STK4011 – Statistical inference. Autumn 2019
Exercises

Emil Aas Stoltenberg

November 4, 2019

Contents

1. Coin tossing

Exercise 1.1. A fair coin is tossed two times. At least one of the tosses came up heads. What is the probability that both came up heads?

Exercise 1.2. In the morning I roll two dice to determine at what time of the day I'll toss my fair coin the first and the second time. Die showing 1 meaning coin flip in the time interval 13:00-13:01, die showing 2 meaning coin flip at 14:00-14:01, and so on. The one minute interval is there in the case that the two dice are equal, giving me time to flip the coin two times. Sometime after 18:01 you're told that at least one of the coins, flipped sometime between 16:00 and 16:01, came up heads. What is the probability that both came up heads?

Exercise 1.3. Helsesøster or bror (school nurse) wants to know the fraction of teenagers who have experienced E, where E is something embarrassing, in fact so embarrassing that the teenagers might not be honest in their 'yes' or 'no' answer when questioned about having experienced E. The school nurse therefore sets up the following anonymisation scheme: Each teenager the school nurse samples are to toss a coin, the outcome of which they keep to themselves. If their coin shows heads, they must answer truthfully. If the coin shows tails, they toss the coin once more. If this second toss shows heads, they answer 'yes', if it shows tails they answer 'no'. Of 17 teenagers 4 answered yes, the remaining ones answered no. The school nurse wants to know the true proportion of 'yes I've experienced E'-teenagers in the population.

(a) State any additional assumptions you feel you need, and provide an unbiased estimator of the true proportion of 'yes'-teenagers. Based on the data given to you by the school nurse, what's your estimate of this proportion?

(b) Explain why the school nurse ought to question more teenagers if (s)he wants a good estimate. That is, explain why, with high probability, it is worthwhile questioning some more teenagers.

(c) Suppose that the school nurse is out of touch with the youth of today, and that the teeneagers are, contrary to the beliefs of the school nurse, fully willing to provide her/him with an honest answer. How much does the school nurse lose by introducing the anonymisation scheme in this situation. *Hint*:

Compare the variance of the estimator you found above, with the variance of the estimator you think the school nurse would have used had (s)he known that (s)he were dealing with honest and upright youth.

Exercise 1.4. A natural way to compare estimators is to look at the loss they incur. By 'loss' we mean a non-negative function $L(\theta, \delta)$, where θ is the parameter of the model, and δ is your estimator (or more generally, you chosen action). We are interested in the performance of an estimator δ when it is repeatedly applied. The long-term average loss of using δ is the risk function $R(\theta, \delta)$, defined by

$$R(\theta, \delta) = \mathrm{E}_\theta\, L(\delta, \theta).$$

The problem of comparing estimators is now cast as the problem of comparing their associated risk functions, and we ought to use the estimator that minimises the risk. As this exercise will indicate, this latter statement is not sufficiently precise to be operational.

Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variables with expectation θ. We are to estimate θ under the squared error loss function

$$L(\delta, \theta) = (\delta - \theta)^2.$$

(a) Find the maximum likelihood estimator of θ and, for $n = 10$, sketch its risk function.

(b) Suppose we have some intuition about where on the unit interval the expectation θ might be located, close to a value $0 < \theta_0 < 1$, say. Consider the estimator

$$\delta_1(X) = \bar{X}_n/2 + \theta_0/2,$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. For $\theta_0 = 1/2$ and $n = 10$, sketch its risk function.

(c) Your task, as was Pierre-Simon Laplace's in 1781 I believe, is to estimate the probability of giving birth to a boy. Which of the two above estimators do you prefer, the maximum likelihood estimator or δ_1 with $\theta_0 = 1/2$?

(d) Consider the estimator

$$\delta_w(X) = w_n \bar{X}_n + (1 - w_n)\frac{1}{2}.$$

Find a function w_n such that the risk function $R(\theta, \delta_{w_n})$ is constant.

(e) For $n = 10$, sketch the risk function of your estimator (that is, draw a line).

(f) Suppose you have absolutely no idea whatsoever about where in the unit interval θ may be located. Which of your three estimators of θ do you prefer?

2. Probabilities, distribution functions, and random variables

Exercise 2.1. Consider the function

$$F(x) = \begin{cases} (1-\theta)x, & 0 \le x < \tau, \\ \theta + (1-\theta)x, & \tau \le x \le 1. \end{cases}, \quad 0 < \theta, \tau < 1, \tag{2.1}$$

and $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 1$.

(a) Show that F is a distribution function.

(b) Sketch how you would simulate data from F.

(c) Find the expectation and the variance of $X \sim F$.

(d) Find the expected sample size needed to identify τ with probability 1.

(e) Suppose that τ is known and that it does not equal $1/2$. Propose two estimators for θ and compare their variances.

Exercise 2.2. (Pólya urn). An urn contains r red balls and b blue balls. A ball is drawn from the urn at random, its colour is noted, and the ball is returned to the urn along with d more balls of the same colour. This is repeated indefinitely.

(a) What's the probability that the second ball drawn is red?

(b) What's the probability that the k'th ball drawn is red?

(c) What's the probability that the first ball drawn was red, given that the third ball drawn is red?

Exercise 2.3. Probability measures are continuous. If $A_1 \subset A_2 \subset \cdots$ is a sequence of events, then

$$P(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n). \tag{2.2}$$

And if $A_1 \supset A_2 \supset \cdots$ is a sequence of events, then

$$P(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n). \tag{2.3}$$

Prove (2.2) and (2.3).

Exercise 2.4. Suppose that P is a finitely additive probability measure with the property that if $A_1 \subset A_2 \subset \cdots$ is a sequence of events, then $P(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n)$. Show that P is countably additive.

Exercise 2.5. Let $F(x)$ be a distribution function. Show that F has at most a countable number of discontinuities, i.e., points x such that $F(x) - F(x-) > 0$, where $F(x-) = \lim_{y\uparrow x} F(x)$.

3. Transformations of random variables

Exercise 3.1. Let X be a standard normal random variable, and set $Y = I\{X \geq c\}$ for constant c.

(a) Find the distribution of Y.

(b) Based on independent draws $Y_1, \ldots, Y_n$, find the maximum likelihood estimator, say $\hat{c}_n$, of c. *Hint:* The maximum likelihood estimator is in this case the estimator your intuition leads you to.

(c) This exercise requires material we have yet to cover in class. Show that

$$\sqrt{n}(\hat{c}_n - c) \xrightarrow{d} \mathrm{N}\big(0, \frac{\Phi(c)(1-\Phi(c))}{\phi(c)^2}\big),$$

as n tends to infinity, where $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ is the standard normal probability density function, and $\Phi(x) = \int_{-\infty}^{x} \phi(y)\,\mathrm{d}y$. *Hint*: Use the delta method. In other words, use the mean value theorem; the fact that $1 - \bar{Y}_n = 1 - n^{-1}\sum_{i=1}^{n} Y_i$ is consistent for $\Phi(c)$; and that $X_n \xrightarrow{p} \eta$, implies $g(X_n) \xrightarrow{p} g(\eta)$, when g is continuous.

Exercise 3.2. Go to the 2014 STK4011 website and do Exercises 1 and 2 ('Transformations of random variables' and 'Transformations of random vectors') from Nils Lid Hjort's 'Exercises and Lecture Notes' `https://www.uio.no/studier/emner/matnat/math/STK4011/h14/exercises_stk4011a.pdf`.

Exercise 3.3. (An exercise from Nils' lecture notes). Let X and Y be independent standard normals, and transform to polar coordinates,

$$X = R\cos\theta, \quad Y = R\sin\theta.$$

Find the distribution of the random length R and the random angle θ, and show that these are independent.

Exercise 3.4. Two more exercises from Nils' lecture notes. Do Exercise 4. Ordering exponentials, and Exercise 5. Ratios of ordered uniforms.

4. Moment generating functions

Exercise 4.1. Let X be a Poisson random variable with mean λ.

(a) Find the moment generating function $M_X(t)$ of X.

(b) Let Y_n be binomial(n, p_n), and assume that np_n tends to λ as $n \to \infty$. Show that the moment generating function of Y_n tends to $M_X(t)$

(c) Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variables with expectations $p_{1,n}, \ldots, p_{n,n}$. Set $Z_n = \sum_{i=1}^n X_i$. Assume that $\sum_{i=1}^n p_{i,n}$ tends to λ and that $\max_{i \le n} p_{i,n}$ tends to 0, as $n \to \infty$. Show that the moment generating function of Z_n tends to $M_X(t)$.

Exercise 4.2. Let X be a standard normal random variable and set $Y = \exp(X)$.

(a) Find the distribution of Y. What's the name of this distribution?

(b) Show that all the moments of Y exist.

(c) Show that the random variable Y does not have a moment generating function.

Exercise 4.3. Let X_n be a sequence of random variables with moment generating functions $M_n(t)$, X be a random variable with moment generating function $M(t)$, and suppose that $P(0 \le X_n \le 1) = P(0 \le X \le 1) = 1$ for all n. Suppose that all the moments of X_n converge to the moments of X, that is $\mathrm{E}\, X_n^k \to \mathrm{E}\, X^k$ for $k = 1, 2, \ldots$. Show that $M_n(t) \to M(t)$.

5. Exponential families

Lemma 5.1. *If $\partial g(x,\theta)/\partial\theta$ exists and is continuous in θ for all x and all θ in an open interval S, and if $|\partial g(x,\theta)/\partial\theta| \le k(x)$ for all $\theta \in S$ for some integrable function $k(x)$, and if $\int g(x,\theta)\,\mathrm{d}\nu(x)$ exists on S, then*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int g(x,\theta)\,\mathrm{d}\nu(x) = \int \frac{\partial}{\partial\theta}\, g(x,\theta)\,\mathrm{d}\nu(x).$$

Exercise 5.1. Prove Lemma 5.1. *Hint*: Use first the mean value theorem, then Dominated convergence.

Exercise 5.2. Let $f_\theta(x)$ be a density function, $\theta \in \mathbb{R}^k$, and suppose that $f_\theta(x)$ and $\partial f_\theta(x)/\partial\theta_j$ satisfy the conditions of Lemma 5.1 for each $j = 1, \ldots, k$.

(a) Define $u_j(\theta; x) = \partial \log f_\theta(x)/\partial\theta_j$, and show that

$$\mathrm{E}_\theta\, u_j(\theta; X) = 0, \quad \text{and} \quad \mathrm{E}_\theta\, u_j(\theta; X) u_l(\theta; X) = -\mathrm{E}_\theta\, \frac{\partial^2}{\partial\theta_j \partial\theta_l} \log f_\theta(X).$$

(b) Suppose that $f_\theta(x)$, $\theta \in \Theta \subset \mathbb{R}^p$ is an exponential family in its natural parametrisation (i.e., $w_j(\theta) = \theta_j$ in Eq. (5.1)). Use (a) to find expressions for the expectation and variance of $t_j(X)$ $j = 1, \ldots, p$.

(c) Let X be Gamma(a, b) with density $b^a/\Gamma(a) x^{a-1} \exp(-bx)$. Show that X has an exponential family distribution and find the expectation of $\log X$.

(d) Consider a sequence of independent Bernoulli trials with success probability p. Let X be the number of failures until the first success. Show that X has an exponential family distribution and find the expectation of X.

Exercise 5.3. Let $X_1, \ldots, X_n$ be independent draws from a distribution with density $f_\theta(x)$ of exponential family form, that is

$$f_\theta(x) = h(x)c(\theta) \exp\Big\{ \sum_{j=1}^k w_j(\theta) t_j(x) \Big\}, \quad x \in \mathcal{X},\ \theta \in \Theta, \tag{5.1}$$

with the sample space $\mathcal{X}$ not depending on θ; $h(x)$ and $c(\theta)$ are non-negative functions; and the $w_j(\theta), t_j(x)$, $j = 1, \ldots, k \ge 1$, are real valued functions. We suppose that $\Theta \subset \mathbb{R}^p$.

(a) Show that the joint distribution of $X_1, \ldots, X_n$ is also an exponential family.

(b) Define $\tilde{t}_j(x) = \sum_{i=1}^n t_j(x_i)$ for $j = 1, \ldots, k$. For the discrete case, that is when $f_\theta(x) = P_\theta(X = x)$, show that the joint distribution of $\tilde{t}_1(X), \ldots, \tilde{t}_k(X)$ is an exponential family with natural parameters $w_1(\theta), \ldots, w_k(\theta)$.

Exercise 5.4. Let $Y_1, \ldots, Y_n$ be independent Bernoulli trials with success probabilities $p_1, \ldots, p_n$. These are linked to covariates $x_1, \ldots, x_n$, regarded as known constants, through

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \quad i = 1, \ldots, n.$$

Show that the joint distribution of $X_1, \ldots, X_n$ is an exponential family.

Exercise 5.5. Let $Y_1, \ldots, Y_n$ be independent $\mathrm{N}(\beta_0 + \beta_1 x_i, 1)$, where $x_1, \ldots, x_n$ are known constants. Show that the joint distribution of $X_1, \ldots, X_n$ is an exponential family.

Exercise 5.6. Let $X_1, \ldots, X_n$ be independent Poisson random variables with mean λ. Show that the joint distribution of X_i, $i = 1, \ldots, n$ is an exponential family, and identify $\tilde{t}(x)$. Find the conditional distribution of $X_1, \ldots, X_n$ given $\tilde{t}(X) = t$. What's noticable about this distribution?

Exercise 5.7. Consider the two-parameter exponential family with $t_1(x) = x$ and $t_2(x) = x^2$, and $h(x) = 1$ on $\{0, 1, 2\}$, and zero outside. Find its natural parameter space and its density.

Exercise 5.8. Consider the one-parameter exponential family $f_\theta(x)$ with $h(x) = 1/x$ on $(0, 1]$, zero elsewhere, and $t(x) = \log x$.

(a) Find its natural parameter space and its density.

(b) Find the distribution of $Y = -\log X$.

(c) Let $X_1, \ldots, X_n$ be independent draws from $f_\theta(x)$. Find the maximum likelihood estimator $\hat{\theta}_n$ of θ and show that it is approximately unbiased.

(d) Find the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$.

6. Sufficiency, ancilliarity and completeness

Exercise 6.1. Suppose $T(X)$ is statistic such that $\theta \mapsto P_\theta(X = x)/P_\theta(T(X) = t)$ is constant for all x. Show that $T(X)$ is sufficient.

Exercise 6.2. Do Exercise 6.9 in Casella and Berger (2002). That is, let $X_1, \ldots, X_n$ be a random sample. For each of the following distributions, find a minimal sufficient statistic for θ.

(a) $f_\theta(x) = (2\pi)^{-1/2} \exp(-(x - \theta)^2/2),\ x, \theta \in \mathbb{R}$;

(b) $f_\theta(x) = \exp(-(x - \theta)),\ x > \theta, \theta \in \mathbb{R}$;

(c) $f_\theta(x) = \exp(-(x - \theta))/\{1 + \exp(-(x - \theta))\},\ x, \theta \in \mathbb{R}$;

(d) $f_\theta(x) = 1/\{\pi(1 + (x - \theta)^2)\},\ x, \theta \in \mathbb{R}$;

(e) $f_\theta(x) = \exp(-|x - \theta|)/2,\ x, \theta \in \mathbb{R}$.

Exercise 6.3. Let $X_1, \ldots, X_n$ be independent draws from the uniform distribution on $(0, \theta),\ \theta > 0$.

(a) Show that $T(X) = T(X_1, \ldots, X_n) = \max_{i \le n} X_i$ is minimal sufficient for θ.

(b) Find a so that $\hat{\theta}_{\max} = aT(X)$ is an unbiased estimator of θ, and compute the variance of this estimator. *Hint*: Start by finding the density of $T(X)$.

(c) Show that the estimator $\hat{\theta}_{\mathrm{mean}} = 2\bar{X}_n$ is unbiased for θ. Compare the variance of $\hat{\theta}_{\mathrm{mean}}$ with the variance of $\hat{\theta}_{\max}$.

(d) Consider the class of estimators $\delta_a(X) = aT(X)$. Find the value of a, say $a^\star$, that minimises the risk function $R(\theta, \delta_a) = \mathrm{E}_\theta(aT(X) - \theta)^2$. Compare the risk functions of $\delta_{a^\star}$ and $\hat{\theta}_{\max}$.

Exercise 6.4. Let g be a positive and integrable function on $(0,\infty)$. Set $c(\theta)^{-1} = \int_\theta^\infty g(x)\,\mathrm{d}x$, and define $f_\theta(x) = c(\theta)g(x)$ for $x > \theta$ and zero otherwise. Suppose $X_1,\ldots,X_n$ are independent draws from $f_\theta(x)$.

(a) Show that $T(X) = \min_{i\le n} X_i$ is sufficient for θ.

(b) Show that $T(X)$ is minimal sufficient.

Exercise 6.5. Let $g(x)$ be a positive integrable function on $(-\infty,\infty)$. For $a < b$, Set $c(a,b)^{-1} = \int_a^b g(x)\,\mathrm{d}x$, and define $f_{(a,b)}(x) = c(a,b)g(x)$ for $a < x < b$, and zero otherwise. If $X_1,\ldots,X_n$ are independent from $f_\theta(x)$, show that $(X_{(1)}, X_{(n)}) = (\min_{i\le n} X_i, \max_{i\le n} X_i)$ is minimal sufficient for (a,b).

Exercise 6.6. (Lehmann and Casella (1999))Let $X_1,\ldots,X_n$ be independent and identically distributed from a contionuous distribution F, that is otherwise unknown. Let $T(X) = (X_{(1)},\ldots,X_{(n)})$ be the order statistics. Show that T is sufficient.

(a) Let $U_1(X) = \sum_{i=1}^n X_i$, $U_2(X) = \sum_{1\le i<j\le n} X_iX_j$, $U_3(X) = \sum_{1\le i<j<k\le n} X_iX_jX_k$, and so on, with, $U_n(X) = X_1\cdots X_n$. Show that $U(X) = (U_1(X),\ldots,U_n(X))$ is sufficient.

