Exercises and Lecture Notes
STK 4090, Spring 2020

Version 0.88, 03-05-2020

Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract

These are Exercises and Lecture Notes for the new course on Statistical Large-Sample Theory, STK 4090 (Master level) or STK 9090 (PhD level), for the spring semester 2020. Some of them are taken from earlier collections, from other courses of mine, but most of the exercises are created during this semester. The internal organisation and sequence of exercises might not be pedagogically optimal (yet), since more exercises are added on dynamically as the course progresses.

A later version of these notes, jfr. the Kioskvelter Project of N.L. Hjort and E.Aa. Stoltenberg, might be finessed and reorganised and polished to land in somewhat separated parts I + II + III + IV + V, where the first four parts roughly correspond to or are correlated with the first four parts of Ferguson (1996), whereas part V will concern the basics of empirical processes.

1. Illustrating the Central Limit Theorem (CLT)

Consider the variable

$$Z_n = (X_1 + \cdots + X_n - n\mu)/(\sqrt{n}\sigma) = \sqrt{n}(\bar{X}_n - \mu)/\sigma,$$

where the X_i are i.i.d. and uniform on the unit interval; here $\mu = 1/12$ and $\sigma = 1/\sqrt{12}$ are the mean and standard deviation, respectively. Your task is to simulate sim $= 10^4$ realisations of the variable Z_n, for say $n = 1, 2, 3, 5, 10, 25$, and display the corresponding histograms. Observe how the distribution of Z_n comes closer and closer to the standard normal, as n increases. To illustrate just how close, consider the case of $n = 6$, for example, and attempt to test the hypothesis that the 10^4 data points you have simulated come from the standard normal. Comment on your findings.

2. Illustrating the Law of Large Numbers (LLN)

Simulate say 10^4 variables $X_1, X_2, \ldots$ drawn from the unit exponential distribution. Compute and display the sequence

$$W_n = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^3 \quad \text{for } n = 1, 2, 3, \ldots,$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Comment on your picture, and show indeed that W_n converges in probability. Generalise your finding.

3. The continuity lemma for convergence in probability

There are actually two 'continuity lemmas' for convergence in probability.

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

(b) Suppose more generally that $X_n \to_{\rm pr} X$, 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 $g(X_n) \to_{\rm pr} g(X)$.

Comments: (i) To prove (b), use uniform continuity over closed and bounded intervals. (ii) In situations of relevance for this course, part (a) will be the more important. The typical application may be that consistency of $\widehat{\theta}_n$ for θ implies consistency of $g(\widehat{\theta}_n)$ for $g(\theta)$.

4. The maximum of uniforms

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

(a) Show that $M_n \to_{\rm pr} \theta$ (i.e. the maximum observation is a consistent estimator of the unknown endpoint).

(b) Find the limit distribution of $V_n = n(\theta - M_n)$, and use this result to find an approximate 95 percent confidence interval for θ.

5. Distribution functions

For a real random variable X, consider its distribution function $F(t) = \Pr\{X \le t\}$. Show that F is right continuous, and that its set of discontinuities is at most countable (in particular, the set of continuity points is dense). Show also that $F(t) \to 1$ when $t \to \infty$ whereas $F(t) \to 0$ when $t \to -\infty$.

6. A 'master theorem' for convergence in distribution

[xx check Ferguson's definition. xx] Let X_n and X be real random variables, with probability distributions P_n and P [so that $P_n(A) = \Pr\{X_n \in A\}$, etc.], and consider the following five statements:

(1) $X_n \to_d X$;

(2) for every open set A, $\liminf P_n(A) \ge P(A)$;

(3) for every closed set B, $\limsup P_n(B) \le P(B)$;

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

(5) for every bounded and continuous g, $\lim \mathrm{E}\, g(X_n) = \mathrm{E}\, g(X)$.

Show that these five statements are in fact all equivalent. Hints: For (1) implies (2), write $A = \cup_{j=1}^{\infty} A_j$ for open sets $A_j = (a_j, b_j)$, where a_j and b_j can be chosen to be among the continuity points for the distribution function F for X. Then show that (2) implies (3) [using that B is closed

if and only if B^c is open], and that (3) implies (4). For (4) implying (5), take g to have its values inside $[0,1]$, without loss of generality, and write

$$\mathrm{E}\, g(X_n) = \int \int_0^1 I\{y \le g(x)\}\, \mathrm{d}y\, \mathrm{d}P_n(x) = \int_0^1 \Pr\{g(X_n) \ge y\}\, \mathrm{d}y,$$

along with a Lebesgue theorem for convergence of integrals. Finally, for (5) implies (1), construct for given F-continuity point x a continuous function g_ε that is close to $g_0(y) = I\{y \le x\}$.

7. The continuity lemma for convergence in convergence

Suppose $X_n \to_d X$ and that h is continuous (and not necessarily bounded). Show that $h(X_n) \to_d h(X)$. [Use e.g. statement (5) of the previous exercise.] Thus $\exp(tX_n) \to_d \exp(tX)$, etc.

8. Convergence in distribution for discrete variables

Let X_n and X take on values in the set of natural numbers, and let $p_n(j) = \Pr\{X_n = j\}$ and $p(j) = \Pr\{X = j\}$ for $j = 0, 1, 2, \ldots$. Show that $X_n \to_d X$ if and only if $p_n(j) \to p(j)$ for each j. To illustrate this, prove the classic 'law of small numbers' (first proven by Ladislaus Bortkiewicz in 1898), that a binomial is close to a Poisson, if the count number is high and the probability is small.

9. Convergence in probability in dimension two (and more)

We have defined $X_n \to_{\mathrm{pr}} X$ to mean that

$$\Pr\{|X_n - X| \ge \varepsilon\} \to 0 \quad \text{for each } \varepsilon > 0.$$

The natural generalisation for the two-dimensional (and higher) case is to say that

$$X_n = (X_{n,1}, X_{n,2}) \to_{\mathrm{pr}} X = (X_1, X_2)$$

provided

$$\Pr\{\|X_n - X\| \ge \varepsilon\} \to 0 \quad \text{for each } \varepsilon > 0,$$

where $\|X_n - X\|$ is the usual Euclidean distance. Prove that $X_n \to_{\mathrm{pr}} X$ (in such a two-dimensional situation) if and only if $X_{n,j} \to_{\mathrm{pr}} X_j$ for $j = 1, 2$ (i.e. ordinary one-dimensional convergence for each component). Generalise.

10. Moment generating functions and convergence in distribution

For a random variable X, its moment generating function (mgf) is

$$M(t) = \mathrm{E} \exp(tX),$$

defined for each t at which the expectation exists. Among its basic properties are the following; attempt to demonstrate these.

1. $M(0) = 1$, and when the mean is finite, then $M'(t)$ exists, with $M'(0) = \mathrm{E}\, X$.
2. More generally, if $|X|^r$ has finite mean, then $M^{(r)}(0) = \mathrm{E}\, X^r$ (the rth derivative of M, at the point zero).

3. 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.

4. If X and Y are two variables with identical mgfs, then their distributions are identical. [There are also 'inversion formulae' in the literature, giving the distribution as a function of M.]

5. If X_n and X have mgfs M_n and M, then $M_n(t) \to M(t)$ for all t in a neighbourhood around zero is sufficient for $X_n \to_d X$.

6. In particular, if $M_n(t) \to \exp(\frac{1}{2}t^2)$ for all t close to zero, then $X_n \to_d \mathrm{N}(0,1)$.

11. Finite moments

Show that if $\mathrm{E}\, X^2$ is finite, then necessarily $\mathrm{E}\, X$ is finite too. Show more generally that $\mathrm{E}\,|X|^q$ is finite, then also $\mathrm{E}\,|X|^p$ is finite for all $p < q$. Prove indeed that $(\mathrm{E}\,|X|^p)^{1/p}$ is a non-decreasing function of p.

12. Proving the CLT (under some restrictions)

Let $X_1, X_2, \ldots$ be i.i.d. with some distribution F having finite variance and mean, and assume for simplicity that the mean is zero.

(a) Show that if the mgf exists, in a neighbourhood around zero, then

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

where σ is the standard deviation of X_i.

(b) Show that $\sqrt{n}\bar{X}_n$ has mgf of the form

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

and conclude that the CLT holds.

13. Characteristic functions

The trouble with the approach to the CLT above is that is has somewhat limited scope, in that some distributions do not have a finite mgf (since $\exp(tX)$ may be too big with too high probability for its mean to be finite). The so-called characteristic functions (chf) provide a more elegant mathematical tool in this regard. For a random variable X, its chf is defined as

$$\phi(t) = \mathrm{E}\, \exp(itX) = \mathrm{E}\, \cos(tX) + i\, \mathrm{E}\, \sin(tX),$$

with $i = \sqrt{-1}$ the complex unit, and $t \in R$.

(a) Show that the chf always exists, and that is is uniformly continuous. Show that the chf for the $\mathrm{N}(0, \sigma^2)$ is $\exp(-\frac{1}{2}\sigma^2 t^2)$.

(b) Assume $X_n \to_d X$. Show that

$$\phi_n(t) = \mathrm{E}\,\exp(itX_n) \to \phi(t) = \mathrm{E}\,\exp(itX) \quad \text{for all } t.$$

(c) The converse is also true (but harder to prove), and it is 'inside the curriculum' to know this: If

$$\phi_n(t) = \mathrm{E}\,\exp(itX_n) \quad \text{converges to some function} \quad \phi(t)$$

for all t in an interval around zero, and this limit function is continuous there, then (i) $\phi(t)$ is necessarily the chf of some random variable X, and (ii) there is convergence in distribution $X_n \to_d X$.

14. When is the sum of Bernoulli variables close to a normal?

Let $X_1, X_2, \ldots$ be independent Bernoulli variables (i.e. taking values 0 and 1 only), with $X_i \sim \mathrm{Bin}(1, p_i)$. We shall investigate when

$$Z_n = \frac{\sum_{i=1}^n (X_i - p_i)}{B_n} \to_d \mathrm{N}(0,1),$$

where $B_n = \{\sum_{i=1}^n p_i(1-p_i)\}^{1/2}$. Show, using mgfs or chfs, that this happens if and only $\sum_{i=1}^\infty p_i = \infty$ – and show, additionally, that this condition is equivalent to $B_n \to \infty$. Thus the cases $p_i = 1/i$ and $p_i = 1/i^2$, for example, are fundamentally different. For this second case, investigate the limit distribution of Z_n (which by the arguments given is not normal).

15. Proving the CLT (again)

Using chfs instead of mgfs gives a more elegant and unified proof of the CLT.

(a) Show that if X has a finite mean ξ, then its chf satisfies

$$\phi(t) = 1 + i\xi t + o(t) \quad \text{for } t \to 0.$$

Also, its derivative exists, and $\phi'(0) = \xi$.

(b) Show similarly that if X has a finite variance σ^2, then

$$\phi(t) = 1 + i\xi t - \tfrac{1}{2}(\xi^2 + \sigma^2 t^2) + o(t^2) \quad \text{for } t \to 0.$$

(c) If $X_1, X_2, \ldots$ are i.i.d. with mean zero and finite variance σ^2, then show that $Z_n = \sqrt{n}\bar{X}_n$ has chf of the form

$$\phi_n(t) = \{1 - \tfrac{1}{2}\sigma^2 t^2/n + o(1/n)\}^n.$$

Prove the CLT from this.

16. More on characteristic functions

Here are some more details and illustrations pertaining to characteristic functions.

(a) Find the characteristic function for a binomial distribution and for a Poisson distribution.

(b) Demonstrate the classical 'Gesetz der kleinen Zahlen' (cf. Exercise 8), that a binomial (n, p_n) tends to the Poisson λ, when $np_n \to \lambda$.

(c) Show that for the Cauchy distribution, with density $f(x) = (1/\pi)(1+x^2)^{-1}$, the chf is equal to $\exp(-|t|)$. Note that this function does not have a derivative at zero, corresponding to the fact that the Cauchy does not have a finite mean (cf. Exercise 15(a)).

(d) Let $X_1, \ldots, X_n$ be i.i.d. from the Cauchy. Show that the chf of $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ is identical to the chf of a single observation. Conclude, by the 'inversion theorem', the amazing fact that $\bar{X}_n =_d X_i$; the average has the same statistical distribution as each single component.

(e) There are several versions of 'inverse theorems', providing a mechanism for finding the distribution of a random variable from its chf; the perhaps primary aspect, defined as an 'inside curriculum fact', is that the chf indeed fully characterises the distribution (if X and Y have identical chfs, then their distributions are identical too). One such inversion formula is as follows: if X has a chf ϕ that is integrable (i.e. $\int |\phi(t)|\,\mathrm{d}t$ is finite), then X has a density f, for which a formula is

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

Write down what this means, in the cases of a normal and a Cauchy, and verify the implied formulae. Show that f in each such case of an integrable $\phi(t)$ necessarily becomes continuous.

(f) Show that the chf for the uniform $[-1,1]$ distribution becomes $\phi(t) = (\sin t)/t$. Deduce that

$$\int \Big|\frac{\sin t}{t}\Big|\,\mathrm{d}t = \infty \quad \text{even though} \quad \int \frac{\sin t}{t}\,\mathrm{d}t = \pi.$$

(g) Point (e) above gives a formula for the density f of a variable, in the case of it having an integrable chf ϕ. One also needs a more general formula, for the case of variables that do not have densities, etc. Let X be any random variable, with cumulative distribution function F and chf ϕ (but with nothing assumed about it having a density), and add on to it a little bit of Gaußian noise:

$$Z_\sigma = X + Y_\sigma, \quad \text{with } Y \sim \mathrm{N}(0,\sigma^2).$$

Then Z 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 Z_σ has cdf of the form

$$F_\sigma(x) = \int F(x-y)\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp(-\tfrac{1}{2}y^2/\sigma^2)\,\mathrm{d}y$$

and chf equal to

$$\phi_\sigma(t) = \phi(t)\exp(-\tfrac{1}{2}\sigma^2 t^2).$$

Hence show that

$$f_\sigma(x) = \frac{1}{2\pi}\int \exp(-itx)\phi(t)\exp(-\tfrac{1}{2}\sigma^2t^2)\,\mathrm{d}t.$$

and that, consequently,

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

(h) Conclude with the following general inversion formula, valid for all continuity points a, b of F:

$$F(b) - F(a) = \lim_{\sigma \to 0} \frac{1}{2\pi} \int \frac{\exp(-itb) - \exp(-ita)}{-it} \phi(t) \exp(-\tfrac{1}{2}\sigma^2 t^2)\, \mathrm{d}t.$$

17. Scheffé's Lemma

There are situations where $g_n(y) \to g(y)$ for all y, for appropriate functions g_n and g, does not imply $\int g_n(y)\,\mathrm{d}y \to \int g(y)\,\mathrm{d}y$. However, it may be shown that this is not a problem when g_n and g are probability densities (due to certain 'dominated convergence' Lebesgue theorems from the theory of measure and integration): if g_n and g are the densities of Y_n and Y, and $g_n(y) \to g(y)$ for (almost) all y, then

$$\int |g_n - g|\,\mathrm{d}y \to 0,$$

and, in particular,

$$\Pr\{Y_n \in [a,b]\} = \int_a^b g_n(y)\,\mathrm{d}y \to \int_a^b g(y)\,\mathrm{d}y = \Pr\{Y \in [a,b]\}$$

for all intervals, and we have $Y_n \to_d Y$. This is Scheffé's Lemma, defined as an inside curriculum fact.

(a) Let $Y_n \sim t_n$, a t distribution with n degrees of freedom. Show that $Y_n \to_d \mathrm{N}(0,1)$, using this lemma. Can you prove this statement in a simpler fashion?

(b) If $X_1, \ldots, X_n$ are i.i.d. from a uniform on $[0,1]$, with $M_n = \max_{i \le n} X_i$, show using the Scheffé Lemma that $n(1 - M_n)$ tends to a unit exponential in distribution.

(c) Suppose $X_n \sim \chi^2_n$, and consider $Z_n = (X_n - n)/\sqrt{2n}$. Prove that $Z_n \to_d \mathrm{N}(0,1)$.

18. The median

'The median isn't the message', said Stephen Jay Gould (when he was diagnosed with a serious illness and looked at survival statistics). Let $X_1, \ldots, X_n$ 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

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

which by the Scheff'e Lemma means that

$$\sqrt{n}(M_n - \theta) \to_d \mathrm{N}(0, \tau^2) \quad \text{with } \tau = \tfrac{1}{2}/f(\theta).$$

Why does this also prove that the sample median is consistent for the population median?

(c) Generalise to the following quantilian result: if $Q_n(p) = F_n^{-1}(p)$ is the pth quantile of the data, then $Q_n(p)$ converges in probability to the corresponding population quantile $\xi_p = F^{-1}(p)$, and

$$\sqrt{n}\{Q_n(p) - \xi_p\} \to_d \mathrm{N}(0, \tau_p^2) \quad \text{with } \tau_p^2 = p(1-p)/f(\xi_p)^2.$$

(d) Constructing a nonparametric confidence interval for an unknown median is not that simple – the 'usual recipe' works, up to a point, and tells us that if we first find a consistent estimator $\widehat{\kappa}$ of the doubly unknown quantity $f(\theta)$ (f is unknown, and so is θ, its median), then we're in business. We would then have

$$Z_n = \frac{\sqrt{n}(M_n - \theta)}{\widehat{\tau}} \to_d \mathrm{N}(0,1), \quad \text{with } \widehat{\tau} = \tfrac{1}{2}/\widehat{\kappa},$$

from which it then follows that

$$I_n = \widehat{\theta} \pm 1.96\,\widehat{\tau}/\sqrt{n} \quad \text{obeys} \quad \Pr\{\theta \in I_n\} \to 0.95.$$

The trouble lies in finding a satisfactory $\widehat{\kappa}$. Try to construct such a consistent estimator.

19. Limiting local power games

This exercise is meant to study a 'prototype situation' in some detail; the type of calculation and results will be seen to rather similar in a long range of different situations. – Let $X_1, \ldots, X_n$ 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

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i \quad \text{and} \quad M_m = \text{median}(X_1, \ldots, X_n).$$

(a) For given value of θ, prove that

$$\begin{aligned} \sqrt{n}(\bar{X}_n - \theta) &\to_d \mathrm{N}(0, \sigma^2), \\ \sqrt{n}(M_n - \theta) &\to_d \mathrm{N}(0, (\pi/2)\sigma^2). \end{aligned}$$

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

$$\begin{aligned} Z_n &= \sqrt{n}(\bar{X}_n - \sigma_0)/\widehat{\sigma} \to_d \mathrm{N}(0,1), \\ Z_n^* &= \sqrt{n}(M_n - \theta_0)/\{(\pi/2)^{1/2}\widehat{\sigma}\} \to_d \mathrm{N}(0,1), \end{aligned}$$

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

– [xx Figure 1: Limiting local power functions for two tests for $\theta \le \theta_0$ 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). xx]

(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;

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

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

(d) 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

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

[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

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

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 $\beta = 0.05$ and $\theta_1 = \theta_0 + \frac{1}{2}\sigma$, when also $\alpha = 0.05$.

(g) Lehmann defines 'the ARE [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.

20. Testing the normal scale

We have essentially covered Exercise 19 in class [xx alter this xx], as a 'prototype illustration' of the themes developed in Chapter 3 [xx change this xx]. Here is another illustration, for you to check that you may develop similar results in a different situation. Data $X_1, \ldots, X_n$ are now taken to be i.i.d. $\mathrm{N}(0, \sigma^2)$, and the object is to construct and compare tests for $H_0 : \sigma = \sigma_0$ vs. $\sigma > \sigma_0$, where σ_0 is some known quantity.

(a) Show that $\mathrm{E}\, X_i^2 = \sigma^2$ and $\mathrm{E}\,|X_i| = b\sigma$, with $b = \sqrt{2/\pi}$. Show that the estimators

$$\widehat{\sigma}_A = \Big\{ n^{-1} \sum_{i=1}^n X_i^2 \Big\}^{1/2} \quad \text{and} \quad \widehat{\sigma}_B = n^{-1} \sum_{i=1}^n |X_i|/b$$

are both consistent for σ.

(b) Find the limit distributions for

$$Z_{n,A} = \sqrt{n}(\widehat{\sigma}_A - \sigma) \quad \text{and} \quad Z_{n,B} = \sqrt{n}(\widehat{\sigma}_B - \sigma),$$

and comment on your findings.

(c) Construct explicit tests A and B, based on respectively $\widehat{\sigma}_A$ and $\widehat{\sigma}_B$, that have asymptotic level $\alpha = 0.01$.

(d) Show that both tests are consistent.

(e) Then we need to compare the two tests in terms of local power. For alternatives of the type $\sigma = \sigma_0 + \delta/\sqrt{n}$, establish limit distributions of the type

$$\begin{aligned} \sqrt{n}(\widehat{\sigma}_A - \sigma_0) &\to_d \mathrm{N}(\delta, \tau_A^2 \sigma^2), \\ \sqrt{n}(\widehat{\sigma}_B - \sigma_0) &\to_d \mathrm{N}(\delta, \tau_B^2 \sigma^2), \end{aligned}$$

with certain values (that you should find) for τ_A and τ_B.

(f) Establish the limiting local power functions $\pi_A(\delta)$ and $\pi_B(\delta)$, and plot them in a diagram (cf. Figure xx of the previous exercise).

(g) Compute the required sample sizes n_A and n_B for tests A and B to achieve detection power 0.99 when the true state of affairs is $\sigma = 1.333\,\sigma_0$.

(h) Compute the ARE for test A w.r.t. test B, and comment.

(i) Could there be other tests for H_0 here that would outperform test A?

21. Algebras of sets

Let $\mathcal{X}$ be a non-empty set, and let $\mathcal{A}$ be a class of subsets of $\mathcal{X}$. We say that $\mathcal{A}$ is an *algebra* if (i) both $\mathcal{X}$ and the empty-set is in $\mathcal{A}$; (ii) each time A is in $\mathcal{A}$, then also its complement A^c is in $\mathcal{A}$; (iii) whem $A_1, \ldots, A_n$ are sets in $\mathcal{A}$, then also their union $\cup_{i=1}^n A_i$ is in $\mathcal{A}$. In other words: an algebra is closed with respect to the formation of complements and finite unions.

(a) Are you yourself closed with respect to compliments?

(b) What's the world's smallest algebra?

(c) Show that an algebra is also closed with respect to finite intersections.

(d) And show that $A - B = A \cap B^c$ is within the algebra if A and B are so.

(e) Construct an example of an algebra.

(f) What was Muhammad ibn Musa al-Khvarizmi [xx fix xx]?

22. Sigma-algebras of sets

A *sigma-algebra* is an algebra $\mathcal{A}$ which is also closed with respect to countably infinite formations of unions, that is, if $A_1, A_2, \ldots$ are in $\mathcal{A}$, then so is $\cup_{i=1}^{\infty} A_i$.

(a) Let $\mathcal{A}$ consist of all those subsets of $\mathcal{R}$, the real numbers, which are themselves either finite or have finite complements. Is $\mathcal{A}$ an algebra? A sigma-algebra?

(b) Show that a sigma-algebra is closed with respect to countably infinite intersection operations.

23. Inverse and direct images of functions

Let $f\colon \mathcal{X} \to \mathcal{Y}$ be an arbitrary function, from set $\mathcal{X}$ to set $\mathcal{Y}$. For subsets A of $\mathcal{X}$, define the *direct image* as $fA = f(A) = \{f(x)\colon x \in A\}$. And for subsets B of $\mathcal{Y}$, define the *inverse image* as $f^1 B = f^{-1}(B) = \{x\colon f(x) \in B\}$.

(a) Let $\{B_i\colon i \in I\}$ be a collection of subsets of $\mathcal{Y}$. Show that $f^{-1}(\cup_i B_i) = \cup_i f^{-1}(B_i)$.

(b) And that $f^{-1}(\cap_i B_i) = \cap_i f^{-1}(B_i)$.

(c) Then show $f^{-1}(\mathcal{Y} - B) = \mathcal{X} - f^{-1}(B)$.

(d) Show that $A \subset f^{-1} f(A)$ for all A.

(e) And that $B \supset f f^{-1} B$ for all B.

(f) For functions $f\colon \mathcal{X} \to \mathcal{Y}$ and $g\colon \mathcal{Y} \to \mathcal{Z}$, show that $(g \circ f)^{-1}(C) = f^{-1} g^{-1} C$.

24. Independence of complements

We say that $A_1, \ldots, A_n$ are independent if

$$P(A_{i_1} \cap \cdots \cap A_{i,m}) = P(A_{i_1}) \cdots P(A_{i_m})$$

for all subsets $\{i_1, \ldots, i_m\}$ of $\{1, \ldots, n\}$. Thus we demand quite a bit more than merely saying that $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$.

Show that if $A_1, \ldots, A_n$ are independent, then so are $A_1^c, \ldots, A_n^c$.

25. The Borel–Cantelli emma

Let $A_1, A_2, \ldots$ denote events with probabilities $P(A_1), P(A_2), \ldots$. We are interested in the event that infinitely many of these A_j occur, i.e.

$$A_{\text{i.o.}} = \cap_{i \geq 1} \cup_{j \geq i} A_j.$$

(a) Show that if $\sum_{i=1}^{\infty} P(A_i) < \infty$, then $P(A_{\text{i.o.}}) = 0$. In other words, it is certain that only a finite number of the A_i will occur.

(b) Show under the additional assumption that the A_j are independent, that the previous result holds in the 'if and only if' sense, i.e. that if $\sum_{i=1}^{\infty} P(A_i) = \infty$, then $P(A_{\text{i.o.}}) = 1$. In particular, under independence, the probability of $A_{\text{i.o.}}$ is either 0 or 1, there is no 'middle ground' possibility.

26. Does this happen infinitely often?

Let $X_1, X_2, \ldots$ be independent with the same Expo(1) distribution, i.e. with density e^{-x} for $x \geq 0$.

(a) Will $X_n > 10 + 0.99 \log n$ infinitely often ?

(b) Will $X_n > 10 + 1.00 \log n$ infinitely often?

(c) Will $X_n > 10 + 1.01 \log n$ infinitely often?

(d) Will $X_n > 10^{12} + \log n$ infinitely often?

27. Normal deviations

Let X be standard normal, and write as usual $\Phi(x)$ for its cumulative distribution function and $\phi(x)$ for its density.

(a) Show that $\Pr\{X > x\} = 1 - \Phi(x) \doteq \phi(x)/x$ for large x.

(b) Let $X_1, X_2, \ldots$ be independent standard normals. Pray, will $X_n > 0.000001\sqrt{n}$ for infinitely many n?

(c) Let $\bar{X}_n$ be the average of the first n of these observations. Show that $|\bar{X}_n| > \varepsilon$ for at most a finite number of n.

(d) If $X_1, X_2, \ldots$ are independent and $\mathrm{N}(\xi, 1)$, what is the probability that $\bar{X}_n$ converges to ξ?

28. If you are sure about infinitely many things

Show that the event $\cap_{n=1}^{\infty} B_n$ is certain (i.e. it takes place with probability 1) if and only if each of the B_n is certain. Construct an example to show that this is *not* the case for uncountably many certain events.

29. At msot countably many discontinuities

Let F be a one-dimensional cumulative distribution function, and let D be the set of its discontinuities. Show that D is either empty, finite, or countably infinite.

30. Borel sets in dimensions one and two

Let $\mathcal{B}$ be the Borel sets in $\mathcal{R}$; it is the smallest sigma-algebra containing all intervals. Define then

$$\mathcal{B} \times \mathcal{B} = \sigma(\mathcal{C}),$$

the smallest sigma-algebra containing all $A \times B$, with A and B in $\mathcal{B}$. (This is the usual definition of a product-sigma-algebra.) Define also

$$\mathcal{B}^2 = \sigma(\mathcal{O}),$$

where $\mathcal{O}$ is the set of all open sets in $\mathcal{R}^2$ (This is the usual definition of a Borel-sigma-algebra.) Show that, luckily & conveniently, $\mathcal{B} \times \mathcal{B} = \mathcal{B}^2$.

31. Measurability of coordinate functions

Let $f, g\colon (\Omega, \mathcal{A}) \to (\mathcal{R}, \mathcal{B})$ be two functions, and let $h\colon \Omega \to \mathcal{R}^2$ be given by

$$h(\omega) = (f(\omega), g(\omega)).$$

Show that h is measurable if & only if both f and g are measurable. Generalise.

32. Normal mixtures

Let first X and Y be independent, with X a standard normal and Y very discrete, $\Pr\{Y = y\} = \frac{1}{2}$ for $y \in \{-1, 1\}$. Note that a sum of a continuous and a discrete variable will have a continuous distribution. Find the density for $X + Y$. Find also its mean and variance.

Generalise to finite normal mixtures, which may be done in several ways, with one path as follows. Start with the density

$$f(x) = \sum_{j=1}^{k} p_j \phi_{\sigma_j}(x - \mu_j),$$

defined via the triples (p_j, μ_j, σ_j) for $j = 1, \ldots, k$. Here the p_j make up a probability vector, i.e. nonnegative with sum 1, and $\phi_\sigma(x-u) = \sigma^{-1}\phi(\sigma^{-1}(x-\mu))$ is the density of the normal (μ, σ). One may now view X, drawn from f, as the result of the two-stage operation where the index $J = j$ is drawn from $\{1, \ldots, k\}$ first, with $\Pr\{J = j\} = p_j$, and $X \,|\, j \sim \mathrm{N}(\mu_j, \sigma_j^2)$. Use this to find $\mathrm{E}\,(X \,|\, j)$ and $\mathrm{Var}\,(X \,|\, j)$, and then the unconditional mean and variance for X.

The class of finite normal mixtures is a large one, and even with say $k \leq 5$ components a broad range of shapes may be attained – play a bit with this on your computer, drawing $f(x)$ curves on your screen, by mixing in different input vectors of p_j, μ_j, σ_j.

Find also a formula for the skewness of f, i.e. $\gamma = \mathrm{E}\,\{(X - \mu)/\sigma\}^3$, in terms of the overall mean and standard deviation μ and σ.

33. The Markov inequality, and bounding tails

Sometimes one wishes to bound tail probabilities, say $\Pr\{X \geq a\} \leq B(a)$, and there are several ways in which to do this.

(a) Let X be a nonnegative random variable, and let $h(x)$ be a nonnegative and nondecreasing function for $x \geq 0$. Demonstrate Неравенство Маркова (Markov's inequality), that

$$\Pr\{X \geq a\} \leq \mathrm{E}\, h(X)/h(a).$$

(b) If X is a random variable with mean ξ, show that

$$\Pr\{|X - \xi| \geq \varepsilon\} \leq \frac{\mathrm{E}\,|X - \xi|^p}{\varepsilon^p} \quad \text{for each } p > 0.$$

For $p = 2$ we have the famous special case of Неравенство Чебышёва (Chebyshov's inequality, from about 1853).

