n\geq 1}$ is uniformly integrable.

Ex. A.17 *Monotone classes and some more.* For many classes of sets it is possible to give an explicit characterisation of its elements. The collection of all half intervals open on the left and closed on the right on the real line, for example, consists of elements $(a, b]$ with $a < b$. [xx fix xx]. With the σ-algebras that interest us in probability and statistics, however, it is not possible to give such 'closed form' characterisations of its elements. This is the motivation for the definitions and results we introduce in this exercise.

(a) A family of sets $\mathcal{M}$ is a *monotone class* if $A_1 \subset A_2 \subset \cdots$ are sets in $\mathcal{M}$, then $\cup_{n\geq 1} A_n \in \mathcal{M}$, and if $B_1 \supset B_2 \supset \cdots$ are sets in $\mathcal{M}$, then $\cap_{n\geq 1} B_n \in \mathcal{M}$. Show that a collection of sets is a σ-algebra if and only if it is a monotone class and a algebra.

(b) Let $\mathcal{C}$ be a family of sets. Show that there is a smallest monotone class, say $m(\mathcal{C})$, such that $\mathcal{C} \subset m(\mathcal{C})$.

Monotone class theorem

(c) 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 $\mathcal{A}_0 \subset \mathcal{M}$. Prove that $\sigma(\mathcal{A}_0) \subset \mathcal{M}$.

The importance of this theorem stems from the fact that certain properties can be proved for sets in an algebra, which are easy to describe, and as long as this algebra is contained in a monotone class, the property holds for all sets in the σ-algebra generated by the algebra. We'll see an example of such a 'monotone class argument' in Ex. A.18.

(d) [xx π-systems, λ-systems, Dynkin's π-λ lemma. If two finite measures agree on a π-system, then they agree on the σ-algebra generated by that π-system. xx]

Ex. A.18 *Fubini and Tonelli.* Let $(\Omega_1, \mathcal{A}_1, \mu_1)$ and $(\Omega_2, \mathcal{A}_2, \mu_2)$ be two measure spaces. The $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mu_1 \times \mu_2)$ is called the *product space*, with $\mathcal{A}_1 \otimes \mathcal{A}_2$ being the smallest sigma algebra generated by the products $A_1 \times A_2$ with $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$, and $\mu_1 \times \mu_2$ is the extension to $\mathcal{A}_1 \otimes \mathcal{A}_2$ (see Ex. abc) of the set funtion $(\mu_1 \times \mu_2)(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$. Product space

(a) (xx check this: If both μ and ν are sigma-finite, then it can be shown using the Carathéodory's Extension Theorem that $\mu_1 \times \mu_2$ is the only measure on $\mathcal{A}_1 \otimes \mathcal{A}_2$ with the property that $\mu_1 \times \mu_2(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ xx)

(b) Let $\mathcal{A}_0$ be the collection of finite disjoint unions of measurable rectangles $A_1 \times A_2$ (that is $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$). Show that $\mathcal{A}_0$ is an algebra.

(c) Fubini's theorem says that if $f\colon \Omega_1 \times \Omega_2 \to \mathbb{R}$ is an integrable function on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mu_1 \times \mu_2)$, then Fubini's theorem

$$\begin{aligned} \int f(x,y)\,\mathrm{d}(\mu_1 \times \mu_2)(x,y) &= \int\int f(x,y)\,\mathrm{d}\mu_1(x)\mathrm{d}\mu_2(y) \\ &= \int\int f(x,y)\,\mathrm{d}\mu_2(y)\mathrm{d}\mu_1(x). \end{aligned} \tag{A.2}$$

Let $B = A_1 \times A_2$ be a set in $\mathcal{A}_1 \otimes \mathcal{A}_2$, and let $f = I_B$ Show that $\int f(x,y)\,\mathrm{d}(\mu_1 \times \mu_2)(x,y) = \int\int f(x,y)\,\mathrm{d}\mu_1(x)\mathrm{d}\mu_2(y)$.

(d) Let $\mathcal{M}$ be the class of sets B such that the claim in A.2 is true for $f = I_B$. By Ex. (c), the algebra $\mathcal{A}_0$ of Ex. (b) is contained in $\mathcal{M}$. Show that $\mathcal{M}$ is a monotone class, and conclude that the claim in A.2 is true for $f = I_B$, with B being any set in $\mathcal{A}_1 \otimes \mathcal{A}_2$. Now, use a bootstrapping argument to finish up the proof of Fubini's theorem.