(b) Let $V_k(X) = X_1^k + \cdots + X_n^k$, and set $V(X) = (V_1(X),\ldots,V_n(X))$. Show that V is sufficient.

Exercise 6.7. Let X be a single observation from $\mathrm{N}(0,\theta)$, $\theta > 0$. Show that both X and $|X|$ are sufficient for θ.

(a) Are they both minimal sufficient?

(b) Let U be Bernoulli(1/2), and set $X' = U|X| - (1-U)|X|$. Show that X' has the same distribution as X.

Exercise 6.8. Let $Y_1,\ldots,Y_n$ be independent $\mathrm{N}(\beta x_i, 1)$, where $x_1,\ldots,x_n$ are fixed constants not all zero.

(a) Show that the least-squares estimator $\hat\beta = \sum_{i=1}^n x_iY_i / \sum_{i=1}^n x_i^2$ is complete and sufficient.

(b) Show that $\hat\beta$ and $\sum_{i=1}^n (Y_i - x_i\hat\beta)^2$ are independent.

Exercise 6.9. Suppose that given λ, the random variables $X_1,\ldots,X_n$ are independent Poisson with mean λ, while λ itself stems from an exponential distribution with mean $1/\theta$. Find a minimal sufficient statistic for θ.

Exercise 6.10. For $i = 1,\ldots,n$, let ε_i and X_i be independent standard normal random variables. For $\theta \in (0,1)$, set

$$Y_i = \theta X_i + \sqrt{1-\theta^2}\varepsilon_i, \quad i = 1,\ldots,n.$$

Suppose we observe the pairs (X_i, Y_i), $i = 1,\ldots,n$.

(a) Find a minimal sufficient statistic for θ.

(b) Is your minimal sufficient statistic complete?

(c) Show that $\sum_{i=1}^n X_i^2$ and $\sum_{i=1}^n Y_i^2$ are ancilliary, but that $(\sum_{i=1}^n X_i^2, \sum_{i=1}^n Y_i^2)$ is not.

Exercise 6.11. Let $Y_1,\ldots,Y_n$ be independent Bernoulli random variables with success probabilites $p_i = 1/\{1+\exp(-\beta_0 - \beta_1 x_i)\}$, $i = 1,\ldots,n$, for fixed and known x_i. Find a minimal sufficient statistic for $\theta = (\beta_0,\beta_1)$.

Exercise 6.12. (Keener (2011)) Let $X_1,\ldots,X_n$ be independent random variables from a Beta (a,b) distribution. Recall that the density of this distribution is

$$f_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in (0,1),$$

for positive parameters a and b. Find a minimal sufficient statistic (i) when a and b vary freely; (ii) when $a = 2b$; and when (iii) $a = b^2$.

Exercise 6.13. Let $X_1, \ldots, X_n$ be independent $\mathrm{N}(\theta, \theta^2)$, $\theta > 0$. Find a minimal sufficient statistic for θ, and show that it is not complete. Explain why Theorem 6.2.25 in Casella and Berger (2002, p. 288) does not apply.

Exercise 6.14. Let $X_1, \ldots, X_n$ be independent $\mathrm{N}(\theta, \sigma^2)$. Show that $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ are independent. *Hint*: Use Basu's theorem.

Exercise 6.15. (Casella and Berger (2002)) Let N be a random variable taking values in $\{1, 2, \ldots, \}$ with known probabilities $p_1, p_2, \ldots$. Given $N = n$, perform n independent Bernoulli trials $X_1, \ldots, X_n$ with success probabilities θ.

(a) Show that $(\sum_{i=1}^N X_i, N)$ is minimal sufficient and that N is ancilliary for θ

(b) Show that $N^{-1} \sum_{i=1}^N X_i$ is unbiased for θ, and find its variance.

Exercise 6.16. (Shao (2003)) Let T and S be two statistics such that $S = g(T)$ for some measurable function g. Show that

(a) If T is complete, then S is complete.

(b) If T is complete and sufficient and g is one-to-one, then S is complete and sufficient.

Exercise 6.17. Let X and Y be independent Poisson random variables with means θ and θ^2, respectively. Find a minimial sufficient statistic. Is the minimal sufficient statistic complete?

Exercise 6.18. Let $X_1, \ldots, X_n$ be independent random variables with density,

$$f_\theta(x) = \frac{1}{\theta} \exp\{-(x-\theta)/\theta\}, \quad \text{for } x > \theta > 0,$$

and zero otherwise. Find a statistic that is minimal sufficient for θ. Is the minimal sufficent statistic complete?

Exercise 6.19. Let $X_1, \ldots, X_n$ be independent exponentials with mean $1/\theta$. Show that $\bar{X}_n$ and $\max_{i \le n} X_i / \min_{i \le n} X_i$ are independent.

Exercise 6.20. Let $X_1, \ldots, X_n$ be independent uniform random variables on (a, b). Show that $T(X) = (\min_{i \le n} X_i, \max_{i \le n} X_i)$ is sufficient and complete.

Exercise 6.21. Let θ be a real-valued parameter that we are to estimate with a loss function $L(\delta, \theta)$ that is convex in δ for each θ. Let $\delta_1(X)$ be an unbiased estimator of θ and suppose that T is a sufficient statistic. Consider

$$\delta_2(X) = \mathrm{E}_\theta[\delta_1(X) \mid T].$$

(a) Explain why δ_2 is an estimator, and show that it is unbiased.

(b) Recall that the risk function of an estimator δ is $R(\delta, \theta) = \mathrm{E}_\theta\, L(\delta(X), \theta)$. Show that

$$R(\delta_2, \theta) \le R(\delta_1, \theta), \quad \text{for all } \theta.$$

What additional assumptions do we need for this inequality to be strict?

(c) Suppose that T is also complete. Show that $R(\delta_2, \theta) \le R(\delta, \theta)$ for all competitors δ.

(d) Go back to Exercise 6.3(c), and explain what you found in view of the current exercise.

7. Miscellaneous exercises

Exercise 7.1. It's raining at Blindern. Is it raining at Huk? The Norwegian Meteorological Institute have pluviometers, or rain gauges, stationed both places, with daily rain measurements in milimeters available on the website yr.no. I found the data for Blindern and Huk here and here, respectively. Here is the rain data for Blindern and for Huk as text-files, use `read.table(path,sep=";")`.

Denote the rain data for the $n = 248$ days from January 1. to September 5., 2019 by,

$$\begin{aligned} &\text{Blindern}: \quad Y_{B,1}, \ldots, Y_{B,n}; \\ &\text{Huk}: \quad Y_{H,1}, \ldots, Y_{H,n}. \end{aligned}$$

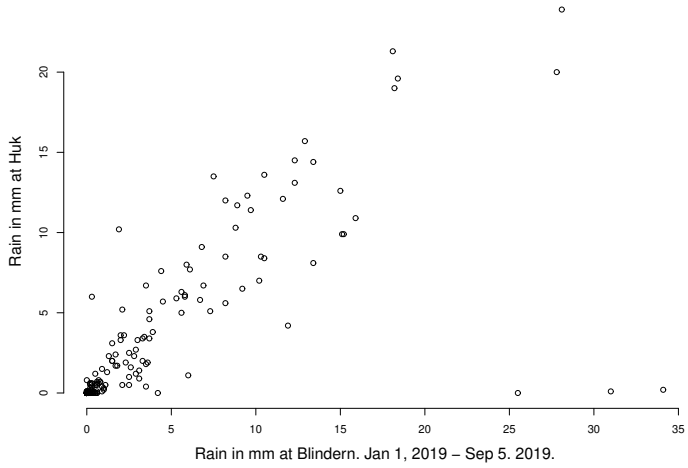

FIGURE 1. Rainfall measured in milimeters at Blindern and Bygdøy weather stations. Data from yr.no, extracted Sept. 5,2019.

In order to predict the weather at Huk by looking out the window at Blindern, we need a statistical model.

(a) All you care about is whether or not it's raining, the amount of rain does not bother you. Sketch how you would estimate Pr(Rain at Huk | Rain at Blindern), and how you would assess the uncertainty of this estimate. Think about what assumptions you're making when using your chosen estimator.

(b) A little rain does not stop you. With the model

$$Y_{H,i} = \beta_0 + \beta_1 Y_{B,i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are independent mean zero random variables with variance σ^2, we can, after having estimated the parameters, say something about quantities such as $\mathrm{E}\,[Y_H \mid Y_B = y]$, $\Pr(Y_H \leq y \mid Y_B = y)$, and so on. Is this a good model for the phenomenon you are interested in? Why, or why not?

(c) Consider the following model. Let $F_{\lambda_j}(y) = 1 - \exp(-\lambda_j y),\ j = B, H$, for $y \geq 0$ and positive parameter λ_j. As (almost) always, let $\Phi(x)$ be the distribution function of the standard normal distribution, and take

$$Y_{j,i} = F_{\lambda_j}^{-1}(\Phi(X_{j,i})), \quad j = H, B,\ i = 1, \ldots, n,$$

where

$$\begin{pmatrix} X_{B,i} \\ X_{H,i} \end{pmatrix} \sim \mathrm{N}(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}), \quad i = 1, \ldots, n,$$

are independent. Denote $\theta = (\lambda_B, \lambda_H, \rho)$, and derive an expression for the log-likelihood function

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(y_{B,i}, y_{H,i}; \theta),$$

where $f(y_{B,i}, y_{H,i}; \theta)$ is the density of (Y_B, Y_H).

(d) Are you comfortable with the assumptions the model in (c) is making about rain in Oslo?

(e) If you program the likelihood function you found in (b) and feed it to an optimiser such as `nlm()` in R, things will go astray (the `nlm()`-function is a minimiser, so give it the negative log-likelihood). Why is that? Propose a modification of the model in (c) that takes care of this problem. *Hint*: Fortunately, there are sunny days.

Exercise 7.2. Let X and Y be random variables with $\mathrm{E}\,X = \xi, \mathrm{Var}\,X = \tau^2$, and $\mathrm{E}\,Y = \mu$, $\mathrm{Var}\,Y = \sigma^2$.

$$P(\{|X - \xi| \geq \varepsilon\tau^2\} \cup \{|Y - \mu| \geq \varepsilon\sigma^2\}) \leq \frac{1 + \sqrt{1 - \rho^2}}{\varepsilon^2}.$$

Hint: If X and Y are mean zero random variables with unit variance, and correlation ρ, establish that $\mathrm{E}\max\{X^2, Y^2\} \leq 1 + \sqrt{1-\rho^2}$. Here you may want to use that $\mathrm{E}\,|XY| \leq (\mathrm{E}\,X^2)^{1/2}(\mathrm{E}\,Y^2)^{1/2}$.

Exercise 7.3. Let X be a random variable with probability distribution $P(X = x_j) = p_j$ for $j = 1, \ldots, r$, with $r \geq 2$. We define the entropy $H = H(p_1, \ldots, p_{r-1})$ of this distribution by

$$H = -\sum_{j=1}^{r} p_j \log(p_j),$$

with $0 \times \log(0) = 0$.

(a) Show that the entropy is maximised when X is uniformly distributed over $\{x_1, \ldots, x_r\}$, and that it is minimised when X takes on one value with probability one. *Hint*: Use Jensen's inequality.

(b) Let $X_1, \ldots, X_n$ be independent Bernoulli(p) random variables, $0 < p < 1$. Show that $Y = \sum_{i=1}^n X_i$ is Binomial(n, p). Let $B_n(\varepsilon)$ be the set of all Bernoulli(p) sequences of length n, such that $|\sum_{i=1}^n X_i/n - p| < \varepsilon$, for some $\varepsilon > 0$. We'll write $B_n(\varepsilon) = \{|Y/n - p| < \varepsilon\}$. Show that the probability of $B_n(\varepsilon)$ tends to 1 as $n \to \infty$.

(c) Let $H(p)$ be the entropy associated with the Bernoulli(p) distribution. Convince yourself of the following

$$\frac{\text{Bernoulli sequences of length } n \text{ giving } Y = y}{\text{All Bernoulli sequences of length } n} = \binom{n}{y} \exp\{-nH(1/2)\}.$$

Show that

$$|\log P(X_1 = x_1, \ldots, X_n = x_n) + nH(p)| \leq \varepsilon n |\log p(1-p)|,$$

when $x_1, \ldots, x_n$ is a sequence in $B_n(\varepsilon)$.

(d) Define $N_n(\varepsilon)$ to be the number of Bernoulli(p) sequences in $B_n(\varepsilon)$. We are going to show that there is an n_0 such that for all $n \geq n_0$,

$$\exp\{n(H(p) - \varepsilon)\} \leq N_n(\eta) \leq \exp\{n(H(p) + \varepsilon)\}, \tag{7.1}$$

where $\eta = -\varepsilon/\{2\log p(1-p)\}$. Write $X^{(n)}$ for Bernoulli sequences of length n, and $x^{(n)} = (x_1, \ldots, x_n)$ for zero-one sequences of length n. Show that for $x^{(n)} \in B_n(\eta)$,

$$P(X^{(n)} = x^{(n)}) \leq \exp\{-n(H(p) - \varepsilon/2)\}, \quad \text{and} \quad P(X^{(n)} = x^{(n)}) \geq \exp\{-n(H(p) + \varepsilon/2)\}.$$

Use these bounds to show that

$$N_n(\eta) \leq \exp\{n(H(p) + \varepsilon/2)\}, \quad \text{and} \quad N_n(\eta) \geq P(B_n(\eta)) \exp\{n(H(p) - \varepsilon/2)\}.$$

Combine this with what you found in (b), and derive (7.1).

8. Convergence concepts

Exercise 8.1. Let $X_1, X_2, \ldots$ be independent Bernoulli variables with success probabilities $p_1, p_2, \ldots$. We shall investigate when

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

where $B_n = \{\sum_{i=1}^n p_i(1-p_i)\}^{1/2}$. Show, using mgf's, that this happens if and only if $\sum_{i=1}^\infty p_i = \infty$. Show also that this condition is equivalent to $B_n \to \infty$. This means that the cases $p_i = 1/i$ and $p_i = 1/i^2$, for example, are very different.

Exercise 8.2. Suppose X_n is Beta($1/n, 1/n$) and X is Bernoulli(1/2). Show that $X_n \xrightarrow{d} X$. What if X_n is Beta($a/n, b/n$)? *Hint*: See Nils exercise 12.

Exercise 8.3. Suppose X_n is uniformly distributed on $\{1/n, 2/n, \ldots, 1\}$. Show that X_n converges in distribution to X, where X is uniform$(0,1)$. Does $X_n \xrightarrow{p} X$?

Exercise 8.4. A few counterexamples.

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

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

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

Exercise 8.5. Let $X_1, X_2, \ldots$ be i.i.d. with density $f(x) = ax^{-(a+1)}$ for $x \in (1, \infty)$, and zero otherwise.

(a) For what values of $a > 0$ is it true that $X_n/n \xrightarrow{p} 0$?

(b) For what values of $a > 0$ and $r > 0$ is it true that $\mathrm{E}\, X_n^r/n \to 0$?

(c) For what values of $a > 0$ is it true that $X_n/n \to 0$ almost surely? *Hint*: Use the Borel–Cantelli lemma.

Exercise 8.6. Let $Y_1, \ldots, Y_n$ be independent Poisson variables with density $P_\theta(X = x) = e^{-\theta}\theta^x/x!$, $x = 0, 1, 2, \ldots$, and let Z_n be the proportion of zeros observed, $Z_n = n^{-1}\sum_{i=1}^n I\{X_i = 0\}$. Show that

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \theta \\ Z_n - e^{-\theta} \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \theta & -\theta e^{-\theta} \\ -\theta e^{-\theta} & e^{-\theta}(1 - e^{-\theta}) \end{pmatrix}\Big).$$

Exercise 8.7. Consider independent observations $Y_1, \ldots, Y_n$ from a normal distribution with expectation μ and variance σ^2.

(a) Write down the log-likelihood function $\ell_n(\mu, \sigma)$, and derive formulae for the maximum-likelihood estimators, say $(\hat{\mu}_n, \hat{\sigma}_n^2)$. Identify also the exact distributions of $\hat{\mu}_n$ and $\hat{\sigma}_n^2$. *Hint*: The maximum-likelihood estimators are the estimators solving $\partial\ell_n(\mu, \sigma)/\partial\mu = 0$ and $\partial\ell_n(\mu, \sigma)/\partial\sigma = 0$.

(b) Show that

$$\sqrt{n}\begin{pmatrix} \hat{\mu}_n - \mu \\ \hat{\sigma}_n - \sigma \end{pmatrix} \xrightarrow{d} \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2/2 \end{pmatrix}\Big).$$

Hint: You can show this by computing it in a straight forward manner, using tools that we covered in class. But, if you want to, take a look at the variance of the score function (the first derivative of the log-likelihood) and its inverse; look back at Exercise 5.2; and Taylor-expand the score function around the true values of the parameters. We'll do all this on Thursday.

(c) Consider the parameter $\gamma = \mu/\sigma$, sometimes called the normalised mean. With $\hat{\gamma}_n = \hat{\mu}_n/\hat{\sigma}_n^2$, show that

$$\sqrt{n}(\hat{\gamma}_n - \gamma) \xrightarrow{d} \mathrm{N}(0, 1 + \gamma^2/2).$$

(d) Let now $h(x) = \sqrt{2}\log(x/\sqrt{2} + \sqrt{1 + x^2/2})$. Show that

$$\sqrt{n}(h(\hat{\gamma}_n) - h(\gamma)) \xrightarrow{d} \mathrm{N}(0, 1).$$

9. Finding estimators

Exercise 9.1. Let $X_1, \ldots, X_n$ be independent Gamma(a, b) random variables with density

$$f_{(a,b)}(x) = \frac{b^a}{\Gamma(a)} x^{a-1}\exp\{-bx\}, \quad x > 0. \tag{9.1}$$

(a) Write down the log-likelihood function $\ell_n(a, b)$.