(c) Let $X_1, X_2, \ldots$ be independent normals $\mathrm{N}(\xi, 1)$, so that $\bar{X}_n \sim \mathrm{N}(\xi, 1/n)$. Writing N for a standard normal, show that

$$\Pr\{|\bar{X}_n - \xi| \geq \varepsilon\} \leq \frac{n^{-p/2}\mathrm{E}\,|N|^p}{\varepsilon^p} \quad \text{for each } p > 0.$$

For $n = 100$ and $\varepsilon = 0.05$, compute the exact probability in question and track the right hand bound as a function of p. Which p gives the sharpest bound, in this case?

(d) Let X have moment generating function $M(t) = \mathrm{E}\exp(tX)$, assumed to be finite for at least $0 \le t \le t_0$. Show that

$$\Pr\{X \ge a\} \le \min_{0 \le t \le t_0} \exp(-ta)M(t).$$

(e) For the case of $\bar{X}_n \sim \mathrm{N}(\xi, 1/n)$ studied above, show that

$$\Pr\{\bar{X}_n - \xi \ge \varepsilon\} \le \exp(-\tfrac{1}{2}n\varepsilon^2).$$

Compare this bound with the one reached via Chebyshov above.

(f) Let $X_1, X_2, \ldots$ be i.i.d. from the χ^2_b distribution, with $\mathrm{E}\,\bar{X}_n = b$ and $\mathrm{Var}\,\bar{X}_n = 2b/n$. Show that with $\varepsilon > 0$ given, there will with probability 1 be only finitely many n with $\bar{X}_n \ge b + \varepsilon$.

(g) [xx invent another application here. xx]

34. Amor's arrows sometimes miss

[From Nils Exam ST 200 December 1989, Exercise 1(e).] Amor shoots her arrows infinitely many times. Her shots are independent of each other, and shot no. n is (X_n, Y_n), measured from origo, where X_n and Y_n are independent and standard normal. The distance from origo is hence $R_n = (X_n^2 + Y_n^2)^{1/2}$, the square-root of a χ^2_2. Show that its density becomes $f(r) = r\exp(-\frac{1}{2}r^2)$. So how often does she miss, and by how much? Find the probabilities for these three events: that $R_n \ge 0.99\sqrt{2\log n}$ infinitely often; that $R_n \ge 1.00\sqrt{2\log n}$ infinitely often; that $R_n \ge 1.01\sqrt{2\log n}$ infinitely often.

35. Twins and paradigm shifts

Let $X_1, X_2, X_3, \ldots$ be an infinite sequence of independent standard normals. Say that X_{i-1} and X_i are *twins* if $|X_i - X_{i-1}| \le c_i$, and that there is a *regime shift* if $|X_i - X_{i-1}| \ge d_i$. Such c_i and d_i will be specified below. Let A be the event that the sequence experiences infinitely many twins, and B the event that the history sees infinitely many regime shifts.

(a) Write up an exact formula for the expected number of twins in the course of the first $n = 10^{12}$ observations. Put up similarly a formula for the expected number of regime shifts over the same period.

(b) Find $P(A)$ for the cases $c_i = 1/i$ and $c_i = 1/i^2$.

(c) Find $P(B)$ for the cases $d_i = 2\sqrt{\log i}$ and $d_i = 2.001\sqrt{\log i}$.

(d) Construct a criterion, expressed in terms of the c_i and d_i, for the history to experience with probability 1 both infinitely many twins and infinitely many regime shifts. Here it many be convenient to first deal with the situations where $\inf_i c_i > 0$ and $\sup_i d_i < \infty$, and then focus on the cases where $c_i \to 0$ and $d_i \to \infty$.

36. Quickness of convergence of average to its mean

Assume that $X_1, X_2, \ldots$ is a sequence of i.i.d. variables with mean zero. Hence $\bar{X}_n$ will converge to 0 in probability, and even with probability 1, by the Law of Large Numbers. But *how fast* will $p_n(a) = \Pr\{\bar{X}_n \ge a\} \to 0$, for fixed $a > 0$?

(a) Assume $\mathrm{Var}\, X_i = \sigma^2$ is finite. Show that $p_n(a) \le \sigma^2/(na^2)$, hence speed of order $1/n$.

(b) Assume that also the fourth order moment is finite, $\mathrm{E}\, X_i^4 < \infty$. Show that $p_n(a) \le K\sigma^2/(n^2a^4)$, for a certain K, which gives speed of order $1/n^2$.

(c) Let us generalise: Assume that $\mathrm{E}\,|X_i|^p < \infty$, for a suitable $p \ge 2$. The central limit theorem says $\sqrt{n}\bar{X}_n/\sigma \to_d \mathrm{N}(0,1)$. One may show that

$$\mathrm{E}\,|\sqrt{n}\bar{X}_n/\sigma|^p \to \mathrm{E}\,|\mathrm{N}(0,1)|^p,$$

see e.g. von Bahr (1965). Show from this that

$$\mathrm{E}\,|\bar{X}_n|^p \le c_p n^{-p/2}\,\mathrm{E}\,|\mathrm{N}(0,1)|^p\,\sigma^p \quad \text{for all } n,$$

for a suitable constant c_p – and one may use $c_p = 1.001$ if 'for all n' is replaced by 'for all large enough n'.

(d) Show that $p_n(a) \le K_p/(n^{p/2}a^p)$ for a suitable constant K_p.

(e) Assume X_i has moments of all orders, such that (d) holds for each p. If you should succeed in proving that $p_n(a) \le 0.999999^n$, is this a sharper result?

(f) Assume that the moment generating function $M(t) = \mathrm{E}\,\exp(tX)$ exists for (at least) $0 \le t \le t_0$. Show that

$$p_n(a) \le \rho^n, \quad \text{where } \rho = \rho(a) = \min_{0\le c\le t_0} \frac{M(c)}{\exp(ac)},$$

and show that $\rho < 1$. (If $\rho = 1$ the result would still hold, but it would be a boring and rather unpublishable one.)

(g) Find $\rho = \rho(a)$ explicitly, when $X_i \sim \mathrm{N}(0,1)$, and when $X_i \sim \mathrm{N}(0,\sigma^2)$.

(h) It is practical to have explicit results also for $p_n(a) = \Pr\{\bar{X}_n \ge \xi + a\}$, of the type above, for the case of $\mathrm{E}\, X_i = \xi$. Establish such results.

(i) Find $\rho = \rho(a)$ explicitly for the cases (1) $X_i \sim \chi^2_m$; (2) $X_i \sim \mathrm{Bin}(1,p)$; and (3) $X_i \sim \mathrm{Pois}(\lambda)$.

37. The discrete and continuous parts of a cumulative distribution function

Let F be an arbitrary cumulative distribution function on $\mathcal{R}$. Show that one always may decompose F into $F = F_c + F_d$, where F_c is continuous and F_d is discrete.

38. A probabilistic excursion into number theory

In this exercise we shall construct certain types of probability distributions on the natural numbers, via placing probabilities on the the exponents in their prime number factorisations. This becomes an excursion into the world of number theory, to show some their results and formulae, but with the probabilist's hat and spectacles. Let $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $p_4 = 7$, $p_5 = 11$, etc., be the prime numbers.

(a) Find, like Gauß did when he was a little kid, all the prime numbers up tp 100. Gauß didn't stop there; as a 15 year old boy in 1792 he had essentially understood the fundamental prime number theorem $\pi(x) \doteq x/\log x$, where $\pi(x)$ is the number of primes below x, see point (xx) below. This was not formally proven until about 1896.

(b) Prove, as Euclid did about 2300 year ago, that there are infinitely many primes! (Later proofs of interest include those of Kummer, Pólya, Euler, Axel Thue, Perott, Auric, Métrod, Washington, and Fürstenberg. Even further proofs flow as corollaries of statements proved below, in points (g) and (k).)

(c) We do have $63 = 3^2{\cdot}7^1$, $104 = 2^3{\cdot}13^1$, $30\,141\,766 = 3^2{\cdot}5^1{\cdot}17^1{\cdot}31^2{\cdot}41$, $702\,958\,333 = 7^1{\cdot}11^4{\cdot}19^3$, right? Make it clear to you that each natural number n may be expressed in a unique prime factorisation fashion, in the form $n = p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$. Here m is the number of the highest prime in n, and $x_1, x_2, \cdots, x_m$ are the exponents. We may also write n as the infinite product $\prod_{j=1}^\infty p_j^{x_j}$, where all x_j from a certain $j_0 + 1$ onwards are equal to zero.

(d) This opens a probabilistic door for us, creating a random natural number N by expressing it as

$$N = p_1^{X_1} p_2^{X_2} \cdots = \prod_{j=1}^\infty p_j^{X_j},$$

where $X_1, X_2, \ldots$ are random variables in $\{0, 1, 2, \ldots\}$, with the property that only a finite number of them are above 1. Let us try: assume the X_j are independent. Show that N is then a well-defined random variable if and only if

$$\sum_{j=1}^\infty \Pr\{X_j \geq 1\} = \sum_{j=1}^\infty [1 - \Pr\{X_j = 0\}] < \infty.$$

The division here is sharp: if the sum diverges, then not only is $N = \infty$ with positive probability, but with probability 1.

(e) As a preliminary example, let the X_j be independent with $X_j \sim \text{Pois}(d_j)$. Show that N is well-defined if and only if $\sum_{j=1}^\infty d_j < \infty$. Find under this condition the expected values of N and $\log N$. Simulate say 10^4 such N, with $d_j = 1/i^{3/2}$.

(f) There's more beauty to be revealed for the case where the X_j are taken independent and geometrically distributed. Let $X_j \sim \text{Geo}(c_j)$, which means

$$\Pr\{X_j = x\} = (1 - c_j)^x c_j \quad \text{for } x = 0, 1, 2, \ldots.$$

Find the mean, the variance, and the generating function for X_j:

$$\text{E}\, X_j = \frac{1 - c_j}{c_j}, \quad \text{Var}\, X_j = \frac{1 - c_j}{c_j^2}, \quad \text{E}\, s^{X_j} = \frac{c_j}{1 - (1 - c_j)s}.$$

Show also that $\Pr\{X_j \geq x\} = (1 - c_j)^x$. Demonstrate that N is well-defined if and only if $\sum_{j=1}^\infty (1 - c_j) < \infty$.

(g) You recall $\sum_{n=1}^\infty 1/n^2 = \pi^2/6$, Euler's sensational finding from about 1734? Consider the choice $c_j = 1 - 1/p_j^2$. Find the probability that N is equal to $1, 11, 63, 103\,141\,766$. Show that

$$\Pr\{N = n\} = \frac{6}{\pi^2} \frac{1}{n^2} \quad \text{for } n = 1, 2, 3, \ldots. \tag{0.1}$$

Then you have also essentially deduced the following intriguing formula:

$$\frac{\pi^2}{6} = \prod_{j=1}^\infty \frac{p_j^2}{p_j^2 - 1} = \frac{4}{3}\frac{9}{8}\frac{25}{24}\frac{49}{48}\frac{121}{120} \cdots.$$

As a low-hanging fruit in this garden: *If* there had been merely a finite number of primes, then π^2 would have been rational. Hence (fill in!).

(h) Show also, conversely, that if N is given the (0.1) distribution, then by necessity this leads to independent X_j which are geometrically distributed with parameters $c_j = 1 - 1/p_j^2$.

(i) With this distribution for N, find the following probabilities:

(i) that N is odd [answer: $\frac{3}{4}$];

(ii) that N is a prime numbers;

(iii) that N is a a 'prime potens', of the form p^y, for some $y \geq 1$;

(iv) that N is a factor in 100;

(v) that 100 is a factor in N [answer: $1/100^2$];

(vi) that N turns out to be a square [answer: $\pi^2/15\,!$];

(vii) invent something yourself.

(j) Find the mean for N and for $\log N$. And their variances, unless your willpower is strong enough to resist.

(k) *Riemann's zeta function* is defined as $\zeta(\alpha) = \sum_{n=1}^{\infty} 1/n^\alpha$, for $\alpha > 1$. Thus $\zeta(2) = \pi^2/6$, $\zeta(4) = \pi^4/90$, $\zeta(6) = \pi^6/945$, etc. Agree to say that N is zeta distributed with parameter α provided

$$\Pr\{N = n\} = \frac{1}{\zeta(\alpha)} \frac{1}{n^\alpha} \quad \text{for } n = 1, 2, 3, \ldots.$$

Assume from this point (k) onwards, up to point (y) below, that N has this distribution. Show that this is equivalent to having the X_j independent and geometric, with $X_j \sim \text{Geo}(1-1/p_j^\alpha)$. Derive in particular the following intriguing representation for the zeta function:

$$\zeta(\alpha) = \prod_{\text{prime}} \frac{p^\alpha}{p^\alpha - 1} = \prod_{j=1}^{\infty} \frac{p_j^\alpha}{p_j^\alpha - 1}.$$

This formula was first derived by Euler. So now we know that

$$\frac{\pi^4}{90} = \frac{16}{15}\frac{81}{80}\frac{625}{624}\frac{2401}{2400}\cdots.$$

Show also that $\zeta(\alpha) \to \infty$ as $\alpha \to 1$, which would not have been true if God had given us only a finite number of prime numbers.

(l) Generalise the questions and solutions from point (i) to the more general situation with parameter α rather than 2. Replace also '100' with an arbitrary $n = p_1^{x_1} \cdots p_m^{x_m}$ for sub-points 4 and 5. [A few answers: (l1) $1-1/2^\alpha$; (l2) $\zeta(\alpha)^{-1} \sum_1^\infty 1/p_j^\alpha$; (l3) $\zeta(\alpha)^{-1} \sum_1^\infty 1/(p_j^\alpha - 1)$; (l4) $\Pr\{N \text{ is a factor in } n\} = \zeta(\alpha)^{-1} n^{-\alpha} \prod_{j=1}^m (1 + p_j^\alpha + \cdots + p_j^{\alpha x_j})$; (l5) $\Pr\{n \text{ is a factor in } N\} = 1/n^\alpha$; (l6) $\zeta(2\alpha)/\zeta(\alpha)$; (l7) go confidently in the direction of your dreams.]

(m) Say that the number n is *modest* if all prime exponents x_j for n are 0 or 1. Show us three modest and three immodest numbers. Show that the probability that N is modest is $\zeta(2\alpha)^{-1}$. Demonstrate also that

$$B(\alpha) = \sum_{n \text{ modest}} \frac{1}{n^\alpha} = \frac{\zeta(\alpha)}{\zeta(2\alpha)} = \prod_{p \text{ primtall}} \frac{p^\alpha + 1}{p^\alpha}.$$

(n) Say that n is *second-order modest* if all prime exponents are less than or equal to 2. Show that the probability that N is such a second-order modest number is $\zeta(3\alpha)^{-1}$.

(o) Show that the events $\{63 \text{ is a factor in } N\}$ and $\{100 \text{ is a factor in } N\}$ are independent, whereas $\{18 \text{ is a factor in } N\}$ and $\{52 \text{ is a factor in } N\}$ are dependent. Generalise – ask the right questions, and find the right answers.

(p) Show, by studying $\mathrm{E}\, N$ for $\alpha = 2$, that $\prod_{p \text{ prime}}(1 + 1/p) = \infty$, and deduce from this that $\sum_{p \text{ prime}} 1/p = \infty$. This was first proven by Euler.

(q) Let $M = \max\{j\colon X_j \geq 1\}$ be the last prime factor present in the random N. FInd the probability distribution of M, and show that it has expected value

$$\sum_{m=1}^{\infty}\Big[1 - \prod_{j=m}^{\infty}\Big(1 - \frac{1}{p_j^{\alpha}}\Big)\Big].$$

(r) Let f and g be functions defined on the natural numbers. Define the *Dirichlet convolution* or *Dirichlet product* $f * g$ by

$$(f * g)(n) = \sum_{d|n} f(d)g(n/d), \quad n \geq 1,$$

with the sum taken over those d in $\{1, \ldots, n\}$ which are factors in n. Show that

$$\sum_{n=1}^{\infty}\frac{f(n)}{n^{\alpha}}\sum_{n=1}^{\infty}\frac{g(n)}{n^{\alpha}} = \sum_{n=1}^{\infty}\frac{(f*g)(n)}{n^{\alpha}}, \quad \text{or} \quad \mathrm{E}\,(f*g)(N) = \zeta(\alpha)\,\mathrm{E}\, f(N)\,\mathrm{E}\, g(N),$$

if the two series converge.

(s) Let $\sigma(n)$ be the number of d in $\{1, \ldots, n\}$ which are factors in n. Show that $\sum_{n=1}^{\infty}\frac{\sigma(n)}{n^{\alpha}} = \zeta(\alpha)^2$; (i) by working with $\mathrm{E}\,\sigma(N)$, (ii) by Dirichlet convolution.

(t) Let $\phi(n)$ be the so-called *Euler totient function*, defined as the number of numbers in $\{1, \ldots, n\}$ which are reciprocally prime with n. It is an important tool in mathematical number theory. Show that $\phi(p) = p - 1$ if p is a prime; that more generally $\phi(p^x) = p^x - p^{x-1}$ if p is a prime; that the function is so-called multiplicative, which means that $\phi(mn) = \phi(m)\phi(n)$ for reciprocally primeish numbers; that $n = \sum_{d|n}\phi(d)$; that $(1 * \phi)(n) = n$; and that $\phi(n) = n\prod_{p|n}(1 - 1/p)$. Prove the formulae

$$\sum_{n=1}^{\infty}\frac{\phi(n)}{n^2} = \infty, \quad \sum_{n=1}^{\infty}\frac{\phi(n)}{n^{\alpha}} = \frac{\zeta(\alpha-1)}{\zeta(\alpha)};$$

(1) by working with $\mathrm{E}\,\phi(N)$, (2) by working with $\mathrm{E}\,\phi(N)/N$; (3) by using Dirichlet convolutions.

(u) Another number theoretic function of importance is the *Möbius function*, defined by $\mu(1) = 1$; $\mu(p_{j_1}\cdots p_{j_r}) = (-1)^r$ if the number is over distinct prime numbers; and $\mu(n) = 0$ for all other n. Show that $\mu(n) \neq 0$ only for the modest numbers studied in point (m). Prove the glamorous formula

$$\sum_{n=1}^{\infty}\frac{\mu(n)}{n^{\alpha}} = \frac{1}{\zeta(\alpha)}, \quad \text{or} \quad \sum_{n=1}^{\infty}\frac{1}{n^{\alpha}}\sum_{n=1}^{\infty}\frac{\mu(n)}{n^{\alpha}} \equiv 1,$$

by working with the mean of the random $\mu(N)$ in a couple of different ways. This point may also be solved by conditioning a zeta distribution on the event that the outcome is modest; check point ($\sqrt{\pi}$).

(v) It follows without too much efforts that $\lim_{\alpha\to 1}\sum_{n=1}^\infty \frac{\mu(n)}{n^\alpha} = 0$; mathematical finesse is however called for to really prove that $\sum_{n=1}^\infty \frac{\mu(n)}{n} = 0$. Attempt to come up with such finesse. Then attempt to attach The Fundamental Prime Number Theorem, which says that if $\pi(x)$ is the number of primes in $\{1, 2, \ldots, x\}$, then $\pi(x) \doteq x/\log x$. [One may prove that this implies and is implied by $\sum_{n=1}^\infty \frac{\mu(n)}{n} = 0$; see Amitsur's 'On arithmetic functions' in *Journal of Analytic Mathematics*, 1956.]

(w) Time has come to introduce the *von Mangholdt function*, defined by $\Lambda(n) = \log p$ for prime potens numbers $n = p^x$ for $x \geq 1$, and $\Lambda(n) = 0$ for all numbers not being prime potenses. Work with $\mathrm{E}\,\Lambda(N)$ and show that

$$\sum_{n=1}^\infty \frac{\Lambda(n)}{n^\alpha} = \sum_{p \text{ primtall}} \frac{\log p}{p^\alpha - 1};$$

(x) and show that

$$\sum_{p \text{ primtall}} \frac{\log p}{p^\alpha - 1} = \sum_{n=1}^\infty \frac{\log n}{n^\alpha} \Big/ \sum_{n=1}^\infty \frac{1}{n^\alpha} = \frac{-\zeta'(\alpha)}{\zeta(\alpha)},$$

by working with $\log N$. Prove also that $(1 * \Lambda)(n) = \log n$.

(y) Find a numerical value for B, the Viggo Brun constant. [Answer: 1.90216054 ...]

(z) Let N_1 and N_2 be independent and zeta distributed with the same parameter α. Find the distribution for the product $N_1 N_2$.

(æ) If $n_1, \ldots, n_k$ are given numbers, let $\gamma\{n_1, \ldots, n_k\}$ be their greatest common divisor; for instance, $\gamma\{20, 30\} = 10$ and $\gamma\{18, 24, 36\} = 6$. If N_1 and N_2 are independent and zeta distributed with parameters α_1 and α_2, show that $\gamma\{N_1, N_2\}$ becomes zeta distributed with parameter $\alpha_1 + \alpha_2$. Generalise.

(ø) Find also the probability distribution for $\lambda\{N_1, N_2\}$, the smallest common multiplum for N_1 and N_2, when $\alpha_1 = \alpha_2$. [The answer is more complicated than for $\gamma\{N_1, N_2\}$.]

(å) Back to semi-reality, or perhaps pseudo-reality, for a little while: The zeta distribution has been applied in certain linguistic studies; it has e.g. been tentatively shown that the frequency of words, in long text corpora, to a certain degree of accuracy follows a zeta distribution. Assume you read V words by Shakespeare, that V_1 words are seen only once, that V_2 words are seen precisely twice, etc. Then the relative frequencies V_n/V should be fitted to the zeta model's $\zeta(\alpha)^{-1}/n^\alpha$. Estimate α for a few of your favourite authors. Who has the lowest α, Anne-Catharine Vestly or Knud Pedersen Hamsun? – The zeta distribution is also partly like a discretised Pareto distribution, and will perhaps fit sufficiently well to distributions of income in different socio-economic groups. Try it out, for a group you know.

(ß) Assume $N_1, \ldots, N_k$ are independent numbers drawn from the zeta distribution with parameter α. Show that the geometric mean $(N_1 \cdots N_k)^{1/k}$ is sufficient and complete. Explain how you can find the maximum likelihood estimator.

(ó͡o) I have simulated 25 realisations from a zeta distribution, using a simple R programme, and found

```
2, 3, 3, 1, 8, 1, 1, 1, 3, 1, 12, 29,

1, 37, 10, 2, 5, 1, 1, 6, 10, 1, 4, 1, 6.
```

Only I know the value of α being used. Estimate this value, and give a confidence interval.

(ă) Show that the maximum likelihood estimator is strongly consistent, and find its limit distribution.

(ç̈) Show that every even number (except 2) can be expressed as a sum of two primes, e.g. by studying the behaviour of an analytic continuation of the zeta function near zero.

($\sqrt{\pi}$) Let us attempt another type of distributions for the X_j than the geometric ones. Let X_j be 0 or 1, with probabilities $1-a_j$ and a_j. Then N is accordingly a random modest number (see point (m)). Show that N is well-defined if and only of $\sum_{j=1}^\infty a_j < \infty$. Show that if a_j is taken to be $1/(p_j^\alpha+1)$, then $\Pr\{N=n\} = B(\alpha)^{-1}/n^\alpha$, for modest n. Show again that $B(\alpha) = \prod_{p \text{ prime}}(p^\alpha+1)/p^\alpha = \zeta(\alpha)/\zeta(2\alpha)$. Show that this model may be characterised as the conditional zeta distribution given that N is modest, and, alternatively, as the conditional zeta distribution given that all the geometric X_j are in $\{0,1\}$. Do a little formula excursion by finding expressions for natural quantities in two ways; in one way, working with the N distribution directly, in another way, using the X_j distributions. You may e.g. impress yourself by showing

$$\sum_{n \text{ modest}} \frac{\log n}{n^\alpha} = \frac{\zeta(\alpha)}{\zeta(2\alpha)} \sum_{p \text{ primtall}} \frac{\log p}{p^\alpha+1},$$

and your surroundings by proving

$$\Pr\Big\{\sum_{j=1}^\infty \text{Bin}\{1, 1/(1+p_j^2)\} \text{ becomes even}\Big\} = 0.70.$$

[Consider $\mathrm{E}\,\mu(N)$.]

(ó͡ı) Then try out Poisson distributed prime number exponents. Say that N is Poisson prime number exponentially distributed with parameters $\{d_1, d_2, d_3, \ldots\}$ provided $X_j \sim \text{Pois}(d_j)$, where these are still independent. Let in particular $d_j = d/p_j^\alpha$, and show that

$$\Pr\{N=n\} = e^{-dA(\alpha)} \frac{d^{s(n)}}{n^\alpha g(n)}, \quad n = 1,2,3,\ldots,$$

where $s(n) = \sum_{j=1}^m x_j$ and $g(n) = \prod_{j=1}^m x_j!$, for given n with factorisation as in (c), and where $A(\alpha) = \sum_{p \text{ primtall}} 1/p^\alpha$. Show, for example, that

$$\sum_{n=1}^\infty \frac{1}{n^\alpha}\frac{1}{g(n)} = \exp\{A(\alpha)\}, \quad \sum_{n=1}^\infty \frac{\log n}{n^\alpha g(n)} = \exp\{A(\alpha)\} \sum_{p \text{ primtall}} \frac{\log p}{p^\alpha}.$$

Show that the probability of having a prime number for N is $A(\alpha)\exp\{-A(\alpha)\}$ when $d_j = 1/p_j^\alpha$. Find some further formulae in the flow created. Show that products of independent Poisson prime number exponentially distributed variables stay being Poisson prime number exponentially distributed. Find a sufficient and complete statistic based on $N_1, \ldots, N_k$ when d and α are unknown parameters. Study the large-sample properties of the maximum likelihood estimators.

(γ) We know that $\prod_p p^2/(p^2-1) = \pi^2/6$, but what is $\prod_p p^2/(p^2-0.99)$? – Allow me to show you my *generalised zeta function*:

$$\zeta_d(\alpha) = \sum_{n=1}^{\infty} \frac{d^{s(n)}}{n^\alpha}, \quad 0 < d \le 2,\ \alpha > 1,$$

where $s(n) = x_1 + x_2 + \cdots$ is the *extravaganza* for the number n. Show taht this de facto exists for $0 < d \le 2$ and $\alpha > 1$. Give probabilistic proofs for the following formulae, which all reduce to previous results when d is set equal to 1:

$$\zeta_d(\alpha) = \prod_{p \text{ primtall}} \frac{p^\alpha}{p^\alpha - d},$$

$$\sum_{n=1}^{\infty} \frac{d^{s(n)}\mu(n)}{n^\alpha} \sum_{n=1}^{\infty} \frac{d^{s(n)}}{n^\alpha} \equiv 1,$$

$$\sum_{n=1}^{\infty} \frac{d^{s(n)}\sigma(n)}{n^\alpha} = \zeta_d(\alpha)^2,$$

$$\sum_{n \text{ beskjeden}} \frac{d^{s(n)}}{n^\alpha} = \prod_{p \text{ primtall}} \frac{p^\alpha + d}{p^\alpha} = \frac{\zeta_d(\alpha)}{\zeta_{d^2}(2\alpha)},$$

$$\sum_{n=1}^{\infty} \frac{d^{s(n)}\phi(n)}{n^\alpha} = \frac{\zeta_d(\alpha-1)}{\zeta_d(\alpha)},$$

$$\sum_{n=1}^{\infty} \frac{d^{s(n)}f(n)}{n^\alpha} \sum_{n=1}^{\infty} \frac{d^{s(n)}h(n)}{n^\alpha} = \sum_{n=1}^{\infty} \frac{d^{s(n)}(f*h)(n)}{n^\alpha},$$

$$\sum_{n=1}^{\infty} \frac{d^{s(n)}\log n}{n^\alpha} = \zeta_d(\alpha) \sum_{n=1}^{\infty} \frac{d^{s(n)}\Lambda(n)}{n^\alpha},$$

$$\Pr\Big\{\sum_{j=1}^{\infty} \text{Bin}\{1, d/(p_j^\alpha + d)\} \text{ becomes even}\Big\} = \frac{1}{2} + \frac{1}{2}\frac{\zeta_{d^2}(2\alpha)}{\zeta_d(\alpha)^2}.$$

Employ as probabilistical tools (1) $X_j \sim \text{Poisson}(d/p_j^\alpha)$; (2) $X_j \sim \text{Bin}\{1, d/(p_j^\alpha + d)\}$; (3) $X_j \sim \text{Geo}(1 - d/p_j^\alpha)$. Discuss relations between these models.

(œ) Investigate consequences for the distribution of primes among the natural numbers, from $\sum_{n=1}^{\infty} d^{s(n)}\mu(n)/n = 0$; as mentioned this statement, for the special case of $d = 1$, implies the glorious prime number distribution theorem.

(α) Put a probability distribution on the modest numbers by taking the X_j to form a time inhomogeneous Markov chain on $\{0, 1\}$. Grei ut.

(ω) Find out a wholde deal on how the prime numbers and their cousins are distributed among the natural numbers, by studying distributions of the type $\mathcal{D}\{N | N \le n_0\}$, where n_0 is big, and by moving this threshold for the α parameter to the left of 1. Meld fra hvor du går.

39. Quartile and quantile differences

One way of assessing the spread of a distribution F, based on data $X_1, \ldots, X_n$, is via the *quartile difference* $Q_3 - Q_1$, the difference between the upper and lower quartiles. Often this difference is

multiplied with a well chosen constant, such that the resulting spread estimate becomes approximately unbiased for the the standard deviation parameter in the case of F being normal.

What is this constant? How clever is this estimator, compared with the usual one under normal conditions? Which cons and pres does the estimator have, compared to others? How do yet other naturally generalised competitors behave, where one uses upper and lower ε quantile, instead of upper and lower 25 percent quantiles? Which of these is best, on Gauß's home turf?

(a) Attempt to make your own exam type exercise, containing progressively more detailed questions, based on the above sentences.

(b) Define $Q_3 = X_{[0.75\,n]}$ and $Q_3 = X_{[0.25\,n]}$, where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics. Speculate a little regarding suitable interpolation tricks to make them better.