(e) Tonelli's theorem is much like Fubini's. The difference is that the function f is assumed nonnegative (but not necessarily integrable), while both μ_1 and μ_2 are assumed to be sigma-finite measures. The conclusion of Tonelli's theorem is that given in (A.2). [xx formulate an exercise xx]. Tonelli's theorem

Ex. A.19 *Conditional probability.* When the sun is shining in the morning, there is a fair chance that it will be shining in the afternoon. Suppose that this inductive insight is based on observing the weather for the n previous days, of which there were m days with a sunny morning, and k days with a sunny morning *and* a sunny afternoon. On this particular day, you wake up in sunshine and want to estimate the probability of a sunny

afternoon. The natural estimate of the probability of a sunny afternoon is k/m. Using the strong law of large numbers (which you will be asked to prove in Ex. abc),

$$\frac{k}{m} = \frac{k/n}{m/n} \to \frac{P(\text{sunny morning \& sunny afternoon})}{P(\text{sunny morning})},$$

as the number of days n grows without bounds. This example goes to show that the definition of the conditional probability is the intuitive one. Here is the definition: Let $(\Omega, \mathcal{A}, P)$ be a probability space, and A and B be events, that is $A, B \in \mathcal{A}$. Provided $P(B) > 0$, the conditional probability of A given B is

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)}, \tag{A.3}$$

The notation $P(A \,|\, B)$ might be unfortunate, because it sort of seems that the events A and B are on 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 other words, $A \mapsto P(A \,|\, B)$ is a probability measure, while $B \mapsto P(A \,|\, B)$ is not a probability measure.

(a) Show that $P(A \,|\, B)$ is a probability measure, providedf $P(B) > 0$.

(b) The function $L(B) = P(A \,|\, B)$ 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 $P(B) > 0$. If A and B are independent, show that $P(A \,|\, B) = P(A)$.

(d) Show that if $P(A) = 0$, then A and B are independent. Also, show that if $P(A) = 1$, then A and B are independent.

Ex. A.20 *Conditional expectation.* Let X and Y be two random variables on a probability space $(\Omega, \mathcal{A}, P_0)$. For events A and B, the probability of X falling in A given that $Y \in B$ is, using the definition in (A.3), $P_0(X \in A \,|\, Y \in B) = P_0(X \in A, Y \in B)/P_0(Y \in B)$, provided $P_0(Y \in B) > 0$. We can then define the conditional expectation of X given $Y \in B$ by

$$\mathrm{E}_0\{X \,|\, Y \in B\} = \frac{\mathrm{E}_0\{X I_{Y \in B}\}}{P_0(Y \in B)} = \int X \mathrm{d}P_0(\omega \mid Y \in B)$$

the conditional c.d.f. $F(x \,|\, Y \in B) = P_0(X \le x \,|\, Y \in B)$, and so on. The point being that conditioning on an event $\{Y \in B\}$ with positive probability is just a matter using the definition in (A.3) in the obvious manner.

(a) Let $X \sim \mathrm{N}(0,1)$ and $Y = I\{X \ge c\}$ for some constant c. Show that

$$\mathrm{E}_0\{X \mid Y = 1\} = \int_{\mathbb{R}} x \frac{\phi(x)}{1 - \Phi(c)} I_{[c,\infty)}(x)\, \mathrm{d}x.$$

If $P^X = P_0 X^{-1}$ is the measure induced on the range of X, and we define the conditional probability $P^{X|Y=1}(A) = P^X(A \,|\, [c,\infty)])/P^X[c,\infty)$, for A in the range of X, explain why this means that $(\mathrm{d}P^{X|Y=1}/\mathrm{d}\lambda)(x) = \phi(x)/\{1 - \Phi(c)\} I_{[c,\infty)}$, where λ is Lebesgue measure on the real line.

(b) In (a) we conditioned on the event $\{Y = 1\} = \{X \geq c\}$. But what if rather want to condition on the random variable Y (as defined in (a)), whatever that might mean? In view of of the above it makes sense to define the conditional expectation of X given Y by

$$\mathrm{E}_0\,(X \,|\, Y)(\omega) = \mathrm{E}_0\,(X \,|\, Y = 0) I_{\{Y=0\}}(\omega) + \mathrm{E}_0\,(X \,|\, Y = 1) I_{\{Y=1\}}(\omega),$$

for $\omega \in \Omega$. Notice that while $\mathrm{E}_0\,(X \,|\, Y = 0)$ and $\mathrm{E}_0\,(X \,|\, Y = 0)$ are constants, $I_{\{Y=1\}}$ and $I_{\{Y=0\}}$ are random variables, entailing that $\mathrm{E}_0\,(X \,|\, Y)$ is a random variable. Show that $\mathrm{E}_0\,\mathrm{E}_0\,(X \,|\, Y) = \mathrm{E}_0\,X$.