(b) Suppose a is known. Find the maximum likelihood estimator $\hat{b}_n$ of b. Why is it unique?

(c) Find the limiting distribution of $\sqrt{n}(\bar{X}_n - a/b)$. Use this to find the limiting distribution of $\sqrt{n}(\hat{b}_n - b)$. Propose an approximate 95% confidence interval for b.

(d) Suppose now that both a and b are unknown. Write down the score functions $u_1(a,b,x) = \partial \log f_{(a,b)}(x)/\partial a$ and $u_2(a,b,x) = \partial \log f_{(a,b)}(x)/\partial b$, and show that

$$\mathrm{E}\begin{pmatrix} u_1(a,b,X) \\ u_2(a,b,X) \end{pmatrix} = 0.$$

Find also the variance matrix $J(a,b)$ of $(u_1(a,b,X), u_2(a,b,X))^{\mathrm{t}}$. *Hint*: Use what you found in Exercise 5.2.

(e) Explain why the maximiser $(\hat{a}_n, \hat{b}_n)$ of $\ell_n(a,b)$ is unique. Download the rain data for Blindern from Exercise 7.1, and remove the days with no rain. Fit a Gamma(a,b) model to these data. Provide point estimates and standard errors of your estimators. Here is a sketch of how to do this in R (there surely are other ways to go about this).

```
path <- "https://www.uio.no/studier/emner/matnat
/math/STK4011/data/rain_blindern2019_tom5sep.txt"
rain <- read.table(path,sep=";")
yy_full <- rain$rain_mm ;yy <- yy_full[yy_full>0] ; nn <- length(yy)
loglik <- function(params){
aa <- params[1] ; bb <- params[2]
ll <- # the log-likelihood here
return(ll)}
min_loglik <- function(params){ # nlm() is a minimiser
return(-loglik(params)) }
fit <- nlm(min_loglik,c(start_aa,start_bb),hessian=TRUE) # provide start values for nlm
```

Exercise 9.2. Consider the model

$$Y_i = \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

for independent $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, and fixed constants $x_1, \ldots, x_n$.

(a) Write down the log-likelihood function of this model.

(b) Find its score function $(u_1(\beta, \sigma^2, Y), u_2(\beta, \sigma^2, y))^{\mathrm{t}}$, and show that $\sum_{i=1}^n (u_1(\beta, \sigma^2, Y_i), u_2(\beta, \sigma^2, Y_i))^{\mathrm{t}}$ can be expressed in terms of a chi-square and normal random variable, and that these are independent. *Hint*: The third central moment of a normal distribution is zero.

(c) Find the observed information matrix $J_n(\beta, \sigma^2)$.

(d) Suppose σ^2 is known. Derive a 95 percent confidence interval for β. *Hint*: Start by identifying the distribution of $\hat{\beta}_n$.

Exercise 9.3. Let $X_1, \ldots, X_n$ be i.i.d. from a distribution with density $f_{\theta_0}(x)$, where $\theta_0 \in \Theta \subset \mathbb{R}$, and θ_0 denotes the true parameter value. Let $U_n(\theta) = \sum_{i=1}^n \partial \log f_\theta(X_i)/\partial\theta$. The maximum likelihood estimator is the value $\hat{\theta}_n$ of θ such that $U_n(\hat{\theta}_n) = 0$. By Taylor's theorem, assuming it is applicable, there exists a (random) value $\tilde{\theta}_n$ between $\hat{\theta}_n$ and θ_0 such that

$$0 = U_n(\hat{\theta}_n) = U_n(\theta_0) + U_n'(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2} U_n''(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^2, \tag{9.2}$$

where $U_n'(\theta) = \partial U_n(\theta)/\partial\theta$ and $U_n''(\theta) = \partial^2 U_n(\theta)/\partial\theta^2$. Under what conditions do you expect this to lead to

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}(0, 1/J(\theta_0)),$$

with $J(\theta_0)$ the Fisher information matrix evaluated in the true value.

10. Decision theory and Bayes

Exercise 10.1. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(θ). We wish to estimate θ, and we are particularly interested in precise estimates of very small and very large values of θ. Therefore, we'll work with the loss function

$$L(\delta, \theta) = \frac{(\delta - \theta)^2}{\theta(1-\theta)}.$$

(a) Compute the risk function of the maximum likelihood estimator. What's noticeable about this risk function?

(b) We now take a Bayesian point of view and give θ a $\text{Beta}(a\theta', a(1-\theta'))$ prior distribution. Compute the expectation and variance of this prior.

(c) With the prior introduced in (b), find the posterior distribution $\pi(\theta \mid x_1, \ldots, x_n)$. Find also the Bayes solution δ_π, i.e., the minimiser of the Bayes risk $\text{BR}(\delta, \theta) = \int R(\delta, \theta)\pi(\theta)\,\mathrm{d}\theta$.

(d) Tweak the parameters of the Beta prior distribution, so that the Bayes solution you found above equals the maximum likelihood estimator from (a). What desirable properties does the maximum likelihood estimator possess?

Theorem 10.1. *Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ be a finite parameter space, and suppose that δ_π is Bayes for the prior $\pi = \{\pi_1, \ldots, \pi_k\}$, where π_j is the prior mass given to θ_j. If $\theta_j > 0$ for $j = 1, \ldots, k$, then δ_π admissible.*

Exercise 10.2. Prove Theorem 10.1.

Exercise 10.3. Suppose $X_1, \ldots, X_n$ are i.i.d. from $\text{N}(0, \sigma^2)$. We are to estimate σ^2 under the loss function

$$L(\delta, \sigma^2) = \frac{(\delta - \sigma^2)^2}{\sigma^2}. \tag{10.1}$$

(a) Find the maximum likelihood estimator and its risk function.

(b) Consider the prior distribution given by density

$$\sigma^2 \sim \frac{b^a}{\Gamma(a)}(1/\sigma^2)^{a+1}\exp(-b/\sigma^2), \quad \sigma^2 > 0,$$

with $a > 1$ and $b > 0$. This is the density of an inverse gamma distribution. Find the prior expectation of σ^2. Find also the prior expectation of $1/\sigma^2$.

(c) Find the posterior distribution $\sigma^2 \mid x_1, \ldots, x_n$, and derive the Bayes solution under the loss function given in (10.1).

(d) Show that the maximum likelihood estimator is inadmissible by exhibiting an estimator, say $\delta^\star$, with uniformly smaller risk. *Hint*: Consider $\delta_\alpha = \alpha\hat{\sigma}^2_{\text{ml}}$.

(e) Is $\delta^\star$ admissible?

Exercise 10.4. (Partly from Nils' Bayes notes) When estimating the price of apples in Oslo, the height of women in Bergen, and the unemployment rate in Trondheim, it is sometimes advantageous to use information about apples is Oslo and women in Bergen to say something about the unemployment rate in Trondheim. The point is that when estimating an ensemble of unrelated things, we can sometimes do better in the estimation by borrowing information across unrelated things. This phenomenon is known as Stein's paradox or the Stein effect. See Stein (1956); James and Stein (1961) for the original articles, and, for example Efron and Morris (1977) and Stigler et al. (1990) for lucid presentations. In the present exercise we'll look at Stein's 1956–1961 result, a result that initiated a whole field of statistical research known as shrinkage estimation.

Let $Y_i \sim \text{N}(\theta_i, 1)$ be independent for $i = 1, \ldots, p$ with $p \geq 3$. We are to estimate $\theta_1, \ldots, \theta_p$ under the combined loss function

$$L(\delta, \theta) = \sum_{i=1}^{p}(\delta_i - \theta_i)^2.$$

The standard approach is to use Y_i as an estimator of θ_i. The estimator Y_i is the maximum likelihood estimator, it is admissible under $(\delta_i - \theta_i)^2$, it is the uniformly minimum variance unbiased estimator, etc.

(a) For obvious reasons, we call $Y = (Y_1, \ldots, Y_p)$ the standard estimator. Compute its risk function.

(b) For a single $Y \sim \mathrm{N}(\theta, 1)$, show that under very mild conditions on the function $b(y)$, one has

$$E_\theta\,(Y-\theta)b(Y) = E_\theta\, b'(Y),$$

where b' is the derivative of b. *Hint*: Use integration by parts.

(c) Let now $b(y) = (b_1(y), \ldots, b_p(y))$. Generalise what you found in (b) to

$$\mathrm{E}_\theta\,(Y_i-\theta_i)b_i(Y) = \mathrm{E}_\theta\, b_{i,i}(Y),$$

where $b_{i,i}(y) = \partial b_i(y)/\partial y_i$.

(d) What you found in (b) and (c) is known as Stein's lemma. We are now going to use Stein's lemma to construct an estimator that uniformly dominates Y. Consider a general competitor to Y of the form $\delta(Y) = (\delta_1(Y), \ldots, \delta_p(Y))$, with

$$\delta_i(Y) = Y_i - b_i(Y). \tag{10.2}$$

Show that the difference in risk between Y and estmators of the form (10.2) can be expressed as

$$R(\delta,\theta) - R(Y,\theta) = \mathrm{E}_\theta\, D(Y),$$

where

$$D(y) = \sum_{i=1}^{p}\{b_i(y)^2 - 2b_{i,i}(y)\}.$$

Then $R(\delta,\theta) = p + \mathrm{E}_\theta\, D(Y)$. The fabulous thing about such a simple lemma as Stein's, is that $D(y)$ does not depend on the unknown $\theta_1, \ldots, \theta_p$. We can therefore try to find a data dependent function $b(y)$ such that $D(y) < 0$ for all y, and consequently an estimator that uniformly dominates the standard estimator. It turns out to be impossible to find such functions $b(y)$ when $p \le 2$, but it is possible for $p \ge 3$.

(e) Try $b_i(y) = ay_i/\|y\|^2$, with $\|y\|^2$ being the squared Euclidian norm $\sum_{i=1}^p y_i^2$, corresponding to

$$\delta(y) = y - b(y) = \big(1 - \frac{a}{\|y\|^2}\big)y.$$

With this choice of $b(y)$, show that

$$D(y) = \frac{1}{\|y\|^2}\{a^2 - 2a(p-2)\}.$$

Show that this is negative for a range of a values provided $p \ge 3$. Demonstrate that the optimal a is $a = p-2$, corresponding to the estimator

$$\delta_{\mathrm{JS}}(Y) = \big(1 - \frac{p-2}{\|Y\|^2}\big)Y. \tag{10.3}$$

This estimator is known as the James–Stein estimator. Show that the risk function of this estimator can be expressed as

$$R(\delta_{\mathrm{JS}},\theta) = p - (p-2)^2\mathrm{E}_\theta\,\frac{1}{\|Y\|^2}.$$

Show that the greatest reduction in risk from using δ_{JS} instead of Y takes place when $\theta_1 = \cdots = \theta_p = 0$, and compute the risk $R(\delta_{\mathrm{JS}}, 0)$ in this point.

(f) We'll now make a connection to empirical Bayes procedures. Start with a prior that takes $\theta_1, \ldots, \theta_p$ independent from $\mathrm{N}(0, \tau^2)$. Show that the Bayes solution is $\delta^B = (\delta_1^B, \ldots, \delta_p^B)$, with

$$\delta_i^B(Y) = \alpha\, Y_i, \quad i = 1, \ldots, p, \quad \text{where} \quad \alpha = \frac{\tau^2}{\tau^2+1}. \tag{10.4}$$

(g) The empirical Bayes approach consists of estimating hyperparameters from data. Hyperparameters are those parameters set by the statistician in a pure Bayesian approach. Show that the marginal distribution of $y_1, \ldots, y_p$ is a product of $\mathrm{N}(0, 1+\tau^2)$ distributions. Find the maximum likelihood estimator of α. Use the maximum likelihood estimator to find an unbiased estimator, say $\tilde{\alpha}$, of α. The empirical Bayes estimator is then $\delta_{\mathrm{EB}}(Y) = \tilde{\alpha}Y$. What's noticeable about this estimator?

11. Testing statistical hypotheses

Exercise 11.1. (STK4011 Exam, Autumn 2014) Suppose θ is some parameter of interest, associated with observations $X_1, \ldots, X_n$. The null hypotheses traditionally dealt with in statistical testing are of the type $\theta = \theta_0$, or $\theta \leq \theta_0$, $\theta \geq \theta_0$, for some pre-specified value θ_0, or even $|\theta - \theta_0| \leq \varepsilon$, for some small positive ε. On this occasion we turn things slightly around, however, and wish to test $H_0 \colon |\theta - \theta_0| \geq \varepsilon$, versus the alternative that $|\theta - \theta_0| < \varepsilon$.

(a) Describe a situation where such a scenario would be fruitful.

(b) To give an illustration of more general constructions of the type pointed to above, suppose now that observations $X_1, \ldots, X_n$ are independent and normal $\mathrm{N}(\theta, 1)$, and assume for simplicity that $\theta \geq 0$ a priori. We shall test the hypothesis H_0 that $\theta \geq \varepsilon$, versus the alternative that $\theta < \varepsilon$, where we for concreteness set $\varepsilon = 1/4$. Consider the test which rejects H_0 if $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i \leq c_n$. Find c_n such that the test has significance level ('type I error') 0.05.

(c) Find the power function for this test, i.e. the probability that H_0 will be rejected, as a function of the parameter. Give a plot of this power function for $n = 100$. Comment on the size of the maximal power.

(d) How big must the sample size be, in order for the above power probability to be above 0.95, if in fact the true θ is equal to $\varepsilon/2$ (i.e. 1/8)?

(e) Show that the test worked with here, rejecting $H_0 \colon \theta \geq 1/4$ vs. the alternative $\theta < 1/4$ when $\bar{X}_n \leq c_n$, is uniformly most powerful, among all tests with significance level 0.05.

(f) Suppose θ is not restricted to be nonnegative a priori, and that one needs a test for $H_0 \colon |\theta| \geq 1/4$ versus the alternative that $|\theta| < 1/4$. Construct a test for this situation, again with significance level 0.05, and draw its power function alongside the one from point (c).

Exercise 11.2. (Casella and Berger (2002)) Suppose that we have two independent random samples: $X_1, \ldots, X_n$ are exponential(θ), and $Y_1, \ldots, Y_m$ are exponential(μ).

(a) Find the likelihood ratio test of $H_0 \colon \theta = \mu$ versus the alternative $\theta \neq \mu$.

(b) Show that the test in part (a) can be based on the statistic

$$T = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i + \sum_{i=1}^m Y_i}.$$

(c) Find the distribution of T when H_0 is true.

Exercise 11.3. (Casella and Berger (2002)) Let X_1, X_2 be i.i.d. uniform on $(\theta, \theta + 1)$. For testing $H_0 \colon \theta = 0$ versus the alternative $\theta > 0$, we have two competing tests:

$$\phi_1(X_1) \colon \text{Reject } H_0 \text{ if } X_1 > 0.95,$$
$$\phi_2(X_1, X_2) \colon \text{Reject } H_0 \text{ if } X_1 + X_2 > c.$$

(a) Find the value of c so that ϕ_2 has the same size as ϕ_1, and sketch the power functions of the two tests.

(b) Is it true that ϕ_2 is more powerful than ϕ_1?

(c) Find a test that has the same size but is more powerful than ϕ_2.

Exercise 11.4. (Keener (2011)) Suppose X has density

$$f_\theta(x) = \frac{\theta}{(1 + \theta x)^2}, \quad x > 0.$$

(a) Show that the derivative of the power function $\beta(\theta) = \mathrm{E}_\theta \, \phi(X)$ of a test ϕ is given by

$$\beta'(\theta) = \mathrm{E}_\theta \, \frac{1 - \theta X}{\theta(1 + \theta X)} \phi(X).$$

(b) Among all tests with $\beta(1) = \alpha$, which one maximises $\beta'(1)$?

Exercise 11.5. (Berger (1985)) Suppose X takes values in $\{1,2,3\}$, and that $\theta \in \{0,1\}$, with X having probability density in each case

$$f(x \mid 0) = 0.005I\{x=1\} + 0.005I\{x=2\} + +0.99I\{x=3\},$$
$$f(x \mid 1) = 0.0051I\{x=1\} + 0.9849I\{x=2\} + +0.01I\{x=3\}.$$

(a) Find the most powerful test of $H_0\colon \theta = 0$ versus the alternative $\theta = 1$ at level $\alpha = 0.01$. Compute the power of this test in θ_1.

(b) Suppose the unlikely event $X = 1$ occurs. Are you comfortable about the conclusion of the most powerful test in this case?

Exercise 11.6. (Lindley and Phillips (1976); Berger (1985)) We are interested in the probability θ of a coin coming up heads, and want to test the hypothesis $H_0\colon \theta \leq 1/2$ versus the alternative $\theta > 1/2$. The coin is tossed 12 times, leading to 9 heads and 3 tails.

(a) Suppose that it was determined in advance that the coin was to be tossed 12 times. A common thing to do is to reject H_0 if the probability of the observed data is smaller than $\alpha = 0.05$ when evaluated under the parameter value(s) of the null hypothesis. Is the null hypothesis rejected?

(b) Now suppose that the coin was tossed until 3 tails were observed, and that the third tail came on the twelfth toss. Use the same procedure as in (a) to test H_0.

(c) Try being Bayesian. Let θ have a prior Beta(a,b) distribution, and set a and b to what you think are reasonable values in this case. It now appears natural to no longer believe that H_0 is true if the null hypothesis has low posterior probability. What's low is up to you. Find the posterior distributions of θ under the models in (a) and (b). Is H_0 rejected?