(c) For a few of the points below we shall take F to be the normal $\mathrm{N}(\xi, \sigma^2)$. Assume for this point only that F is strictly increasing with a continuous density f. Show that $Q_3 - Q_1$ converges almost surely to $q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25)$. With which constant do we need to multiply $Q_3 - Q_1$ in order to get a consistent estimator of σ, in the case where F is a normal?

(d) Show that

$$\begin{pmatrix} \sqrt{n}(Q_1 - q_1) \\ \sqrt{n}(Q_3 - q_3) \end{pmatrix} \to_d \begin{pmatrix} (F^{-1})'(0.25)\, U \\ (F^{-1})'(0.75)\, V \end{pmatrix},$$

where

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3/16 & 1/16 \\ 1/16 & 3/16 \end{pmatrix} \Big).$$

(e) Let $z(\varepsilon) = \Phi^{-1}(1-\varepsilon)$ be the upper ε quantile for the standard normal, and let

$$\widetilde{\sigma} = \frac{Q_3 - Q_1}{2z(0.25)} = \frac{Q_3 - Q_1}{1.349}.$$

Show that $\sqrt{n}(\widetilde{\sigma} - \sigma)$ tends to $\mathrm{N}(0, \kappa^2)$, with $\kappa = 1.1664\,\sigma$.

(f) Here it is natural to compare with the traditional estimator $\widehat{\sigma}$, the empirical standard deviation. Show (which is more standard, right?) that $\sqrt{n}(\widehat{\sigma} - \sigma) \to_d \mathrm{N}(0, (0.7071\,\sigma)^2)$.

(g) Then generalise! That is, consider

$$\widetilde{\sigma}(\varepsilon) = \frac{X_{[(1-\varepsilon)n]} - X_{[\varepsilon n]}}{2z(\varepsilon)} = \frac{F_n^{-1}(1-\varepsilon) - F_n(\varepsilon)}{2z(\varepsilon)},$$

where F_n is the empirical cumulative distribution function, and find the limit distribution for $\sqrt{n}(\widetilde{\sigma} - \sigma)$ under normal conditions. The answer should becomes $\mathrm{N}(0, \kappa(\varepsilon)^2)$, where

$$\kappa(\varepsilon) = \frac{\sqrt{2\pi}}{2\varepsilon} \sqrt{2\varepsilon(1-\varepsilon)} \exp\{\tfrac{1}{2} z(\varepsilon)^2\}\, \sigma.$$

(h) Investigate how the precision of $\widetilde{\sigma}(\varepsilon)$ changes when ε varies between 0 and $\frac{1}{2}$. Show in particular that the asymptotically speaking very best estimator of this type, under normality, is

$$\sigma^* = \frac{F_n^{-1}(0.931) - F_n^{-1}(0.069)}{2.9666},$$

with limit distribution $\mathrm{N}(0, (0.8755\,\sigma)^2)$, a loss of 1.2382 compared with the optimal value $\sigma/\sqrt{2}$.

(i) Investigate the behaviour of such estimators outside normality.

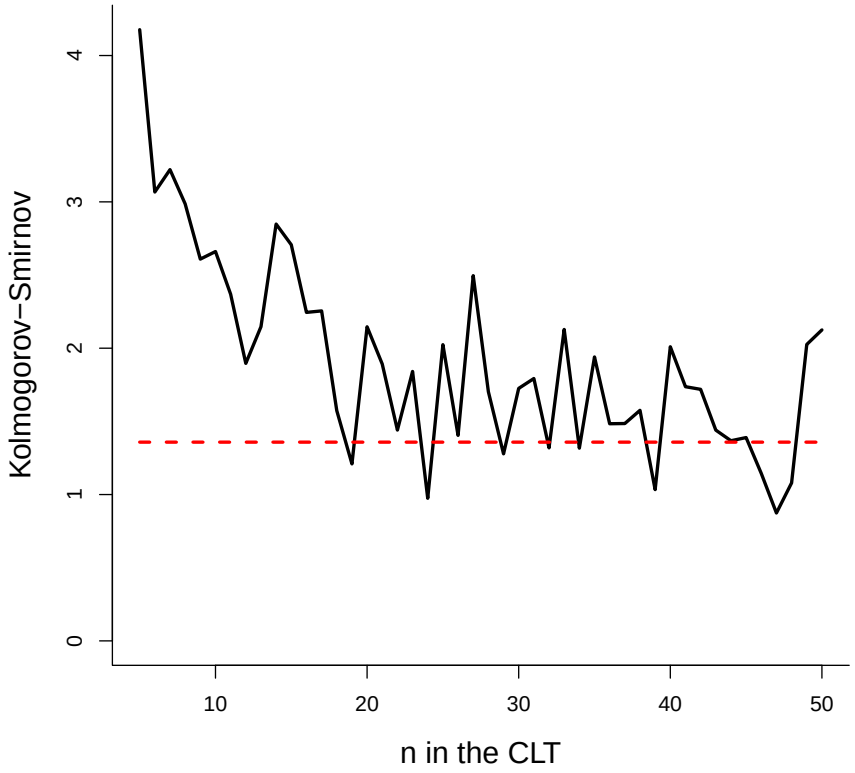

Figure 0.1: For each n, from 5 to 50, I have simulated $\mathrm{sim} = 10^4$ realisations of Z_n of Exercise 41, and then computed the Kolmogorov–Smirnov test statistic $D_{\mathrm{sim}} = \mathrm{sim}^{1/2} \max_t |F_{\mathrm{sim}}(t) - \Phi(t)|$ to check whether the Z_n distribution is close to the limiting standard normal. The red horizontal line is at 1.358, the 0.95 point of the null distribution.

40. Checking out the CLT

This is a cousin exercise to Exercise 1, using simulation to check whether the variable

$$Z_n = (X_1 + \cdots + X_n - n\mu)/(\sqrt{n}\sigma) = \sqrt{n}(\bar{X}_n - \mu)/\sigma$$

has a distribution decently close to the limiting standard normal, nor not. This is a function of both the underlying distribution and the size of n, of course. One learned in Exercise 1 that if the start distribution of a single X_i is the uniform, then the histograms of say 10^4 realisations of Z_n succeed in getting pretty close to the normal, for pretty low n. This might be classified as 'disappointing' or 'encouraging', avhengig av dagsformen – at any rate, a key reason why this happens is that the start distribution is symmetric.

To investigate different scenarios, with skewness on board, and where convergence towards limiting normality is decidedly slower, let's make the Beta distribution the start distribution, with

parameters $(a, b) = (1, 5)$. Display the density of this distribution; use the formulae

$$\mathrm{E}\, X = \xi = \frac{a}{a+b} \quad \text{and} \quad \mathrm{Var}\, X = \frac{\xi(1-\xi)}{a+b+1}$$

ti find the mean and standard deviation, and compute the skewness $\gamma_3 = \mathrm{E}\,(X-\xi)^3/\sigma^3$. Show also that

$$\mathrm{skew}(Z_n) = \gamma_3/\sqrt{n}.$$

Your task is now to simulate $\mathrm{sim} = 10^4$ realisations of the variable Z_n, for say $n = 5, 6, \ldots, 50$. For each such n, you might check the corresponding histogram, and observe how these become steadily 'more normal'; you may also use `plot(density(zz))` to look at the empirical densities based on the sim realisations. Also, for each such simulated dataset of Z_n, carry out two tests for standard normality, in order to see how 'far off' from the limit one might still be. These tests are first the Kolmogorov–Smirnov one, from 1933, and then the Karl Pearson one, from 1900, see Figures 0.1 and 0.2. The first is

$$D_{\mathrm{sim}} = \sqrt{\mathrm{sim}}\, \max_t |F_{\mathrm{sim}}(t) - \Phi(t)|,$$

with $F_{\mathrm{sim}}(t)$ the empirical distribution function of the simulated data. The Pearson chi-squared statistic is

$$K_{\mathrm{sim}} = \sum_{j=1}^{m} \frac{(N_j - \mathrm{sim}\, p_{0,j})^2}{\mathrm{sim}\, p_{0.j}},$$

with N_j the number of datapoints landing in cell j, and $p_{0,j}$ the standard normal probability for that cell. The cells can be constructed as one pleases, but here I have taken $(\Phi^{-1}((j-1)/m), \Phi^{-1}(j/m))$, so that each of these have probability $p_{0,j} = 1/m$ under standard normality.

Observe how the distribution of Z_n comes closer and closer to the standard normal, as n increases, but rather slowly, and much more slowly than for Exercise 1, due to the skewness $\gamma_3/\sqrt{n}$ tending slowly to zero. With 10^4 datapoints we observe that the distributions underlying the data are in fact not really normal, yet, for $n \leq 40$, say, but for larger n we would need even more data to be able to statistically see that they are not really from the standard normal.

Feel free to build in your own extra test for normality, and make a figure corresponding to Figures 0.1–0.2. You may also play around with the (a, b) parameters of the Beta distribution you sample from, to check more extreme behaviour, in the sense of the Z_n needing larger sample sizes n in order to have a distribution closer to the standard normal.

41. The Strong Law of Large Numbers: Basics

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 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. As usual $\bar{X}_n$ is the sample average of the n first datapoints. I will tend to various steps to eventually demonstrate this statement, which is the Strong Law of Large Numbers (first proven by Колмогоров in 1933). We may for simplicity and without loss of generality take $\xi = 0$ below.

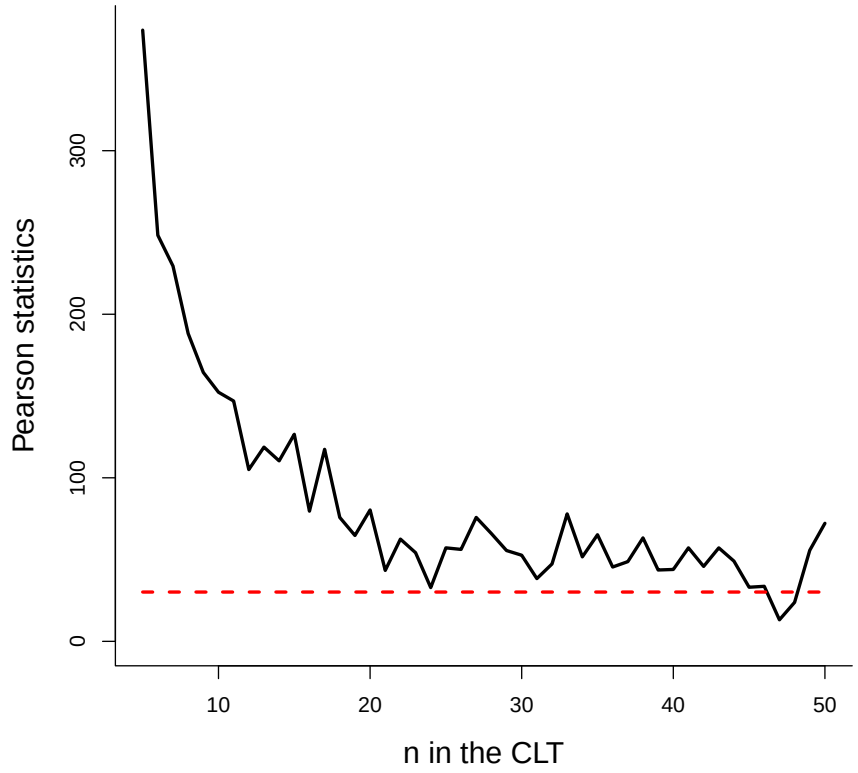

Figure 0.2: For each n, from 5 to 50, I have simulated 10^4 realisations of Z_n of Exercise 41, and then computed the Pearson chi-squared test statistic $K_n = \sum_{j=1}^{20}(N_j - 10^4 p_{0,j})^2/(10^4\, p_{0,j})$, for closeness of N_j, the number of points in cell j, namely $(\Phi^{-1}((j-1)/20), \Phi^{-1}(j/20))$, to $10^4\, p_{0,j}$, with $p_{0,j} = 1/20$. The red horizontal line is at 30.144, the 0.95 point of the null distribution.

(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.

(a) Show that if $\Pr(A_N) = 1$ for all N, then $\Pr(\cap_{N\geq 1} A_n) = 1$ – if your fully certain about a countable number of events, then you're also fully certain about all of them, jointly. This is actually not true with a bigger index set: if $X \sim \mathrm{N}(0,1)$, then you're 100 percent sure that $B_x = \{X \text{ is not } x\}$ takes place, for each single x, but from this does it *not* follow that you should be sure about $\cap_{\text{all } x} B_x$. Explain why.

(c) Show that $\Pr(A) = 1$ if and only if $\Pr(B_{n_0}) \to 0$, for each $\varepsilon > 0$, where

$$B_{n_0} = \cup_{n \geq n_0} \{|\bar{X}_n| \geq \varepsilon\}.$$

In words: for a given ε, the probability should be very low that there is *any* $n \geq n_0$ with $|\bar{X}_n| \geq \varepsilon$.

(d) A simple bound is of course

$$\Pr(B_{n_0}) \le \sum_{n \ge n_0} \Pr\{|\bar{X}_n| \ge \varepsilon\},$$

so it suffices to show, if possible, under appropriate conditions, that $\sum \Pr\{|\bar{X}_n| \ge \varepsilon\}$ is a convergent series. With finite variance σ^2, show that the classic simple Chebyshov bound does *not* solve any problem here.

(e) Show, however, that if the fourth moment is finite, then

$$\Pr\{|\bar{X}_n| \ge \varepsilon\} \le \frac{1}{\varepsilon^4} \mathrm{E}\,|\bar{X}_n|^4 \le \frac{c}{\varepsilon^4}\frac{1}{n^2},$$

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 I indicate here, without full details. Assume $\mathrm{E}\,|X_i|^r$ is finite, for some $r > 2$, like $r = 2.02$. Then one may show, via arguments in von Bahr (1965), that the sequence $\mathrm{E}\,|\sqrt{n}\bar{X}_n|^r$ is bounded. This leads to the bound

$$\Pr\{|\bar{X}_n| \ge \varepsilon\} \le \frac{1}{(\sqrt{n}\varepsilon)^r} \mathrm{E}\,|\sqrt{n}\bar{X}_n|^r,$$

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 exercise.

42. The Strong Law of Large Numbers: nitty-gritty details

This exercise goes through the required extra technical details, along with a few intermediate lemmas, to secure a full proof of the full LLN theorem: as long as $\mathrm{E}\,|X_i|$ is finite, the infinite sequence of sample means $\bar{X}_n$ will with probability equal to a hundred percent converge to $\xi = \mathrm{E}\,X_i$.

(a) We start with Kolmogorov's inequality: Consider independent zero-mean variables $X_1, \ldots, X_n$ with variances $\sigma_1^2, \ldots, \sigma_n^2$, and with partial sums $S_i = X_1 + \cdots + X_i$. Then

$$\Pr\{\max_{i \le n} |S_i| \ge \varepsilon\} \le \frac{\mathrm{Var}\,S_n}{\varepsilon^2} = \frac{1}{\varepsilon^2}\sum_{i=1}^{n} \sigma_i^2.$$

Note that this is a much stronger result than the special case of caring only about $|S_n|$, with $\Pr\{|S_n| \ge \varepsilon\} \le \mathrm{Var}\,S_n/\varepsilon^2$, which is the Chebyshov inequality. To prove it, work with the disjoint decomposition

$$A_i = \{|S_1| < \varepsilon, \ldots, |S_{i-1}| < \varepsilon, |S_i| \ge \varepsilon\} \quad \text{and} \quad A = \cup_{i=1}^n A_i = \{\max_{i \le n} |S_i| \ge \varepsilon\}.$$

Show that

$$\mathrm{E}\,S_n^2 \ge \mathrm{E}\,S_n^2 I(A) = \sum_{i=1}^{n} \mathrm{E}\,S_n^2 I(A_i),$$

that

$$\mathrm{E}\,S_n^2 I(A_i) = \mathrm{E}\,(S_i + S_n - S_i)^2 I(A_i) \ge \varepsilon^2 \Pr(A_i),$$

and that this leads to the inequality asked for.

(b) Consider a sequence of independent $X_1, X_2, \ldots$ with means zero and variances $\sigma_1^2, \sigma_2^2, \ldots$. Show that if $\sum_{i=1}^{\infty} \sigma_i^2$ is convergent, then $\sum_{i=1}^{\infty} X_i$ is convergent with probability 1. – It suffices to show that the sequence of partial sums $S_n = X_1 + \cdots + X_n$ is Cauchy with probability 1. Show that this is the same as

$$\lim_{n\to\infty} \Pr\bigl[\cup_{i,j\geq n}\{|S_i - S_j| \geq \varepsilon\}\bigr] = 0 \quad \text{for each } \varepsilon > 0.$$

Use the Kolmogorov inequality to show this.

(c) A quick example to illustrate this result is as follows. Consider

$$X = \frac{X_1}{10} + \frac{X_2}{100} + \frac{X_3}{1000} + \cdots,$$

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 Exercise 42 – 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 need characterisations of the tails of a distribution with finite mean. Show that if $X \geq 0$, with distribution function F, then $\mathrm{E}\, X = \int_0^{\infty} \{1 - F(x)\}\, \mathrm{d}x$. Show more generally that for any X,

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

(h) Then show that if X has finite mean, then

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \int_{(-i,i)} x^2\, \mathrm{d}F(x) < \infty.$$

(i) I note that upon examining the arguments needed to prove (h), one learns that this is an if-and-only-if result. More generally, attempt to prove that

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

(j) 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 Borel–Cantelli 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 $\xi_i = \mathrm{E}\, Y_i$.

(j) 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:

$$\Pr\{\limsup_{n\to\infty} \bar{X}_n = \infty\} = 1.$$

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.

43. Yes, we converge with probability one

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

$$\widehat{\sigma}_n = \Big\{ \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \Big\}^{1/2}$$

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

$$\widehat{\gamma}_{3,n} = \frac{1}{n} \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^3}{\widehat{\sigma}^3}$$

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.

44. Exam STK 201 1989, #1

Determine for each of the following statements whether it is true or not. If it is correct, give a short proof; if it is incorrect, construct a counterexample.

(a) If X and Y are two real random variables defined on the same probability space, and

$$\phi_X(t) = \mathrm{E}\,\exp(itX) = \mathrm{E}\,\exp(itY) = \phi_Y(t) \quad \text{for all } t,$$

then $X = Y$ with probability 1.

(b) If (X, Y) is a random pair, with the property that

$$\mathrm{E}\,\exp\{i(sX + tY)\} = \mathrm{E}\,\exp(isX)\,\mathrm{E}\,\exp(itY) \quad \text{for all } s \text{ and } t,$$

then X and Y are stochastically independent.

(c) If X_n and X are real random variables, and X_n converges in distribution to X, then

$$\lim_{n\to\infty} \Pr\{X_n = x\} = 0$$

for each continuity point x for the cumulative distribution function for X.

(d) If X_n and X are real random variables, and X_n converges in distribution to X, and a certain set A has the property that $\Pr\{X_n \in A\} = 1$ for every n, then $\Pr\{X \in A\} = 1$ too.

45. Exam STK 201 1989, #2

One wants to estimate the position of a parameter point (a, b) in the plane. For this task one obtains n independent pairs of measurements $(X_1, Y_1), \ldots, (X_n, Y_n)$. These come from the same unknown distribution, but it is known that the X_i have expected value a and standard deviation 1, and that the Y_i have expected value b and standard deviation 1. Finally, X_i and Y_i are uncorrelated.

(a) Introduce $\widehat{a}_n = (1/n)\sum_{i=1}^n X_i$ and $\widehat{b}_n = (1/n)\sum_{i=1}^n Y_i$. Find the simultaneous (joint) limit distribution for

$$\begin{pmatrix} \sqrt{n}(\widehat{a}_n - a) \\ \sqrt{n}(\widehat{b}_n - b) \end{pmatrix}.$$

(b) Construct an asymptotic 90 percent simultaneous (joint) confidence region for (a, b). What is the shape of this region?

(c) Itis often useful to give the position of (a, b) in *polar coordinates*, that is, by the length $\rho = (a^2 + b^2)^{1/2}$ and the angle $\theta = \arctan(b/a)$. [This is equivalent to $a = \rho\cos\theta$ and $b = \rho\sin\theta$.] Let

$$\widehat{\rho}_n = (\widehat{a}_n^2 + \widehat{b}_n^2)^{1/2} \quad \text{and} \quad \widehat{\theta}_n = \arctan(\widehat{b}_n/\widehat{a}_n).$$

Find the simultaneous (joint) limit distribution for

$$\begin{pmatrix} \sqrt{n}(\widehat{\rho}_n - \rho) \\ \sqrt{n}(\widehat{\theta}_n - \theta) \end{pmatrix},$$

and comment on thie result. [The derivative of the $\arctan x$ function is $1/(1 + x^2)$.]

46. Exam STK 201 1989, #3

Let $X_1, X_2, X_3, \ldots$ be a sequence of independently and identically distributed real random variables. The common distribution of X_i is continuous. Agree to say that if

$$X_n > \max\{X_1, \ldots, X_{n-1}\},$$

then 'X_n has set a new record'. Let

$$R_n = \begin{cases} 1, & \text{if } X_n \text{ has set a new record;} \\ 0, & \text{if } X_n \text{ has not set a new record.} \end{cases}$$

We count X_1 as a 'new record', so that $R_1 = 1$.

(a) Show, by direct arguments, that

$$\Pr\{R_n = 1\} = 1/n \quad \text{for } n \geq 1.$$

Note: One can also prove that the R_n become stochastically independent. You do not have to show this (during exam hours), but you can use the result in the rest of the present exercise.

(b) Let Y_n be the number of new records during the first n observations. Introduce

$$a_n = \sum_{i=1}^{n} \frac{1}{i} \quad \text{and} \quad \sigma_n^2 = \sum_{i=1}^{n} \frac{1}{i}\Big(1 - \frac{1}{i}\Big).$$

Show that

$$\frac{Y_n - a_n}{\sigma_n} \to_d \mathrm{N}(0,1).$$

(c) Then use this result to reach the following:

$$\frac{Y_n - \log n}{\sqrt{\log n}} \to_d \mathrm{N}(0,1).$$

Here $\log n$ is the natural logarithm (the one with the Ibsen-Tolstoy base number e), and the following mathematical results are at your disposal:

$$\sum_{i=1}^{n} \frac{1}{i} - \log n \to \gamma = 0.5772..., \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6} = 1.6449...$$

(d) I wonder: about how many new records will be set during the first million observations? Construct an interval that with probability approximately 95 percent contains $Y_{1\,000\,000}$.

(e) Let Z_n be the number of new records among the observations $X_{n+1}, \ldots, X_{2n}$. Prove that Z_n converges in distribution to a Poisson with parameter $\lambda = \log 2$.

47. Exam STK 201 1989, #4

The following situation was studied in Exercise 4 of the ST 001 exam in May 1989 (yesterday, actually). Certain measurements $X_1, \ldots, X_n$ are independent and have the same probability density f, with expected value ξ and standard deviation σ. The parameters are unknown. Introduce

$$\widehat{\xi}_n = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \widehat{\sigma}_n^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The ST 001 students were among other things asked to answer this question:

(a) Explain briefly how you by counting the number of observations in the intervals $(\bar{X}-s, \bar{X}+s)$, $(\bar{X}-2s, \bar{X}+2s)$, $(\bar{X}-3s, \bar{X}+3s)$ may get a rough idea of whether the observations $X_1, \ldots, X_n$ are normally distributed or not.

The present ST 201 exercise takes a closer look at the intuitive arguments that were expected of the ST 001 students. Assume in what follows that $X_1, X_2, \ldots$ really are independent and normal (ξ, σ^2), so that the common underlying cumulative distribution function is

$$F(t) = \Pr\{X_i \le t\} = \Pr\Big\{\mathrm{N}(0,1) \le \frac{t-\xi}{\sigma}\Big\} = \Phi\Big(\frac{t-\xi}{\sigma}\Big).$$

(a) Let $F_n(t) = (1/n)\sum_{i=1}^n I\{X_i \le t\}$ be the empirical cumulative distribution function. What can you say about the behaviour of F_n for large n?

(b) Assume that you have succeeded in proving the following statement: For each given c will

$$F_n(\widehat{\xi}_n + c\widehat{\sigma}_n) \to_{\text{a.s.}} F(\xi + c\sigma).$$

Show that this leads to

$$A_n = \frac{1}{n}\sum_{i=1}^n I\Big\{a < \frac{X_i - \widehat{\xi}_n}{\widehat{\sigma}_n} \le b\Big\} \to_{\text{a.s.}} \Pr\{a < \mathrm{N}(0,1) \le b\} = \Phi(b) - \Phi(a).$$

(c) Explain why this gives an answer to the ST 001 exam question quoted above!

(d) Finally, prove the result given in (b). *Note:* There are several ways of proving this result. If you should choose a method of proof that leads to convergence in probability, and not convergence almost surely, then you will still be awarded full score by the examination censors & markers.

48. Exam STK 201 1989, cont., #1

Determine for each of the following four statements whether it is correct or wrong. If it is correct, give a brief argument for this; if not, give a counterexample.

(a) Dersom X_n converges in distribution to the normal $\mathrm{N}(0,1)$, then the mean of X_n converges to zero.

(b) Hvis X_n converges to a in probability, then X_n will also converge to a almost surely.

(c) Såfremt $X_n \to_d X$ and $Y_n \to_d Y$, then $X_n + Y_n \to_d X + Y$.

(d) Ifall $X_n = (X_{n,1}, \ldots, X_{n,p})^{\mathrm{t}}$ converges in distribution to $X = (X_1, \ldots, X_p)^{\mathrm{t}}$ in distribution, where the components of the latter are independent and standard normal, then $\sum_{i=1}^p X_{n,i}^2$ will converge in distribution to the χ_p^2.

49. Exam STK 201 1989, cont., #2

Let $X_1, X_2, X_3, \ldots$ be a sequence of independently and identically distributed real random variables. The common distribution of X_i is continuous. Agree to say that if

$$X_n > \max\{X_1, \ldots, X_{n-1}\},$$

then 'X_n has set a new record'. Let

$$R_n = \begin{cases} 1, & \text{if } X_n \text{ sets a new record,} \\ 0, & \text{if } X_n \text{ does not set a new record.} \end{cases}$$

We count X_1 as a 'new record', so that $R_1 = 1$.

(a) Show via direct arguments that

$$\Pr\{R_n = 1\} = 1/n \quad \text{for } n \geq 1.$$

(b) Explain what it means that a sequence of random variables are stochastically independent. Show explicitly that R_1, R_2, R_3 are independent. – *Note:* One may show that the full sequence of $R_1, R_2, R_3, \ldots$ are indeed independent, but you need not show this during exam hours. You may however use this fact for the points below.

(c) Let's push the records aside for two minutes, but formulate and prove the so-called Borel–Cantelli lemma.

(d) What is the probability that the sequence $X_1, X_2, X_3, \ldots$ will produce infinitely many records?

50. Exam STK 201 1989, cont., #3

Make the following statement precise, and then prove it: A binomial (n, p) variable is approximately a Poisson, when n is large and p is small.

51. Exam STK 201 1989, cont., #4

The following result is to taken as known: If $Y_1, Y_2, \ldots$ are independent and come from the same distribution, of the parametric form $f(y, \theta)$, and $\widehat{\theta}_n$ is the rimelighetsfunksjonsmaksimeringsestimatoren, then, under appropriate and mild regularity conditions, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to_d \mathrm{N}_p(0, J(\theta)^{-1}).$$

Here p is the dimension of θ, and

$$J(\theta) = \mathrm{E}_\theta\, u(Y, \theta) u(Y, \theta)^{\mathrm{t}} = -\mathrm{E}_\theta\, \frac{\partial^2 \log f(Y, \theta)}{\partial \theta \partial \theta^{\mathrm{t}}}$$

is Fisher's information matrix, involving also the score function $u(y, \theta) = \partial \log f(y, \theta)/\partial \theta$. Finally E_θ signals expectation under the distribution $f(y, \theta)$.

(a) Assume the parameter θ is one-dimensional. Show that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to_d \tau(\theta)\mathrm{N}(0, 1),$$

where

$$\tau(\theta) = \frac{1}{\sqrt{-\mathrm{E}_\theta \partial^2 \log f(Y, \theta)/\partial \theta^2}}.$$

(b) Apply this to the exponential model, where $f(y, \theta) = \theta \exp(-\theta y)$ for positive y and θ is a positive parameter.

(c) It is often important to estimate the underlying density behind the observations, say $f(y)$. In the parametric case, where $f(y) = f(y, \theta)$, it is natural to use the simple plug-in estimator $\widehat{f}(y) = f(y, \widehat{\theta}_n)$. Show, in the general but still one-dimensional case, that

$$\sqrt{n}\{f(y, \widehat{\theta}_n) - f(y, \theta)\} \to_d f(y, \theta)u(y, \theta)\tau(\theta)\mathrm{N}(0, 1).$$

(d) An often used measure of quality for a density estimator $\widehat{f}$ for f is the integrated squared error

$$\mathrm{ise}_n = \int \{f(y, \widehat{\theta}_n) - f(y, \theta)\}^2 \, \mathrm{d}y.$$

Show, still for the general but one-dimensional case, that

$$n \, \mathrm{ise}_n \to_d c(\theta)\chi^2_1,$$

where the proportionality factor involved is

$$c(\theta) = \tau(\theta)^2 \int f(y, \theta)^2 u(y, \theta^2) \, \mathrm{d}y.$$

(e) Show that mean integrated squared error,

$$\mathrm{mise}_n = \mathrm{E}_\theta \int \{f(y, \widehat{\theta}_n) - f(y, \theta)\}^2 \, \mathrm{d}y,$$

with a first-order approximation, is equal to $\theta/(4n)$ for the exponential distribution case.

(f) Then establish the following intriguingly simple, general, and informative result concerning iwse_n and miwse_n, the $1/f$ weighted versions of ise_n and mise_n:

$$n \, \mathrm{iwse}_n = n \int \frac{\{f(y, \widehat{\theta}_n) - f(y, \theta)\}^2}{f(y, \theta)} \, \mathrm{d}y \to_d \chi^2_p, \quad \mathrm{miwse}_n \doteq p/n.$$

Again, p is the number of parameters in the model. Note that this result does not depend on *which* parametric model is used, or on the sample space for the observations (or, for that matter, on the dominating measure used to define the densities $f(y, \theta) = \mathrm{d}P_\theta(y)/\partial\mu$).

52. Exam STK 201 1995, #1

Here are some questions from the core curriculum of the course.

(a) Explain what a probability spacae $(\Omega, \mathcal{A}, P)$ is. List the demands for P being a probability measure.