(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 $B_k = \{\omega \in \Omega \mid Y(\omega) = y_k\}$ for $k = 1, 2, \ldots$. As a mild extension of (b) we define the conditional expectation of X given Y as

$$\mathrm{E}_0\,(X \,|\, Y)(\omega) = \sum_{k \geq 1} \mathrm{E}_0\,(X \,|\, B_k) I_{B_k}(\omega). \tag{A.4}$$

Show that $\mathrm{E}_0\,\mathrm{E}_0\,(X \,|\, Y) = \mathrm{E}_0\,X$, and also $\mathrm{E}_0\,I_{B_j}\mathrm{E}_0\,(X \,|\, Y) = \mathrm{E}_0 I_{B_j}\,X$ for any B_j as defined above. In fact, show that

$$\mathrm{E}_0\,I_C \mathrm{E}_0\,(X \,|\, Y) = \mathrm{E}_0 I_C\,X, \tag{A.5}$$

for any event C in the σ-algebra generated by Y. Note that $\{\emptyset, B_1, B_2, \ldots\}$ is a π-system generating $\sigma(Y)$, so we can use Ex. A.17(d) here, and, for simplicity you may assume that $X \geq 0$ so that both $C \mapsto \mathrm{E}_0\,I_C \mathrm{E}_0\,(X \,|\, Y)$ and $C \mapsto \mathrm{E}_0 I_C\,X$ are finite measures.

(d) So far we have only considered discrete random variables Y. But what if the random variable we want to condition on, say Y, is continuous, so that $P_0(Y = y) = 0$ for all y. Then the definition of $\mathrm{E}_0\,(X \,|\, Y)$ given in (A.4) does not make sense, because it would involve division by zero. The solution to this problem is to take (A.5) as the definition of conditional expectation: On the probability space $(\Omega, \mathcal{A}, P_0)$ let X be an integrable random variable. Let $\mathcal{C} \subset \mathcal{A}$. Then any $\mathcal{C}$-measurable random variable Z such that

conditional expectation

$$\mathrm{E}_0\,I_C Z = \mathrm{E}_0\,I_C X, \quad \text{for all } C \in \mathcal{C}, \tag{A.6}$$

is called the conditional expectation of X given $\mathcal{C}$, denoted $\mathrm{E}_0(X \mid \mathcal{C})$. This definition entails that the conditional expectation is only defined up to sets of measure zero, and we call each Z satisfying (A.6) a *version* of the conditional expectation. Suppose that $X \geq 0$ and that $Z_1 \geq 0$ and $Z_2 \geq 0$ are random variables, both satisfying (A.6). Show that $P_0(Z_1 \neq Z_2) = 0$.

(e) Comparing the definition in (A.6) with the result in (A.5), we see that $\mathcal{C}$ corresponds to the σ-algebra generated by Y, i.e., $\sigma(Y)$. In fact, when $\mathcal{C}$ is a σ-algebra generated by a random variable Y, we typically write $\mathrm{E}_0(X \mid Y)$ instead of the more cumbersome $\mathrm{E}\,(X \,|\, \mathcal{C})$ or $\mathrm{E}\,\{X \,|\, \sigma(Y)\}$. Let's go 'backwards', and see that (A.6) leads back to the definition we started out with. Suppose that $B_1, B_2, \ldots$ are disjoint sets of positive probability whose union equals Ω.

(f) (xx properties, Jensen's, DCT, and the like for conditional expectation xx)

(g) (xx Conditional variance. Show that (xx modulo simple general assumptions xx) xx) $\operatorname{Var} X = \operatorname{E} \operatorname{Var} (X \mid Y) + \operatorname{Var} \operatorname{E} (Y \mid X)$.