Exercise 11.7. Let $X \sim f_\theta(x)$ and consider the simple hypothesis $H_0\colon \theta = \theta_0$ versus the simple alternative $\theta = \theta_1$. The statistical tests ϕ, with $\phi(x) = 1$ meaning 'reject H_0, and $\phi(x) = 0$ 'keep H_0', are to be evaluated under the loss function

$$L(\phi,\theta_0) = \begin{cases} 0, & \text{if } \phi(x)=0, \\ K_1, & \text{if } \phi(x)=1, \end{cases} \qquad L(\phi,\theta_1) = \begin{cases} K_2, & \text{if } \phi(x)=0, \\ 0, & \text{if } \phi(x)=1. \end{cases}$$

(a) Let $0 < \pi_0 < 1$ be your prior probability of H_0 being true. Derive an expression for the posterior expected loss, and show that the Bayes solution ϕ_π is of the likelihood ratio type

$$\phi_\pi(x) = \begin{cases} 1, & \text{if } f(x \mid \theta_1) > k_\pi f(x \mid \theta_0), \\ 0, & \text{if } f(x \mid \theta_1) < k_\pi f(x \mid \theta_0). \end{cases}$$

Find k_π and relate this quantity to the level of a test.

(b) Let now $X \mid \theta$ be N$(\theta,1)$. We want to test $H_0\colon \theta = 0$ versus $\theta_1 = 1/2$ using the Bayes solution when the prior is $\pi_0 = 1/2$. Find K_1 and K_2 such that $\mathrm{E}_{\theta_0}\,\phi_\pi(X) = 0.05$.

(c) Show that any Bayesian test with a prior giving weight to both the null- and the alternative hypothesis, is the most powerful test of its size. *Hint*: Use what you know about Bayes solutions and admissibility.

---

12. Solutions

If you find any mistakes, which surely are scattered around with a far from negligible Poisson rate, please send me an email at emilas@math.uio.no so that I can inform the others, and update this file.

Ex. 1.1. The sample space is $\{HH, HT, TH, TT\}$. It cannot be TT, hence the probability is 1/3. Can use Bayes,

$$\begin{aligned} P(HH \mid \text{at least one H}) &= \frac{P(\text{at least one H} \mid HH)P(HH)}{P(\text{at least one H})} \\ &= \frac{P(HH)}{P(HH) + P(HT) + P(TH) + 0 \times P(HH)} = \frac{1}{3}. \end{aligned}$$

Ex. 1.2. We use Bayes theorem.

$$P(HH \mid H16) = \frac{P(H16 \mid HH)P(HH)}{P(H16)},$$

where $P(HH) = 1/4$ and the probability of getting at least one heads at 16:00 is

$$P(H16 \mid HH) = \frac{1}{6} \times \frac{5}{6} + \frac{5}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} = \frac{11}{36}.$$

Moreover,

$$\begin{aligned} P(H16) &= \frac{1}{4}(P(H16 \mid HH) + P(H16 \mid TH) + P(H16 \mid HT) + 0) \\ &= \frac{1}{4}(\frac{11}{36} + 1/6 + 1/6 + 0) = \frac{1}{4}(\frac{11}{36} + 1/6 + 1/6 + 0) = \frac{1}{4} \times \frac{23}{36}. \end{aligned}$$

Plug this into Bayes formula and obtain $P(HH \mid H16) = 11/23$, which is bigger than 1/3!.

Ex. 1.3. **(a)** Assume that all the coins are fair, that all the teenagers act independently when questioned, and that that the probability of each teenager being a 'yes' is the same, say θ. Moreover, we assume that the teenagers whose first toss come up heads do indeed answer truthfully. For each teenager there is an (unobservable, we suspect) Bernoulli random variable Y with success probability θ. Our anonymisation device creates a new random variable,

$$X_i = W_i Y_i + (1 - W_i)W_i',$$

where W_i and W_i' are independent Bernoulli (1/2) random variables corresponding to the two coin tosses (both independent of Y_i). Then

$$P(\text{answer yes}) = P(\text{answer yes} \mid H)P(H) + P(\text{answer yes} \mid T)P(T) = \frac{\theta}{2} + \frac{1}{4},$$

or

$$\mathrm{E}\, X_i = \mathrm{E}\,(W_i)\mathrm{E}\,(Y_i) + \mathrm{E}\,(1 - W_i)\mathrm{E}\,(W_i') = \frac{\theta}{2} + \frac{1}{4},$$

by independence. The expectation of $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is $\mathrm{E}\,\bar{X}_n = \theta/2 + 1/4$, and an unbiased estimator of θ is

$$\hat{\theta}_n = 2\bar{X}_n - \frac{1}{2}.$$

Based on the number collected by helsesøster, our estimate is $\hat{\theta}_n = 2 \times 4/17 - 1/2 = -0.2647$, which is meaningless. See the next exercise.

(b) With the estimator $\hat{\theta}_n$ it is possible to obtain estimates of θ outside of $(0, 1)$. The probability of this happening gets smaller and smaller when the sample size increases. Set $\gamma = \theta/2 + 1/4 \in (1/4, 3/4)$. Using Chebyshev's inequality,

$$P(|\bar{X}_n - \gamma| \geq \varepsilon) \leq \varepsilon^{-2}\,\mathrm{E}\,(\bar{X}_n - \gamma)^2 = \frac{\gamma(1-\gamma)}{\varepsilon^2 n},$$

which converges to zero as $n \to \infty$. Can also compute the probability of bad estimates directly,

$$P(\hat{\theta}_n \text{ bad}) = 1 - P(\hat{\theta}_n \in (1/4, 3/4)),$$

and show that this probability converges to zero.

(c) In the all-answer-truthfully case the rv Y_i is observable, and a natural estimator for θ is $\bar{Y}_n$, with variance $\mathrm{Var}\,\bar{Y}_n = \theta(1-\theta)/n$. The variance of X_i is

$$\mathrm{Var}\, X_i = (\theta/2 + 1/4)(1 - \theta/2 - 1/4) = \theta(1-\theta)/4 + 3/16.$$

so that

$$\operatorname{Var}\hat{\theta}_n = 4\operatorname{Var}\bar{X}_n = n^{-1}\{\theta(1-\theta)+3/4\}.$$

A natural way to compare unbiased estimators is by the ratio of the variances,

$$\frac{\operatorname{Var}\hat{\theta}_n}{\operatorname{Var}\bar{Y}_n} = 1 + \frac{3}{4\theta(1-\theta)},$$

which attains its minimum in $\theta = 1/2$, where the ratio is 4.

Ex. 2.1. **(a)** Right-continuous, non-decreasing, $F(0) = 0$ and $F(1) = 1$. Importantly, note that there is a jump at τ,

$$P(X = \tau) = F(\tau) - F(\tau-) = \theta.$$

(b) Since $P(X = \tau) = \theta$, generate B_i Bernoulli(θ) independent, and independent of $U_i \sim \text{unif}(0,1)$. Then set

$$X_i = B_i\tau + (1-B_i)U_i, \quad i = 1,\dots,n.$$

(c) Let $X = B\tau + (1-B)U$ as in (b). Then

$$\operatorname{E} X = \theta\tau + (1-\theta)\frac{1}{2},$$

and the variance is

$$\operatorname{Var} X = \{1/4 - \tau(1-\tau)\}\theta(1-\theta) + (1-\theta)/12. \tag{12.1}$$

(d) Show that it is two times the expectation of a Geometric experiment. So

$$\operatorname{E}\{\text{trials until two } X = \tau\} = 2/\theta.$$

(e) Here are two estimators of θ, (provided $\tau \neq 1/2$),

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} I\{X_i = \tau\}, \quad \text{and} \quad \hat{\theta}_2 = (\tau - 1/2)^{-1}n^{-1}\sum_{i=1}^{n}(X_i - 1/2).$$

Compare the variances of these two estimators. $\operatorname{Var}\hat{\theta}_1 = \theta(1-\theta)/n$, and $\operatorname{Var}\hat{\theta}_2 = (\tau-1/2)^{-2}n^{-1}\operatorname{Var} X$, where $\operatorname{Var} X$ is given in (12.1). Then

$$\begin{aligned}\frac{\operatorname{Var}\hat{\theta}_2}{\operatorname{Var}\hat{\theta}_1} &= \frac{1}{(\tau-1/2)^2}\big\{1/4 - \tau(1-\tau) + 1/(12\theta)\big\} \\ &= \frac{1}{(\tau-1/2)^2}\big\{(\tau-1/2)^2 + 1/(12\theta)\big\} = 1 + \frac{1}{12(\tau-1/2)^2\theta} > 1,\end{aligned}$$

which shows that $\hat{\theta}_1$ is a better estimator. Can relate this to sufficiency of $\hat{\theta}_1$ for the θ in $\sum_{i=1}^{n} I\{X_i = \tau\} \sim \text{Binom}(n,\theta)$, perhaps?

Ex. 1.4. $X_1,\dots,X_n$ are independent Bernoulli(θ) trials. **(a)** The maximum likelihood estimator of θ is the empirical mean $\bar{X}_n$. This is an unbiased estimator of θ, with variance $\theta(1-\theta)/n$. Thus, under squared error loss, $R(\bar{X}_n,\theta) = \theta(1-\theta)/n$. See Figure 2 for its risk function. **(b)** The risk of an estimator of the form

$$\delta_w(X) = w\bar{X}_n + (1-w)\theta^\star,$$

where $\theta^\star$ is a prior guess at θ (I changed θ_0 to $\theta^\star$, so that we don't confuse it with the true value of θ), is

$$\begin{aligned}R(\delta_w,\theta) &= \operatorname{E}_\theta\,(w\bar{X}_n + (1-w)\theta^\star - \theta)^2 = w^2\frac{\theta(1-\theta)}{n} + (w\theta + (1-w)\theta^\star - \theta)^2 \\ &= w^2\frac{\theta(1-\theta)}{n} + (1-w)^2(\theta^\star - \theta)^2.\end{aligned}$$

In Fig. 2 I draw the risk function of $\delta_1(X) = \bar{X}_n/2 + 1/4$, for $n = 10$. **(c)** We have a good reasons to believe that the true value of θ is close to $1/2$, so from Fig. 2 we see that $\delta_1(X)$ is a better choice than $\bar{X}_n$. **(d)** The risk function of this estimator is

$$R(w\bar{X}_n + (1-w)/2,\theta) = w^2\frac{\theta(1-\theta)}{n} + (1-w)^2(1/2-\theta)^2.$$

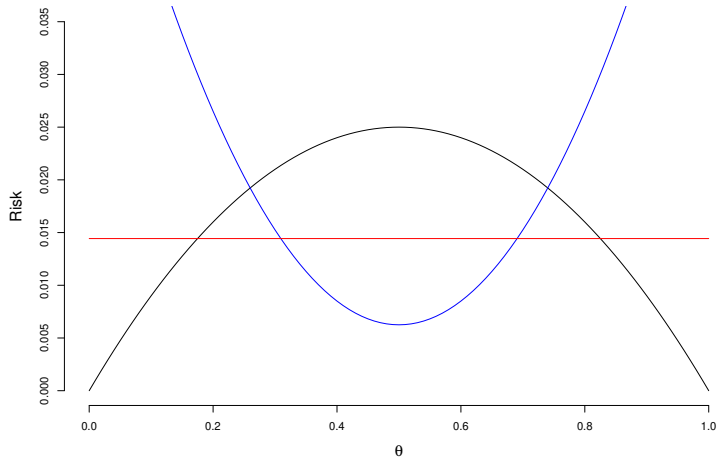

FIGURE 2. Risk function of the three estimators in Exercise 1.4. Black curve is risk of $\bar{X}_n$, blue curve is the risk of $\bar{X}/2+1/4$, and red curve the risk of the minimax estimator $\delta^\star$.

To be constant (as a function of θ), the derivative of this function must be zero for all θ, that is

$$0 = w^2\frac{1-2\theta}{n} - 2(1-w)^2(1/2-\theta) = w^2\frac{1-2\theta}{n} - (1-w)^2(1-2\theta).$$

So

$$w^2/n = (1-w)^2 \quad \Rightarrow \quad w/(1-w) = \sqrt{n} \quad \Rightarrow \quad w = \frac{\sqrt{n}}{\sqrt{n}+1}.$$

The estimator

$$\delta^\star(X) = \frac{\sqrt{n}}{\sqrt{n}+1}\bar{X}_n + (1-\frac{\sqrt{n}}{\sqrt{n}+1})\frac{1}{2},$$

has constant risk

$$\begin{aligned} R(\delta^\star,\theta) &= \frac{n}{(\sqrt{n}+1)^2}\frac{\theta(1-\theta)}{n} + \frac{1}{(\sqrt{n}+1)^2}(1/2-\theta)^2 = \frac{1}{(\sqrt{n}+1)^2}\{\theta(1-\theta)+(1/2-\theta)^2\} \\ &= \frac{1}{(\sqrt{n}+1)^2}\{\theta-\theta^2+1/4-\theta+\theta^2\} = \frac{1}{4(\sqrt{n}+1)^2}. \end{aligned}$$

We soon have the tools to prove that $\delta^\star(X)$ is minimax. That's for another exercise. **(f)** If you have no idea about where the true value of θ is, then use $\delta^\star$.

Ex. 2.3. To show (2.2): Let $A_1 \subset A_2 \subset A_3 \subset \cdots$ be measurable sets. Append the set $A_0 = \emptyset$ and set $B_n = A_n \setminus A_{n-1} = A_n \cap A_{n-1}^c$ for $n = 1, 2, \ldots$. Then $B_1, B_2, \ldots$ is a disjoint sequence, and

$$\begin{aligned} P(\cup_{n=1}^\infty A_n) = P(\cup_{n=1}^\infty B_n) &= \lim_{N\to\infty}\sum_{n=1}^N P(B_n) \\ &= \lim_{N\to\infty}\sum_{n=1}^N P(A_n \setminus A_{n-1}) = \lim_{N\to\infty}\sum_{n=1}^N \{P(A_n) - P(A_{n-1})\} = \lim_{N\to\infty} P(A_N). \end{aligned}$$

The second equality uses countable additivity; the fourth uses that $P(A_n \setminus A_{n-1}) = P(A_n) - P(A_n \cap A_{n-1}) = P(A_n) - P(A_{n-1})$ because $A_{n-1} \subset A_n$; and the last one uses that the sum is telescoping.

To show (2.3): Let $A_1 \supset A_2 \supset \cdots$ be a decreasing sequence of measurable sets. Notice that $A_1 \setminus A_n \subset A_1 \setminus A_{n+1}$ for all n, so by the previous result

$$P(\cup_{n=1}^\infty A_1 \setminus A_n) = \lim_{n\to\infty} P(A_1 \setminus A_n),$$

but $\cup_{n=1}^\infty A_1 \setminus A_n = A_1 \setminus \cap_{n=1}^\infty A_n$, so we have

$$P(A_1) - P(\cap_{n=1}^\infty A_n) = P(A_1) - \lim_{n\to\infty} P(A_n).$$

Notice that we here use that $P(A_1) \leq 1$ since P is a probability measure. For a general measure μ we need that $\mu(A_1) < \infty$ (or that there is a k such that $\mu(A_k) < \infty$).

Ex. 2.5. Show that a distribution function $F(x)$ has at most a countable number of discontinuities. If x is a discontinuity point of F, then $F(x) - F(x-) = F(x) - \lim_{y \uparrow x} F(y) = P(\{x\}) > 0$. But since P is a probability measure we can only have countably many such points.

Ex. 3.1. **(a)** Y is a Bernoulli distributed random variable with success probability

$$P(Y = 1) = P(g(X) = 1) = P(X \in g^{-1}(\{1\})) = P(X \in (c, \infty)) = 1 - \Phi(c).$$

(b) The expectation of $\bar{Y}_n$ is $1 - \Phi(c)$, so a natural estimator of c is

$$\hat{c}_n = \Phi^{-1}(1 - \bar{Y}_n).$$

This estimator is the maximum likelihood estimator.

(c) Recall that if a function f has an inverse f^{-1}, and f is differentiable at x, then

$$\frac{\mathrm{d}}{\mathrm{d}x} f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}.$$

By the central limit theorem $\sqrt{n}(1 - \bar{Y} - 1 + \Phi(c))$ converges to a mean zero normal with variance $\Phi(c)(1 - \Phi(c))$. Since $\mathrm{d}\Phi^{-1}(c)/\mathrm{d}c = 1/\phi(\Phi^{-1}(c))$, an application of the delta method then gives the result. We can be more careful and use the mean value theorem

$$\begin{aligned}\sqrt{n}(\hat{c}_n - c) &= \sqrt{n}(\Phi^{-1}(1 - \bar{Y}_n) - \Phi^{-1}(\Phi(c))) \\ &= \frac{1}{\phi(\Phi^{-1}(\Phi(\tilde{c})))} \sqrt{n}(1 - \bar{Y}_n - \Phi(c)) \xrightarrow{d} \frac{1}{\phi(c)} \mathrm{N}\big(0, \Phi(c)(1 - \Phi(c))\big),\end{aligned}$$

where $\Phi(\tilde{c})$ lies between $1 - \bar{Y}_n$ and $\Phi(c)$. Since $1 - \bar{Y}_n$ is consistent for $\Phi(c)$, it must be the case that $\Phi(\tilde{c})$ converges in probability to $\Phi(c)$, using that Φ is a continuous function.

Here is a script with simulations illustrating the asymptotics

```
nn <- 5*10^3
cc <- 0.567
sims <- 4*10^3
zz <- 0*1:sims
for(jj in 1:nn){
  xx <- rnorm(nn,0,1)
  yy <- 1*(xx >= cc)
  cc_hat <- qnorm(1 - mean(yy))
  zz[jj] <- sqrt(nn)*(cc_hat - cc)
}
hist(zz,freq=FALSE)
avar <- pnorm(cc)*(1 - pnorm(cc))/dnorm(cc)^2
curve(dnorm(x,0,sqrt(avar)),add=TRUE,col="green")
```

Ex. 3.3. Think of $R, \theta)$ has already transformed rv's, i.e., $h^{-1}(r, \theta) = (r\cos\theta, r\sin\theta) = (x, y)$. The determinant of the Jacobian is $\det(J) = r\cos^2\theta + r\sin^2\theta = r$. Take it from there and you find that

$$g(r, \theta) = \frac{1}{2\pi} r \exp(-r^2/2), \quad r > 0, \theta \in (0, 2\pi).$$

So R and θ are independent and θ is uniform on $(0, 2\pi)$.