(b) From the definitions in (a), show that if $B_1, B_2, \ldots$ are arbitrary sets in $\mathcal{A}$, then we have $P(\cup_{i=1}^n B_i) \leq \sum_{i=1}^n P(B_i)$, and also $P(\cup_{i=1}^\infty B_i) \leq \sum_{i=1}^\infty P(B_i)$.

(c) Formulate and prove the so-called Borel–Cantelli lemma.

53. Exam STK 201 1995, #2

This exercise concerns the use of characteristic functions to, well, characterise distributions.

(a) Define the characteristic function ϕ for a real random variable X. Show that this function is bounded and uniformly continuous.

(b) Assume X has mean zero and finite variance σ^2. Show that

$$\phi(t) = 1 - \tfrac{1}{2}\sigma^2 t^2 + o(t^2).$$

[Here I wish for 'direct arguments using the definitions'; simply saying this is inside the curriculum is not sufficient, on this particular occasion.]

(c) Let in this point X and X' be independent and normal $(0, \sigma^2)$ variables. Show, using characteristic functions, that $(X+X')/\sqrt{2}$ has the same distribution as each of the two observations. Give a generalisation.

(d) Let X be as in point (b), and assume that its distribution has the invariance property from point (c), i.e. that if X and X' are independent with this same distribution, then $(X+X')/\sqrt{2}$ has the same distribution as each of X and X'. Show that this leads to

$$\phi\Big(\frac{t}{2^{k/2}}\Big)^{2^k} = \phi(t) \quad \text{for all natural numbers } k \text{ and all real } t.$$

(e) Show that the assumption of point (d) implies that X by necessity must be normally distributed, or equal to zero. – The zero-mean normal is hence the only distribution in this universe with the $(X + X')/\sqrt{2} \sim X$ property.

54. Exam STK 201 1995, #3

This exercise works itself towards the construction of a certain natural test for the hypothesis that different groups of normally distributed data have the same standard deviation. Such a test is important also since many standard techniques use such an equality of spread parameters as a basic working assumption.

(a) Let $Y_1, \ldots, Y_n$ be independent with the same distribution, and assume this distribution has a finite fourth moment. Let mean and standard deviation be μ and σ, and let $\gamma_4 = \mathrm{E}\,(Y - \mu)^4/\sigma^4 - 3$ be the so-called kurtosis. Construct a consistent estimator for γ_4.

(b) The usual empirical variance is $\widehat{\sigma}_n^2 = (1/n)\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, where $\bar{Y}_n$ is the sample mean $(1/n)\sum_{i=1}^n Y_i$. Show that

$$\sqrt{n}(\widehat{\sigma}_n^2 - \sigma^2) \to_d \mathrm{N}(0, \sigma^4(2 + \gamma_4)).$$

(c) Find the limit distribution for $\sqrt{2n}(\log \widehat{\sigma}_n - \log \sigma)$. Show in particular that the limit is the standard normal $\mathrm{N}(0, 1)$ in the case where the X_i are normal.

(d) Construct a confidence interval with coverage approximately 90 percent for σ, which ought to be valid also outside normal conditions.

(e) Assume now that there are n observations for each of five normally distributed populations, with standard deviations $\sigma_1, \ldots, \sigma_5$. Let further $\widehat{\sigma}_{n,j}^2$ be the empirical variance for group j, for $j = 1, \ldots, 5$. Find the limit distribution for

$$\begin{pmatrix} \sqrt{2n}(\log \widehat{\sigma}_{n,1} - \log \sigma_1) \\ \vdots \\ \sqrt{2n}(\log \widehat{\sigma}_{n,5} - \log \sigma_5) \end{pmatrix}.$$

(f) Construct a test for the hypothesis $H_0\colon \sigma_1 = \cdots = \sigma_5$, using the result from the previous point, and which should have limiting significance level 5 percent. [For simplicity it is assumed that there are equally many observations in each group here. It is however not difficult to generalise this to the case of sample sizes $n_1, \ldots, n_5$ being different. You may do this after exam hours.]

55. Exam STK 201 1995, #4

This exercise concerns estimation in the so-called truncated Poisson model.

(a) Assume that a certain Y_0 has a Poisson distribution with parameter θ, but that X_0 can only be observed if its value is at least 1. Let Y be such an observation. Show that its probability distribution is

$$\Pr\{Y = y\} = f(y, \theta) = \frac{\exp(-\theta)\theta^y/y!}{1 - \exp(-\theta)} \quad \text{for } y = 1, 2, 3, \ldots.$$

(b) Assume $Y_1, Y_2, \ldots$ are independent observations from such a truncated Poisson distribution. Put up an equation to determine the rimelighetsfunksjonsmaksimeringsestimatoren $\widehat{\theta}_n$ for θ.

(c) Describe the large-sample behaviour of $\widehat{\theta}_n$, e.g. by using results about the rimelighetsfunksjonsmaksimeringsestimatorsekvensen from the course curriculum.

(d) Suppose now that one cannot necessarily trust the parametric modelling assumption of (a), but that there is a certain underlying true data generating mechanism, on $\{1, 2, 3, \ldots\}$. Assume that this true distribution has a finite mean ξ and standard deviation τ. Explain what the rimelihetsfunksjonsmaksimeringsestimatoren $\widehat{\theta}_n$ converges towards, under these wider assumptions. Express your answers in terms of ξ and τ.

56. Exam STK 201 1995, #5

The usual ingredients in so-called linear-normal statistical theory are as follows: (i) observations are independent; (ii) they have the same variance; (iii) the mean structure is linear in certain explanatory variables, or covariates; and (iv) the underlying distribution is normal. Under these assumptions there is as we know built a broad, very frequently applied, and exact theory.

This particular exercise is meant to illustrate that one also might come a long way also in the absence of the exact normality condition (iv). Assume that

$$Y_i = \beta x_i + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the x_i are given, and where the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent from the same distribution, with mean zero and standard deviation σ (i.e. without the traditional extra words 'and their distribution is normal'). The parameters β and σ are unknown and need to be estimated.

(a) Show that the least squares estimator for β is $\widehat{\beta}_n = \sum_{i=1}^n x_i Y_i / M_n$, where $M_n = \sum_{i=1}^n x_i^2$. Give an estimator also for σ.

(b) Under the exact normality assumption it holds that $Z_n = M_n^{1/2}(\widehat{\beta}_n - \beta)$ is normal $(0, \sigma^2)$, and the classical inference methods are based on this fact. Your task is now to demonstrate that the limit distribution of Z_n is indeed this $\mathrm{N}(0, \sigma^2)$, under certain conditions, but without assuming that the ε_i follow a normal distribution.

(c) Construct a confidence interval for β with coverage converging to 0.90, and make your assumptions and arguments clear.

57. How large is the last time?

Let $Y_1, Y_2, \ldots$ be an infinite sequence of independent normal (ξ, σ^2) variables, and let $\widehat{\xi}_n, \widehat{\sigma}_n$ be the maximum likelihood estimators.

(a) Find these, by all means & for all del.

(b) Show that

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

(c) Results from Hjort and Fenstad (1992) may be applied here, to show that the following. Let $N_{1,\varepsilon}$ is the very last time $|\widehat{\xi}_n - \xi| \geq \varepsilon$, and $N_{2,\varepsilon}$ the very last time $|\widehat{\sigma}_n - \sigma| \geq \varepsilon$. Why are $N_{1,\varepsilon}$ and $N_{2,\varepsilon}$ well-defined random variables? Then

$$\begin{pmatrix} \varepsilon^2 N_{1,\varepsilon} \\ \varepsilon^2 N_{2,\varepsilon} \end{pmatrix} \to_d \begin{pmatrix} \sigma^2 W_{1,\max}^2 \\ \frac{1}{2}\sigma^2 W_{2,\max}^2 \end{pmatrix}$$

when ε marches to zero, where $W_{1,\max}$ and $W_{2,\max}$ are the maximal absolute values ot two independent Brownian motions over the $[0,1]$ interval. (You are not yet supposed to show this.) Let N_ε tbe the very last n where either $|\widehat{\xi}_n - \xi| \geq \varepsilon$ or $|\widehat{\sigma}_n - \sigma| \geq \varepsilon$. Show that

$$\varepsilon^2 N_\varepsilon \to_d \sigma^2 \max\{W_{1,\max}^2, W_{2,\max}^2\}.$$

Attempt to finds its distribution.

(d) Generalise.

58. Bernshteĭn and Weierstraß

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 have 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.

In the present exercise we give a *constructive and relatively straightforward* proof of the Weierstraß theorem, involving so-called Bernshteĭn polynomials. Let $g\colon [0,1] \to \mathcal{R}$ be continuous, and construct

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

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

(a) Show that $B_n(p) \to_{\mathrm{pr}} g(p)$, for each p.

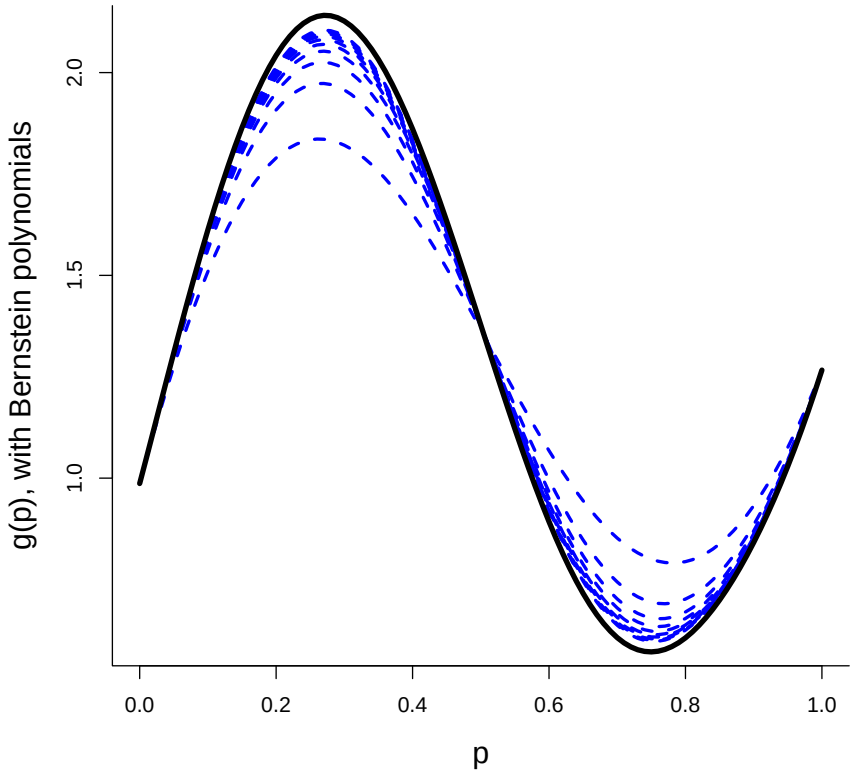

Figure 0.3: The given non-polynomial function $g(p)$, along with approximating Bernshteĭn polynomials, of order $10, 20, \ldots, 90, 100$.

(b) Then show that the convergence is actually uniform. 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 fill in the required arguments for the following:

$$\begin{aligned} |B_n(p) - g(p)| &\leq \mathrm{E}_p\,|g(X_n/n) - g(p)| \\ &\leq \mathrm{E}_p\,|g(X_n/n) - g(p)|\,I\{|X_n/n - p| < \delta\} \\ &\qquad + \mathrm{E}_p\,|g(X_n/n) - g(p)|\,I\{|X_n/n - p| \geq \delta\} \\ &\leq \varepsilon + 2M\,\Pr\{|X_n/n - p| \geq \delta\}, \end{aligned}$$

with M a bound on $|g(x)|$.

(c) Show from this that

$$\max_p |B_n(p) - g(p)| \to 0 \quad \text{as } n \to \infty.$$

(d) 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. Attempt to construct a version of Figure 0.3,

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?

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

(f) Speculation, Your Honor: a distribution F *is* completely specified by its characteristic function

$$\phi(t) = \mathrm{E}\,\exp(itX) = \int \cos(tx)\,\mathrm{d}F(x) + i\int \sin(tx)\,\mathrm{d}F(x).$$

This can be proven in various ways, see earlier Exercises 15–16. But it may be attacked afresh, in the spirit of Weierstraß type approximations etc. It is sufficient to show that with two distributions F and G with the same $\phi(t)$, we must have $\int h\,\mathrm{d}F = \int h\,\mathrm{d}G$ for each continuous bounded h (cf. the master theorem of Exercise 6). From the assumption we know that

$$\int h^*(x)\,\mathrm{d}F(x) = \int h^*(x)\,\mathrm{d}G(x) \quad \text{for all } h^*(x) = \sum_{j=1}^{m} a_j\{\cos(t_j x) + i\sin(t_j x)\}.$$

So try to show that for the given continuous and bounded h, and for each bounded interval $[-c, c]$ and $\varepsilon > 0$, there must exist such a function h^* with $\max_{x\in[-c,c]} |h(x) - h^*(x)| \leq \varepsilon$. Prove that this would be sufficient to prove that $F = G$ (once again). Could there be a Bernshteĭn type result lurking here?

59. Even more on characteristic functions

Here we go into a couple of helpful intermediate results for characteristic functions. Let $\phi(t) = \mathrm{E}\,\exp(itX)$, for X with a distribution F.

(a) Show that $|\exp(it) - 1| \leq |t|$ for all t, and that this implies

$$|\phi(t) - 1| \leq \int |tx|\,\mathrm{d}F(x) = |t|\,\mathrm{E}\,|X|.$$

(b) Show that $|\exp(it) - 1 - it| \leq \frac{1}{2}|t|^2$ for all t, and with $\xi = \mathrm{E}\,X$ show that this implies

$$|\phi(t) - 1 - it\xi| \leq \tfrac{1}{2}|t^2|\,\mathrm{E}\,|X|^2.$$

(c) Generalise further to

$$|\exp(it) - 1 - it - \tfrac{1}{2}(it)^2| \leq \tfrac{1}{6}|t|^3 \quad \text{for all } t.$$

Assume $\xi = \mathrm{E}\,X = 0$ and that $\mathrm{Var}\,X = \sigma^2$ is finite. Show that if also the third moment is finite, then

$$|\phi(t) - 1 - \tfrac{1}{2}(it)^2\sigma^2| = |\phi(t) - (1 - \tfrac{1}{2}t^2\sigma^2)| \leq \tfrac{1}{6}|t|^3\,\mathrm{E}\,|X|^3.$$

In particular,

$$\phi(t) = 1 - \tfrac{1}{2}\sigma^2 t^2 + O(|t|^3).$$

(d) Show that we may rid ourselves with the finite third moment assumption here, by proving that

$$\phi(t) = 1 - \tfrac{1}{2}\sigma^2 t^2 + o(|t|^2),$$

under only zero mean and finite σ conditions. Specifically, the task is to show that

$$\frac{1}{t^2}\int \{\exp(itx) - 1 - itx - \tfrac{1}{2}(it)^2 x^2\}\,\mathrm{d}F(x) \to 0 \quad \text{as } t \to 0.$$

This is also related to the fact that when $\mathrm{E}\,|X|^2$ is finite, then

$$\phi''(t) = \mathrm{E}\,(iX)^2 \exp(itX) = \int (ix)^2 \exp(itx)\,\mathrm{d}F(x)$$

exists and is a continuous function in t.

(e) Use induction to show that

$$|\exp(it) - 1 - it - \tfrac{1}{2}(it)^2 - \cdots - (1/m!)(it)^m| \le |t|^{m+1}/(m+1)! \quad \text{for all } t,$$

and that this implies

$$|\phi(t) - 1 - it\,\mathrm{E}\,X - \tfrac{1}{2}(it)^2\,\mathrm{E}\,X^2 - \cdots - (1/m!)(it)^m\,\mathrm{E}\,X^m| \le \frac{|t|^{m+1}\,\mathrm{E}\,|X|^{m+1}}{(m+1)!}.$$

Show also, without a finite $\mathrm{E}\,|X|^{m+1}$, that if $\mathrm{E}\,|X|^m$ is finite, then

$$\phi^{(m)}(t) = \mathrm{E}\,(iX)^m \exp(itX) = \int (ix)^m \exp(itx)\,\mathrm{d}F(x),$$

and that this function is continuous in t.

60. A tail inequality & tightness & limits

Let X have distribution F and characteristic function ϕ. The aim of this exercise is to establish the useful tail inequality

$$\Pr\Big\{|X| \ge \frac{2}{\varepsilon}\Big\} \le \frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} \{1 - \phi(t)\}\,\mathrm{d}t.$$

So, tail probabilities for X are tied to the behaviour of ϕ close to zero.

(a) Use the Fubini theorem (you know, interchanging the order of integration) to demonstrate that

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

In particular, the integral of $\phi(t)$ over an interval symmetric around zero is really a real number (i.e. the complex component disappears).

(b) Deduce that

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

with the value $c = 2$ yielding the inequality given above.

(c) For the case of X being standard normal, check the precision of the tail inequality. (The answer appears to be: no, it's rather unsharp, and is utterly conservative in its tail probability assessment.) From the simple approximation $\phi(t) \doteq 1 - \frac{1}{2}\sigma^2 t^2$, for t small, for a variable with zero mean and standard deviation σ, work out that $\Pr\{|X| \geq 2/\varepsilon\} \leq (1/3)\sigma^2\varepsilon$. Explain why this is blunter, as in less sharp, than with e.g. the Chebyshov inequality.

(c) If we now have a collection of random variables, where their characteristic functions have approximately the same level of smoothness around zero, then we should get *tightness*, a guarantee there is no runaways with mass escaping from the crowd. Assume that X_n has characteristic function ϕ_n, with $\phi_n(t)$ converging pointwise to some $\phi(t)$, continuous at zero, on some $[-\varepsilon, \varepsilon]$. For a given ε', find ε such that $|1 - \phi(t)| \leq \varepsilon'$ for $|t| \leq \varepsilon$. Show that

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

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

(d) It's somewhat technical, but the following argument can be understood even without the finest nitty-gritty details. With the situation as in point (c), there is always *some* subsequence, say $X_{n'}$ for some subsequence n' running to infinity, such that their cumulative distribution functions $F_{n'}$ tends to some appropriate nondecreasing right-continuous F on the latter's continuity points – but technically speaking we do not know yet that F is a proper cumulative distribution function; it could be degenerate. With the tightness, however, we're guaranteed that F is bona fide, with $F(-\infty) = 0$ and $F(\infty) = 1$. Hence $X_{n'} \to_d X$, for the X having this F as its cumulative. But that again implies $\phi_{n'}(t) \to \phi_X(t)$, pointwise, and the limit function $\phi(t)$ is identical to $\phi_X(t)$ and hence a bona fide characteristic function.

(e) Verify that all of this implies the following highly useful device: Suppose X_n is such that its characteristic function $\phi_n(t)$ converges to *some* $\phi(t)$, in a neighbourhood around zero, and that the limit function $\phi(t)$ is continuous at zero. Then (1) the limit is a characteristic function, for some appropriate X, and, lo & behold, $X_n \to_d X$. – The point is also that in some cases, one discovers and then proves the existence of a new probability distribution in this fashion.

(f) Suppose you just arrived at this planet this morn' and first invented the super-simple two-point distribution with values ± 1 with equal probabilities $\frac{1}{2}$ and $\frac{1}{2}$ – show that its characteristic function is $\phi(t) = \cos t$. Then you wonder what happens if you sum outcomes of that distribution, and form $Z_n = \sum_{i=1}^n X_i/\sqrt{n}$. Then you deduce that this variable's characteristic function is $\cos(t/\sqrt{n})^n$, and then that it converges ... to $\exp(-\frac{1}{2}t^2)$. You would then have discovered, and proven the existence of, the standard normal distribution, from the proverbial scratch.

61. The Liapunov and Lindeberg theorems: main story

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.

So 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 $\phi_1, \phi_2, \cdots$. The question is when we can rest assured that the normalised sum,

$$Z_n = \frac{X_1 + \cdots + X_n}{B_n} = \frac{\sum_{i=1}^n X_i}{(\sum_{i=1}^n \sigma_i^2)^{1/2}},$$

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. The lemma is stated, proven, and used in Hjort (1990, Appendix).

(b) Show that Z_n has characteristic function

$$\kappa_n(t) = \mathrm{E}\,\exp(itZ_n) = \phi_1(t/B_n)\cdots\phi_n(t/B_n).$$

(c) We know that $\phi_i(s) \doteq 1 - \frac{1}{2}\sigma_i^2 s^2$ 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) $\sum_{i=1}^n |1 - \phi_i(t/B_n)|$ staying bounded.

Show that

$$\begin{aligned} |\phi_i(s) - (1 - \tfrac{1}{2}\sigma_i^2 s^2)| &= \Big| \int \{\exp(isx) - 1 - isx - \tfrac{1}{2}(isx)^2\}\,\mathrm{d}F_i(x) \Big| \\ &\le \int |\exp(isx) - 1 - isx - \tfrac{1}{2}(isx)^2|\,\mathrm{d}F_i(x) \\ &\le \tfrac{1}{6}|s|^3\,\mathrm{E}\,|X_i|^3. \end{aligned}$$

(d) This leads to the условие Ляпунова version of the Lindeberg theorem: show that if the variables all have finite third order moments, with $B_n \to \infty$ and

$$\sum_{i=1}^n \mathrm{E}\Big|\frac{X_i}{B_n}\Big|^3 \to 0,$$

then $\kappa_n(t) \to \exp(-\frac{1}{2}t^2)$, which 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.

(e) The issue waits however for an 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 условие Ляпунова is in force, then the Lindeberg condition holds (so farmer Lindeberg assumes less than Lyapunov).

(f) 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 Emil Stoltenberg (2019) has done, in a technical note to the STK 4011 course (he's incidentally too modest when he writes that his note is an epsilon-extension of Inlow's 2010 paper; the extension is harder than several ε). Check his note, on the course website, and make sure you understand his main tricks and steps.

62. The Lindeberg theorem: nitty-gritty details

The essential story, regarding Lyapunov and Lindeberg, has been told in the previous exercise. Here we tend to the smaller-level but nevertheless crucial remaining details, in order for the ball to be shoven across the finishing line after all the preliminary work. You may also check partly corresponding details in Stoltenberg's note (2019). Again, let $X_1, X_2, \ldots$ be independent, with distributions $F_1, F_2, \ldots$, standard deviations $\sigma_1, \sigma_2, \ldots$, and characteristic functions $\phi_1, \phi_2, \ldots$. 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 условие Линдеберга, 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

$$\alpha_n = \max_{i \le n} \frac{\sigma_i^2}{B_n^2} \to 0.$$

From this in particular follows

$$|\phi_i(t/B_n) - 1| \le \int |\exp(itx/B_n) - 1 - itx/B_n| \, \mathrm{d}F_i(x) \le \tfrac{1}{2} t^2 \int (x/B_n)^2 \, \mathrm{d}F_i(x) \le \tfrac{1}{2} t^2 \alpha_n,$$

so all $\phi_i(t/B_n)$ are eventually inside radius say $\frac{1}{2}$ of 1, which means we're in a position to take the logarithm and work with

$$\kappa_n(t) = \log \mathrm{E} \exp(itZ_n) = \sum_{i=1}^n \log \phi_i(t/B_n)$$

etc.; see the start lemma of the preceding exercise.

(b) In continuation and refinement of arguments above, show that

$$\begin{aligned}
|\phi_i(t/B_n) - (1 - \tfrac{1}{2}\sigma_i^2 t^2/B_n^2)| &= \Big| \int \{\exp(itx/B_n) - 1 - itx/B_n - \tfrac{1}{2}(itx/B_n)^2\} \, \mathrm{d}F_i(x) \Big| \\
&\le \int |\exp(itx/B_n) - 1 - itx/B_n - \tfrac{1}{2}(itx/B_n)^2| \, \mathrm{d}F_i(x) \\
&\le \int_{|x|/B_n \le \varepsilon} \tfrac{1}{6} \frac{|t|^3 |x|^3}{B_n^3} \, \mathrm{d}F_i(x) \\
&\qquad + \int_{|x|/B_n > \varepsilon} \Big(\tfrac{1}{2} \frac{|t|^2 |x|^2}{B_n^2} + \tfrac{1}{2} \frac{|t|^2 |x|^2}{B_n^2} \Big) \, \mathrm{d}F_i(x) \\
&\le \tfrac{1}{6} |t|^3 \varepsilon \frac{\sigma_i^2}{B_n^2} + t^2 \, \mathrm{E} \Big| \frac{X_i}{B_n} \Big|^2 I\Big\{ \Big| \frac{X_i}{B_n} \Big| \ge \varepsilon \Big\}.
\end{aligned}$$

(c) Show that this leads to

$$\sum_{i=1}^n \big| \phi_i(t/B_n) - (1 - \tfrac{1}{2}\sigma_i^2 t^2/B_n^2) \big| \le \tfrac{1}{6} |t|^3 \varepsilon + t^2 \, L_n(\varepsilon),$$

and via the start lemma of the previous exercise that this secures what we were after, that $\prod_{i=1}^n \phi_i(t/B_n) \to \exp(-\frac{1}{2}t^2)$ and hence triumphantly $Z_n \to_d \mathrm{N}(0,1)$, under the Lindeberg condition only.

63. Convergence in Euclidean space

[xx spelling out the basics for $X_n \to_d X$ in $\mathcal{R}^k$. The Portmanteau Lemma holds, with the required modifications. Also, $X_n \to_d X$ is equivalent to

$$\phi_n(t) = \mathrm{E} \exp(it^{\mathrm{t}} X_n) \to \phi(t) = \mathrm{E} \exp(it^{\mathrm{t}} X) \quad \text{for all } t \in \mathcal{R}^k.$$

show that if $X \sim \mathrm{N}_k(0, \Sigma)$, then

$$\phi(t) = \exp(-\tfrac{1}{2} t^{\mathrm{t}} \Sigma t).$$

a simple example or two. xx]

64. The Cramér–Wold device

Consider random vectors X_n and X in $\mathcal{R}^k$. Using the characterisations of convergence of distributions via characteristic functions, show that $X_n \to_d X$ if and only if all linear combinations converge appropriately, i.e. $a^{\mathrm{t}} X_n \to_d a^{\mathrm{t}} X$ for all a. This is called the Cramér–Wold device, from Harald Cramér and Herman Wold (1936).

(a) Prove the k-dimensional Central Limit Theorem: if $X_1, X_2, \ldots$ are i.i.d. in $\mathcal{R}^k$ with finite variance matrix $\Sigma = \mathrm{E}\,(X-\xi)(X-\xi)^{\mathrm{t}}$, then

$$Z_n = \sqrt{n}(\bar{X}_n - \xi) \to_d \mathrm{N}(0, \Sigma).$$

(b) Let $X_1, X_2, \ldots$ be i.i.d. from the unit exponential distribution. Find first the limit distributions of $\sqrt{n}(n^{-1}\sum_{i=1}^n X_i - 1)$ and $\sqrt{n}(n^{-1}\sum_{i=1}^n X_i^2 - 2)$. Then find the joint limit distribution of

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - 1) \\ \sqrt{n}(W_n - 2) \end{pmatrix},$$

with $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $W_n = n^{-1}\sum_{i=1}^n X_i^2$, and also the limit distribution of $\sqrt{n}(W_n/\bar{X}_n - 2)$.

(c) Suppose $X_1, X_2, \ldots$ are independent with mean zero and variance matrices $\Sigma_1, \Sigma_2, \ldots$; their distributions are here not assumed to be equal. Find suitable conditions, of the Lyapunov or Lindeberg type, which secure limiting normality of $\sum_{i=1}^n X_i$, suitably normalised.

65. The sample mean and standard deviation

Consider i.i.d. data $X_1, \ldots, X_n$, from which we compute the classical

$$\widehat{\xi} = \bar{X} = n^{-1}\sum_{i=1}^n X_i \quad \text{and} \quad \widehat{\sigma} = \Big\{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2\Big\}^{1/2}.$$

These are of course estimators for the underlying mean ξ and standard deviation σ. Here we derive their joint limit distribution, after which the delta method may be called upon to deduce approximate distributions for several quantities of interest.

(a) Make sure you understand and can prove that $\widehat{\xi}$ and $\widehat{\sigma}$ are strongly consistent for ξ and σ, assuming only that the standard deviation is finite.

(b) Assume now that also the fourth order moment is finite. Use

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \xi)^2 - (\bar{X} - \xi)^2$$

to deduce that

$$\sqrt{n}(S_n^2 - \sigma^2) \quad \text{and} \quad \sqrt{n}\Big\{n^{-1}\sum_{i=1}^n (X_i - \xi)^2 - \sigma^2\Big\}$$

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

$$\gamma_4 = \mathrm{E}\,\{(X_i - \xi)/\sigma\}^4 - 3.$$

The 'subtract 3' is merely a thing of mild convenience, making the kurtosis equal to zero for normal distributions.

(c) A minor kjepphest of mine is that statisticians should work with and tell stories about standard deviations, not variances – nobody should say 'my variance is 64 square metres' when the point, regarding interpretation and communication, is that the standard deviation is 8 metres. So let's transform the above, from variance to its square root, getting back to the real scale of the measurements: show that

$$\sqrt{n}(\widehat{\sigma}-\sigma) \to_d \mathrm{N}(0, (\tfrac{1}{2}+\tfrac{1}{4}\gamma_4)\sigma^2).$$

(d) Show that

$$\widehat{\gamma}_4 = n^{-1}\sum_{i=1}^{n}\Big(\frac{X_i-\bar{X}}{\widehat{\sigma}}\Big)^4 - 3$$

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.

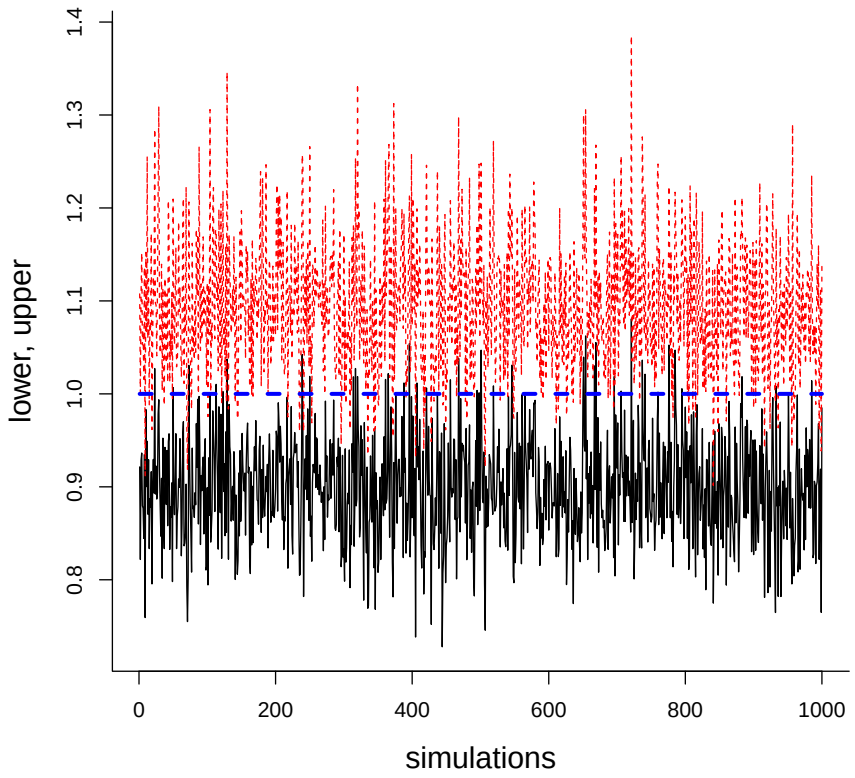

Figure 0.4: Simulations, with datasets of size $n = 500$ from the unit exponential, displaying lower and upper confidence points; the intervals attempt to cover the true value $\sigma = 1$.