(h) (xx conditional probability conditioned on σ-algebra, regular conditional probability xx)

(i) (xx conditional densities xx)

(j) (xx Borel pardox, examples xx)

Ex. A.21 *Conditional expectation.* red(xx drop this old ex xx) In Ex. A.20 we defined the conditional expectation $\operatorname{E}[X \mid Y]$ in the case that Y is discrete, or (X, Y) has a joint density. In this exercise we proceed to the general case, and define the conditional expectation of a random variable given a σ-algebra, instead of a random variable. [xx some more intro xx] We concentrate on expectations because probabilities are expectation over indicators functions, and if these probabilities are dominated, then there is density (per the Radon–Nikodym theorem). In other words, one can argue that among expectations, probabilities, and densities, the first is the fundamental concept.

Let X and Y be two random variables on the probability space $(\Omega, \mathcal{A}, P_0)$, with joint distribution P. Recall that $\sigma(Y)$ is the σ-algebra generated by Y [xx not defined xx].

(a) As in Ex. A.20, suppose that Y is discrete or (X, Y) has a joint density. Define $\operatorname{E}[X \mid \sigma(Y)] = \varphi(Y)$, where $\varphi(y)$ is as defined in (**??**) or (**??**). Show that (i) $\operatorname{E}[X \mid \sigma(Y)]$ is $\sigma(Y)$ measurable, and (ii) that $\int_A \operatorname{E}[X \mid \sigma(Y)] \, dP_0 = \int_A X \, dP_0$ for all $A \in \sigma(Y)$.

(b) Properties (i) and (ii) in Ex. (a) are the characterising features of conditional expectation: Any random variable with these properties is a conditional expectation. Here is the definition: Let $(\Omega, \mathcal{A}, P_0)$ be a probability space, X an integrable random variable, and let $\mathcal{G} \subset \mathcal{A}$. A random variable Z is the conditional expectation of X given $\mathcal{G}$ if Z is $\mathcal{G}$-measurable and $\int_A Z \, dP_0 = \int_A X \, dP_0$ for all $A \in \mathcal{G}$. We write $\operatorname{E}[X \mid \mathcal{G}]$ for Z, and when we want to be pedantic, call $\operatorname{E}[X \mid \mathcal{G}]$ a version of the conditional expectation of X given $\mathcal{G}$. It is a version of this conditional expectation, because any other random variable that is almost surely equal to Z also is also the conditional expectation of X given $\mathcal{G}$. Suppose that Z and Y are the conditional expectation of X given $\mathcal{G}$, show that $P_0(Z = Y) = 1$.

Conditional expectation

(c) Conditional expectation exists. Let $(\Omega, \mathcal{A}, P_0)$ be a probability space, and X an integrable random variable. Let $\mathcal{G} \subset \mathcal{A}$. Then there exists a $\mathcal{G}$-measurable function Z such that

$$\int_A Z \, dP_0 = \int_A X \, dP_0, \quad \text{for all } A \in \mathcal{G}.$$

Use Ex. A.4 along with the Radon–Nikodym theorem to prove this claim.

(d) [xx properties of conditional expectation xx]

(e) Let $X \colon \Omega \to \mathbb{R}$ and $Y \colon \Omega \to \mathcal{Y}$ be random variables on a probability space $(\Omega, \mathcal{A}, P_0)$. Assume that X is integrable, that the σ-algebra $\mathcal{C}$ on $\mathcal{Y}$ contains all singletons, and that

P_Y is the probability measure on $(\mathcal{Y},\mathcal{C})$ induced by Y (see Ex. A.10(a)). Prove that there exists a function $\varphi\colon \mathcal{Y}\to\mathbb{R}$ such that

$$\int_{Y^{-1}(C)} X\,\mathrm{d}P_0 = \int_C \varphi(y)\,\mathrm{d}P_Y(y),$$

for all $C\in\mathcal{C}$. It is the function $\varphi(y)$ we typically denote by $\mathrm{E}\,[X\,|\,Y=y]$. *Hint*: Combine the Radon–Nikodym theorem with Ex. A.10 (c).

(f)

Ex. A.22 *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 $\int_0^\infty(1-F)\,\mathrm{d}x$.