Ex. 3.2. See Section 13 with proposed solutions to the Nils exercises.

Ex. 4.1. The rv X is Poisson(λ). **(a)** $M_X(t) = \mathrm{E}\, e^{tX} = \sum_{k=0}^{\infty} e^{tk}\lambda^k e^{-\lambda}/x! = e^{-\lambda}\sum_{k=0}^{\infty}(\lambda e^t)^k/x! = e^{-\lambda}e^{\lambda e^t} = \exp(\lambda(e^t-1))$ is the mgf of a Poisson random variable.

(b) If X_n is Binomial (n, p_n), then $M_n(t) = \{1 + p_n(e^t - 1)\}^n$. Assume that $np_n \to \lambda > 0$. Then $\sqrt{n}p_n = np_n/\sqrt{n} \to 0$, so $p_n = o(n^{-1/2})$. For a sequence a_n

$$\log(1+a_n) = a_n - a_n^2/2 + a_n^3/3 - \cdots = a_n + a_n^2\{-1/2 + a_n/3 - \cdots\} = a_n + a_n^2 R(a_n),$$

where $R(a_n) \to -1/2$ when $a_n \to 0$, and for $|a_n| \le 1/2$, $|R(a_n)| \le (1/2) + (1/3)(1/2) + (1/4)(1/2)^2 \cdots = \sum_{k=1}^{\infty} 1/(k+1)2^{-(k-1)} \le 1/2\sum_{k=1}^{\infty} 2^{-(k-1)} = 1$. Given $\varepsilon > 0$, since $p_n \to 0$ we can find n_0 such that $|np_n - \lambda| < \varepsilon$ for all $n \ge n_0$, and an n_1 so that $np_n^2 \le \varepsilon$ for all $n \ge n_1$. For all $n \ge \max(n_0, n_1)$ then

$$\begin{aligned}|\log M_n(t) - \lambda(e^t-1)| &= |n\log(1+p_n(e^t-1)) - \lambda(e^t-1)| \\ &\le |np_n - \lambda|(e^t-1) + |np_n^2(e^t-1)^2 R(p_n)| \le \varepsilon(e^t-1)\{1-(e^t-1)\}.\end{aligned}$$

(e) Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variable with success probabilities $p_{n,1}, \ldots, p_{n,n}$. Set $Z_n = \sum_{i=1}^n X_i$. If $\sum_{i=1}^n p_{n,i} \to \lambda > 0$ and $\max_{i\le n} p_{n,i} \to 0$, then Z_n converges in distribution to a Poisson(λ). By what we did in the previous exercise, and independence of the X_i's, write

$$\log M_n(t) = \sum_{i=1}^n \log(1 + p_{n,i}(e^t-1)) = \sum_{i=1}^n \{p_{n,i} + p_{n,i}^2 R(p_{n,i})\}$$

Is suffices to show that $\sum_{i=1}^n p_{n,i}^2 R(p_{n,i}) \to 0$. But for n big enough

$$|\sum_{i=1}^n p_{n,i}^2 R(p_{n,i})| \le \sum_{i=1}^n p_{n,i}^2 |R(p_{n,i})| \le \max_{i\le n} p_{n,i} \sum_{i=1}^n p_{n,i}|R(p_{n,i})| \le \max_{i\le n} p_{n,i} \sum_{i=1}^n p_{n,i},$$

where we have used that for big enough n all the $p_{n,i} \le 1/2$. The right hand side tends to zero by the assumptions.

Ex. 4.2. **(a)** An exercise in the transformation of random variables,

$$f_Y(y) = \phi(\log y)1/y = \frac{1}{\sqrt{2\pi}}\frac{1}{y}\exp\{-\frac{1}{2}(\log y)\}, \qquad y > 0.$$

This is the pdf of a log-normal distribution. **(b)** Note that $\mathrm{E}\, Y^k = \mathrm{E}\, e^{kX} = e^{k^2/2}$ is the mgf of a standard normal random variable. **(c)** We are to show that $M_Y(t)$ does not exist. See Definition 2.3.4 p. 62 in C&B. The mgf must exist in a nbhd of zero. Since $e^{-x} \le 1$ for all $x > 0$, we have that $\mathrm{E}\, e^{tY} \le 1$ for all $t < 0$. Show that for $t > 0$, $M_Y(t) = \infty$.

$$M_Y(t) = \mathrm{E}\, e^{tY} = \frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{e^{ty}}{y} e^{-\frac{1}{2}(\log y)}\,\mathrm{d}y = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{te^x - x^2/2}\,\mathrm{d}x,$$

changing variables. Using that for $x > 0$, $e^x \ge 1 + x + x^2/2 + x^3/6$, we see that for large enough x, $te^x - x^2/2 \ge t + tx$, so $M_Y(t) \ge \int_K^\infty e^{t+tx}\,\mathrm{d}x$, for some $K > 0$. But the right hand side diverges for all K. The log-normal is therefore an example of a distribution where all the moments exist, but the moment generating function does not.

Ex. 4.3. Fix t and define the sequence Y_K of random variables given by $Y_K = \sum_{k=0}^K t^k/k!(X_n^k - X^k)$. By assumption $|X_n^k - X^k| \le 1$. Thus, $|Y_K| \le \sum_{k=0}^K |t|^k/k!|(X_n^k - X^k)| \le \sum_{k=0}^K |t|^k/k! \le e^{|t|}$, and $\mathrm{E}\, e^{|t|} = e^{|t|} < \infty$. This shows that the sequence Y_K is bounded by an integrable random variable (a constant one), and we can apply the dominated convergence theorem in the fourth equality below. Given $\varepsilon > 0$, there is, by assumption, an n_0 such that $|\mathrm{E}\, X_n^k - \mathrm{E}\, X^k| < \varepsilon\,, k = 1, \ldots, K$ for all $n \ge n_0$. For n

big enough

$$|M_n(t) - M(t)| = \big|\mathrm{E}\sum_{k=0}^{\infty}\frac{t^k}{k!}X_n^k - \mathrm{E}\sum_{k=0}^{\infty}\frac{t^k}{k!}X^k\big| = \big|\mathrm{E}\sum_{k=0}^{\infty}\frac{t^k}{k!}(X_n^k - X^k)\big|$$
$$= \big|\mathrm{E}\lim_{K\to\infty}\sum_{k=0}^{K}\frac{t^k}{k!}(X_n^k - X^k)\big| = \big|\lim_{K\to\infty}\sum_{k=0}^{K}\frac{t^k}{k!}\mathrm{E}\,(X_n^k - X^k)\big|$$
$$\le \lim_{K\to\infty}\sum_{k=0}^{K}\frac{|t|^k}{k!}\big|\mathrm{E}\,[X_n^k] - \mathrm{E}\,[X^k]\big| \le \varepsilon\lim_{K\to\infty}\sum_{k=0}^{K}\frac{|t|^k}{k!} = \varepsilon e^{|t|}.$$

Ex. 5.1. Proof of Lemma 5.1. By the mean value theorem

$$\big|\frac{g(x,\theta+h)-g(x,\theta)}{h}\big| = \big|\int_0^1 \frac{\partial}{\partial\theta}g(x,\theta+ht)\,\mathrm{d}t\big| \le \int_0^1 \big|\frac{\partial}{\partial\theta}g(x,\theta+ht)\big|\,\mathrm{d}t \le k(x).$$

The result then follows from the dominated convergence theorem by taking the limit as $h\to 0$ on both sides of

$$\frac{1}{h}\Big\{\int g(x,\theta+h)\,\mathrm{d}\nu(x) - \int g(x,\theta)\,\mathrm{d}\nu(x)\Big\} = \int \frac{1}{h}\big\{g(x,\theta+h)-g(x,\theta)\big\}\,\mathrm{d}\nu(x).$$

Ex. 5.2. (Note: There were some errors and insufficiencs in the previous statement of this exercise. They are now fixed.) **(a)** A density function $f_\theta(x)\,\theta\in\mathbb{R}^k$, and we assume that $f_\theta(x)$ satisfies Lemma 5.1. The mean of the score function $u_j(\theta;x) = \partial\log f_\theta(x)/\partial\theta_j$ is zero,

$$\mathrm{E}_\theta\, u_j(\theta;X) = \int \frac{\partial\log f_\theta(x)}{\partial\theta_j}\,f_\theta(x)\,\mathrm{d}x = \int\frac{\partial f_\theta(x)}{\partial\theta_j}\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta_j}\int f_\theta(x)\,\mathrm{d}x = 0.$$

The variance of the score equals minus the expectation of its second derivative: First,

$$\frac{\partial^2}{\partial\theta_j\partial\theta_l}\log f_\theta(x) = \frac{\big\{\frac{\partial^2}{\partial\theta_j\partial\theta_l}f_\theta(x)\big\}f_\theta(x) - \frac{\partial}{\partial\theta_j}f_\theta(x)\frac{\partial}{\partial\theta_l}f_\theta(x)}{f_\theta(x)^2} = \frac{\big\{\frac{\partial^2}{\partial\theta_j\partial\theta_l}f_\theta(x)\big\}}{f_\theta(x)} - u_j(\theta;x)u_l(\theta;x).$$

Now, the claim follows by taking expectation on both sides, because $\int\partial f_\theta(x)/\partial\theta_j\partial\theta_l\,\mathrm{d}x = 0$, where $\partial f_\theta/\partial\theta_j$, $j=1,\dots,k$ satisfy the conditions of Lemma 5.1, by assumption. What we have now proved are known as the first and the second Bartlett identity. An important thing to notice is that in in general $\mathrm{E}_\theta u(\theta',X)\neq 0$ when $\theta\neq\theta'$. **(b)** Suppose $f_\theta(x)$ is a exponential family in its natural parametrisation,

$$f_\theta(x) = h(x)c(\theta)\exp\Big\{\sum_{j=1}^{k}\theta_j t_j(x)\Big\}.$$

By the first Bartlett identity

$$0 = \mathrm{E}_\theta\frac{\partial}{\partial\theta_j}\log f_\theta(x) = \frac{\partial}{\partial\theta_j}\log c(\theta) + \mathrm{E}_\theta\, t_j(X),$$

so that $\mathrm{E}_\theta\, t_j(X) = -\partial\log c(\theta)/\partial\theta_j$. For the variance

$$\mathrm{E}_\theta\,\big(\frac{\partial}{\partial\theta_j}\log f_\theta(x)\big)^2 = \mathrm{E}_\theta\,\big(\frac{\partial}{\partial\theta_j}\log c(\theta) + t_j(X)\big)^2 = \mathrm{E}_\theta\,\big(t_j(X) - \mathrm{E}_\theta t_j(X)^2\big)^2 = \mathrm{Var}_\theta\, t_j(X).$$

(c) The rv X has a Gamma(a,b) density. On exponential family form

$$f_{a,b}(x) = \frac{b^a}{\Gamma(a)}\frac{1}{x}\exp\{a\log x - bx\}, x>0.$$

Then

$$\mathrm{E}\,\log(X) = -\frac{\partial}{\partial a}\big(a\log b + \log\Gamma(a)\big) = -\log b - \psi(a),$$

where $\psi(a)$ is the digamma function $\partial\log\Gamma(a)/\partial a$. **(d)** Let X be the number of Bernoulli(p) failures until the first success, then X is geometrically distributed, it has probability mass function

$$P(X=k) = (1-p)^k p = p\exp\{\log(1-p)k\}, \qquad k=0,1,2,\dots.$$

Note that this is not in its natural parametrisation. From the first Bartlett identity, $1/p - \mathrm{E}_p\, X/(1-p) = 0$, so that $\mathrm{E}_p\, X = (1-p)/p$.

Ex. 5.8. A one-parameter exponential family is in our notation a family of densities on the form $f_\theta(x) = h(x)c(\theta)\exp\{\theta t(x)\}$, with $\theta \in \mathbb{R}$. **(a)** If $h(x) = 1/x$ $t(x) = \log x$ on $(0, 1]$, then

$$\int_0^1 \frac{1}{x}\exp(\theta \log x)\,\mathrm{d}x = \int_0^1 x^{\theta-1}\,\mathrm{d}x < \infty,$$

when $\theta > 0$. The natural parameter space is therefore $(0, \infty)$. For $\theta > 0$, we must have

$$1/c(\theta) = \int_0^1 x^{\theta-1}\,\mathrm{d}x = 1/\theta,$$

so $c(\theta) = \theta$, and its density is $f_\theta(x) = \theta x^{\theta-1}$ on $(0, 1]$. **(b)** If $X \sim f_\theta(x) = \theta x^{\theta-1}$, then $Y = -\log X$ has a density on $(0, \infty)$ given by

$$g_\theta(Y) = f_\theta(e^{-y})e^{-y} = \theta e^{-(\theta-1)y}e^{y} = \theta e^{-\theta}, \qquad y > 0,$$

which is an exponential distribution. **(c)** With $X_1, \ldots, X_n$ i.i.d. from $f_\theta(x)$, the log-likelihood function is

$$\ell_n(\theta) = n\log\theta + (\theta - 1)\sum_{i=1}^n \log x_i$$

The estimating equation is $n/\theta + \sum_{i=1}^n \log x_i = 0$, note also that $\partial^2\ell_n(\theta)/\partial\theta^2 = -n/\theta^2 < 0$ for all $\theta > 0$, so $\ell_n(\theta)$ is everywhere concave. The maximum likelihood estimator is

$$\hat\theta_n = 1/\{n^{-1}\sum_{i=1}^n(-\log X_i)\}.$$

We must show that $\mathrm{E}_\theta\hat\theta_n = \theta + \varepsilon_n$, where ε_n is something small (that might depend on θ). From (b) we know that $-\log X_1, \ldots, -\log X_n$ are independent exponentials with mean $1/\theta$. The mgf of an exponential(θ) rv Y is

$$M_Y(t) = \mathrm{E}_\theta\, e^{tY} = \int_0^\infty e^{ty}\theta e^{-\theta y}\,\mathrm{d}y = \int_0^\infty \theta e^{-(\theta-t)y}\,\mathrm{d}y = \theta/(\theta - t) = (1 - t/\theta)^{-1}, \quad \text{for } t < \theta.$$

From this we see that the mgf of a sum of independent exponentials is

$$M_{\sum_{i=1}^n Y_i}(t) = (1 - t/\theta)^{-n}, \quad \text{for } t < \theta,$$

which we recognise as the mgf of a Gamma(n, θ) random variable. Let $Z_n \sim$ Gamma(n, θ), and the maximum likelihood estimator is $\hat\theta_n =_d n/Z_n$. In general, the expectation of an inverse Gamma(a, b) is

$$\mathrm{E}\, X^{-1} = \int_0^\infty \frac{b^a}{\Gamma(a)}x^{(a-1)-1}e^{-bx}\,\mathrm{d}x = \frac{b^a}{\Gamma(a)}\frac{\Gamma(a-1)}{b^{a-1}} = \frac{b}{a-1},$$

provided $a > 1$. For us a is the sample size n, so this is not a problem. We then see that

$$\mathrm{E}_\theta\,\hat\theta_n = \mathrm{E}_\theta\,\frac{n}{Z_n} = \frac{n}{n-1}\theta = \theta + \frac{\theta}{n-1},$$

which means that $\hat\theta_n$ is approximately unbiased for large enough sample size n. **(d)** From (b) we know that $Y_i = -\log X_i$ are i.i.d. random variables with mean $1/\theta$ and variance $1/\theta^2$. The central limit theorem then tells us that

$$\sqrt{n}(\bar Y_n - 1/\theta) \xrightarrow{d} \mathrm{N}(0, 1/\theta^2).$$

Use the delta-method with $g(x) = 1/x$, whose first derivative is $g'(x) = -1/x^2$. Then

$$\sqrt{n}(\hat\theta_n - \theta) = \sqrt{n}(g(\bar Y_n) - g(1/\theta)) \xrightarrow{d} g'(1/\theta)\mathrm{N}(0, 1/\theta^2) =_d \mathrm{N}(0, \theta^2).$$

Ex. 7.2. First the hint. We can write $\max\{x,y\} = 1/2(x+y+|x-y|)$, which can be shown using that $|x| = \max\{x,-x\}$. Then

$$\mathrm{E}\,\max\{X^2,Y^2\} = \frac{1}{2}\mathrm{E}\,(X^2+Y^2+|X^2-Y^2|) = 1+\frac{1}{2}\mathrm{E}\,|X^2-Y^2|.$$

So we must show that $\mathrm{E}\,|X^2-Y^2| \le 2\sqrt{1-\rho^2}$. Using Hölder's inequality (which in this case is the Cauchy–Schwarz inequality),

$$\begin{aligned}(\mathrm{E}\,|X^2-Y^2|)^2 &= (\mathrm{E}\,|(X-Y)(X+Y)|)^2 \le \mathrm{E}\,(X-Y)^2\mathrm{E}\,(X+Y)^2\\ &= \mathrm{E},(X^2+Y^2-2XY)\mathrm{E},(X^2+Y^2+2XY)\\ &= 4(1-\mathrm{Cov}(X,Y))(1+\mathrm{Cov}(X,Y)) = 4(1-\mathrm{Cov}(X,Y)^2) = 4(1-\rho^2).\end{aligned}$$

[xx latex the remainder of the exercise xx]

13. Proposed solutions to Nils' exercises

The exercises referred to in this section can be found in Nils' Lecture Notes and Exercises, version as of 12/x/14, from when Nils gave the course in 2014. `https://www.uio.no/studier/emner/matnat/math/STK4011/h14/exercises_stk4011a.pdf`