(e) Ok, let's bother enough to do it, it's a useful and not too hard simulation exercise. 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

(e.g. via `rexp` in `R`). 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. Make suitable diagrams to summarise what you find, and examine in particular the coverage of your intervals – how often do they contain the correct σ? See Figure 0.4.

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

$$\begin{pmatrix} \sqrt{n}(\bar{X}-\xi) \\ \sqrt{n}(S_n^2-\sigma^2) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^3\gamma_3 \\ \sigma^3\gamma_3 & \sigma^4(2+\gamma_4) \end{pmatrix}\Big),$$

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 $n = 444$ 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.

66. Functions of the sample mean and standard deviation

With full large-sample control for the joint behaviour of sample mean and standard deviation, from the previous exercise, we may deduce approximations for a long list of interesting functions of these.

(a) In the situation above, with $X_1, \ldots, X_n$ being i.i.d. from some distribution with finite fourth moment, show that if $g(\xi, \sigma)$ is any smooth function of these two parameters, then

$$\sqrt{n}\{g(\widehat{\xi},\widehat{\sigma}) - g(\xi,\sigma)\} \to_d Z = \frac{\partial g(\xi,\sigma)}{\partial \xi} A + \frac{\partial g(\xi,\sigma)}{\partial \xi} B,$$

in which

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3 & \frac{1}{2}+\frac{1}{4}\gamma_4 \end{pmatrix}\Big).$$

Why is Z a zero-mean normal?

(b) Consider the parameter $\delta = \xi/\sigma$, with estimator $\widehat{\delta} = \widehat{\xi}/\widehat{\sigma}$. Find the limit distribution for $\sqrt{n}(\widehat{\delta}-\delta)$, and construct a confidence interval for δ.

(c) For this point assume that the distribution is normal, and verify that

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

Find the limit distribution for $\sqrt{n}(\widehat{\delta}-\delta)$ in this case, and check how your confidence construction simplifies. Comment on the off-diagonal zero for the covariance matrix.

(d) Still under normality, consider the threshold probability

$$p = \Pr\{X_{n+1} \le x_0\} = \Phi\Big(\frac{x_0-\xi}{\sigma}\Big)$$

for some x_0. Find the limit distribution for $\sqrt{n}(\widehat{p}-p)$, with $\widehat{p} = \Phi((x_0-\widehat{\xi})/\widehat{\sigma})$. Compare your result to that for the simple binomial procedure which does not care about normality, but merely takes $\widetilde{p} = F_n(x_0)$, the relative frequency of data points below x_0. Comment on your findings.

67. Are your count data overdispersed?

Everyone in the room knows that for the Poisson distribution, the variance is equal to the mean. It is not uncommon for count data to display a bit more variability than what the Poisson assumption points to, however. In this exercise we construct a test for Poisson-ness of a dataset, by checking if the empirical variance is too big compared to the empirical mean.

(e) For X having a Poisson distribution with parameter θ, show that

$$\begin{aligned} \mathrm{E}\, X &= \theta, \\ \mathrm{E}\, X(X-1) &= \theta^2, \\ \mathrm{E}\, X(X-1)(X-2) &= \theta^3, \\ \mathrm{E}\, X(X-1)(X-3)(X-4) &= \theta^4, \end{aligned}$$

tenk det, and deduce from these formulae not merely for $\mathrm{E}\, X = \theta$ and $\mathrm{Var}\, X = \theta$, but also for

$$\gamma_3 = \mathrm{E}\Big(\frac{X-\theta}{\sqrt{\theta}}\Big)^3 = \frac{1}{\theta^{1/2}} \quad \text{and} \quad \gamma_4 = \mathrm{E}\Big(\frac{X-\theta}{\sqrt{\theta}}\Big)^4 - 3 = \frac{1}{\theta}.$$

(f) Suppose $X_1, \ldots, X_n$ are i.i.d. from the Poisson, and compute from the sample the usual $\bar{X}$ and empirical variance S_n^2. Show that

$$\begin{pmatrix} \sqrt{n}(\bar{X}-\theta) \\ \sqrt{n}(S_n^2-\theta) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \theta & \theta \\ \theta & 2\theta^2+\theta \end{pmatrix}\Big).$$

What is the limiting correlation, between $\bar{X}$ and S_n^2?

(g) There is often *overdispersion* in count data, with variance somewhat bigger than the mean (see Hjort's FocuStat Blog Post, 2018b). Show that if the data really come from a Poisson, then

$$\sqrt{2n}(S_n^2/\bar{X} - 1) \to_d \mathrm{N}(0,1),$$

and use this to build a test for Poisson-ness against overdispersion.

68. Correlation measures

Ferguson's book has a separate section with analysis of the classical empirical correlation coefficient R_n, yielding the limit distribution of $\sqrt{n}(R_n - \rho)$, etc. The present exercise considers a couple of simpler related situations, with simpler in the sense of adding more modelling assumptions. In the following, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. from some joint distribution, where X_i and Y_i have finite fourth moments.

(a) For deriving certain moment formulae, for the case where the (X_i, Y_i) have a binormal distribution, the following is useful. Assume (X_0, Y_0) has the binormal distribution with means zero and standard deviations one, and correlation $\rho = \mathrm{corr}(X_0, Y_0)$. Show that $Y_0 \,|\, x_0 \sim \mathrm{N}(\rho x_0, 1-\rho^2)$. Use this to show that

$$\mathrm{E}\, X_0^2 Y_0^2 = \mathrm{E}\, \mathrm{E}\,(X_0^2 Y_0^2 \,|\, X_0) = \mathrm{E}\, X_0^2(\rho^2 X_0^2 + 1 - \rho^2) = 1 + 2\rho^2,$$

and find with similar type of efforts formulae for

$$\mathrm{E}\, X_0^3 Y_0, \quad \mathrm{E}\, X_0 Y_0^3, \quad \mathrm{E}\, X_0^4 Y_0, \quad \mathrm{E}\, X_0 Y_0^4, \quad \mathrm{E}\,(X_0 - Y_0)^3, \quad \mathrm{E}\,(X_0 - Y_0)^4.$$

(b) Assume first that the means ξ_1, ξ_2 are zero and the standard deviations σ_1, σ_2 are one. With $\widehat{\rho}_b = n^{-1}\sum_{i=1}^n X_i Y_i$ a natural estimator of $\rho = \mathrm{E}\, XY$, show that $\sqrt{n}(\widehat{\rho}_a - \rho)$ has a $\mathrm{N}(0, \tau_b^2)$ limit distribution. Give a suitable expression for τ_b^2, and find what τ_b^2 is in the case of the underlying distribution for (X_i, Y_i) being binormal.

(c) Next consider the setup where the means are known to be zero, the standard deviations taken to be equal, but unknown. The natural correlation estimator is then

$$\widehat{\rho}_c = n^{-1}\sum_{i=1}^n \frac{X_i Y_i}{\widetilde{\sigma}_c^2}, \quad \text{with} \quad \widetilde{\sigma}_c^2 = \tfrac{1}{2}(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2),$$

in terms of $\widetilde{\sigma}_1^2 = n^{-1}\sum_{i=1}^n X_i^2$ and $\widetilde{\sigma}_2^2 = n^{-1}\sum_{i=1}^n Y_i^2$. Show that this estimator is strongly consistent for ρ, and find the limit distribution $\mathrm{N}(0, \tau_c^2)$ for $\sqrt{n}(\widehat{\rho}_c - \rho)$, both under general conditions and under the specific extra assumption of binormality. Comment on τ_c versus τ_b.

(d) Then work out what happens in the more general situation where the means are known and equal to zero, but where the correlation ρ as well as the standard deviations σ_1 and σ_2 are unknown. The natural estimator is then

$$\widehat{\rho}_d = n^{-1}\sum_{i=1}^n \frac{X_i Y_i}{\widetilde{\sigma}_1 \widetilde{\sigma}_2},$$

with $\widetilde{\sigma}_1$ and $\widetilde{\sigma}_2$ as above. In other words, find expressions for the limiting standard deviation in for $\sqrt{n}(\widehat{\rho}_d - \rho) \to_d \mathrm{N}(0, \tau_d^2)$, both under general conditions and under binormality.

(e) Finally do the Full General Story, where the five parameters in question are unknown, and where everyone uses the classic

$$R_n = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = n^{-1}\sum_{i=1}^n \frac{(X_i - \widehat{\xi}_1)(Y_i - \widehat{\xi}_2)}{\widehat{\sigma}_1 \widehat{\sigma}_2},$$

in terms of the usual empirical means and standard deviations. Show that this situation is actually not genuinely more complicated than under (d), so in a sense the work has been done; one does not earn precision, for large n, by knowing the means.

(f) Conclude from your efforts above that $\sqrt{n}(R_n - \rho) \to_d \mathrm{N}(0, (1-\rho^2)^2)$ under binormality. Use this to also show that

$$\sqrt{n}(\widehat{\zeta} - \zeta) \to_d \mathrm{N}(0,1), \quad \text{where} \quad \zeta = \tfrac{1}{2}\log\frac{1+\rho}{1-\rho} \text{ and } \widehat{\zeta} = \tfrac{1}{2}\log\frac{1+R_n}{1-R_n}.$$

This is Fisher's *variance stabilising transformation* for the correlation coefficient. Once upon a time, Florence Nightingale Davis carried out numerical work to ascertain that this transformation also affords better approximation to normality, for moderate to low sample sizes; her approximation is $\widehat{\zeta} \approx_d \mathrm{N}(\zeta, 1/(n-3))$. This makes statistical inference for the binormal correlation parameter easy.

(g) [xx nils puts in a bit more here, in a little while. xx]

69. The Karl Pearson statistic and the chi-squared

Isn't it a glorious & rather informative title, for a journal article? In 1900, Karl Pearson (1857–1936) published the deservedly famous *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 in *Philosophical Magazine, Series 5*. (a) He invents a very useful general test, to check whether a probability vector is equal to a set of specified values; (b) he shows that the test statistic can be approximated with a new distribution, which is the first-ever published chi-squared distribution, which conveniently does not depend on the specified probability vector, but only the number of boxed under consideration; and (c) he develops logically sound arguments for when should keep one's theory, and when one should reject it. In yet other words, he invents the notion of statistical testing, via a test statistic, which he shows has a limit distribution, and he almost touches on p-values. In one of perhaps several nutshells, Pearson builds a full apparatus *to test a given theory*.

The notes below are supplements to Ferguson's brief treatment. Let $N = (N_1, \ldots, N_k)$ be a multinomial vector, with n independent draws for k given boxes, and probability vector $p = (p_1, \ldots, p_k)$. A favourite example to point to is to roll your die n times, count the numbers $(N_1, \ldots, N_6)$ of the different outcomes 1, 2, 3, 4, 5, 6; if your die is fair, this is a multinomial vector with $p = (1/6, \ldots, 1/6)$.

(a) Show that each N_j is binomial, with $N_j \sim \mathrm{Bin}(n, p_j)$. Hence $\mathrm{E}\, N_j = np_j$ and $\mathrm{Var}\, N_j = np_j(1-p_j)$.

(b) It's actually not necessarily important to know the formula for the joint distribution of the $(N_1, \ldots, N_k)$, but please check that you both understand and may derive the formula

$$f(N_1, \ldots, N_6) = \frac{n!}{N_1! \cdots N_6!} p_1^{N_1} \cdots p_k^{N_k} \quad \text{for } N_1 \geq 0, \ldots, N_k \geq 0, N_1 + \cdots + N_k = n.$$

(c) Write

$$\begin{aligned} N_1 &= Y_{1,}1 + \cdots + Y_{1,n}, \\ N_2 &= Y_{2,}1 + \cdots + Y_{2,n}, \\ &\cdots \\ N_k &= Y_{k,}1 + \cdots + Y_{k,n}, \end{aligned}$$

or more compactly $N = Y_1 + \cdots + Y_n$ with Y_j the vector of length k, with 0-s and 1-s, for trial j. It can take the values $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$, with probabilities $p_1 \ldots, p_k$. Show that

$$\mathrm{E}\, Y_j = p \quad \text{and} \quad \mathrm{Var}\, Y_j = \Sigma,$$

where Σ is a matrix of size $k \times k$, with elements $p_j(1-p_j)$ on the diagonal and $-p_j p_l$ outside. It is convenient to write the (j, l) element as $\delta_{j,l} p_j - p_j p_l$, where $\delta_{j,l}$ is the Leopold Kronecker delta ("Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk"), equal to 1 if $j = l$ and 0 if else.

(d) Write $\widehat{p} = N/n = \bar{Y}_n$, with components $\widehat{p}_j = N_j/n$. With

$$X_n = \sqrt{n}(\bar{Y}_n - p) = \sqrt{n}(\widehat{p} - p),$$

show that $X_n \to_d X \sim \mathrm{N}_k(0, \Sigma)$. Note that Σ is not invertible, since $p_1 + \cdots + p_k = 1$, and show that indeed $\sum_{j=1}^k X_j = 0$.

(e) Now introduce the super-famous Pearson statistic,

$$K_n = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} = \sum_{j=1}^k \frac{(\text{obs}_j - \text{exp}_j)^2}{\text{exp}_j} = \sum_{j=1}^k \frac{n(\widehat{p}_j - p_j)^2}{p_j},$$

with the familiar ratios of $(\text{obs}_j - \text{exp}_\text{j})^2/\text{exp}_j$, involving 'observed' and 'expected' numbers. Show that

$$K_n = \sum_{j=1}^k \frac{X_{n,j}^2}{p_j} \to_d K = \sum_{j=1}^k \frac{X_j^2}{p_j},$$

with the $X \sim \text{N}_k(0, \Sigma)$ above. This is 'the main job' (now accomplished); the rest of the story is to demonstrate that this K has a χ^2_{k-1} distribution. Show, directly, that $\text{E}\, K = k-1$.

(f) For the case of $k = 3$ boxes, start with the smaller 2×2 submatrix Σ_0, and show that

$$\begin{aligned}\Sigma_0^{-1} &= \begin{pmatrix} p_1(1-p_1) & -p_1p_2 \\ -p_1p_2 & p_2(1-p_2) \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 1/p_1 + 1/p_3 & 1/p_3 \\ 1/p_3 & 1/p_2 + 1/p_3 \end{pmatrix} = \begin{pmatrix} 1/p_1 & 0 \\ 0 & 1/p_2 \end{pmatrix} + \frac{1}{p_3} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.\end{aligned}$$

Show that

$$X_1^2/p_1 + X_2^2/p_2 + X_3^2/p_3 = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}^\text{t} \Sigma_0^{-1} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}^\text{t}.$$

Show from this that $K \sim \chi^2_2$, for this case of $k = 3$ boxes.

(g) Generalise the arguments above. For the $(k-1) \times (k-1)$ submatrix Σ_0, show that

$$\Sigma_0 = D_0 - p_0 p_0^\text{t},$$

where D_0 is diagonal with $p_0 = (p_1, \ldots, p_{k-1})^\text{t}$ on its diagonal. Use this to show that

$$\Sigma_0^{-1} = D_0^{-1} + (1/p_k)\mathbf{1}\mathbf{1}^\text{t},$$

where $\mathbf{1}$ is the $(k-1)$-length vector $(1, \ldots, 1)^\text{t}$. Show that $K = X_0^\text{t} \Sigma_0^{-1} X_0$, and conclude, as Pearson did some 120 years ago, but with other words and symbols and arguments and thoughts that presently in your head, that $K \sim \chi^2_{k-1}$.

(h) An alternative to the classic K_n is

$$K_n' = \sum_{j=1}^k \frac{(N_j - np_j)^2}{N_j} = \sum_{j=1}^k \frac{(\text{obs}_j - \text{exp}_j)^2}{\text{obs}_j} = \sum_{j=1}^k \frac{n(\widehat{p}_j - p_j)^2}{\widehat{p}_j},$$

i.e. using observed and not expected in the denominator. Show that K_n' and K_n must have identical limit distributions; hence $K_n' \to_d \chi^2_{k-1}$ too. [xx nils: check reference Laake et al. book, what they write about such matters. xx]

(i) Presumably 'all students' in beginning statistics courses are told to memorise the $(\text{obs}_j - \text{exp}_\text{j})^2/\text{exp}_j$ formula. If tired with this, why not do the presumably also clever root variant,

$$L_n = \sum_{j=1}^k \frac{|\text{obs}_j - \text{exp}_\text{j}|}{\sqrt{\text{exp}_j}} = \sum_{j=1}^k \frac{|N_j - np_j|}{\sqrt{np_j}} = \sum_{j=1}^k \frac{\sqrt{n}|\widehat{p}_j - p_j|}{p_j^{1/2}}.$$

Show that $L_n \to_d L = \sum_{j=1}^k |X_j|/p_j^{1/2}$. Find an expression for its mean. Speculate on useful alternatives.

(j) With $h(p) = 2\arcsin\sqrt{p}$, show that the transformation stabilises the variance, in the sense that

$$\sqrt{n}\{h(\widehat{p}_j) - h(p_j)\} \to_d h'(p_j)X_j \sim \mathrm{N}(0,1).$$

This is at least very fine for each individual p_j. What are the limit distributions for

$$\sum_{j=1}^{k} \sqrt{n}|h(\widehat{p}_j) - h(p_j)| \quad \text{and} \quad \sum_{j=1}^{k} n\{h(\widehat{p}_j) - h(p_j)\}^2,$$

as n increases?

70. Estimating f

Suppose $X_1, \ldots, X_n$ are i.i.d. from some density f? Well, if a parametric model is thought to fit well, we may use the ensuing $f(x, \widehat{\theta})$, but without such additional assumptions it's not entirely clear how to do it, nor how well the nonparametric job can be done.

(a) It's in several ways easier to estimate the cumulative F nonparametrically, where the natural method is that of the empirical distribution function (try `ecdf` in `R`),

$$F_n(t) = n^{-1} \sum_{i=1}^{n} I\{X_i \le t\}.$$

This is simply the binomial estimator, counting the number of $X_i \le t$. Show that $\mathrm{E}\, F_n(t) = F(t)$, that its variance is $n^{-1}F(t)\{1 - F(t)\}$, and also that

$$Z_n(t) = \sqrt{n}\{F_n(t) - F(t)\} \to_d Z(t) \sim \mathrm{N}(0, F(t)\{1 - F(t)\}).$$

Later on we shall learn more about this empirical process and its full convergence to a full stochastic process $Z = \{Z(t) \colon t \in \mathcal{R}\}$.

(b) Since f is the derivative of F, consider

$$f_n(t) = \frac{F_n(t+h) - F_n(t-h)}{2h},$$

for a 'suitably small' h. Find expressions for the precise mean and variance of $f_n(t)$.

(c) It's not enough to say 'let $h \to 0$' above, since the variance will then explode. Show in fact that if $h \to 0$ and $nh \to \infty$, then both the bias and variance go to zero, and that this impies $f_n(t) \to_{\mathrm{pr}} f(t)$ for each t.

(d) Try it out – simulate $n = 500$ points from e.g. $f = 0.50\,\mathrm{N}(-1,1) + 0.50\,N(1,1)$, then compute and plot the density estimate function $f_n(t)$, as above, with $\varepsilon = c/\sqrt{n}$, where you can attempt to finetune the c in question. Incidentally, don't cheat, please, when you simulate 500 points from the bimixture here; don't just take 250 points from each of the two components.

71. Density estimation: more!

Here I spell out a bit more regarding the problem of estimating the density f underlying an observed sample $x_1, \ldots, x_n$. The topic of density estimation is a very large one, see e.g. Hjort and Glad (1995), Hjort and Jones (1996). First, there are many methods out there, and yet to be invented, for estimating f, and, secondly, each of these methods have smoothing or finetuning parameters, and the accurate setting of these is often complicated and delicate. The intention here is to show 'the basics', for the kernel density estimation method, with easy conditions for consistency.

(a) Let $K(u)$ be a density, symmetric around zero, e.g. the standard normal, with finite values of

$$k_2 = \int u^2 K(u)\,\mathrm{d}u \quad \text{and} \quad R(K) = \int K(u)^2\,\mathrm{d}u.$$

For the normal choice $K = \phi$, show that $k_2 = 1$ and $R(K) = \phi(0)/\sqrt{2} = 1/(2\sqrt{\pi}) = 0.2821$. The K is our kernel function.

(b) Our kernel density estimator is

$$f_n(x) = n^{-1}\sum_{i=1}^{n} K_h(x_i - x), \quad \text{where} \quad K_h(u) = h^{-1}K(h^{-1}u).$$

The idea is to let h tend slowly to zero with increasing n. Show that f_n is a density function, and work out that

$$\mathrm{E}\, f_n(x) = \int K_h(x' - x) f(x')\,\mathrm{d}x' = \int K(u) f(x + hu)\,\mathrm{d}u.$$

Show that the bias of the estimator tends to zero if $h \to 0$.

(c) Then, assuming f has at least two continuous derivatives, use Taylor expansion $f(x + hu) \doteq f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2$ to show that

$$\mathrm{E}\, f_n(x) = f(x) + \tfrac{1}{2}k_2 h^2 f''(x) + O(h^3).$$

Show similarly that

$$\mathrm{Var}\, f_n(x) = \frac{R(K)}{nh} f(x) + O(h/n).$$

(d) Show that $f_n(x)$ is consistent for $f(x)$ provided $h \to 0$ and $nh \to \infty$. So, if $h = cn^{-a}$, we need $a \in (0, 1)$.

(e) So what's the wisest choice of the bandwidth h? That's a somewhat tricky question to sort out fully (there are a few hundred technical journal articles about that topic), but start by working with the approximate mean squared error at position x, say

$$\mathrm{mse}(x) = \mathrm{bias}^2(x) + \mathrm{var}(x) \doteq \tfrac{1}{4}k_2^2 h^4 f''(x)^2 + \frac{R(K)}{nh} f(x).$$

Show in general terms that $ah^4 + b/(nh)$ becomes smallest for h of rate $n^{-1/5}$, with minimal value of size $n^{-4/5}$. This is already an important finding, that one cannot achieve the usual $1/n$ rate, for parametric problems, but must be content with $n^{-0.80}$ for smooth nonparametric problems.

(f) Show, with more detail, that the best bandwidth, using the approximate mse above, becomes

$$h_n^*(x) = \Big\{\frac{R(K)}{k_2^2}\Big\}^{1/5}\Big\{\frac{f(x)}{f''(x)}\Big\}^{1/5}\frac{1}{n^{1/5}}.$$

Compute and display this $h_n^*(x)$ for the case of f being the standard normal.

(g) It is fully ok to use a perhaps complicated bandwidth $h = h_n(x)$ depending on the position x, but with a complicated rule for this one risks messing up the overall performance. Also for that reason it is customary to select one h to be used for all x. Consider the mean integrated squared error

$$\text{mise} = \int \text{mse}(x)^2 \, \mathrm{d}x = \int \{\text{bias}(x)^2 + \text{var}(x)\} \, \mathrm{d}x,$$

and show that this may be approximated with

$$\text{mise} \doteq \tfrac{1}{4} k_2^2 h^4 R(f'') + \frac{R(K)}{nh},$$

with $R(f'') = \int (f'')^2 \, \mathrm{d}x$ sometimes called the roughness of the density.

(h) Show that the best bandwidth, in the sense of minimising the approximate mise, is

$$h_n^* = \Big\{ \frac{R(K)}{k_2^2} \Big\}^{1/5} R(f'')^{-1/5} \frac{1}{n^{1/5}}.$$

Find also an expression for the corresponding best possible mise, and note the crucial aspect that this quantity goes to zero with n at the speed of $1/n^{4/5} = 1/n^{0.80}$. This is the price to pay for being nonparametric, compared to the parametric rates $1/n$.

(i) For the case of f being a classic normal $\mathrm{N}(\xi, \sigma^2)$, show that

$$R(f'') = \frac{3}{8\sqrt{\pi}} \sigma^{-5}.$$

This leads to a 'rule of thumb' density estimator: use the kernel density estimator f_n, with the normal kernel, and bandwidth

$$h = (4/3)^{1/5} \widehat{\sigma} / n^{1/5} = 1.0592 \, \widehat{\sigma} / n^{1/5},$$

with a suitable robust estimate for the standard deviation of the data.

72. Convergence of means

well

73. The last time for estimator functionals

[xx point to Steffen Grønneberg's master thesis and later paper, and also Hjort and Fenstad (1992). xx]

74. Confidence ellipsoids

well

75. The arctan estimator

[xx the exercise from emil stoltenberg's exam set, 2016, with a little more. xx]

76. The score function, the information function, and the Bartlett identity

Consider a parametric density model $f(y, \theta)$, where $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{t}}$, the parameter of the model, is contained in some open parameter region Ω. Introduce

$$u(y, \theta) = \frac{\partial \log f(y, \theta)}{\partial \theta} \quad \text{and} \quad i(y, \theta) = \frac{\partial^2 \log f(y, \theta)}{\partial \theta \partial \theta^{\mathrm{t}}},$$

called the *score function*, with p components, and the *information function*, a $p \times p$ matrix. These partial derivatives are assumed to exist and be continuous; note that this concerns smoothness in the parameter θ, not necessarily smoothness in y. We also 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 $[0, \theta]$.

(a) Show that the score function has mean zero, i.e.

$$\mathrm{E}_\theta\, u(Y,\theta) = \int f(y,\theta) u(y,\theta)\,\mathrm{d}y = 0.$$

(b) Let next

$$J(\theta) = -\mathrm{E}_\theta\, i(Y,\theta) \quad \text{and} \quad K(\theta) = \mathrm{Var}_\theta\, u(Y,\theta),$$

and show that indeed $J(\theta) = K(\theta)$, the so-called Bartlett identity. This matrix is often called *Fisher's information matrix* for the model. Note that the calculation of both $J(\theta)$ and $K(\theta)$ is taking place under the assumption that the model is actually correct.

(c) For the exponential model, with density $\theta \exp(-\theta y)$, find the score function, and compute the Fisher information function in two ways.

(d) For the normal $\mathrm{N}(\xi, \sigma^2)$ model, show that the score function can be expressed as

$$u(y,\xi,\sigma) = \begin{pmatrix} \frac{1}{\sigma}\frac{y-\xi}{\sigma} \\ \frac{1}{\sigma}\left\{\left(\frac{y-\xi}{\sigma}\right)^2 - 1\right\} \end{pmatrix} = \frac{1}{\sigma}\begin{pmatrix} z \\ z^2-1 \end{pmatrix},$$

writing $z = (y-\xi)/\sigma$, which is a standard normal when y comes from the model. Demonstrate that the Fisher information matrix becomes

$$J(\xi,\sigma) = \mathrm{Var}_{\xi,\sigma}\, u(Y,\xi,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

(e) 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.

77. Behaviour of the maximum likelihood estimator, under model conditions

Let $Y_1, \ldots, Y_n$ be independent from the same density $f(y,\theta)$, where $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{t}}$. As in the previous exercise, let $u(y,\theta)$ 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 $A(\theta)$ 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.

(b) Taylor-expand $U_n(\widehat{\theta})$ around θ_0 to show

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \{-n^{-1}I_n(\widetilde{\theta})\}^{-1}n^{-1/2}U_n(\theta_0),$$

where $\widetilde{\theta}$ is somewhere between θ_0 and $\widehat{\theta}$. Why does

$$n^{-1/2}U_n(\theta_0) \to_d U \sim \mathrm{N}_p(0, J(\theta_0),$$

and why will $-n^{-1}I_n(\theta_0) \to_{\mathrm{pr}} J(\theta_0)$?

(c) Deduce that

$$\sqrt{n}(\widehat{\theta}_n-\theta_0) \to_d J(\theta_0)^{-1}U \sim \mathrm{N}_p(0, J(\theta_0)^{-1}).$$

This is the celebrated theorem on the large-sample behaviour of ML estimates (under model conditions). – What regularity conditions do you need to construct a rigorous proof?

(d) Check that you understand (and can use) the delta method consequence of the above: if $\gamma = g(\theta)$ is some parameter of interest, a smooth function of the basic model parameter vector, then $\widehat{\gamma} = g(\widehat{\theta})$ is the ML estimator, and

$$\sqrt{n}(\widehat{\gamma}-\gamma) \to_d c^{\mathrm{t}}J(\theta_0)^{-1}U \sim \mathrm{N}(0,\tau^2),$$

with $\tau^2 = c^{\mathrm{t}}J(\theta_0)^{-1}c$. Here $c = \partial g(\theta_0)/\partial\theta$.

(e) 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.

(f) Construct an approximate 90 percent confidence ellipsoïd for the unknown parameter vector. [Recall that if $X \sim \mathrm{N}_p(\mu,\Sigma)$, then $(X-\mu)^{\mathrm{t}}\Sigma^{-1}(X-\mu)$ is χ^2 distributed with p degrees of freedom.] Can you prove that your chosen region has the minimal possible volume, among all asymptotic 90 percent confidence regions for θ?

78. 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

$$d(g,f) = \int g\log\frac{g}{f}\,\mathrm{d}y.$$

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), 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) 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.

(e) The above few examples led to KL distances being symmetric, between the two densities in question, but this is more unntak than regel. Compute the KL distance from $\mathrm{N}(\xi, \sigma_1^2)$ to $\mathrm{N}(\xi, \sigma_2^2)$, and compare to the reciprocal case.

(f) For densities which are not far from each other, start from

$$d(g,f) = -\int g \log\Big\{1 + \Big(\frac{f}{g} - 1\Big)\Big\},$$

and use Taylor expansion to find

$$d(g,f) \approx \tfrac{1}{2} \int g(f/g-1)^2 \, \mathrm{d}y = \tfrac{1}{2}\Big(\int f^2/g \, \mathrm{d}y - 1 \Big).$$

– 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.