(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.23 *Convolutions.* If μ and ν are two finite measures on $\mathbb{R}$ (equipped with the Borel σ-algebra), the *convolution* $\mu*\nu=\nu*\mu$ is the unique finite Borel measure on $\mathbb{R}$ such that for any bounded continuous function $f\colon\mathbb{R}\to\mathbb{R}$

$$\int f\,\mathrm{d}(\mu*\nu)=\int\int f(x+y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)=\int\int f(x+y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x). \tag{A.7}$$

Notice that if $\mu\times\nu$ is the product measure on $\mathbb{R}\times\mathbb{R}$, and $g\colon\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is the function $g(x,y)=x+y$, then g induces the measure $(\mu\times\nu)\circ g^{-1}$ on $\mathbb{R}$, and $(\mu\times\nu)\circ g^{-1}=\mu*\nu$, because

$$\int f(z)(\mu\times\nu)\circ g^{-1}(\mathrm{d}z)=\int f(g(x,y))\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)=\int f(x+y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y).$$

In this sense, the definition in (A.7) is just a reformulation of Fubini's theorem.

(a) Let g be a real-valued non-negative integrable function. Define the measure $\nu(B)=\int_B g(x)\,\mathrm{d}x$, and let μ be a finite measure on $\mathbb{R}$. Show that

$$\int f\,\mathrm{d}(\mu\times\nu)=\int f(y)(g*\mu)(y)\,\mathrm{d}y,$$

where $(g*\mu)(y)$ is the function

$$(g*\mu)(y)=\int g(y-x)\,\mathrm{d}\mu(x).$$

(b) Let X and Y be two independent random variables with distributions P_X and P_Y, respectively. Show that the distribution P_Z of the sum $Z=X+Y$ is given by the convolution of P_X and P_Y, defined by

$$(P_X*P_Y)(A)=\int\int I_A(x+y)P_X(\mathrm{d}x)P_Y(\mathrm{d}y).$$

(c) Suppose now that X has density $f_X(x)$. Show that $Z = X + Y$ has density

$$f_Z(z) = \int f_X(z-y)\, P_Y(\mathrm{d}y).$$

Show also that if Y has density $f_Y(y)$, then

$$f_Z(z) = \int f_X(z-y) f_Y(y)\,\mathrm{d}y = \int f_X(x) f_Y(z-x)\,\mathrm{d}x,$$

which we may write as $f_Z(z) = \mathrm{E}\, f_X(z-Y) = \mathrm{E}\, f_Y(z-X)$.

(d) Suppose that g is bounded and continuously differentiable, whose derivative g' is bounded. Show that for any finite measure μ, the convolution $(g * \mu)'$ is also bounded and continuously differentiable, and that

$$(g*\mu)'(x) = (g'*\mu)(x) = \int g'(x-y)\,\mathrm{d}\mu(y).$$

Generalise to k times continuously differentiable functions with compact support. [xx check details here xx]

Ex. A.24 *More convolutions.* (xx we shall see, perhaps something easier than the previous very general exercise, for sums. xx) [xx we show basic convolution formulae; useful for later. xx] Let X and Y be independent real random variables with cumulative distribution functions F and G, and consider $Z = X + Y$.

(a) Show that Z has distribution function

$$H(z) = \int F(z-y)\,\mathrm{d}G(y) = \int G(z-x)\,\mathrm{d}F(x).$$

(b) When both F and G have densities, say f and g, show that H has density

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

(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)

Ex. A.25 *Something More.* (xx not yet an exercise, but a place to jot down a few comments, also as of 13-Aug-2023. 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)

1.C 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. Billingsley (1968); Royden and Fitzpatrick (2010). 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. xx)

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.

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.

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.

Billingsley, P. (1961). *Statistical Inference for Markov Processes*. Chicago University Press, Chicago.

Billingsley, P. (1968). *Convergence of Probability Measures*. 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.

Brown, L. D. (1986). Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lecture Notes-Monograph Series*, 9:i–279.

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. (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. (2023). 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.

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.

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.

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. (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.

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. U. (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. (1927). The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, 19:240–245.

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.

Hary, A. (1960). *10,0*. Copress, München.

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. (1976). The Dirichlet Process Applied to Nonparametric Estimation Problems. Technical report, University of Tromsø, University of Tromsø. Cand. real. thesis, Universities of Tromsø and Oslo.

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. (1988). The eccentric part of the non-central chi square. *The American Statistician*, 42:130–132.

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. (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 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.

Irwin, J. O. (1927). On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson's type II. *Biometrika*, 19:225–239.

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.

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.

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. (1933). 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.

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.

Phadia, E. G. (1973). Minimax estimation of a cumulative distribution function. *Annals of Statistis*, 6:1149–1157.

Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Books, Toronto.

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.

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. (2017). Bayesian Analysis: Always and Everywhere? *FocuStat Blog, University of Oslo*, iii.

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.

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 18551955*. 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. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155.

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.

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