13.1. **Transformation of random variables. (a)** $X \sim f(x)$, and $Y = h(X)$, where h is smooth and monotone. The density of Y is then

$$g(y) = f(h^{-1}(y))\big|\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y)\big|.$$

To see this,

$$G(y) = P(Y\le y) = P(h(X)\le y) = \begin{cases} P(X\le h^{-1}(y)) = F(h^{-1}(y)), & \text{if } h \text{ is increasing,}\\ 1-P(X\le h^{-1}(y)) = 1-F(h^{-1}(y)), & \text{if } h \text{ is decreasing.}\end{cases}$$

If h is increasing, then the first derivative of $h^{-1}(y)$ is positive, if it is decreasing, then the first derivative of $h^{-1}(y)$ is negative. We get

$$g(y) = \frac{\mathrm{d}}{\mathrm{d}y}G(y) = \begin{cases} f(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y), & h \text{ increasing,}\\ -f(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y), & h \text{ decreasing,}\end{cases} = f(h^{-1}(y))\big|\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y)\big|.$$

(b) Let $X\sim \mathrm{N}(0,1)$, the density of $Y = \exp(X)$ is easily seen to be

$$g(y) = \frac{\sqrt{2\pi}y}{\exp}(-(\log y)^2/2), y>0.$$

If $X\sim \mathrm{N}(\xi,\sigma^2)$, then

$$g_{\xi,\sigma}(y) = \frac{1}{\sigma}\phi\big(\frac{\log y-\xi}{\sigma}\big)\frac{1}{y} = \frac{1}{\sqrt{2\pi}\sigma y}\exp\big(-\frac{1}{2\sigma^2}(\log y-\xi)^2\big).$$

The easiest way to find its mean, variance, and skewness is by using stuff we already know about the normal moment generating function. Let Z be a standard normal rv.

$$\mathrm{E}\,Y = \mathrm{E}\,e^X = \mathrm{E}\,e^{\sigma Z+\mu} = e^\mu M_Z(\sigma) = e^{\mu+\frac{1}{2}\sigma^2}.$$

$$\mathrm{Var}(Y) = \mathrm{E}\,e^{2X} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}(e^{\sigma^2}-1).$$

What we are after here is the skewness, which is a measure of asymmetry of a random variable about its mean. Negative skew: The left tail is longer, so the probability mass is concentrated to the right; Positive skew: The right tail is longer, so the mass of the distribution is concentrated on the left. Zero skew means symmetry about the mean. The normal disitrbution has zero skew. With $\xi = \mathrm{E}\,X$ and $\sigma^2 = \mathrm{Var}(X)$, the skew is the third standardised moment,

$$\mathrm{skew}(X) = \mathrm{E}\,\big(\frac{X-\xi}{\sigma}\big)^3 = \frac{1}{\sigma^3}\big(\mathrm{E}\,X^3 - 3\xi\sigma^2 - \xi^3\big).$$

Compute the third moment of the log-normal,

$$\mathrm{E}\, Y^3 = \mathrm{E}\, e^{3\xi+3\sigma Z} = \exp(3\xi + \frac{9}{2}\sigma^2).$$

Plug the expectation, variance, and the third moment into the formula, and obtain

$$\mathrm{skew}(Y) = (e^{\sigma^2}+2)(e^{\sigma^2}-1)^{1/2},$$

after some tedious algebra.

(c) U is uniform$(0,1)$. The density of $V = U/(1-U)$. Introduce $h(u) = u/(1-u)$, which is monotone increasing, and has inverse

$$h^{-1}(v) = \frac{v}{1+v}.$$

The density of V is

$$g(v) = \frac{1}{(1+v)^2}, \quad v \geq 0,$$

with cdf $G(v) = v/(1+v)$. The median of V is the value m satisfying $G(m) = 1/2$, which is clearly $m = 1$. The random variable V does not, however, have mean. To see this, change of variable $x = 1+v$,

$$\begin{aligned}\int_0^K vg(v)\,\mathrm{d}v &= \int_0^K v\frac{1}{(1+v)^2}\,\mathrm{d}v \\ &= \int_1^{K+1} \big(\frac{1}{x} - \frac{1}{x^2}\big)\,\mathrm{d}x \geq \int_1^{K+1} \frac{1}{x}\,\mathrm{d}x = \log(K+1),\end{aligned}$$

which diverges when $K \to \infty$.

(d) Let U is uniform$(0,1)$, and set $W = -\log U$. Using the probability integral transform we see that $W \sim \mathrm{expo}(1)$.

(e) Suppose X is Weibull with cdf

$$F(x) = 1 - \exp(-(x/a)^b), \quad \text{for } x \geq 0.$$

The distribution of $V = (X/a)^b$ is then that of a unit exponential. We can simulate Weibull random variables X by drawing unit exponentials, then setting $X = aV^{1/b}$.

13.2. **Transformations of random vectors. (a)**: Prove the formula. **(b)**: X, Y independent unit exponentials. Set $U = X/(X+Y)$ and $V = X+Y$. I like to write

$$h(x,y) = (x/(x+y), x+y) = (u,v), \qquad h^{-1}(u,v) = (uv, v(1-u)).$$

The Jacobian is

$$J(y) = \begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix}, \qquad \det J(y) = v(1-u) + vu = v.$$

The joint density of (U,V) is then

$$g(u,v) = ve^{-uv}e^{-v(1-u)} = 1 \times ve^{-v},$$

which we recognise as the product of the uniform$(0,1)$ and a Gamma$(2,1)$ density. Hence independent. Note for the next exercise, a Beta(1,1) rv is uniform$(0,1)$.

(c): Let $X \sim \mathrm{Gamma}(a,1)$ and $Y \sim \mathrm{Gamma}(b,1)$ independent. We have already done more than half the job in (b). Set $U = X/(X+Y)$ and $V = X+Y$, with inverse and Jacobian as above. The joint density is

$$\begin{aligned} g(u,v) &= \frac{1}{\Gamma(a)}(uv)^{a-1}e^{-uv}\frac{1}{\Gamma(b)}[v(1-u)]^{b-1}e^{-v(1-u)}v \\ &= \frac{1}{\Gamma(a)\Gamma(b)}u^{a-1}(1-u)^{b-1}v^{a+b-1}e^{-v}, \quad u \in (0,1),\ v \geq 0.\end{aligned}$$

Integrate out v and get the Beta(a,b) density,

$$\frac{\Gamma(a+v)}{\Gamma(a)\Gamma(b)}u^{a-1}(1-u)^{b-1}, \quad 0 \leq u \leq 1.$$

(d): Let $X_1, \ldots, X_n$ be independent Gamma$(a_i, 1)$, $i = 1, \ldots, n$. We are to find the distribution of the vector

$$(Y_1, \ldots, Y_n) = (\frac{X_1}{S}, \ldots, \frac{X_n}{S}),$$

where $S = X_1 + \cdots + X_n$. Consider the transformation

$$h(x_1, \ldots, x_{n-1}, x_n) = (\frac{x_1}{s}, \ldots, \frac{x_{n-1}}{s}, s) = (y_1, \ldots, y_{n-1}, v),$$

with inverse

$$h^{-1}(y_1, \ldots, y_{n-1}, v) = (y_1 v, \ldots, y_{n-1} v, v(1 - y_1 - \cdots - y_{n-1})).$$

The Jacobian of this inverse function is (here shown for the case where $n = 4$)

$$J(y, v) = \begin{pmatrix} v & 0 & 0 & 0 \\ 0 & v & 0 & 0 \\ 0 & 0 & v & 0 \\ -v & -v & -v & 1 \end{pmatrix}, \qquad \text{with} \qquad \det J(y) = v^{n-1}.$$

Plug this into the transformation formula,

$$\begin{aligned} g(y_1, \ldots, y_{n-1}, v) &= \{ \prod_{i=1}^{n-1} \frac{1}{\Gamma(a_i)} (y_i v)^{a_i - 1} e^{-y_i v} \} \frac{1}{\Gamma(a_n)} (v(1 - y_1 - \cdots - y_{n-1}))^{a_n - 1} e^{-v(1 - y_1 - \cdots - y_{n-1})} \, v^{n-1} \\ &= \frac{1}{\Gamma(a_1) \cdots \Gamma(a_n)} \{ \prod_{i=1}^{n-1} y_i^{a_i - 1} \} (1 - y_1 - \cdots y_{n-1})^{a_n - 1} v^{a_1 + \cdots a_n - 1} e^{-v}. \end{aligned}$$

Integrate out v to get the result, that is, use that

$$\int_0^\infty v^{a_1 + \cdots a_n - 1} e^{-v} \, \mathrm{d}v = \Gamma(a_1 + \cdots a_n).$$

Find the expectation, variance, and covariance of Dirichlet rv's. Several ways to go about this. The following is straightforward.

$$\mathrm{E}\, Y_j = \int \frac{\Gamma(a)}{\prod_{i=1}^n \Gamma(a_i)} \{ \prod_{i \neq j} y_i^{a_i - 1} \} y_j^{(a_j + 1) - 1} \, \mathrm{d}y = \frac{\Gamma(a)}{\prod_{i=1}^n \Gamma(a_i)} \frac{\prod_{i \neq j} \Gamma(a_i) \Gamma(a_j - 1)}{\Gamma(a + 1)} = \frac{a_j}{a}.$$

Proceed in the same manner, and do some algebra to find the variance and covariance. Can also show and use that the marginals are

$$Y_j \sim \mathrm{Beta}(a_j, a - a_j).$$

13.3. **A pair of normals.** See Exercise 3.3.

13.4. **Ordering exponentials.** X_1, X_2, X_3 are independent with distribution funtion $F(x) = 1 - \exp(-x)$, $x > 0$. Order statistics $X_{(1)}, X_{(2)}, X_{(3)}$. Form

$$Y_1 = X_{(1)}, \quad Y_2 = X_{(2)} - X_{(1)}, \quad Y_3 = X_{(3)} - X_{(2)}.$$

The joint pdf of n order statistics is (see Casella and Berger (2002, p.230)),

$$f_{x_{(1)}, \ldots, x_{(n)}}(x_1, \ldots, x_n) = n! \, f(x_1) \cdots f(x_n),$$

for $-\infty < x_1 < \ldots x_n < \infty$, and zero otherwise. Transformation

$$h(x_1, x_2, x_3) = (x_1, x_2 - x_1, x_3 - x_2) = (y_1, y_2, y_3),$$

with inverse

$$h^{-1}(y_1, y_2, y_3) = (y_1, y_2 + y_1, y_3 + y_2 + y_1),$$

whose Jacobian is

$$J = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

The determinant of a lower (and of an upper) diagonal matrix is the product of its diagonal elements. So, $\det(J) = 1$. Let $f(x) = e^{-x}$, the probability density function of an unit exponential. The joint density of Y_1, Y_2, Y_3 is then

$$\begin{aligned} g(y_1, y_2, y_3) &= 3!\, f(y_1) f(y_2 + y_1) f(y_3 + y_2 + y_1) \det(J) \\ &= 3!\, e^{-y_1} e^{-(y_2+y_1)} e^{-(y_3+y_2+y_1)} = 3e^{-3y_1} 2e^{-y_2} e^{-y_3}, \end{aligned}$$

which is a product of exponentials with means $1/3$, $1/2$ and 1, respectively. This also shows that Y_1, Y_2, Y_3 are independent. This generalises to $X_1, \ldots, X_n$ independent unit exponentials, with order statistics $X_{(1)}, \ldots, X_{(n)}$, with $Y_1 = X_{(1)}, Y_2 = X_{(2)} - X_{(1)}, \ldots, Y_{n-1} = X_{(n-1)} - X_{(n-2)}, Y_{n-1} = X_{(n)} - X_{(n-1)}$, whose joint density is seen to be

$$g(y_1, \ldots, y_n) = ne^{-ny_1}(n-1)e^{-(n-1)y_2} \cdots 2e^{-2y_{n-1}} e^{-y_n},$$

that is the product of $\mathrm{expo}(n)\mathrm{expo}(n-1)\cdots\mathrm{expo}(2)\mathrm{expo}(1)$. The random variables

$$V_1 = nY_1, \quad V_2 = (n-1)Y_2, \quad \ldots \quad V_{n-1} = 2Y_{n-1}, \quad V_n = Y_n,$$

are then seen to be scaled so that they are i.i.d. unit exponentials. If $X \sim \mathrm{expo}(\theta)$, then $\theta X \sim \mathrm{expo}(1)$. Now, set $M_n = \max_{i \le n} X_i = X_{(n)}$, and note that

$$M_n = X_{(n)} = V_n + \frac{V_{n-1}}{2} + \cdots + \frac{V_1}{n}.$$

Using this representation, the expectation of M_n is seen to be the harmonic series

$$\mathrm{E}\, M_n = \sum_{i=1}^{n} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots \approx \log n + \gamma,$$

where $\gamma = 0.5772...$ is the Euler constant. By independence and the fact that $\mathrm{Var}\, V_i = 1$,

$$\mathrm{Var}\, M_n = \sum_{i=1}^{n} \frac{1}{i^2}$$

which is close to $\pi^2/6$ for large n, for $\lim_{n\to\infty} \sum_{i=1}^{n} 1/i^2 = \pi^2/6$. Find the limit distribution of $W_n = M_n - \log n$. Since the X_i's are independent,

$$\begin{aligned} P(W_n \le x) = P(X_{(n)} \le \log n + x) &= P(\text{all } X_i \le \log n + x) = \prod_{i=1}^{n} P(X_i \le \log n + x) \\ &= F(\log n + X)^n = (1 - e^{-\log n - x})^n = (1 - n^{-1}e^{-x})^n \to \exp(-e^{-x}), \end{aligned}$$

as $n \to \infty$. Here $G(x) = \exp(-e^{-x})$ is the distribution function of the standard Gumbel distribution, often used to model phenomena in extreme value statistics.

13.5. **Ratios of ordered uniforms.** [xx latex it xx]

13.6. **The multinormal distribution.** $X = (X_1, \ldots, X_k)^{\mathrm{t}}$ is multinormal with mean vector ξ and variance matrix Σ (a positive definite $k \times k$ matrix), if its density is

$$f(x) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\big\{ -\frac{1}{2}(x-\xi)^{\mathrm{t}} \Sigma^{-1} (x-\xi) \big\}, \quad x \in \mathbb{R}^k. \tag{13.1}$$

where $|\Sigma| = \det(\Sigma)$. Write $X \sim \mathrm{N}_k(\xi, \Sigma)$. **(a)** If $X \sim \mathrm{N}_k(\xi, \Sigma)$, and A is $k \times k$ of full rank (its rows and columns are independent and it is invertible), and b is a $k \times 1$ vector, then

$$Y = AX + b \sim \mathrm{N}_k(A\xi, A\Sigma A^{\mathrm{t}}).$$

Show this using what we know about transformation of random variables: $h(X) = AX + b = Y$, so that $h^{-1}(y) = A^{-1}(y - b)$, where h^{-1} exists because A is of full rank. The Jacobian of $h^{-1}(y)$ is $J(y) = A^{-1}$, and recall that $\det(A^{-1}) = \det(A)^{-1}$. Then

$$g(y) = f(A^{-1}(y-b))|A|^{-1},$$

is the density of a $\mathrm{N}_k(A\xi, A\Sigma A^{\mathrm{t}})$ because the exponent in $f(A^{-1}(y-b))$ is

$$\begin{aligned}(A^{-1}(y-b)-\xi)^{\mathrm{t}}\Sigma^{-1}(A^{-1}(y-b)-\xi) &= (A^{-1}((y-b)-A\xi))^{\mathrm{t}}\Sigma^{-1}A^{-1}((y-b)-A\xi)\\ &= (y-(A\xi+b))(A\Sigma A^{\mathrm{t}})^{-1}(y-(A\xi+b)).\end{aligned}$$

(b) We are to show that if $X \sim \mathrm{N}_k(\xi,\Sigma)$, then indeed $\mathrm{E}\,X=\xi$ and $\mathrm{Var}\,X=\Sigma$. To avoid integration w.r.t. (13.1), we do as follows: Introduce the matrix $\Sigma^{1/2}$ that is such that $\Sigma^{1/2}\Sigma^{1/2}=\Sigma$, and let $Z\sim \mathrm{N}_k(0,I_k)$, with I_k the k-dim. identity matrix. The Z has density

$$f(z) = (2\pi)^{-k/2}\exp\{-\frac{1}{2}z^{\mathrm{t}}z\} = \prod_{j=1}^{k}\phi(z_j),$$

where $\phi(z)=(2\pi)^{-1/2}e^{-z^2/2}$ is the standards normal density. The density above is easy to integrate, and clearly, $\mathrm{E}\,Z=0$ and $\mathrm{Var}\,Z=I_k$. Define $X=\Sigma^{1/2}Z+\xi$, and it follows from (a) that $\mathrm{E}\,X=\xi$ and $\mathrm{Var}\,X=\Sigma$.

(c) From linear algebra, we know that for a given positive definite symmetric $k\times k$ matrix Σ, there is an orthonormal matrix P (i.e., $PP^{\mathrm{t}}=I_kP^{\mathrm{t}}P$), such that $P\Sigma P^{\mathrm{t}}=D$, where D is a diagonal matrix with the eigenvalues $\lambda_1,\ldots,\lambda_k$ along its diagonal. If $X\sim\mathrm{N}_k(0,\Sigma)$, then $Y=PX$ has independent components: From (a), $Y=PX\sim\mathrm{N}_k(0,P\Sigma P^{\mathrm{t}})=\mathrm{N}_k(0,D)$, and D is diagonal, that is $\mathrm{Cov}(Y_i,Y_j)=0$ for all $i\neq j$, while $\mathrm{Var}(Y_i)=\lambda_i$. Here we use that $\mathrm{Cov}(Y_i,Y_j)=0$ implies that Y_i and Y_j are independent when Y_i and Y_j are normal. This implication does not hold in general (but the other way is always true, i.e., independence $\Rightarrow$ Cov $=0$). From (a) we also see that the components of $Y=PX$ are independent when X has a non-zero mean.