79. What is the maximum likelihood aiming for?

Assume independent observations $Y_1, Y_2, \ldots$ become available, from a certain data generating mechanism g, the famous true but typically unknown data density. With a parametric model f_θ, with $f_\theta(y) = f(y,\theta)$, what it the maximum likelihood method aiming for? We learn here that there is a clear answer, intimately connected to the Kullback–Leibler distance from truth to approximation: ML ♡ KL and KL ♡ ML.

(a) Consider the usual log-likelihood function $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i, \theta)$. The framework of Exercise 77 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

$$A_n(\theta) = n^{-1}\ell_n(\theta) \to_{\rm pr} A(\theta) = {\rm E}_g \log f(Y,\theta) = \int g \log f_\theta \, {\rm d}y$$

for each θ.

(b) Note that this involves the Kullback–Leibler distance, since $d(g, f_\theta) = \int g \log g \, {\rm d}y - A(\theta)$. Under reasonable regularity conditions, which we'll be coming back to during lectures, it will then be the case that the maximiser of A_n, which is the ML estimator $\widehat{\theta}$, will tend to the maximiser θ_0 of A, 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) $A_n(\theta) \to_{\rm pr} A(\theta)$ for each θ to (ii) argmax$(A_n) \to_{\rm pr}$ argmax(A). You may also check with Hjort and Pollard (1993), to see simple conditions via convexity, but in many cases the convexity condition is not met.

– So we've uncovered what goes on in the mindset of the maximum likelihood 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

$$d_{\rm min} = \min d(g, f_\theta) = d(g, f(\cdot, \theta_0))$$

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 maximum likelihood 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

$$d(g, f(\cdot, \xi, \sigma)) = \int g \log g \, {\rm d}y + \log \sigma + \tfrac{1}{2} \frac{\sigma_0^2 + (\xi - \xi_0)^2}{\sigma^2},$$

and that this is being minimised, over all (ξ, σ) pairs, for precisely $\xi = \xi_0 = {\rm E}\, Y$ and $\sigma = \sigma_0 = ({\rm Var}\, Y)^{1/2}$.

(d) Let's do a few exercises where the point is to set up a real 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 maximum likelihood 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.

(i) Let $g = 0.33\,\mathrm{N}(-1, 1) + 0.67\,\mathrm{N}(1, 1)$. Find the best normal approximation.

(ii) The Gamma distribution with parameters (a, b) has density $f = \{b^a/\Gamma(a)\} y^{a-1} \exp(-by)$, and the Weibull distribution [note the Swedish pronunciation] with parameters (c, d) has cumulative distribution $F(y) = 1 - \exp\{-(y/c)^d\}$. Let g be a Gamma with parameters $(2.22, 3.33)$. Find the best Weibull approximant, and also the best log-normal approximant.

(iii) Let $g = 0.95\,\text{Expo}(1) + 0.05\,\text{Expo}(0.01)$, 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.

(iv) Invent your own test case.

(e) Suppose data really come from $\text{N}(0.333, \sigma_1^2)$, with $\sigma_1 = 1.111$, where a statistician fits the simpler $\text{N}(0, \sigma^2)$ model. First, find out what happens to the maximum likelihood 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 $A(\sigma)$. Comment on your findings.

80. Behaviour of the maximum likelihood estimator, under agnostic conditions

Luckily, it might be fair to say, maximum likelihood 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 $\{f_\theta : \theta \in \Omega\}$ 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 $\widehat{\theta} \to_{\text{pr}} \theta_0$, the *least false parameter value* (a term invented by Hjort, Hjort believes, see Hjort 1986b, 1992, and now used somewhat frequently in the literature), as judged by the Kullback–Leibler distance $d(g, f_\theta)$. With terms and notation from Exercise 77, establish that the score function has mean zero, at this true parameter value:

$$\text{E}_g\, u(Y, \theta_0) = \int g(y) u(y, \theta_0)\, \text{d}y = 0.$$

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

$$\begin{aligned} J &= -\text{E}_g\, i(Y, \theta_0) = -\int g(y) \frac{\partial^2 \log f(Y, \theta_0)}{\partial\theta\partial\theta^{\text{t}}}, \\ K &= \text{Var}_g\, u(Y, \theta_0) = \int g(y) u(y, \theta_0) u(y, \theta_0)^{\text{t}}\, \text{d}y, \end{aligned}$$

assumed to be finite and positive definite. Verify (again) that under model conditions, these are identical.

(c) In extension of the previous 'under the model' exercise, show that

$$n^{-1/2} U_n(\theta_0) = n^{-1/2} \ell_n'(\theta_0) = n^{-1/2} \sum_{i=1}^n u(Y_i, \theta_0) \to_d U \sim \text{N}_p(0, K),$$

(d) Use arguments similar to an in fact extending those of the previous 'under the model' exercise, to learn that the basic Taylor expanion consequence

$$\sqrt{n}(\widehat{\theta}-\theta_0)=\{-n^{-1}I_n(\widetilde{\theta})\}^{-1}n^{-1/2}U_n(\theta_0),$$

still holds, where $\widetilde{\theta}$ is somewhere between θ_0 and $\widehat{\theta}$.

(e) Show from this that

$$\sqrt{n}(\widehat{\theta}-\theta_0)\to_d J^{-1}U\sim \mathrm{N}_p(0,J^{-1}KJ^{-1}),$$

with the 'sandwich matrix' as the limit distribution variance matrix.

(f) Natural estimators for J and K, needed for estimating the sandwich from data, are

$$\widehat{J}=-\frac{1}{n}\sum_{i=1}^n\frac{\partial^2\log f(y_i,\widehat{\theta})}{\partial\theta\partial\theta^{\mathrm{t}}}\quad\text{and}\quad\widehat{K}=\frac{1}{n}\sum_{i=1}^n u(y_i,\widehat{\theta})u(y_i,\widehat{\theta})^{\mathrm{t}}.$$

Attempt to show in general terms that $n^{-1}\sum_{i=1}^n h(Y_i,\widehat{\theta})\to_d h_0=\mathrm{E}_g\,h(Y,\theta_0)$, which is what required to prove that $\widehat{J}$ and $\widehat{K}$ are consistent for J and K. This is also yields what we need, a consistent estimator of the sandwich.

81. 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

$$\text{skew}=\mathrm{E}\,Z^3\quad\text{and}\quad\text{kurt}=\mathrm{E}\,Z^4-3$$

are finite, with $Z=(Y-\mathrm{E}\,Y)/\mathrm{sd}(Y)=(Y-\xi_0)/\sigma_0$; see Exercise 65. Working with the score function, and the second order derivatives, show that

$$J=\frac{1}{\sigma_0^2}\begin{pmatrix}1,&0\\0,&2\end{pmatrix}\quad\text{and}\quad K=\frac{1}{\sigma_0^2}\begin{pmatrix}1,&\gamma_3\\\gamma_3,&2+\gamma_4\end{pmatrix}.$$

(c) For the ML estimators $\widehat{\xi}$ and $\widehat{\sigma}$, show from this that

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

Note that this is a 'rediscovery' of what we found in Exercise 65, 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]

82. Extension to regression setups

[xx nils spells out, in time, that the essential stories for ML, told above for i.i.d. setups, extend very nicely and conveniently to regression setups, with $f(y_i \,|\, x_i, \theta)$ etc. xx]

83. A log-likelihood function process

Consider i.i.d. observations $Y_1, Y_2, \ldots$ from some density g, with a model $f(y, \theta)$ fitted via maximum likelihood. 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 $\theta_0 = 3.333$. 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) Use Taylor expansion to find

$$A_n(s) = U_n^{\mathrm{t}} s - \tfrac{1}{2} s^{\mathrm{t}} J_s + \varepsilon_n,$$

where

$$U_n = n^{-1/2} \frac{\partial \ell_n(\theta_0)}{\partial \theta} = n^{-1/2} \sum_{i=1}^{n} u(Y_i, \theta_0) \to_d U \sim \mathrm{N}_p(0, K)$$

and

$$J_n = -n^{-1} \frac{\partial^2 \ell_n(\theta_0)}{\partial \theta \partial \theta^{\mathrm{t}}} = -n^{-1} \sum_{i=1}^{n} i(Y_i, \theta_0) \to_d J,$$

and give conditions under which the remained term $\varepsilon_n \to_{\mathrm{pr}} 0$.

(c) Use the 'argmax to argmax principle' to argue that

$$\operatorname{argmax} A_n \to_d \operatorname{argmax} A,$$

and translate this to

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d J^{-1} U.$$

This gives the limiting normality results for maximum likelihood once again, with the sandwich matrix $J^{-1} K J^{-1}$ in general and the Fisher information matrix inverse $J(\theta_0)^{-1}$ under model conditions.

(d) Then use the 'max to max principle to argue that

$$\max A_n \to_d \max A,$$

and translate this to the nice result

$$D_n(\theta_0) = 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d W = U^{\mathrm{t}} J^{-1} U.$$

Show that the mean of the limit is $p^* = \mathrm{Tr}(J^{-1}K)$, and that $W \sim \chi^2_p$ under model conditions. The $D_n(\theta)$ is called the deviance function, and the result reached is called a *Wilks theorem* (there are several of them, hence the 'a'), and we extend this later on to more general setups.

(e) Go back to your ten simulated versions of $A_n(s)$ for the exponential case, where the true $\theta_0 = 3.333$. Use the above results to test the hypothesis that $\theta = 4.444$.

84. The deviance function and the confidence curve

Here we illustrate the Wilks theorem (or slightly more precisely 'the first Wilks theorem) and uses of the deviance function, including construction of *confidence curves*, with more to come.

(a) Assume $Y_1, \ldots, Y_n$ are modelled via the usual normal $\mathrm{N}(\xi, \sigma^2)$. Find a formula for $D_n(\xi_0, \sigma_0)$, and verify that indeed $D_n(\xi_0, \sigma_0) \to_d \chi^2_2$ if these parameters are the correct ones. Note that with a given data set, fitted to the normal, you would be able to compute $D_n(\xi_0, \sigma_0)$ just from working numerically with the log-likelihood function, i.e. you would not necessarily need the explicit formula I ask for here.

(b) Since $D_n(\xi_0, \sigma_0) \to_d \chi^2_2$ at the true value of the parameter pair, we can construct the set

$$C_{0.90} = \{(\xi, \sigma) \colon D_n(\xi, \sigma) \le 4.605\}.$$

Show that the probability that this set will cover the true (ξ, σ) converges to 0.90; it's hence a confidence set (here in dimension two, since the model parameter is two-dimensional). Simulate a set of $n = 50$ normal data, from $(\xi_{\text{true}}, \sigma_{\text{true}}) = (3.33, 0.77)$, and construct and display this confidence set. If I have 100 clever students all doing this, how many of them will have made confidence sets covering $(3.33, 0.77)$?

(c) Consider the simple dataset

```
0.038 0.075 0.091 0.185 0.190 0.347 0.378 0.423 0.482 0.735 0.898 0.933
```

with values in the unit interval. Fit these to the model with cumulative distribution $F(y, \theta) = y^\theta$ on the unit interval. Compute not only the maximum likelihood estimate, but also the full deviance curve

$$D_n(\theta) = 2\{\ell_{n,\max} - \ell_n(\theta)\}.$$

Then construct and plot the *confidence curve*,

$$\mathrm{cc}(\theta, \text{data}) = \Gamma_1(D_n(\theta_0)),$$

with $\Gamma_1(x)$ the `pchisq(x,1)`, the cumulative χ^2_1 function; in other words, produce a version of Figure 0.5. Show that the magical property

$$\Pr_\theta\{\mathrm{cc}(\theta, \text{data}) \le \alpha\} \to \alpha \quad \text{for all } \alpha \in (0, 1).$$

holds at the true model. Hence

$$C_{0.90} = \{\theta \colon \mathrm{cc}(\theta, \text{data}) \le 0.90\}$$

has in the limit probability 0.90 of covering the true θ, etc. So confidence intervals at all levels may be read off from Figure 0.5.

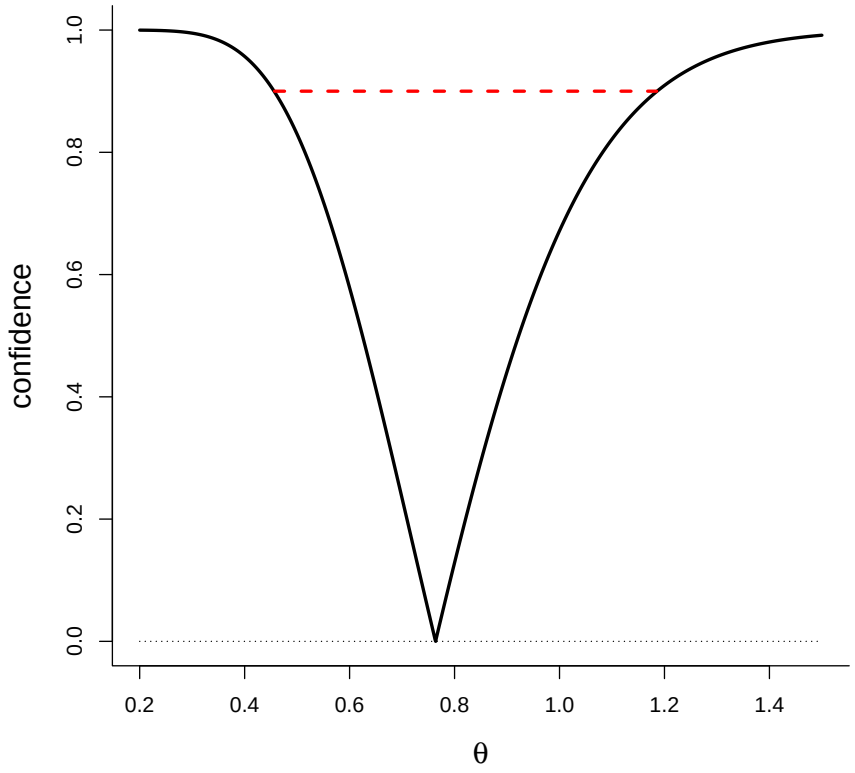

Figure 0.5: Confidence curve $\mathrm{cc}(\theta)$ for the simple dataset of Exercise 84(b). It points to the maximum likelihood estimate 0.764, and confidence intervals at all levels may be read off. The 0.90-line gives $[0.457, 1.185]$.

85. Maximum likelihood analysis in practice!

We have seen above that for a given dataset and parametric model, with a model parameter θ of dimension p, we can carry out maximum likelihood analysis, with (at least) two very useful and versatile results. Under model conditions, with θ_0 denoting the true parameter. these are as follows:

(i) For the maximum likelihood estimator, $\widehat{\theta} \approx_d \mathrm{N}_p(\theta_0, \widehat{J}^{-1})$, where

$$\widehat{J}_{\mathrm{total}} = -\partial^2 \ell_n(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$$

is the Hessian matrix associated with the log-likelihood maximisation, also called the observed Fisher information matrix. Here $\widehat{J} = \widehat{J}_{\mathrm{total}}/n$ is estimating the Fisher information matrix $J(\theta_0)$, so think about $\widehat{J}_{\mathrm{total}}$ as the information matrix for the total dataset, and it grows in size with n. We can hence read off confidence intervals for all model parameters, etc.

(ii) The deviance function

$$D_n(\theta) = 2\{\ell_{n,\max} - \ell_n(\theta)\}$$

can be computed, for any θ, and $D_n(\theta_{\mathrm{true}}) \to_d \chi^2_p$. This can be used for testing, for finding a *confidence set* for the full parameter, etc. For the one-dimensional case this also leads to an easy to construct confidence curve, the

$$\mathrm{cc}_n(\theta, \mathrm{data}) = \Gamma_1(D_n(\theta)),$$

as in Figure 0.5.

We need machinery for handling *a given focus parameter*. In exercises below we come to the *profiled log-likelihood*, and a *generalised Wilks theorem*, but in the present exercise we keep to the structurally simpler ways associated with the delta method. So consider such a focus parameter, say $\gamma = g(\theta) = g(\theta_1, \ldots, \theta_p)$. For a Gamma distribution (a, b), this could be the mean a/b or the standard deviation $a^{1/2}/b$; for the normal $\mathrm{N}(\xi, \sigma^2)$ it could be the quantile $\xi + 1.645\,\sigma$; in a regression setup it could be $\beta_{\mathrm{norway}}/\beta_{\mathrm{sweden}}$, etc.

(a) We do have our friend the delta method on board. It is always useful, but occasionally a bit too rough, and the log-likelihood profiling with Wilks will often be better. But show indeed, or perhaps you've done it before, that with $\widehat{\gamma} = g(\widehat{\theta})$, aiming for the true $\gamma_0 = h(\theta_0)$, we have

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d Z \sim c^{\mathrm{t}}\,\mathrm{N}(0, J(\theta_0)^{-1})$$

where $c = \partial g(\theta_0)/\partial\theta$. Hence $Z \sim \mathrm{N}(0, \tau^2)$, with $\tau^2 = c^{\mathrm{t}} J(\theta_0)^{-1} c$, and we estimate $\tau/\sqrt{n}$ using

$$\widehat{\tau}^2/n = \widehat{c}^{\,\mathrm{t}}\, \widehat{J}_{\mathrm{total}}^{-1} \widehat{c},$$

with $\widehat{c} = \partial g(\widehat{\theta})/\partial\theta$.

(b) Use two minutes to behold the beauty, the general versatility, and the generic practicality of what is told in (a). For each application, as long as you manage to programme the log-likelihood function, it's essentially plain sailing from there: (i) you find the maximiser, the ML; (ii) in the same operation, you find $\widehat{J}_{\mathrm{total}}$, via a suitable `hessian` operation; (iii) when required you also find $\widehat{c}$, via a `grad` operation. The R package `numDeriv` has such `hessian` and `grad` procedures, so you don't need to crank out first and second order derivatives of the log-likelihood function. My own default method for finding the ML in the first place, for a given dataset and model (perhaps one I've just invented, and for which there is no package doing things for me), is to first programme `logL`, then define `minuslogL`, then do

```
nils = nlm(minuslogL,starthere,hessian=T)
```

requiring also a start position `starthere` for the iterative `nlm` algorithm to start working. After this, I can do my pretty generic

```
ML = nils$estimate
Jtotalhat = nils$hessian
se = sqrt(diag(solve(Jtotalhat)))
showme = cbind(ML,se)
print(round(showme,4))
```

With a focus parameter $\gamma = g(\theta)$ I can then more or less start with properly defining my `gg` as a function, and then do

```
gammahat = gg(ML)
chat = grad(gg,ML)
kappahat <- sqrt(chat %*% solve(Jtotalhat) %*% chat)
```

giving me $\widehat{\gamma}$ and its estimated standard deviation $\widehat{\kappa} = \widehat{\tau}/\sqrt{n}$, leading if I wish to $\widehat{\gamma} \pm 1.96\,\widehat{\kappa}$ etc. I can also construct the very useful *first order normal approximation confidence curve*,

which is

$$\mathrm{cc}_n(\gamma,\mathrm{data}) = \Big|1-2\Phi\Big(\frac{\gamma-\widehat{\gamma}}{\widehat{\kappa}}\Big)\Big| = \Big|1-2\Phi\Big(\frac{\gamma-\widehat{\gamma}}{\widehat{\tau}/\sqrt{n}}\Big)\Big|.$$

Show indeed that the two solutions to $\mathrm{cc}_n(\gamma,\mathrm{data})=0.95$ is the familiar $\widehat{\gamma}\pm 1.96\,\widehat{\tau}/\sqrt{n}$, etc., so this is a simple and good diagram from which all confidence intervals can be read off. The (second) Wilks theorem in the next exercise gives another recipe, which tends to be better for smaller sample sizes, when distributions for estimators are skewed, etc.

(c) Get hold of the `egypt-data` set from the course website, comprising life-lengths from Roman era Egypt, a century B.C., for 82 men and 59 women. Use maximum likelihood to fit these data to the Gamma (a,b) model, with density

$$f(y,a,b) = \frac{1}{\Gamma(a)}y^{a-1}\exp(-by) \quad \text{for } y>0.$$

Present parameter estimates and their estimated standard deviations, assuming at least initially that the model is adequate. Produce histograms with the estimated gamma densities on top, to check this.

(d) We then take an interest in $\gamma=\mathrm{E}\,Y$, the expected or mean life-length in ancient Egypt, for men and for women. With the Gamma model this means $\gamma=a/b$. Use the delta method to find estimated standard errors. You should find something like this:

```
men:
  1.4457  0.2056 a
  0.0424  0.0072 b
 34.1203  3.1412 mean a/b
 women:
  2.0632  0.3544 a
  0.0796  0.0155 b
 25.9237  2.3526 mean a/b
```

(e) Test the hypothesis that the mean life-length is the same for men and for the women. If the men had a significantly higher mean life expectancy, attempt to find a plausible explanation. Use also the normal approximations to produce the confidence curves, say $\mathrm{cc}_m(\gamma_m)$ and $\mathrm{cc}_w(\gamma_w)$, from the setup of (b) above. You should hence produce a version of the two full curves of Figure 0.6. I find 95 percent confidence intervals $[27.97, 40.27]$ for men and $[21.32, 30.53]$ for women. – Figure 0.6 also has to accompanying confidence curves, shown as dashed curves. These are produced by the Wilks theorem, using the recipe in the next exercise (see Schweder and Hjort, 2016, and Hjort and Schweder, 2018). As we see the delta method and the Wilks theorem based method yield very similar results, for this particular dataset, and this model.

(f) With the same gamma model, carry out similar analysis for two more focus parameters, namely the standard deviation, $\sigma=a^{1/2}/b$, and the median, $\mu=F^{-1}(\frac{1}{2},a,b)$.

(g) Then attempt to redo these analyses with a different model, namely the Weibull, with cumulative distribution function $F(y)=1-\exp\{-(y/c)^d\}$.

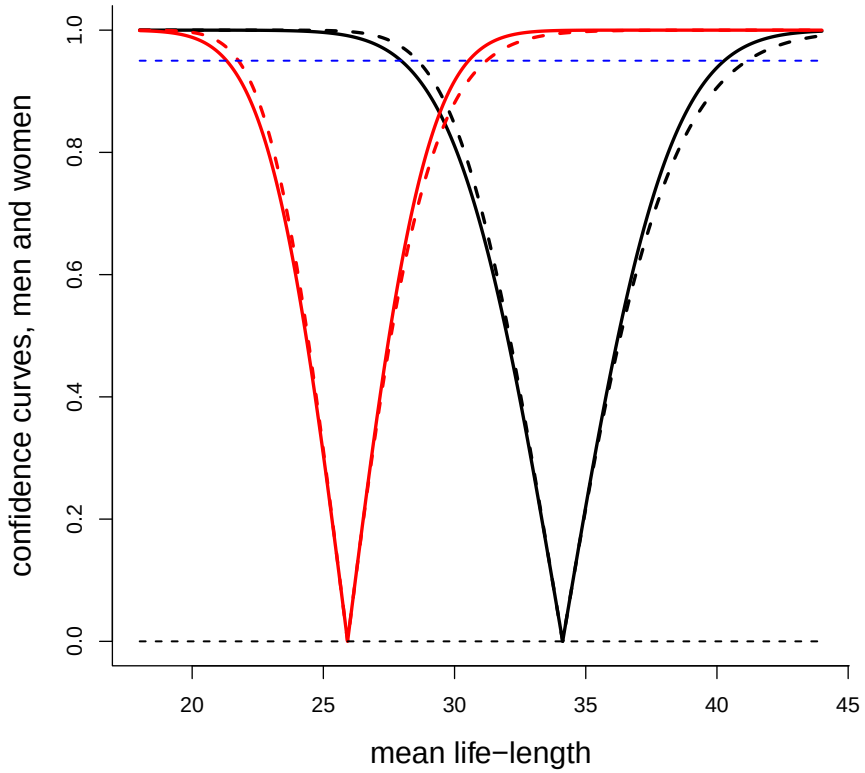

Figure 0.6: Confidence curves $\text{cc}_m(\gamma_m)$ (black) and $\text{cc}_w(\gamma_w)$ (red), for the mean life-length distributions of Roman Era Egypt, based on the gamma model. Full curves: normal approximation; dashed curves: via the chi-squared approximation.

86. The profiled log-likelihood function and the Wilks theorem

In addition to the already rather well-working delta method things, spelled out and illustrated in the previous exercise, it turns out to be very fruitful to generalise the deviance function and the (first) Wilks theorem to the case of a one-dimensional focus parameter $\gamma = g(\theta_1, \ldots, \theta_p)$. This leads to the second and more general Wilks theorem, and to a general recipe for constructing a confidence curve $\text{cc}_n(\gamma, \text{data})$, which often will be more accurate than what the first order normal approximation will provide. An 'early warning' here is that with good data, high enough n, and so on, estimators will be close to normally distributed, limit distribution variances will be well estimated, etc., implying that the first-order normal approximations apparatus of the previous exercise will be well-working. In particular, the simple confidence curve, essentially saying that post-data knowledge on the unknown focus parameter γ, corresponds in an almost Bayesian but actually not-at-all-Bayesian manner to $\gamma \,|\, \text{data} \approx_d \text{N}(\widehat{\gamma}, \widehat{\tau}^2/n)$. The chi-squared approximations worked with below will tend to work better in 'less clear situations', with moderate n, difficult parameters, skewed distributions, etc.

(a) For the given focus parameter $\gamma = g(\theta) = g(\theta_1, \ldots, \theta_p)$, start with the full log-likelihood function $\ell_n(\theta) = \ell_n(\theta_1, \ldots, \theta_p)$, and then do *profiling*,

$$\ell_{n,\text{prof}}(\gamma) = \max\{\ell_n(\theta) \colon g(\theta) = \gamma\}.$$

[xx simple illustration here. xx]

(b) Note that the ML for γ is simply $\widehat{\gamma} = g(\widehat{\theta})$, and that the maximum value of $\ell_{n,\text{prof}}(\gamma)$ is the same as the maximum value of $\ell_n(\theta)$:

$$\ell_{n,\max} = \max_{\text{all } \theta} \ell_n(\theta) = \ell_n(\widehat{\theta}) \quad \text{is the same as} \quad \ell_{n,\max} = \max_{\text{all } \gamma} \ell_{n,\text{prof}}(\gamma) = \ell_{n,\text{prof}}(\widehat{\gamma}).$$

The deviance function for the γ parameter is defined as

$$D_n(\gamma) = 2\{\ell_{n,\max} - \ell_{n,\text{prof}}(\gamma)\}.$$

[xx for the example, compute and display. xx]

– The (second) Wilks Theorem now very nicely says that

$$D_n(\gamma_0) \to_d \chi^2_1$$

at the true value γ_0. This is a proper generalisation of the first version, where there is no profiling. We come back to proofs and conditions, but at the moment we learn what the deviance is, and note, with admiration and gratitude, its simple chi-squared limit distribution. Part of the story is of course that this works for (almost) any smooth parametric model and for (almost) any smooth focus parameter $\gamma = g(\theta)$.

(c) The deviance function is one-dimensional and can be displayed and inspected, even if θ is seven-dimensional. We may also read off confidence intervals. Show that the true γ_0 is covered by the set $C_n = \{\gamma\colon D_n(\gamma) \le 3.941\}$ with probability converging to 0.95.

(d) It's convenient and fruitful to present the *confidence curve* instead, a simple transformation of the deviance curve, namely

$$\text{cc}_n(\gamma, \text{data}) = \Gamma_1(D_n(\gamma)).$$

Show that $\text{cc}_n(\gamma_0, \text{data}) \to_d \text{unif}$ at the true parameter value.

(e) For the simple setup with $y_1, \ldots, y_n$ coming from the Beta$(\theta, 1)$ model, with density $\theta\, y^{\theta-1}$ on the unit interval, generate a dataset with e.g. $n = 25$ and $\theta = 0.333$. Compute and display the $\text{cc}_n(\theta, \text{data})$, and check the height of the confidence curve at the true value. Generate perhaps $N = 10$ or $N = 50$ datasets, from the same model and the same θ_0, display all the confidence curves in the same diagram, and give a histogram of the N attained values of $\text{cc}_n(\theta_0, \text{data})$. These should follow the uniform distribution.

(f) Show the magical property

$$\Pr_{\theta_0}\{\theta\colon \text{cc}_n(\theta, \text{data}) \le \alpha\} \to \alpha \quad \text{for all } \alpha \in (0,1).$$

This means precisely that confidence intervals at all levels may be read off, in a simple and clear fashion.

(g) Going back to Ancient Egypt, a century B.C., see the previous exercise, carry out such log-likelihood profile computations for the focus parameter $\gamma = \text{E}\, Y = a/b$ with the two-parameter gamma model. You should then attempt to reproduce Figure 0.6. In particular, read off 0.95 intervals for the mean, for men and for women. I find quite similar results for

the direct normal approximation (i.e. the delta method) and for the Wilks based method of Schweder and Hjort (2016):

	trad		schweder-hjort	
men	27.97	40.27	28.61	41.15
women	21.32	30.53	21.77	31.21

87. More on the Wilks theorems

[xx to come here. it's somewhat technical, whether one attempts path 1 or path 2 or path 3. point to Schweder and Hjort (2016, Appendix). linear algebra things come into the bargain. i decide that in the course it's good to understand the basic ideas also for the proof, and it isn't very mysterious, but it's even more important to understand it and use it in practice, with the $\text{cc}_n(\gamma, \text{data})$ among its features. xx] The setup here is that of i.i.d. observations $y_1, \ldots, y_n$ from a data-generating mechanism $g(y)$, fitted to a parametric model $f(y, \theta)$, with the associated log-likelihood function $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i, \theta)$. We know then that the ML estimator $\widehat{\theta}$ tends to the least false parameter θ_0 minimising the Kullback–Leibler distance $d(g, f_\theta)$. We have also seen, in Exercise 83, that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d J^{-1}U \sim \text{N}_p(0, J^{-1}KJ^{-1}),$$

with $U \sim \text{N}_p(0, K)$ and $J = -\text{E}_g\, \partial^2 \log f(y, \theta_0)/\partial\theta\partial\theta^{\text{t}}$. Here we analyse associated *deviance functions* and their limits.

(a) Consider first

$$D_n(\theta_0) = 2\{\ell_{n,\max} - \ell_n(\theta_0)\}.$$

We have seen in or via Exercises 80 and 83 that

$$D_n(\theta_0) \to_d W = U^{\text{t}}J^{-1}U,$$

a quadratic function of a multivariate vector, with mean $p^* = \text{Tr}(J^{-1}K)$. Show that under model conditions, $J = K$, and $W \sim \chi^2_p$.

(b) Now consider a one-dimensional focus parameter $\phi = h(\theta)$, where the ML estimator is $\widehat{\phi} = h(\widehat{\theta})$. Our task is to examine the random deviance function

$$D_n(\phi) = 2\{\ell_{n,\max} - \ell_{n,\text{prof}}(\phi)\}.$$

Note that this is a function which can be computed and displayed, with minimum value zero at the precise location $\widehat{\phi}$. Explain indeed that

$$\max_{\text{all } \theta} \ell_n(\theta) = \max_{\text{all } \phi} \ell_{n,\text{prof}}(\phi),$$

which also can be expressed as

$$\ell_{n,\max} = \ell_n(\widehat{\theta}) = \ell_{n,\text{prof}}(\widehat{\phi}).$$

(c) It is useful to work with a Taylor approximation of $\ell_n(\theta)$ around the ML estimator $\widehat{\theta}$. Let

$$J_n = -\frac{1}{n}\frac{\partial^2 \ell_n(\widehat{\theta})}{\partial\theta\partial\theta^{\text{t}}}$$

be the normalised observed information matrix, for which we have $J_n \to_{\rm pr} J$. Verify that for θ in the vicinity of $\widehat{\theta}$,

$$\ell_n(\theta) = \ell_n(\widehat{\theta}) - \tfrac{1}{2} n(\theta - \widehat{\theta})^{\rm t} J_n(\theta - \widehat{\theta}) + O_{\rm pr}(n\|\theta - \widehat{\theta}\|^3),$$

under natural regularity conditions. A picture to have in mind is that the log-likelihood surface is approximately a negative quadratic near its maximum point.

(d) Show that this leads to an alternative view of the log-likelihood profiling, namely to

$$D_n^*(\phi) = \min\{Q_n(\theta)\colon h(\theta) = \phi\}, \quad \text{with} \quad Q_n(\theta) = n(\theta - \widehat{\theta})^{\rm t} J_n(\theta - \widehat{\theta}).$$

Give arguments showing that $D_n(\phi)$ and $D_n^*(\phi)$ are large-sample equivalent, i.e. their difference tends to zero in probability, at the position $\phi_0 = h(\theta_0)$.

(e) Assume now that $\phi = b^{\rm t}\theta = b_1\theta_1 + \cdots + b_p\theta_p$, i.e. that the focus parameter is a simple linear combination of the θ components (with known coefficients). In this case we can find an explicit expression for the minimum of $Q_n(\theta)$ under the side condition that $b^{\rm t}\theta = \phi$. This essentially becomes a linear algebra question, of finding the minimum of a quadratic form $x^{\rm t}Ax$ under the side condition that $b^{\rm t}x = \phi$, with A a fixed symmetric positive definite $p \times p$ matrix, b a fixed p-vector, and ϕ a given number. Show, perhaps using Lagrange multiplicators, that the minimiser in this problem is

$$x_0 = \frac{A^{-1}b}{b^{\rm t}A^{-1}b}\phi,$$

with

$$\min\{x^{\rm t}Ax\colon b^{\rm t}x = \phi\} = x_0^{\rm t}Ax_0 = \frac{\phi^2}{b^{\rm t}A^{-1}b}.$$

Use this, or show it directly, again with Lagrange multipliers being the presumably best mathematical path, that minimisation of $Q_n(\theta) = n(\theta - \widehat{\theta})^{\rm t} J_n(\theta - \widehat{\theta})$ under constraint $b^{\rm t}\theta = \phi$ takes place for ..., with minimum

$$D_n^*(\phi) = \min\{n(\theta - \widehat{\theta})^{\rm t} J_n(\theta - \widehat{\theta})\colon b^{\rm t}\theta = \phi\} = \frac{n(\phi - \widehat{\phi})^2}{b^{\rm t}J_n^{-1}b}.$$

(f) For the situation of point (e), show that

$$\sqrt{n}(\widehat{\phi} - \phi_0) \to_d b^{\rm t}J^{-1}U \sim {\rm N}(0, b^{\rm t}J^{-1}KJ^{-1}b),$$

at the least false position ϕ_0. Conclude from this that at $\phi = \phi_0 = h(\theta_0)$,

$$D_n^*(\phi_0) \to_d \frac{b^{\rm t}J^{-1}U}{b^{\rm t}J^{-1}b} = \frac{b^{\rm t}J^{-1}KJ^{-1}b}{b^{\rm t}J^{-1}b}\frac{b^{\rm t}J^{-1}U}{b^{\rm t}J^{-1}KJ^{-1}b} \sim k\chi_1^2,$$

with

$$k = \frac{b^{\rm t}J^{-1}KJ^{-1}b}{b^{\rm t}J^{-1}b}.$$

(g) This result, on the log-likelihood-ratio under agnostic conditions, is perhaps not so well known, see Schweder and Hjort (2016, Appendix). Under model conditions, however, we have

$$D_n(\phi_0) \to_d \chi_1^2 \quad \text{and} \quad D_n^*(\phi_0) \to_d \chi_1^2,$$

and these results are decidedly and deservedly statistically famous. The 'Wilks theorem' has come to mean a little portmanteau bag of things, and one of these is the $D_n(\phi_0) \to_d \chi^2_1$ under model conditions. Explain why and how this result can be used in at least two ways. The first is to test the null hypothesis that ϕ is equal to some given and sufficiently interestinug ϕ_0. The second is the confidence curves of Schweder and Hjort (2016): show that with

$$\mathrm{cc}(\phi) = \Gamma_1(D_n(\phi_0)),$$

we have

$$\Pr_{\theta_0}\{\phi\colon \mathrm{cc}(\phi) \le \alpha\} \to \alpha \quad \text{for all } \alpha \in [0,1],$$

which means that confidence sets can be read off, at any desired level.

(h) Explain how the k quantity of point (f) can be estimated consistently, and give a recipe for such a $\widehat{k}$. Under model conditions, we would have $\widehat{k} \to_{\mathrm{pr}} 1$. Explain how a model-robust confidence curve can be constructed from this, say $\mathrm{cc}^*(\phi)$, with the property that

$$\Pr_g\{\phi\colon \mathrm{cc}^*(\phi) \le \alpha\} \to \alpha \quad \text{for all } \alpha \in [0,1].$$

now to be seen as a confidence curve in *the least false parameter value* $\phi_0 = h(\theta_0)$, rather than in 'the true parameter value'.

(i) The reasoning above has so far been limited to the simpler case of $\phi = b^{\mathrm{t}}\theta$ being a linear function of θ. Of course we need the above apparatus also for general $\phi = h(\theta)$, for any smooth parameter function h. But this can be dealt with too, essentially 'by linearisation'. Since $\widehat{\theta}$ is near θ_0 with high probability, for growing n, actually as near as $O_{\mathrm{pr}}(1/\sqrt{n})$, we have

$$\phi = h(\theta) = h(\theta_0) + b^{\mathrm{t}}(\theta - \theta_0) + O(\|\theta - \theta_0\|^2),$$

with $b = \partial h(\theta_0)/\partial\theta$. Try to squeeze a proper proof for $D_n(\phi_0) \to_d \chi^2_1$ out of this.

88. Yet more on Wilks

An alternative route for understanding and handling some of the finer mathematical details in the various parts of longer proofs in the previous Wilks Theorems exercise is as follows. This is also useful for certain generalisations, as with yet-to-come glorious insights for the II-CC-FF meta-fusion setup of Cunen and Hjort (2020).

(a) We start off by 'translating' the starting insight

$$\ell_n(\theta) = \ell_n(\widehat{\theta}) - \tfrac{1}{2}n(\theta - \widehat{\theta})^{\mathrm{t}} J_n(\theta - \widehat{\theta}) + O_{\mathrm{pr}}(n\|\theta - \widehat{\theta}\|^3)$$

to the $1/\sqrt{n}$ scale, via the random process

$$B_n(s) = \ell_n(\widehat{\theta} + s/\sqrt{n}) - \ell_n(\widehat{\theta}).$$

Show that

$$B_n(s) = -\tfrac{1}{2}s^{\mathrm{t}} J_n s + O_{\mathrm{pr}}(\|s\|^3/\sqrt{n}).$$

(b) Show next that the deviance function

$$D_n(\phi) = 2\{\ell_{n,\max} - \ell_{n,\text{prof}}(\phi)\},$$

for a given smooth focus parameter $\phi = h(\theta)$, can be written

$$D_n(\phi) = \min\{s^{\rm t} J_n s + O_{\rm pr}(\|s\|^3/\sqrt{n})\colon h(\widehat{\theta} + s/\sqrt{n}) = \phi\}.$$

With

$$h(\widehat{\theta} + s/\sqrt{n}) = h(\widehat{\theta}) + \widehat{b}^{\rm t} s/\sqrt{n} + O_{\rm pr}(\|s\|^2/n),$$

writing $\widehat{b} = \partial h(\widehat{\theta})/\partial\theta$, show that $D_n(\phi)$ is large-sample equivalent to

$$\begin{aligned} D_n^*(\phi) &= \min\{s^{\rm t} J_n s\colon \widehat{\phi} + \widehat{b}^{\rm t} s/\sqrt{n} = \phi\} \\ &= \min\{s^{\rm t} J_n s\colon \widehat{b}^{\rm t} s = \sqrt{n}(\phi - \widehat{\phi})\} \\ &= \frac{n(\phi - \widehat{\phi})^2}{\widehat{b}^{\rm t} J_n^{-1} \widehat{b}}, \end{aligned}$$

where these few steps involve 'only' the linear algebra results handled in the previous exercise, about minimisation of a quadratic form under lineaer constraints. This large-sample equivalence need to be shown at the true of least false position $\phi_0 = h(\theta_0)$, but can incidentally be proven to hold [xx nils thinks, at the moment xx] also for $O(1/\sqrt{n})$ neighbourhoods around θ_0, which is important for some further developments.

(c) Conclude, as with the previous exercise, that

$$D_n(\phi_0) \to_d k\chi_1^2 = \frac{b^{\rm t} J^{-1} K J^{-1} b}{b^{\rm t} J^{-1} b} \chi_1^2,$$

where $k = 1$ under model conditions. Again, all of this leads to the splendidly useful confidence curves

$$\text{cc}(\phi) = \Gamma_1(D_n(\phi)) \quad \text{and} \quad \text{cc}^*(\phi) = \Gamma_1(D_n(\phi)/\widehat{k}),$$

the first so to speak the classical one, under model conditions, the second valid also outside model conditions, but then to be interpreted as the confidence curve for the least falso parameter $\phi_0 = h(\theta_0)$. For more discussion related to these matters, see Schweder and Hjort (2016, Chs. 3, 4, 7).

(d) We learn here that under natural and broad regularity conditions, there's 'Wilks behaviour' – a somewhat tentative term, perhaps introduced in Schweder and Hjort (2016), to indicate that deviance type functions exhibit approximate chi-squared behaviour – of the deviance function. Specifically,

$$D_n(\phi) = 2\{\ell_{n,\max} - \ell_{n,\text{prof}}(\phi)\} = \frac{n(\phi - \widehat{\phi})^2}{\widehat{b}^{\rm t} J_n^{-1} \widehat{b}} + o_{\rm pr}(1),$$

at and close to the $\phi_0 = h(\theta_0)$ position in the parameter space. This again entails the χ_1^2 limit under model conditions. There are many generalisations and extensions, also when it comes to results concerning the accuracy of approximations, etc. Here I point to just a few key points.

(i) The setting above has been that of i.i.d. observations from a data generating density g, modelled via some parametric $f(y,\theta)$, fitted via maximum likelihood. Conceptually and operationally, it's a significant lift from i.i.d. to regression models for (x_i, y_i) type data, where $y_i \,|\, x_i$ is modelled as coming from some $f(y_i \,|\, x_i, \theta)$ model. Here the θ in question would contain both regression coefficients and perhaps just a few general parameters; for linear regression, we would have $\theta = (\beta, \sigma)$, etc. Crucially, *most of the arguments and technical details* carry over, from i.i.d. to regression, with the required amount of extra bureacracy, book-keeping, the use of Lindeberg instead of the plain CLT, etc. The key point for Wilks theorems to hold is as above, that

$$\ell_n(\theta) = \ell_n(\widehat{\theta}) - \tfrac{1}{2} n(\theta - \widehat{\theta})^{\mathrm{t}} J_n (\theta - \widehat{\theta}) + O_{\mathrm{pr}}(n\|\theta - \widehat{\theta}\|^3)$$

describes the behaviour of the log-likelihood function in the vicinity of its maximiser $\widehat{\theta}$. So Wilks theorems may be used for say Poisson regression, where some appropriate deviance function

$$D_n(\beta_3) = 2\{\ell_{n,\max} - \ell_{n,\mathrm{prof}}(\beta_3)\}$$

may be used for at least two purposes. First, to test whether $\beta_3 = 0$ or not, by comparing $D_n(0)$ to the χ^2_1. Second, to compute and display a full confidence curve $\mathrm{cc}(\beta_3)$ for that regression parameter.

(ii) Sometimes data do not come from only one homogeneous group, of course; imagine e.g. that we have n_j data points for group j, for $j = 1, \ldots, k$. If there is a parametric model binding these data together, perhaps with some α_j parameters specific to group j and other parameters common to all, there would be a log-likelihood function

$$\ell_{\mathrm{grand}}(\theta) = \sum_{j=1}^{k} \ell_{\mathrm{group}\ j}(\theta),$$

and an ML estimator $\widehat{\theta}$; here θ denotes the full parameter vector comibing those for the individual groups. The crucial point is that maximum likelihood theory still works, after appropriate checking of all steps of all arguments, so to speak. We would still have

$$\ell_{\mathrm{grand}}(\theta) = \ell_{\mathrm{grand}}(\widehat{\theta}) - \tfrac{1}{2}(\theta - \widehat{\theta})^{\mathrm{t}} J_{\mathrm{grand}} (\theta - \widehat{\theta}) + O_{\mathrm{pr}}(N\|\theta - \widehat{\theta}\|^3),$$

under suitable and not hard conditions, with $J_{\mathrm{grand}} = -\partial^2 \ell_{\mathrm{grand}}(\widehat{\theta})/\partial\theta\partial\theta^{\mathrm{t}}$ (this time not normalised by sample size), and $N = \sum_{j=1}^{k} n_j$ the combined sample size. The necessary requirement is that the diagonal elements of J_{grand} become bigger with combined data volume: then Wilks theorems and their consequent chi-squared approximations still work.

For the Roman Era Egypt data, for instance, with $n_m = 82$ men and $n_w = 59$ women, and with a model saying y_i is $\mathrm{Gamma}(a, b_m)$ for men and $\mathrm{Gamma}(a, b_w)$ for women, there's a complete fine three-parameter log-likelihood function

$$\ell(a, b_m, b_w) = \sum_{i=1}^{n_m} \log f(y_{m,i}, a, b_m) + \sum_{i=1}^{n_w} \log f(y_{w,i}, a, b_w),$$

where ML theory and Wilks theory still work, complete with tests and confidence curves etc. The key requirement is that the 3×3 Fisher information matrix has diagonal

elements that are big enough. It wouldn't work, however, if we attempted a model with say 50 parameters for these 82 + 59 data values.

(iii) Above we've been discussing 'Wilks behaviour' up & down, in settings where the dimension of the parameter vector with fixed dimension, say p, and with sample size n growing. This is the classical large-sample setup that we've been exploring for most of the STK 4090 course. Sometimes one wishes to work with models with a growing number of parameters, however. The tentative point to make here, without full precision is that extending the full ML theory to $\widehat{\theta}-\theta_0$ is partly difficult, since the dimension of the θ is growing under its feet, but that the *Wilks theorem about focus parameters* might still be shown to hold, under not to strict conditions. In yet other words, we might still have

$$D_n(\phi) = 2\{\ell_{n,\max} - \ell_{n,\mathrm{prof}}(\phi)\} = \frac{n(\phi-\widehat{\phi})^2}{\widehat{b}^{\mathrm{t}} J_n^{-1}\widehat{b}} + o_{\mathrm{pr}}(1)$$

even when the profiling in $\phi = h(\theta_1,\ldots,\theta_p)$ might be over a long parameter vector. In Schweder and Hjort (2016, Ch. 14) there's a multi-billion-dollar lawsuit analysis of 48 court famous 2×2 tables, where we use a model with 49 parameters, basically one for each table and one crucial extra parameter to monitor the alleged extra deadliness of the medicine in question. When we profiled the log-likelihood function $\ell(\theta_1,\ldots,\theta_{48},\gamma)$ over the first 48 parameters, we discovered a clear χ^2_1 behaviour of the deviance function, i.e. what we term Wilks behaviour. So the very carefully constructed and computed optimal confidence curve $\mathrm{cc}(\gamma)$ was surprisingly very close to the 'simpler' thing we also computed via profiling and the Wilks theorem.

89. Wilks Theorem for k-dim subsets of p-dim parameter space

Material on Wilks Theorems for courses such as this one 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 $k<p$; perhas this corresponds to having $\theta_j=0$ for $p-k$ of the components. Then we may define and compute

$$\begin{aligned}\ell_{\max,\mathrm{all}} &= \max\{\ell_n(\theta)\colon \theta\in\Omega\},\\ \ell_{\max,H_0} &= \max\{\ell_n(\theta)\colon \theta\in\Omega_0\},\end{aligned}$$

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

$$\Delta_n = 2(\ell_{\max,\mathrm{all}} - \ell_{\max,H_0}).$$

Then the splendidly useful Wilks theorem, going all the way back to his 1938 paper, says that under H_0 conditions,

$$\Delta_n \to_d \chi^2_{\rm df}, \quad \text{with df} = p - k.$$

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. What I've just summarised here is also presented and proved in Ferguson's Chapter 22 – and there's of course more to say about it, other ways to express and prove parts of it, to extend it further, etc.

(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 Exercise 83.

(b) Assume next that H_0 correspons 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)

90. Confidence curves, A

Here we visit Roman Era Egypt again, aiming for confidence curves for natural focus parameters. In this exercise we fit the lifelengths for men and for women using Gamma distributions, with say (a_m, b_m) for men and (a_w, b_w) for women. You should work out details, and display results, both with 'Recipe One', the direct normal approximation, and 'Recipe Two', with the Wilks theorem with log-likelihood profiling and the chi-squared approximation.

(a) First, fit the two Gamma models for the men and the women data, and give 90 percent confidence intervals for the parameter (a_m, b_m) and (a_w, b_q).

(b) Find confidence curves for the two mean parameters $\xi_m = a_m/b_m$ and $\xi_w = a_w/b_w$.

(c) Find confidence curves for the two standard deviation parameters $\xi_m = a_m^{1/2}/b_m$ and $\xi_w = a_w^{1/2}/b_w$.

(d) We're learning that men overall lived longer than women then. One of many ways in which to express this statistically is to work with the ratio parameter

$$\rho = \xi_m/\xi_w = \frac{a_m/m_m}{a_w/b_w}.$$

So find a confidence curves for this parameter.

(e) Produce a plot with the estimated quantile ratio

$$\gamma(q) = \frac{F_m^{-1}(q, a_m, b_m)}{F_w^{-1}(q, a_m, b_m)}, \quad \text{for } q = 0.05, 0.06, \ldots, 0.94, 0.95,$$

along with a 90 percent confidence band. Comment on what you find.

(f) Re-do all of the above with the Weibull distribution in lieu of the Gamma.

91. Confidence curves, B

[xx a few examples, expo case with moderate n, etc. cc(ψ) for the win. xx]

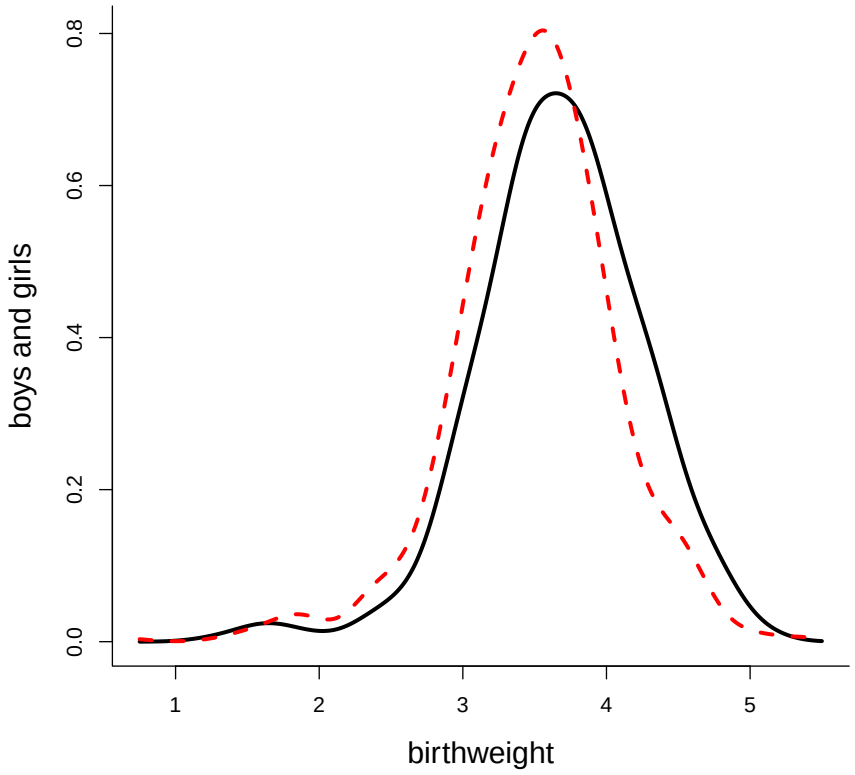

Figure 0.7: Density estimates for the distribution of birthweights, for boys (black, full) and girls (red, dashed), using the Voldner et al. data from Rikshospitalet, Oslo, in the 2001–2008 period.

92. Birthweights for boys and girls

Get hold of the `birthweigth-boys` and `birthweight-girls` datasets from the course website, with data from 480 girls and 548 boys, all born at Rikshospitalet, Oslo, in the 2001–2008 period; I've been given these data, from bigger data files from the STORK study of Voldner et al. (2008).

(a) Use kernel density estimation to produce a version of Figure 0.7.

(b) Check via suitable tests if the data for the two groups can reasonably be seen as normally distributed.

(c) Check with a t-test if the means of the two populations are the same.

(d) Assuming normality (regardless of what you find in (a)), test whether $\sigma_b = \sigma_g$, for the two standard deviations, and also find a confidene interval for $\rho = \sigma_b/\sigma_g$.

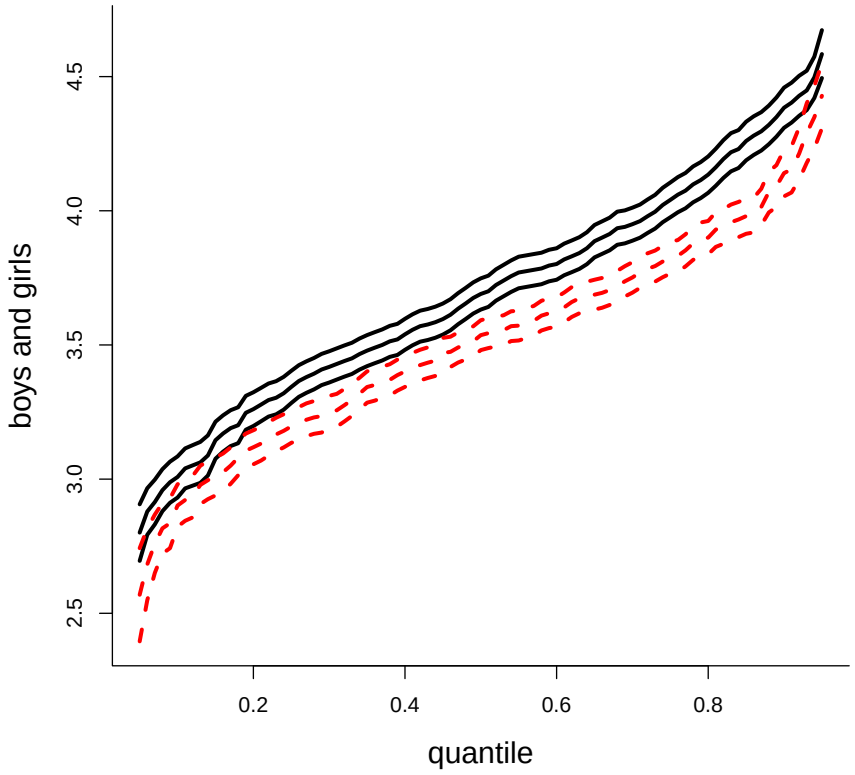

Figure 0.8: Density estimates for the distribution of birthweights, for boys (black, full) and girls (red, dashed), using the Voldner et al. data from Rikshospitalet, Oslo, in the 2001–2008 period.

(e) Then carry out nonparametric quantile inference, producing a version of Figure 0.8. It involves estimating

$$\mu_q = F^{-1}(q),$$

for each quantile level q, perhaps from 0.05 to 0.95, and also estimating the standard deviations, using the large-sample formula $(1/n)q(1-q)/f(\mu_q)^2$ for the approximate variance.

(f) Then flex your delta method muscles, to produce a plot of the curve $\widehat{\rho}(q)$, with an approximate 95 percent confidence band around it, where $\rho(q) = F_b^{-1}(q)/F_g^{-1}(q)$, the ratio of the boy quantile to the girl quantile. Comment on what you find.

(g) Ask your parents how much you weighed when you were born, and estimate which quantile you belonged to.

h) Do a similar but perhaps simpler analysi for $\rho(q)$, assuming normality for the two populations, and compare with your nonparametric analysis.

93. Brownian motion

Time has come for us to consider *stochastic processes*, as opposed to 'only' random variables and vectors. I'm willing to argue that the two most fantastic and crucial distributions, in the one-dimensional situation, are (i) the central normal, the famous $\mathrm{N}(0,1)$, and (ii) the Poisson. They live, they are there, they serve as limits and approximations and Lego-brikker for a long list of other things, in all of probability theory and statistics. Similarly, but slightly more tentatively, the two central and deservedly superfamous stochastic processes are (i') the *Brownian motion* and (ii') the *Poisson process*. In this exercise we learn the basics about Brownian motion, also called the Wiener process (Robert Brown 1773–1858 is the British botanist, who once had a cup of tea; Norbert Wiener 1894–1964 the father of cybernetics etc.; also Albert Einstein 1879–1955 belongs on the list of famous scientists who early on worked with this process).

(a) Have a cup of tea, complete with the classical tea leaves (English Breakfast is slightly better than Earl Grey). Study them for a minute. You're observing Brownian motion. Try to describe, in probability terms, what is going on, and simplify such descriptions to the one-dimensional case.

(b) Consider then $W = \{W(t)\colon t \geq 0\}$, a random process evolving over time t, with the following properties: (i) $W(0) = 0$; (ii) increments $W(t) - W(s)$ have normal $\mathrm{N}(0, t-s)$ distributions, i.e. with variance equal to the length of the interval in question; (iii) increments over disjoint intervals are independent. Show that $\mathrm{cov}\{W(s), W(t)\} = \min(s,t)$, and for the random triple $(W(s), W(t), W(u))$, with $s < t < u$, find the 3×3 covariance matrix.

(c) It is incidentally fruitful and useful to think of the W process via cumulative small movements, as in $W(t)$ being the sum of many tiny and independent

$$\mathrm{d}W(s) = W(s + \mathrm{d}s) - W(s) \sim \mathrm{N}(0, \mathrm{d}s).$$

Express $W(1)$ as the sum of 1000 tiny such $W((i+1)/n) - W(i/n)$, and show that you get the right variance.

(d) Suppose a Crazy Probabilist tries to define a process $W^* = \{W^*(t)\colon t \geq 0\}$ by putting up (i), (ii), (iii) above, but now with $\mathrm{Var}\,\{W(t) - W(s)\} = (t-s)^{1/2}$, rather than $t - s$. Show that it would all backfire solidly. – So it's not enough to put up 'something' for the variance or covariance function, as it may lead from Kapitol to the Trojan Cliffs. One needs to check for logical coherency.

(e) Show that with $t_1 < \cdots < t_k$, then by necessity the vector $(W(t_1), \ldots, W(t_k))$ must be multinormal, and give its $k \times k$ covariance matrix.

– Note that *the existence* of the Brownian motion process is not entirely obvious, and there is no easy way to put down a full joint probability density for the full thing – so the pure existence of a stochastic process is a more delicate and complex matter than when working in one- and finite-dimensional cases. But fear not, Brownian motion exists, which may be proved in many ways, including via the Donsker theorem below, which says that a certain well-defined sequence $X_n = \{X_n(t)\colon t \geq 0\}$ has a well-defined limit, and this limit behaves according to (i), (ii), (iii) given above.

94. Partial-sum processes and the Donsker theorem

Let $U_1, U_2, \ldots$ be i.i.d., with mean zero and variance one (though we do not need to say anything more regarding their distribution). Then we know from the CLT that $n^{-1/2}(U_1 + \cdots + U_n) \to_d \mathrm{N}(0,1)$. But we also see that e.g. $n^{-1/2}(U_1 + \cdots + U_{[n/2]})$ must have a normal limit, and similarly with other partial sums. This exercise goes through the basics full process of partial sums, and leads to the famous, fundamental, and very useful Donsker theorem. Define indeed the random process

$$X_n = \{X_n(t) \colon t \geq 0\}$$

via

$$X_n(t) = n^{-1/2} \sum_{i \leq [nt]} U_i = \frac{U_1 + \cdots + U_{[nt]}}{\sqrt{n}}.$$

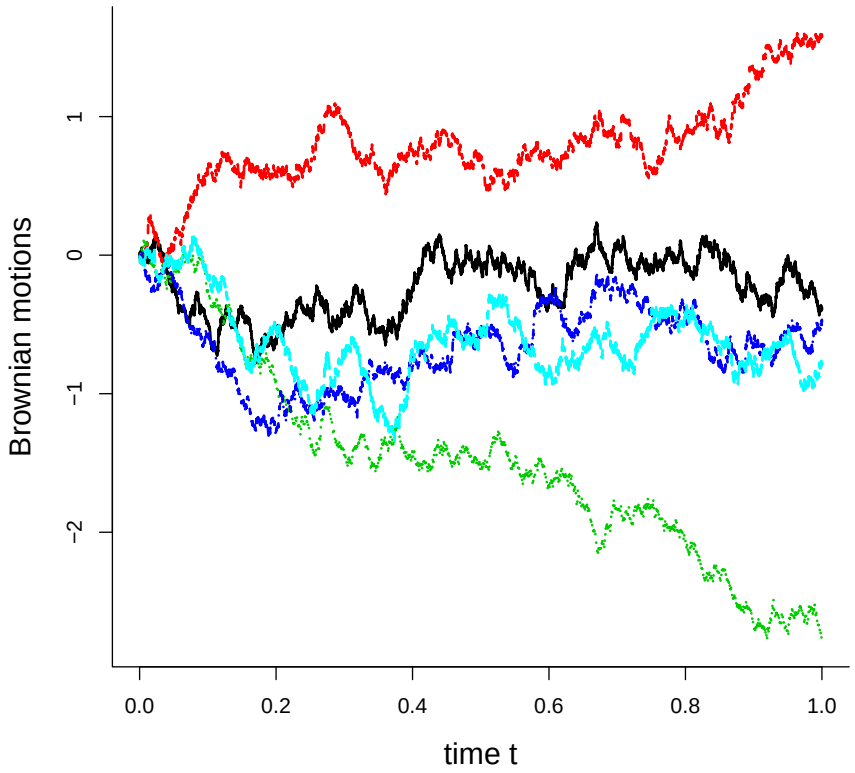

Figure 0.9: Simulated Brownian motion paths. Well, actually, these are simulated partial-sum processes, the $X_n(t)$ of Exercise 90, with $n = 10^4$, and the eye can barely see the difference between these and the Real Brown McCoy.