(d) Show that X is multinormal iff all linear combinations $a^{\mathrm{t}}X=a_1X_1+\cdots a_kX_k$ are normal. In particular, $a^{\mathrm{t}}X\sim\mathrm{N}(a^{\mathrm{t}}\xi,a^{\mathrm{t}}\Sigma a)$. Note that in some textbooks this is taken as the definition of a multinormal random vector. ($\Rightarrow$): Suppose $X\sim N_k(\xi,\Sigma)$. Set $Y=PX$, where $P^{\mathrm{t}}\Sigma P=D$. Then $Y\sim\mathrm{N}_k(P\xi,D)$, with independent components. For any $k\times 1$ vector a, define $b=b_a=Pa$, and consider the mgf

$$\begin{aligned}M_{a^{\mathrm{t}}X}(t) &= \mathrm{E}\,e^{ta^{\mathrm{t}}X} = \mathrm{E}\,e^{ta^{\mathrm{t}}P^{\mathrm{t}}Y} = \mathrm{E}\,e^{tb^{\mathrm{t}}Y} = \mathrm{E}\,e^{t\sum_{j=1}^k b_jY_j}\\ &= \prod_{j=1}\mathrm{E}\,e^{tb_jY_j} = \prod_{j=1}e^{tb_j(P\xi)_j+\frac{1}{2}t^2b_j^2\lambda_j}\\ &= e^{t\sum_{j=1}^k b_j(P\xi)_j+\frac{1}{2}t^2\sum_{j=1}^k b_j^2\lambda_j} = e^{tb^{\mathrm{t}}P\xi+\frac{1}{2}t^2b^{\mathrm{t}}Db} = e^{ta^{\mathrm{t}}P^{\mathrm{t}}P\xi+\frac{1}{2}t^2a^{\mathrm{t}}P^{\mathrm{t}}DPa}\\ &= e^{ta^{\mathrm{t}}\xi+\frac{1}{2}t^2a^{\mathrm{t}}\Sigma a}\end{aligned}$$

($\Leftarrow$): Suppose X is a random vector such that $a^{\mathrm{t}}X$ is normal $(a^{\mathrm{t}}\xi,a^{\mathrm{t}}\Sigma a)$ for all $k\times 1$ vectors a. The mgf of $X\sim\mathrm{N}_k(\xi,\Sigma)$ is

$$M_X(u) = \mathrm{E}\,e^{u^{\mathrm{t}}X} = e^{u^{\mathrm{t}}\xi+\frac{1}{2}u^{\mathrm{t}}\Sigma u}. \tag{13.2}$$

We have seen that $M_{a^{\mathrm{t}}X}(u)=\exp\{ua^{\mathrm{t}}\xi+\frac{1}{2}u^2a^{\mathrm{t}}\Sigma a\}$. So if $a^{\mathrm{t}}X$ is normal for all vectors a, then

$$M_X(u) = \mathrm{E}\,e^{u^{\mathrm{t}}X} = \mathrm{E}\,e^{u^{\mathrm{t}}X} = M_{u^{\mathrm{t}}X}(1) = e^{u^{\mathrm{t}}\xi+\frac{1}{2}u^{\mathrm{t}}\Sigma u},$$

which show that X is multinormal(ξ,Σ).

(e) A is $m\times k$, and $X\sim\mathrm{N}_k(\xi,\Sigma)$. Use (13.2) on $Y=AX$; for any $u=(u_1,\ldots,u_m)^{\mathrm{t}}$,

$$\begin{aligned}M_Y(u) &= \mathrm{E}\,e^{u^{\mathrm{t}}Y} = \mathrm{E}\,e^{u^{\mathrm{t}}AX} = \mathrm{E}\,e^{(A^{\mathrm{t}}u)^{\mathrm{t}}X} = e^{(A^{\mathrm{t}}u)^{\mathrm{t}}\xi+\frac{1}{2}(A^{\mathrm{t}}u)^{\mathrm{t}}\Sigma(A^{\mathrm{t}}u)}\\ &= e^{u^{\mathrm{t}}A\xi+\frac{1}{2}u^{\mathrm{t}}(A\Sigma A^{\mathrm{t}})u},\end{aligned}$$

and we conclude that $Y=AX\sim\mathrm{N}_m(A\xi,A\Sigma A^{\mathrm{t}})$.

13.7. **Multinormal conditional distributions.** [xx latex it xx]

13.8. **Distribution associated with a normal sample.** Suppose $X_1, \ldots, X_n$ are i.i.d. standard normal, and let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, and $Z = \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Write $X = (X_1, \ldots, X_n)^{\mathrm{t}}$.

(a) P is an orthonormal $n \times n$ matrix. By Nils Exercise 6(a) ("The multinormal distribution"), we have that

$$Y = PX \sim \mathrm{N}_n(0, PI_nP^{\mathrm{t}}),$$

but $PI_nP^{\mathrm{t}} = I_k$, so $Y_1, \ldots, Y_n$ are also independent standard normals. $Y = PX$. Then $\|Y\|^2 = \mathrm{E}\, Y^{\mathrm{t}}Y = \mathrm{E}\,(PX)^{\mathrm{t}}PX = \mathrm{E}\, X^{\mathrm{t}}P^{\mathrm{t}}PX = \mathrm{E}\, X^{\mathrm{t}}X = \|X\|^2$.

(b) The $n \times n$ matrix P given by

(13.3)

$$P = \begin{pmatrix} 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n} \\ 1/\sqrt{2\times 1} & -1/\sqrt{2\times 1} & 0 & 0 & \cdots & 0 \\ 1/\sqrt{3\times 2} & 1/\sqrt{3\times 2} & -2/\sqrt{3\times 2} & 0 & \cdots & 0 \\ 1/\sqrt{4\times 3} & 1/\sqrt{4\times 3} & 1/\sqrt{4\times 3} & -3/\sqrt{4\times 3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & \cdots & -(n-1)/\sqrt{n(n-1)} \end{pmatrix},$$

is orthonormal. For when we may need it, here is an R-script constructing this matrix.

```
nn <- 10
PP <- matrix(0,nn,nn)
PP[1,] <- 1/sqrt(nn)*rep(1,nn)
for(jj in 2:nn){
PP[jj,1:(jj-1)] <- 1/(sqrt(jj*(jj-1)))
PP[jj,jj] <- -(jj - 1)/(sqrt(jj*(jj-1)))
}
# Check
t(PP)%*%PP ; PP%*%t(PP)
```

The components of $X = (X_1, \ldots, X_n)^{\mathrm{t}}$ still $\mathrm{N}(0,1)$. Set $Y = PX$. From the matrix above

$$Y_1 = (PX)_1 = X_1/\sqrt{n} + \cdots + X_n/\sqrt{n} = \sqrt{n}\bar{X}_n.$$

Moreover,

$$\begin{aligned} \sum_{i=2}^n Y_i^2 &= \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=1}^n Y_i^2 - n\bar{X}_n^2 = Y^{\mathrm{t}}Y - n\bar{X}_n^2 \\ &= (PX)^{\mathrm{t}}(PX) - n\bar{X}_n^2 = X^{\mathrm{t}}P^{\mathrm{t}}PX - n\bar{X}_n^2 = X^{\mathrm{t}}X - n\bar{X}_n^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 = Z. \end{aligned}$$

The $Y_1, \ldots, Y_n$ are independent $\mathrm{N}(0,1)$ by (a). Since $Y_1 = \sqrt{n}\bar{X}_n$, and $Y_2, \ldots, Y_n$ are independent of Y_1, the mean $\bar{X}_n$ must be independent of Z. Z is a sum of $n-1$ squared independent standard normals, hence $Z \sim \chi^2_{n-1}$.

(c) Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$. Show that

$$\hat{\mu}_n = \bar{X}_n, \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

are independent. To show this we could use Basu's theorem (Casella and Berger, 2002, p. 287), as in Exercise 6.14. From the previous exercise we know that when $Z_1, \ldots, Z_n$ are independent standard normals, $\bar{Z}_n$ and $\sum_{i=1}^n (Z_i - \bar{Z}_n)^2$ are independent. Since $X_i = \sigma Z_i + \mu$ for $i = 1, \ldots, n$,

$$\hat{\mu}_n = \bar{X}_n = \sigma\bar{Z}_n + \mu,$$

and

$$\hat{\sigma}_n^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^n (Z_i - \bar{Z}_n)^2,$$

from which we see that $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are independent. The mean $\bar{X}_n$ is a linear combination of normals, so by Nils exercise 6(d) it is normal. By linearity $\mathrm{E}\,\bar{X}_n = \mu$, and by independence $\mathrm{Var}\,\bar{X}_n = \sigma^2/n$.

Let P be the big matrix in (13.3), and set $Y = PZ$. Then

$$\hat{\sigma}_n^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 = \frac{\sigma^2}{n-1}\sum_{i=2}^{n} Y_i^2,$$

where the $Y_2, \ldots, Y_n$ are independent standard normals, and the previous exercise gives that $(n-1)\hat{\sigma}_n^2/\sigma^2$ is chi-square with $n-1$ degrees of freedom.

(d) Show that $t = \sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}$ is t-distributed with $n-1$ degrees of freedom.

$$t = \Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\Big)/\Big(\frac{(n-1)\hat{\sigma}_n^2}{\sigma^2(n-1)}\Big)^{1/2} =_d \frac{\mathrm{N}(0,1)}{(\chi^2_{n-1}/(n-1))^{1/2}},$$

which is t_{n-1} distributed.

13.9. **Convergence in probability.** **(a)** Suppose that $V_n \xrightarrow{p} a$, for a constant a, and let h be a function that is continuous in a. Then $h(V_n) \xrightarrow{p} h(a)$. Given $\varepsilon > 0$, we can find $\delta > 0$, such that

$$|v - a| < \delta \quad \Rightarrow \quad |h(v) - h(a)| < \varepsilon.$$

It follows that

$$P(|V_n - a| < \delta) \le P(|h(V_n) - h(a)| < \varepsilon) \le 1,$$

and since the left hand side tends to 1 as $n \to \infty$, the probability on the right must also tend to 1.

(b) Prove that $V_n \xrightarrow{p} V \Rightarrow h(V) \xrightarrow{p} h(V)$, when h is continuous on the domain of V, i.e., on the set C that is such that $P(V \in C) = 1$. If h is *uniformly* continuous on the domain of V, the we could just use the proof from (a). The hint in Exercise 5.39 Casella and Berger (2002, p. 262) might make one think that the proof in (a) works for both cases, that is, both for a limiting constant *and* for a limiting random variable. That's not correct, unless h is uniformly continuous.

Here is a proof for a function h that is continuous on C, where $P(V \in C) = 1$. Given $\varepsilon > 0$, for each $k > 0$,

$$\{|h(V_n) - h(V)| \ge \varepsilon\} = \{|h(V_n) - h(V)| \ge \varepsilon\} \cap \{|V| \le k\} \cup \{|V| > k\}.$$

Since h is continuous, it is uniformly continuous on $[-k, k]$. So, given $\varepsilon > 0$, there is a $\delta > 0$ such that $|x - y| < \delta \Rightarrow |h(x) - h(y)| < \varepsilon$ for all $x, y \in [-k, k]$. But then

$$\{|h(V_n) - h(V)| \ge \varepsilon\} \cap \{|V| \le k\} \subset \{|V_n - V| > \delta\} \cap \{|V| \le k\} \subset \{|V_n - V| > \delta\}.$$

Combining this with the equality above yields

$$\{|h(V_n) - h(V)| \ge \varepsilon\} \subset \{|V_n - V| \ge \varepsilon\} \cup \{|V| > k\}.$$

Then

$$P(|h(V_n) - h(V)| \ge \varepsilon) \le P(|V_n - V| \ge \varepsilon) + P(|V| > k),$$

by subadditivity of P. The sequence $\{|V| > k\}$ tends to the empty set as k increases, so $\lim_{k\to\infty} P(|V| > k) = 0$. Therefore, for any given $\eta > 0$, we can choose k so that $P(|V| > k) < \eta$. For fixed k, we pick a ε-dependent $\delta > 0$, so that

$$P(|h(V_n) - h(V)| > \varepsilon) \le P(|V_n - V| > \delta) + \eta,$$

where the right hand side tends to η as $n \to \infty$. Since $\eta > 0$ was arbitrary, the result follows.

(c) We can copy the proof from (a). Suppose $h : \mathbb{R}^p \to \mathbb{R}^q$, and that h is continuous at the point $x \in \mathbb{R}^p$. Then, given $\varepsilon > 0$, we can find $\delta > 0$, such that

$$\|x - y\|_p < \delta \quad \Rightarrow \quad \|h(x) - h(y)\|_q < \varepsilon,$$

where $\|x\|_p = (\sum_{j=1}^p x_j^2)^{1/2}$ is the Euclidian norm. If $X_n = (X_{1,n}, \ldots, X_{p,n})^{\mathrm{t}}$ is a sequence of random vectors in $\mathbb{R}^p$ converging in probability to a constant $x = (x_{1,n}, \ldots, x_{p,n})^{\mathrm{t}}$, and $h : \mathbb{R}^p \to \mathbb{R}^q$ is continuous in x, then, given $\varepsilon > 0$, we can find $\delta > 0$, such that

$$P(\|X_n - x\|_p < \delta) \le P(\|h(X_n) - h(x)\|_q < \delta) \le 1,$$

and the right hand side tends to 1 and $n \to \infty$. Both $h(x,y) = x + y$ and $h(x,y) = xy$ are continuous functions.

[xx perhaps also include the proof thought of in the exercise xx]

13.10. **The law of large numbers.** $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with $\mathrm{E}\, X_i = \xi$ and $\mathrm{Var}\, X_i = \sigma^2 < \infty$.

(a) Since the second moment is finite, $\mathrm{E}\, X_i^2 < \infty$, we can use Chebyshov's inequality. For an arbitrary $\varepsilon > 0$,

$$P(|\bar{X}_n - \xi| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathrm{E}\,(\bar{X}_n - \xi)^2 = \frac{1}{\varepsilon^2}\frac{\sigma^2}{n} \to 0,$$

as $n \to \infty$. Thus, $\bar{X}_n \xrightarrow{p} \xi$. The mean is *consistent* for the expected value.

(b) If you try Chebyshov's inequality here, you quickly get stuck, because there is no assumption on the fourth moment of the X_i's. Use instead what we found in Nils exercise 9(a) and (c), i.e. continuous mapping and that if $A_n \xrightarrow{p} a$ and $B_n \xrightarrow{p} b$, then $A_n + B_n \xrightarrow{p} a + b$: Since $\bar{X}_n \xrightarrow{p} \xi$, continuous mapping gives $\bar{X}_n^2 \xrightarrow{p} \xi^2$. Now, if we can prove that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathrm{E}\, X_1 = \xi^2 + \sigma^2, \tag{13.4}$$

then we can use Nils exercise 9(c) to conclude that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \xrightarrow{p} (\xi^2 + \sigma^2) - \xi^2 = \sigma^2.$$

We know that the first moment of X_i^2 is finite and equals $\xi^2 + \sigma^2$, but that's it. We need to dispense with the second moment assumption on the Law of large numbers to conclude that (13.4) is true.

Theorem 13.1. (Weak law of large numbers). *Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with finite expectation, $\mathrm{E}\, X_i = \xi$. Then $\bar{X}_n \xrightarrow{p} \xi$.*

Proof. Let X be a random variable with $\mathrm{E}\, X = \xi < \infty$. Write

$$X = XI_{|X| \leq M} + XI_{|X| > M}.$$

Then $\mathrm{E}\, XI_{|X|>M} \to 0$ by the Monotone convergence theorem. And $\mathrm{E}\, XI_{|X|\leq M} \to \xi$ as $M \to \infty$ because $|\mathrm{E}\, I_{|X|\leq M} - \mathrm{E}\, X| = |\mathrm{E}\, XI_{|X|>M}| \leq \mathrm{E}\,|X|I_{|X|>M}$. Moreover, $\mathrm{Var}\, XI_{|X|\leq M} = \mathrm{E}\, X^2 I_{|X|\leq M} - (\mathrm{E}\, XI_{|X|\leq M})^2 \leq \mathrm{E}\, X^2 I_{|X|\leq M} + (\mathrm{E}\, XI_{|X|\leq M})^2 \leq 2\mathrm{E}\, X^2 I_{|X|\leq M} \leq 2M\,\mathrm{E}|X| \leq \infty$, using Jensen's inequality. We can write

$$\begin{aligned}\bar{X}_n - \xi &= \frac{1}{n}\sum_{i=1}^{n} X_i I_{|X_i|\leq M} + \frac{1}{n}\sum_{i=1}^{n} X_i I_{|X_i|>M} - \xi \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \xi)I_{|X_i|\leq M} + \frac{1}{n}\sum_{i=1}^{n}(X_i - \xi)I_{|X_i|>M} = A_n + B_n,\end{aligned}$$

by which we define A_n and B_n (where I drop the dependence on M from the notation). For arbitrary $\varepsilon > 0$,

$$P(|A_n| \geq \varepsilon) \leq \frac{1}{\varepsilon n}\mathrm{Var}\, XI_{|X|\leq M} \leq \frac{2M}{\varepsilon^2 n}\mathrm{E}\,|X|,$$

and the second term,

$$P(|B_n| \geq \varepsilon) \leq \frac{\mathrm{E}\, XI_{|X|>M}}{\varepsilon},$$

by Chebyshov's ineq. and Markov's ineq., respectively. We have the inclusion

$$\{|A_n| + |B_n| \geq \varepsilon\} \subset \{|A_n| \geq \varepsilon/2\} \cup \{|B_n| \geq \varepsilon/2\},$$

(I find $\{|A_n| < \varepsilon/2\} \cap \{|B_n| < \varepsilon/2\} \subset \{|A_n| + |B_n| < \varepsilon\}$ easier to immediately accept), so using the triangle inequality and subadditivity of probability measures

$$\begin{aligned}P(|A_n + B_n| \geq \varepsilon) &\leq P(|A_n| + |B_n| \geq \varepsilon) \\ &\leq P(|A_n| \geq \varepsilon/2) + P(|B_n| \geq \varepsilon/2) \leq \frac{8M}{\varepsilon^2 n}\,\mathrm{E}\,|X| + \frac{2}{\varepsilon}\,\mathrm{E}\, XI_{|X|>M}.\end{aligned}$$

With an appropriate choice of M and n, the right hand side can be made arbitrarily small. □

(c) and (d) Assume that $\mathrm{E}\, X_i^3 < \infty$. Write

$$T_n = \frac{1}{n}\sum_{i=1}^n X_i^3 - 3\bar{X}_n \frac{1}{n}\sum_{i=1}^n X_i^2 + 4\bar{X}_n,$$

use Theorem 13.1, and what we found in Nils excercise 9(a) and (c). Same thing for $\hat{\kappa}_3$, as well as for the generalisation to higher moments.