(a) Use one or two of your favourite distributions for U_i, with mean zero and variance one, to generate some sample paths of X_n. I've done this for Figure 0.9, actually using $U_i = V_i - 1$, with the V_i being unit exponential; the point is to use something with a skewness, though that aspect disappears with X_n in the limit.

(b) Fix t positive. Show that $X_n(t)$ has mean zero and variance $[nt]/n$ converging to t. Show also that $X_n(t) \to_d \mathrm{N}(0,t)$. In view of the Brownian motion process defined above, we may write $X_n(t) \to_d W(t)$.

(c) Show more generally that if $s < t < u < v$, then

$$(X_n(t) - X_n(s), X_n(v) - X_n(u)) \to_d (W(t) - W(s), W(v) - W(u)),$$

namely two independent normal pieces with variances $t - s$ and $v - u$.

(d) Show that with $t_1 < \cdots < t_k$ we have

$$(X_n(t_1), \ldots, X_n(t_k)) \to_d (W(t_1), \ldots, W(t_k)).$$

(e) Show then that for $s < t < u$, we have

$$\mathrm{E}\,|X_n(t) - X_n(s)|^2\,|X_n(u) - X_n(t)|^2 \to (t-s)(u-t) \le (u-s)^2.$$

(f) We have established that $X_n \to_d W$ in the sense of finite-dimensional distributions, *and* the bound (d) may be used to establish that the X_n sequence is *tight* (perhaps 'stram' på norsk). The technical definition is that for each $\varepsilon > 0$, there is a sufficiently big compact set K such that

$$\Pr\{X_n \in K\} > 1 - \varepsilon \quad \text{for all big } n.$$

The opposite of tightness, a sequence which is not stram, would then be that there is no such big compact K holding on to the X_n sequence, which means, somehow, that 'part of the probability is escaping to infinity'. These delicate things by necessity involve even more details, namely *what is a compact set*, in this space where the random processes live. Briefly, the space is $D[0,1]$, all right-continuous functions $x\colon [0,1] \to \mathcal{R}$ with left-hand limits, and the natural topology is that of the *Skorokhod metric*; see the classic Billingsley (1968) for all details. The point, at present, is that with (i) finite-dimensional convergence and (ii) tightness, via the sufficient condition in (d), we really have full, glorious, splendid, fruitful convergence in distribution, of the X_n process to the W process:

$$X_n \to_d W \quad \text{in } D[0,1].$$

This is Donsker's theorem, from 1951.

– There are of course other conditions securing the crucial tightness of a sequence of processes X_n, but I refrain from going too deeply in that correction. One general sufficient condition, which quite often can be established, is the following extension of what was used in (e) above: suppose that for all $s < t < u$ we have

$$\mathrm{E}\,|X_n(t) - X_n(s)|^2\,|X_n(u) - X_n(t)|^2 \le k\,|G_n(u) - G_n(s)|^{1+\delta},$$

for some $\delta > 0$, and some big enough k, where G_n converges pointwise to some continuous and monotone G. Then $\{X_n\colon n = 1, 2, 3, \ldots\}$ is tight, and if in addition X_n tends to X for all finite-dimensional distributions, then gloriously & triumphantly, $X_n \to_d X$ in the $D[0,1]$ space. [xx nils, check with Billingsley 1968 section 15 or so, when i get to the office. xx]

95. Applying the Donsker theorem

Modulo some technicalities, which we do not have the proper time to go sufficiently deeply into here, we have established that $X_n \to_d W$ above: the natural partial-sum process tends to the Brownian motion process. These technicalities are of course important, and when depth and precision are called for one needs to deal with them. First of all, we need a proper definition for $X_n \to_d X$, in the space $D[0,1]$ of such random processes, and it is that

$$\mathrm{E}\, h(X_n) \to \mathrm{E}\, h(X) \quad \text{for all bounded, continuous } h\colon D[0,1] \to \mathcal{R}.$$

(a) Show from these definitions that if $X_n \to_d X$ and $g\colon D[0,1] \to \mathcal{R}$ is a continuous functional, then $g(X_n) \to_d g(X)$. Here the g does not need to be bounded. Examples are $h_1(x) = x(t_0)$, a simple projection; $h_2(x) = \max |x(t)|$; $h_3(x) = m\{t \in [0,1]\colon x(t) > 0\}$, the amount of time $x(t)$ has been above zero; $h_4(x) = \int_0^1 x(t)\,\mathrm{d}t$, etc. For each of these cases, we then have

$$X_n \to_d X \quad \text{implies} \quad h(X_n) \to_d h(X).$$

The point here is also that it might be quite hard to prove $h(X_n) \to_d h(X)$ directly, or separately; the 'natural way' to prove it is via the master lemma $X_n \to_d X$ first. So such a master lemma has a long list of consequences.

(b) Further, to the details of convergence in distribution apparatus for random processes: these relate to (i) the precise understanding of the Skorokhod metric, used to set up a clear distance $d(x_1, x_2)$ between functions $x_1, x_2\colon [0,1] \to \mathcal{R}$; (ii) how compact sets then can be characterised and recognised; (iii) setting up good enough criteria for tightness; (iv) knowing and showing the basic Prokhorov theorem, that if $X_n \to_d X$ for all finite-dimensional vectors, and if there is tightness, then we're really guaranteed $X_n \to_d X$. – In this light, think through the full $X_n \to_d W$ again, the partial-sums process and their convergence to Brownian motion.

(c) A very simple continuous function is $h(x) = x(1)$, reading off the value of $x(t)$ at the endpoint $t = 1$. It is continuous. Hence $X_n(1) \to_d X(1)$. Go back to Donsker and deduce the CLT. – In this light, the Donsker theorem is really a much bigger brother to the CLT; the Donsker has a thousand corollaries, some simple, many complex, and one of these is the CLT.

(d) Consider the Donsker theorem setup, and define $h(x) = \max_{0\le t\le 1} x(t)$. It is a continuous functional. Deduce that with $M_n = \max_{i\le n} |S_i|$, where $S_i = U_1 + \cdots + U_i$, we have

$$M_n/\sqrt{n} = \max_{0\le t\le 1} X_n(t) \to_d M = \max_{0\le t} W(t).$$

Note that this limit ensues regardless of the distribution of the i.i.d. sequence U_i, as long as these have mean zero and variance one. The limit is the same, with U_i being standard normal, or symmetric ± 1 variables, of $U_i = V_i - 1$ with unit exponential V_i, etc. A separate matter

(e) We play a long game, you and I. My winning in game i is U_i, with a distribution symmetric around zero and variance one; if it's positive, good for me, it it's negative, good for you. After i games, my bank account has $U_1 + \cdots + U_i$, and your bank account has minus this sum. We play a few thousand times. – How much of the time have I been in the lead? This is food for the Donsker theorem, since this is about

$$T_n = (1/n) \sum_{i=1}^{n} I\{U_1 + \cdots + U_i > 0\} = m\{t \in [0,1]\colon X_n(t) > 0\},$$

the Lebesgue measure of how much time the process has been above zero. This is a continuous functional, so

$$T_n \to_d T = m\{t \in [0,1] \colon W(t) > 0\}.$$

The problem hence a clear-cut general solution: I've been in the lead, over you, a portion T_n of the time, and T_n tends to T in distribution. So it's 'only' a matter of finding the distribution of T. This is non-trivial, but the solution is the Beta($\frac{1}{2}, \frac{1}{2}$), intriguingly, with U-shaped density

$$f(t) = \frac{1}{\pi} \frac{1}{\sqrt{t(1-t)}} \quad \text{for } t \in (0,1).$$

Thus there's a high chance that *one of the two of us* has been leading for a very long time; the least likely outcome is that we've each been leading about half the time.

96. The Brownian bridge

We say that a process $X = \{X(t) \colon t \in [0,1]\}$ is normal, or Gaußian, if all its finite-dimensional distributions are normal. This is the same as saying that all linear combinations are normal. Thus the Brownian motion process is normal, for example.

(a) To define and describe a normal process X, show that it is sufficient to give (i) the mean function $\xi(t) = \mathrm{E}\, X(t)$ and (ii) the covariance function $K(s,t) = \mathrm{cov}\{X(s), X(t)\}$. So nothing more is required than these two functions.

(b) Describe the Brownian motion process via its mean and covariance functions.

(c) Then consider the process

$$W^0(t) = W(t) - tW(1) \quad \text{for } t \in [0,1],$$

with W being the Brownian motion. Show that W^0 is normal, with zero mean, and covariance function $\min(s,t) - st$, i.e. $s(1-t)$ for $s \le t$. In particular, its variance is $t(1-t)$. The W^0 is called the Brownian bridge.

(d) For two disjoint time windows, say $[s_1, s_2]$ and $[t_1, t_2]$, with $s_1 < s_2 < t_1 < t_2$, find the covariance and correlation between the two bridge increments $W^0(s_2) - W^0(s_1)$ and $W^0(t_2) - W^0(t_1)$.

(e) Simulate ten paths of W, and transform these to ten paths of W^0.

(f) Show that W^0 also can be characterised as W conditional on $W(1) = 0$.

97. The empirical distribution process

Consider $U_1, U_2, \ldots$ being i.i.d. from the uniform distribution. We study the empirical distribution process,

$$G_n(t) = \frac{1}{n} \sum_{i=1}^{n} I\{U_i \le t\} \quad \text{for } t \in [0,1].$$

(a) Simulate $n = 100$ datapoints, and plot two processes: first the $G_n(t)$, called the empirical cumulative distribution function, and then the normalised and scaled process

$$Z_n(t) = \sqrt{n}\{G_n(t) - t\}.$$

(b) Show that $G_n(t)$ has mean t and variance $t(1-t)/n$. Hence show that $Z_n(t)$ has mean zero and variance $t(1-t)$. Note indeed that Z_n starts and ends at zero.

(c) Show that $Z_n(t) \to_d \mathrm{N}(0, t(1-t))$. Show also that for $s < t$, $(Z_n(s), Z_n(t)) \to_d (A, B)$, say, a binormal zero-mean with variances $s(1-s)$ and $t(1-t)$, and covariance $s(1-t)$.

(d) As an interlude, which turns out to be relevant in a minute, consider a trinomial situation, with (X, Y) having the trinomial distribution with sample size n and probability parameter (p, q). In other words,

$$\Pr\{X = x, Y = y\} = \frac{n!}{x!\, y!\, z!} p^x q^y r^z$$

for $x \geq 0, y \geq 0, x + y \leq n$, with $r = 1 - p - q$ and $z = n - x - y$. We know from very classical binomial analyses that

$$\sqrt{n}(X/n - p) \to_d \mathrm{N}(0, p(1-p)) \quad \text{and} \quad \sqrt{n}(Y/n - q) \to_d \mathrm{N}(0, q(1-q)).$$

Now the ambition level is slightly higher, however, as you have to find the joint binormal limit distribution of $(\sqrt{n}(X/n - p), \sqrt{n}(Y/n - q))$. Try to find the answer in two ways: via the two-dimensional CLT, and via general limit distribution results for maximum likelihood estimators.

(e) Now back to G_n and the empirical process Z_n. There is actually full convergence in distribution here,

$$Z_n \to_d W^0 \quad \text{in } D[0,1],$$

to the Brownian bridge. Show this – which means demonstrating (i) finite-dimensional convergence and (ii) tightness. For the latter you might need to work with a suitable upper bound for

$$\mathrm{E}\,|Z_n(t) - Z_n(s)|^2\,|Z_n(u) - Z_n(t)|^2,$$

which might entail some efforts for trinomial probabilities.

(f) So how much can $G_n(t)$ deviate from its mean function t? One answer is to apply the max functional. Start with

$$D_n = \max_{0 \leq t \leq 1} |G_n(t) - t| = \max_{i \leq n} \{|G_n(i/n-) - i/n|, |G_n(i/n) - i/n|\},$$

the maximum distance from $G_n(t)$ to t. Then deduce

$$\sqrt{n} D_n = \max_{0 \leq t \leq 1} |Z_n(t)| \to_d D = \max_{0 \leq t \leq 1} |W^0(t)|.$$

The distribution of this limit D is 'somewhat famous' and has been tabulated; it is sometimes called the Kolmogorov–Smirnov distribution. I only remember one of the numbers, from these tables, namely that 1.358 is the upper 0.05 point. So $\max_t |G_n(t) - t| \leq 1.358/\sqrt{n}$, with probability 0.95. If you have data and compute D_n, with $\sqrt{n} D_n$ bigger than 1.358, you might be skeptical about the assumption that your data are uniform.

(g) Pretend that we have lost the Kolmogorov–Smirnov tables, and that we don't have the time to derive its (rather complicated) distribution. Simulate 10^5 paths in your computer, and read off the 0.95 point, which should be close to the 1.358 number I remember from these old tables. [xx nils check this. xx]

98. The Kolmogorov–Smirnov test

The above story can be nicely generalised, without many efforts. Suppose you have i.i.d. data $y_1, \ldots, y_n$ and need to test the hypothesis H_0 that their distribution F is equal to some given F_0, like the standard normal. Consider the e.c.d.f, the empirical cumulative distribution function

$$F_n(t) = (1/n) \sum_{i=1}^{n} I\{y_i \le t\}.$$

, and form fom this the empirical process

$$Z_n(t) = \sqrt{n}\{F_n(t) - F(t)\}.$$

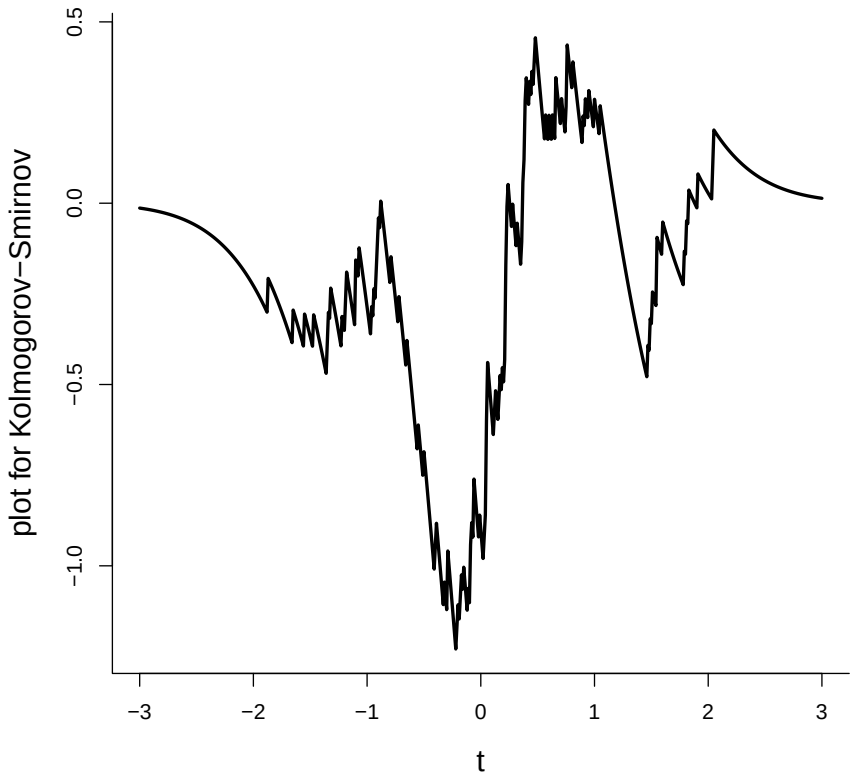

Figure 0.10: Plot of the Kolmogorov–Smirnov related process $Z_n = \sqrt{n}(F_n - F)$, to check if 100 data points from the standard normal look nonnormal or not. Here the $\max_t |Z_n(t)|$ value is 1.229, which is smaller than the 0.05 upper point of the $\max_s |W^0(s)|$ distribution.

(a) Show that $F_n(t)$ has mean $F(t)$ and variance $F(t)\{1 - F(t)\}/n$.

(b) Show that $Z_n(t) \to_d \mathrm{N}(0, F(t)\{1-F(t)\})$.

(c) Show that

$$Z_n \to_d Z = W^0(F(\cdot)),$$

i.e. with $Z(t) = W^0(F(t))$, the Brownian bridge time-transformed via $F(t)$. You may prove this via results in the previous simpler setup with uniforms, using that $U_i = F(Y_i)$ is uniform. Figure 0.10 shows a plot of Z_n, for $n = 100$ simulated points from the standard normal.

(d) Deduce that

$$\sqrt{n} \max_t |F_n(t) - F(t)| \to_d D = \max_t |W^0(F(t))| = \max_s |W^0(s)|.$$

(e) Show that

$$\Pr\{F(t) \in F_n(t) \pm 1.358/\sqrt{n}, \text{ for all } t\} \to 0.95.$$

This is the famous Kolmogorov–Smirnov simultaneous confidence band.

(f) Consider the test for $F = F_0$ consisting in rejecting if $F_0(t)$ for some part of the t domain is outside the Kolmogorov–Smirnov band. Show that this test has level 0.05 for large n. Show also that it is equivalent to rejecting if $D_n = \max_t |F_n(t) - F_0(t)| > 1.358/\sqrt{n}$.

(g) Note in particular that since $\sqrt{n}D_n$ has a limit distribution, we must have $D_n \to_{\mathrm{pr}} 0$. For large n, the maximal difference, between the e.c.d.f. F_n and the true F, goes to zero. There's a slightly stronger version of this statement, which says that $D_n \to 0$ almost surely, or $\Pr\{D_n \to 0\} = 1$. This is the Glivenko–Centelli theorem, from 1933 – and yes, it's famous, and it's your cultural duty to know about it, it tells us that with lots of data, we can uncover any tiny little aspect of their underlying distribution, we can be as sophisticated as we might wish to. The result above is however more informative, since we learn how quickly it goes to zero, namely with speed $1/\sqrt{n}$. This is incidentally impossible for density estimators, where the best one can hope for is that the maximal distance $\max_t |f_n(t) - f(t)|$ goes to zero with speed $1/n^{2/5}$.

99. The Poisson process

well

100. Integrate and display your integrity

well

101. Regression models

well

102. Mrs. Jones is pregnant

Access the `smallchildren-data` dataset from the course website, with data (x_1, x_2, x_3, x_4, y) on $n = 189$ mothers and their newborns, from a wider research project carried out at a hospital in Massachusetts, the US, in 1980ies. Here x_1 is age of mother; x_2 is weight (in kg) prior to pregnancy; x_3 is 0-1 for nonsmoker and smoker; x_4 is 0-1 for white and nonwhite; and finally y is 0-1 for 'normal weight' and 'small weight' for the newborn (with small birthweight defined as less than 2500 g). The task is to find which covariates influence the chance of $y = 1$ and in which ways.

(a) Check the distribution of the four covariates, including correlations between them.

(b) Show that with any model for

$$p_i = p_i(\beta) = \Pr\{y_i = 1 \,|\, x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}\},$$

the log-likelihood function may be written

$$\ell_n(\beta) = \sum_{i=1}^{n} \{y_i \log p_i + (1 - y_i) \log(1 - p_i)\}.$$

(c) Everyone's favourite model (apparently) for such data is the *logistic regression model*, with

$$p_i = H(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}),$$

with $H(u) = \exp(u)/\{1 + \exp(u)\}$ the logistic transform. Programme the log-likelihood function and find the ML estimators $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_4)$.

(d) Show that the normalised Fisher observation matrix can be written

$$J_n = -\frac{1}{n} \frac{\partial^2 \ell_n(\widehat{\beta})}{\partial\beta\partial\beta^{\mathrm{t}}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{p}_i(1 - \widehat{p}_i) x_i x_i^{\mathrm{t}},$$

writing here x_i for the 5×1 vector $(1, x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})^{\mathrm{t}}$. Compute the J_n matrix and the estimated standard deviations, the square roots of the diagonal elements of J_n^{-1}/n. Make a little table with the ML estimators, these standard errors, and the Wald ratios $\widehat{\beta}_j/\mathrm{se}_j$. Which of the four covariates can be seen to influence the small-baby probability?

(e) Mrs. Jones is pregant! She's 28, weights 58 kg, is white, and has never smoked a cigarette in her life. Estimate p_{jones}, and find a 90 percent confidence interval, the probability that her child will be below the 2500 g threshold. Then invent her cousin, Mrs. Smith, who is also 28, with the same weight, but she's a smoker, and is non-white (an euphemism for being either black or if First Americans descent), and analyse similarly p_{smith}. Attempt to construct full confidence curves

$$\mathrm{cc}(p_{\mathrm{jones}}) \quad \text{and} \quad \mathrm{cc}(p_{\mathrm{smith}}).$$

(f) *Parts* of the above computations can incidentally be carried out in half a second using

```
glm(yy ~ x1 + x2 + x3 + x4, family="binomial")
```

in `R`. It is however very useful to be able to programme such log-likelihoods 'from scratch', as models you might stumble into, or create yourself, might not at all be on the list of super-famous already-implemented models. Complement the above analysis with using

$$p_i = H(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}),$$

now using H equal to the cumulative distribution function for the standard Cauchy, or `pcauchy` in `R`. Note how strangely easy it is to go from one model to another model, operationally speaking, as it often might amount to changing a few lines in a computer programme. The formula for the J_n matrix above is however valid only for the logistic regression model, but in general one may use minus the Hessian matrix from the computer maximisation.

103. Yet other things to come

[xx We'll see what I manage or decide to put in, in this growing collection of both exercises and lecture notes. There must be empirical processes, some empirical likelihood, confidence curves, something with nonstandard limits, and the Aalen–Nelson and Kaplan–Meier estimators. With applications. And Cramér–Wold. And Hjort and Fenstad (1992) for the last n, and Hjort and Pollard (1994) for asymptotics for minimisers. xx]

References

Amitsur (1956). On arithmetic functions. *Journal of Analytic Mathematics*.

von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Annals of Mathematical Statistics* **36**, 808–818.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Clauset, A. (2017). The enduring threat of a large interstate war. Technical Report, One Earth Foundation.

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances* **4**, xx-xx.

Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. *Journal of the London Mathematical Society* **11**, 290–294.

Cunen, C. (2019). Confidence curves for dummies. In *FocuStat 2014-2018, Final Report*, available online at the FocuStat website.

Cunen, C. and Hjort, N.L. (2020). Confidence curves for dummies. FocuStat Blog Post, April 2020.

Cunen, C. and Hjort, N.L. (2020). Combining information across diverse sources: the II-CC-FF paradigm. *Scandinavian Journal of Statistics* [to appear].

Cunen, C., Hjort, N.L., and Nygård, H. (2020). Statistical sightings of better angels. *Journal of Peace Research*.

Donsker, M.D. (1951). An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society*.

Doxiadis, A.K. (1992). *Uncle Petros and Goldbach's Conjecture: A Novel of Mathematical Obsession*.

Ferguson, T.S. (1996). *A Course in Large Sample Theory*. Text in Statistical Science, Chapman & Hall, Madras.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.

Ferguson, T.S. and Klass, M.J. (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics* **43**, 1634–1643.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge.

Ghosh, A. (2017). Robust inference under the Beta regression model with application to health care studies. `arXiv`

Gould, S.J. (1995). The median isn't the message. In *Adam's Navel and other essays*. Penguin.

Heger, A. (2007). Jeg og jordkloden. Dagsavisen.

Hjort, N.L. (1976). *The Dirichlet Process Applied to Some Nonparametric Problems.* Cand. real. thesis [in Norwegian], Department of Mathematics, Nordlysobservatoriet, University of Tromsø.

Hjort, N.L. (1979). *S 205: Sannsynlighetsregning III, med statistiske anvendelser.* Kompendium, Department of Mathematics, University of Oslo.

Hjort, N.L. (1985). Discussion contribution to P.K. Andersen and Ø. Borgan's 'Counting process models for life history data: A review'. *Scandinavian Journal of Statistics* **12**, xx–xx.

Hjort, N.L. (1985). An informative Bayesian bootstrap. Technical Report, Department of Statistics, Stanford University.

Hjort, N.L. (1986a). Discussion contribution to P. Diaconis and D. Freedman's paper 'On the consistency of Bayes estimators'. *Annals of Statistics* **14**, 49–55.

Hjort, N.L. (1986b). *Notes on the Theory of Statistical Symbol Recognition.* Statistical Research Monograph, Norwegian Computing Centre, Oslo.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics* **18**, 1259–1294.

Hjort, N.L. (1991). Bayesian and empirical Bayesian bootstrapping. Statistical Research Report, Department of Mathematics, University of Oslo.

Hjort, N.L. (1992). On inference in parametric survival data models. *International Statistical Review* **xx**, 355–387.

Hjort, N.L. (2003). Topics in nonparametric Bayesian statistics [with discussion]. In *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort, S. Richardson). Oxford University Press, Oxford.

Hjort, N.L. (2010). An invitation to Bayesian nonparametrics. In *Bayesian Nonparametrics* (by Hjort, N.L., Holmes, C.C., Müeller, P., and Walker, S.G.), 1–21.

Hjort, N.L. (2018a). Towards a More Peaceful World [Insert '!' or '?' Here]. FocuStat Blog Post.

Hjort, N.L. (2018b). Overdispersed children. FocuStat Blog Post.

Hjort, N.L. and Fenstad, G.U. (1992). On the last time and the number of times an estimator is more than ε from its target value. *Annals of Statistics* **20**, 469–489.

Hjort, N.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Annals of Statistics* **23**, 882–904.

Hjort, N.L., Holmes, C.C., Müller, P., and Walker, S.G. (2010). *Bayesian Nonparametrics.* Cambridge University Press, Cambridge.

Hjort, N.L. and Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics* **24**, 1619–1647.

Hjort, N.L. and Kim, Y. (2013). Beta processes and their applications and extensions. Statistical Research Report, Department of Mathematics, University of Oslo.

Hjort, N.L. and Ongaro. A. (2005). Exact inference for random Dirichlet means. *Statistical Inference for Stochastic Processes* **8**, 227–254.

Hjort, N.L. and Ongaro. A. (2006). On the distribution of random Dirichlet jumps. *Metron* **LXIV**, 61–92.

Hjort, N.L. and Petrone, S. Nonparametric quantile inference using Dirichlet processes. In *Festschrift for Kjell Doksum* (ed. V. Nair).

Hjort, N.L. and Pollard, D.B. (1994). Asymptotics for minimisers of convex processes. [xx fill in xx]

Hjort, N.L. and Schweder, T. (2018). Confidence distributions and related themes. General introduction article to a Special Issue of the *Journal of Statistical Planning and Inference* dedicated to this topic, with eleven articles, and with Hjort and Schweder as guest editors; **195**, 1–13.

Hjort, N.L. and Stoltenberg, E.Aa. (2020). Monitoring the Level and the Slope of the Corona. FocuStat Blog Post, April 2020.

Hjort, N.L. and Walker, S.G. (2009). Quantile pyramids for Bayesian nonparametrics. *Annals of Statistics* **37**, 105–131.

Hveberg, K. (2019). *Lene din ensomhet langsomt mot min.* Aschehoug, Oslo.

Inlow, M. (2010). A moment generating function proof of the Lindeberg–Lévy central limit theorem. *The American Statistician* **64**, 228–230.

Kolmogorov A. N. (1933a). *Grundbegriffe der Wahrscheinlichkeitrechnung*, in Ergebnisse der Mathematik, Berlin.

Kolmogorov A. N. (1933b). Sulla determinazione empirico di una legge di distribuzione. *Giornale dell 'Istituto Italiano degli Attuari* **4**, 83–91.

Lehmann, E.L. (1951). Notes on the Theory of Point Estimation. (Mimeographed by C. Blyth.) Department of Statistics, University of Berkeley, California.

Müller, O., Quintana, F.A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis.* Springer-Verlag, Berlin.

Ottosen, K. (1983). *Theta Theta.*

Ottosen, R. (1959). *ML.*

Owen, A. (2001). *Empirical Likelihood.* Chapman & Hall, New York.

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* **50**, 157–175.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics* **9**, 130–134.

Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood, Probability.* Cambridge University Press, Cambridge.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sethuraman, J. and Tiwari, R. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In: *Proceedings of the Third Purdue Symposium on Statistical Decision Theory and Related Topics* (eds. S.S. Gupta and J. Berger), 305–315. Academic Press, New York.

Slutsky, E. (1925). Über stochastische Asymptoten und Grenzwerte. *Metron* **5**, 3–89.

Stoltenberg, E. (2019). A moment generating function proof of the central limit theorem. Note, related to his STK 4011 teaching, autumn semester 2019, Department of Mathematics, University of Oslo.

Stoltenberg, E.Aa. and Hjort, N.L. (2019a). Simultaneous estimation of Poisson parameters. *Journal of Multivariate Analysis*, in its way.

Stoltenberg, E.Aa. and Hjort, N.L. (2019b). Modelling and analysing the Beta- and Gamma Police Tweetery data. [Manuscript, in progress.]

Voldner, N., Frøslie, K.F., Haakstad, L., Hoff, C., Godang, K., Bollersleiv, J., and Henriksen, T. (2008). Modifiable determinants of fetal macrosomia: role of lifestyle-related factors. *Acta Obstreticia et Gynecologica Scandinavia* **87**, 423–429.

Walker, S.-E. and Hjort, N.L. (2020). Estimation and model selection via weighted likelihoods. Manuscript.

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**, 60–62.

Wolpert, R.L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 251–267.

Aalen, O.O., Borgan, Ø., and Gjessing, H. (2008). *Survival and Event History Analysis: a Process Point of View*. Springer-Verlag, Berlin.