13.11. **Convergence in distribution.** Write C_F for the continuity points of F.

(a) If $V_n \to_d V$, then $F_n(x) \to F(x)$ for all $x \in C_F$. Here $F_n(x) = P(V_n \le x)$ and $F(x) = P(V \le x)$. Then

$$P(V_n \in (a,b]) = F_n(b) - F_n(a) \to F(b) - F(a) = P(V \in (a,b]),$$

provided $a, b \in C_F$.

(b) Let $U_1, \ldots, U_n$ be i.i.d. uniform on $(0,1)$. Set $M_n = \max_{i\le n} U_i = U_{(n)}$. The distribution function of M_n is

$$F(m) = P(M_n \le m) = P(\max_{i\le n} U_i \le m) = P(U_i \le m, \text{ for } i = 1, \ldots, n) = m^n,$$

by independence of the U_i's. Set $V_n = n(1 - M_n)$,

$$P(V_n \le v) = P(-M_n \le v/n - 1) = 1 - P(M_n \le 1 - v/n) = 1 - (1 - v/n)^n \to 1 - e^{-v},$$

for $v > 0$. Which shows that V_n converges in distribution to a unit exponential.

(c) X_n and X are random variables with distributions on the integers $0, 1, 2, \ldots$, with probabilities $P(X_n = j) = f_n(j)$ and $P(X = j) = f_n(j)$ for $j = 0, 1, 2, \ldots$. We shall prove that $X_n \to_d X$ is equivalent to $f_n(j) \to f(j)$ for each j.

First, suppose that $X_n \to_d X$, that is $F_n(j) \to F(j)$ for all $j = 0, 1, 2, \ldots$. We are to show that $|f_n(j) - f(j)| \to 0$ for $j = 0, 1, 2 \ldots$. Clearly, $|f_n(0) - f(0)| = |F_n(0) - F(0)| \to 0$. And for any $j \ge 1$,

$$\begin{aligned} |f_n(j) - f(j)| &= |F_n(j) - F_n(j-1) - (F(j) - F(j-1))| \\ &\le |F_n(j) - F(j)| + |F_n(j-1) - F(j-1)| \to 0. \end{aligned}$$

Now, suppose that $|f_n(j) - f(j)| \to 0$ for $j = 0, 1, 2 \ldots$. For each j we can find $n_j \ge 1$ such that $|f_n(j) - f(j)| < \varepsilon/2^j$ for all $n \ge n_j$. For any $k = 0, 1, 2, \ldots$, we have $|P(X_n \le k) - P(X \le k)| \le \sum_{j=0}^k |f_n(j) - f(j)| < \sum_{j=0}^k \varepsilon/2^j$, for $n \ge \max\{n_1, \ldots, n_k\}$. Hence, $X_n \to_d X$.

(d) and (e) See solution to Exercise 4.1.

13.12. **Convergence of densities.** Suppose X_n and X have densities f_n and f.

(a) If $f_n(x) \to f(x)$ for all x, we have $X_n \to_d X$. Write

$$|f_n(x) - f(x)| = f_n(x) - f(x) + 2\max\{f(x) - f_n(x), 0\}.$$

Then

$$\begin{aligned} |F_n(x) - F(x)| &\le \int_{-\infty}^x |f_n(y) - f(y)|\,\mathrm{d}y \le \int_{-\infty}^\infty |f_n(y) - f(y)|\,\mathrm{d}y \\ &\le \int_{-\infty}^\infty (f_n(y) - f(y))\,\mathrm{d}y + 2\int_{-\infty}^\infty \max\{f(y) - f_n(y), 0\}\,\mathrm{d}y \\ &= 2\int_{-\infty}^\infty \max\{f(y) - f_n(y), 0\}\,\mathrm{d}y, \end{aligned}$$

because $\int_{-\infty}^\infty (f_n(y) - f(y))\,\mathrm{d}y = 1 - 1 = 0$. The function $\max\{f(y) - f_n(y), 0\} \le f(y)$ and $f(y)$ is integrable, it's a probability density function. Therefore, we can use the dominated convergence theorem (in second equality here),

$$\begin{aligned} \lim_{n\to\infty} |F_n(x) - F(x)| &= \lim_{n\to\infty} 2\int_{-\infty}^\infty \max\{f(y) - f_n(y), 0\}\,\mathrm{d}y \\ &= 2\int_{-\infty}^\infty \lim_{n\to\infty} \max\{f(y) - f_n(y), 0\}\,\mathrm{d}y = 0. \end{aligned}$$

This shows that pointwise convergence of densities implies convergence of the cumulatives, hence convergence in distribution.

(b) The same argument as in (a) can be used to show that pointwise convergence of densities implies L_1 convergence, that is $\int |f_n(y) - f(y)|\,\mathrm{d}y \to 0$. L_1 convergence is also equivalent to convergence of $\Delta(P_n, P)$, where

$$\Delta(P_n, P) = \sup_A |P_n(A) - P(A)|.$$

Assume L_1 convergence of the densities, then

$$\begin{aligned}\Delta(P_n, P) = \sup_A |P_n(A) - P(A)| &= \sup_A \Big|\int_A (f_n(y) - f(y))\,\mathrm{d}y\Big| \\ &\le \sup_A \int_A |f_n(y) - f(y)|\,\mathrm{d}y = \int |f_n(y) - f(y)|\,\mathrm{d}y \to 0,\end{aligned}$$

by the L_1 convergence. Assume that $\Delta(P_n, P) \to 0$. Define $B_n = \{y\colon f_n(y) \ge f(y)\}$. Note that $P_n(B_n) \ge P(B_n)$ by monotonicity of the integral, i.e., if $0 \le h(x) \le g(x)$ for all $x \in B$, then $\int_B h(x)\,\mathrm{d}x \le \int_B g(x)\,\mathrm{d}x$. For any event A,

$$\begin{aligned}\int_A |f_n(y) - f(y)|\,\mathrm{d}y &= \int_{A\cap B_n} |f_n(y) - f(y)|\,\mathrm{d}y + \int_{A\cap B_n^c} |f_n(y) - f(y)|\,\mathrm{d}y \\ &= \int_{A\cap B_n} (f_n(y) - f(y))\,\mathrm{d}y + \int_{A\cap B_n^c} (f(y) - f_n(y))\,\mathrm{d}y \\ &= P_n(A\cap B_n) - P(A\cap B_n) + P(A\cap B_n^c) - P_n(A\cap B_n^c) \\ &= |P_n(A\cap B_n) - P(A\cap B_n)| + |P(A\cap B_n^c) - P_n(A\cap B_n^c)| \le 2\Delta(P_n, P),\end{aligned}$$

where the monotonicity property is used in the fourth equality. Since this holds for any event A, it must also hold for the sure event, the sample space.

(c) The probability density function of the t_m, the Student's t-distribution with m degrees of freedom is

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

Stirlings's approximation for the gamma function is

$$\Gamma(z) \doteq \sqrt{\frac{2\pi}{z}} z^z e^{-z}.$$

Using this, and rearranging,

$$\frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\Gamma(m/2)} \doteq \big(\frac{m}{m+1}\big)^{1/2}\big(\frac{m+1}{2\pi m}\big)^{1/2}\big(1 + \frac{1}{m}\big)^{m/2} e^{-1/2} \to \frac{1}{(2\pi)^{1/2}},$$

as $m \to \infty$, where we have used that $(1 + 1/m)^{m/2} \to e^{1/2}$. The second factor is

$$(1 + \frac{x^2}{m})^{-(m+1)/2} = (1 + \frac{x^2}{m})^{-m/2}(1 + \frac{x^2}{m})^{-1/2} \to e^{-x^2/2},$$

as $m \to \infty$. Putting this together $f_m(x) \to e^{-x^2/2}/\sqrt{2\pi}$ for each x, and by (a) we have convergence in distribution of a sequence of t_m random variables to a standard normal. One lesson of this story is the following: If you have independent normal data $X_1, \ldots, X_n$ and want to test things using the statistic

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n},$$

which is distributed t_{n-1} (see Nils exercise 8(d)), then it's okay to treat this statistic as standard normal provided n is big enough. The same goes for the regression setting, etc.

(d) This is an exercise in the oblig.

13.13. **The Portmanteau theorem for convergence in distribution.** Here is the statement of the theorem given in Nils' exercise set.

Theorem 13.2. (Portmanteau theorem). *X_n and X are random variables with probability measures $P_n(A) = P(X_n \in A)$ and $P(A) = P(X \in A)$, and distribution functions $F_n(x) = P_n(-\infty, x]$ and $F(x) = P(-\infty, x]$. The following statements are equivalent.*

(i) *$X_n \to_d X$, i.e., $F_n(x) \to F(x)$ for all continuity points of F.*
(ii) *$\liminf_n P_n(O) \geq P(O)$ for all open sets O.*
(iii) *$\limsup_n P_n(C) \leq P(C)$ for all closed sets C.*
(iv) *$\lim_n P_n(A) = P(A)$ for all sets A whose boundary $\partial A = \bar{A} \setminus A^o$ has $P(\partial A) = 0$.*
(v) *$\mathrm{E}\, h(X_n) \to \mathrm{E}\, h(X)$ for all continuous and bounded functions $h : \mathbb{R} \to \mathbb{R}$.*

Proof. (i)⇒(ii) Use that every non-empty open set in $\mathbb{R}$ is a disjoint union of open intervals (a_i, b_i). For and open set $O = \cup_k I_k$, with $I_k = (a_k, b_k)$, $a_k < b_k$, choose $I'_k = (a'_k, b'_k] \subset I_k$ such that a'_k,b'_k are continuity points of F and $P(I'_k) \leq P(I_k) + \varepsilon/2^k$, for an arbitrary $\varepsilon > 0$. In Exercise 2.5 we saw that there are only countably many discontinuity points, so there is no problem finding such subintervals. Since the $I_1, I_2, \ldots$ interval are disjoint

$$P_n(O) = \sum_k P(I_k) \geq \sum_k P_n(I'_k).$$

Fatou's lemma (i.e, $\liminf_n \int f_n \,\mathrm{d}\mu \geq \int \liminf_n f_n \,\mathrm{d}\mu$), our condition on the I'_k intervals, and assuming (i), give

$$\begin{aligned}
\liminf_n P_n(O) &\geq \liminf_n \sum_k P_n(I'_k) \geq \sum_k \liminf_n P_n(I'_k) \\
&= \sum_k \liminf_n \{F_n(b'_k) - F_n(a'_k)\} = \sum_k \{F(b'_k) - F(a'_k)\} \\
&= \sum_k P(I'_k) = \sum_k \{P(I'_k) - \varepsilon/2^k\} = P(O) - \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, (i) is seen to imply (ii).

(ii)⇒(iii) If C is closed, then its complement C^c is open. By (ii)

$$\limsup_n P_n(C) = \limsup_n \{1 - P_n(C^c)\} = 1 - \liminf_n P_n(C^c) \leq 1 - P(C^c) = P(C).$$

This also gives (iii)⇒(ii), since the complement of an open set is closed.

(iii)⇒(iv) The boundary ∂A of a set A is $\partial A = \bar{A} \setminus A^o$, where $\bar{A}$ is the closure of A, which is always closed, and A^o is the interior of A, which is always open. Moreover, $A^o \subset A$, and $A \subset \bar{A}$. Also, if A is closed $A^o = A$, and if A is open,

$$P(\bar{A}) = P(A^o \cup \partial A) = P(A^o) + P(\partial A) = P(A),$$

since A^o and ∂A are disjoint, and $P(\partial A) = 0$ by assumption. Then

$$\limsup_n P_n(A) \leq \limsup_n P_n(\bar{A}) \leq P(\bar{A}) = P(A),$$

and

$$\liminf_n P_n(A) \geq \limsup_n P_n(A^o) \leq P(A^o) = P(A),$$

so $\lim_n P_n(A) = P(A)$ for all set A whose boundary have P-probability zero.

(iv)⇒(v). Let $a \leq f(x) \leq b$ be a continuous and bounded function. Define

$$h(x) = \frac{f(x) - a}{b - a},$$

so that $0 \leq h(x) \leq 1$. We can write (make a drawing)

$$h(x) = \int_0^1 I\{y \leq h(x)\} \,\mathrm{d}y.$$

Then

$$\int_0^1 h(x)\,\mathrm{d}P(x) = \int_0^1 \int_0^1 I\{y \le h(x)\}\,\mathrm{d}y\,\mathrm{d}P(x)$$
$$= \int_0^1 \int_0^1 I\{y \le h(x)\}\,\mathrm{d}P(x)\,\mathrm{d}y = \int_0^1 P(h(X) \ge y)\,\mathrm{d}y.$$

by Fubini's theorem. We can therefore write $\mathrm{E}\,h(X_n) = \int_0^1 P(h(X_n) \ge y)\,\mathrm{d}y = \int_0^1 P_n(A_y)\,\mathrm{d}y$ and $\mathrm{E}\,h(X) = \int_0^1 P(h(X) \ge y)\,\mathrm{d}y = \int_0^1 P(A_y)\,\mathrm{d}y$, where $A_y = \{x\colon h(x) \ge y\}$. If we can show that $P_n(A_y) \to P(A_y)$ for all continuity points of $P(h(X) \le y)$, then since $0 \le P_n(A_y) \le 1$, the bounded convergence theorem gives $\mathrm{E}\,h(X_n) = \int_0^1 P_n(A_y)\,\mathrm{d}y \to \int_0^1 P(A_y)\,\mathrm{d}y = \mathrm{E}\,h(X)$. Since h is continuous $\partial A_y \subset \{x\colon h(x) = y\}$, and also $P(\{x\colon h(x) = y\}) > 0$, for at most countably many y, because h is continuous. This means that $P(\partial A_y) = 0$ for all $y \in [0,1]$, except for at, at most, countably many points. By (iv), this gives $P_n(A_y) \to P(A_y)$.

(v)⇒(i). For a given y, consider the functions that for some $\varepsilon > 0$ are given by

$$h_{1,\varepsilon}(x) = \begin{cases} 1, & x \le y - \varepsilon, \\ \frac{y-x}{\varepsilon}, & y - \varepsilon \le x \le y, \\ 0, & x \ge y. \end{cases} \quad \text{and} \quad h_{2,\varepsilon}(x) = \begin{cases} 1, & x \le y, \\ \frac{y+\varepsilon-x}{\varepsilon}, & y \le x \le y + \varepsilon, \\ 0, & x \ge y + \varepsilon. \end{cases}$$

Both these functions are continuous and bounded by one, so that (v) applies. Note that

$$h_{1,\varepsilon}(x) \le I\{x \le y\} \le h_{2,\varepsilon}(x), \quad \text{for all } x.$$

Then

$$\liminf_n F_n(y) \ge \liminf_n \mathrm{E}\,h_{1,\varepsilon}(X_n) = \mathrm{E}\,h_{1,\varepsilon}(X),$$

and

$$\limsup_n F_n(y) \le \limsup_n \mathrm{E}\,h_{2,\varepsilon}(X_n) = \mathrm{E}\,h_{2,\varepsilon}(X),$$

so that

$$\mathrm{E}\,h_{1,\varepsilon}(X) \le \liminf_n F_n(y) \le \limsup_n F_n(y) \le \mathrm{E}\,h_{2,\varepsilon}(X).$$

Since both $h_{1,\varepsilon}$ and $h_{2,\varepsilon}$ are bounded by one and integrable, dominated convergence ensures that $\mathrm{E}\,h_{1,\varepsilon}(X)$ and $\mathrm{E}\,h_{2,\varepsilon}(X)$ both converge to $F(y)$ when $\varepsilon \to 0$. □

13.14. **The continuity theorem.** [xx latex it xx]

13.15. **Slutsky–Cramér rule.** [xx latex it xx]

13.16. **The Central Limit Theorem.** [xx latex it xx]

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis. Second Edition.* Springer.

Billingsley, P. (1995). *Probability and Measure. Third Edition.* Wiley.

Casella, G. and Berger, R. L. (2002). *Statistical Inference. Second Edition.* Duxbury Pacific Grove, CA.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236:119–127.

Ferguson, T. S. (1996). *A course in large sample theory.* Chapman & Hall/CRC.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379.

Keener, R. W. (2011). *Theoretical statistics: Topics for a core course.* Springer, New York.

Lehmann, E. L. and Casella, G. (1999). *Theory of point estimation. Second edition.* Springer, New York.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses. Third edition.* Springer, New York.

Lindley, D. V. and Phillips, L. (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, 30:112–119.

Romano, J. P. and Siegel, A. F. (1986). *Counterexamples in Probability and Statistics.* Wadsworth & Brooks/Cole.

Schervish, M. J. (1995). *Theory of Statistics.* Springer, New York.

Shao, J. (2003). *Mathematical Statistics. Second Edition*. Springer.

Shiryaev, A. N. (1990). *Probability. Second Edition*. Springer.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206.

Stigler, S. M. et al. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press.