STK 4011-9011 Autumn 2023
Statistical Inference Theory: the Oblig

This is The Oblig, the mandatory assignment, for STK 4011-9011, Statistical Inference Theory, Autumn 2023. It is made available at the course website Monday October 9, and the submission deadline is Monday October 23, 13:58, *via the Canvas system*. Reports may be written in nynorsk, bokmål, riksmål, English, or Latin, should preferably be text-processed (for instance with TeX or LaTeX), and must be submitted as a single pdf file. The submission must contain your name, the course, and assignment number.

The Oblig set contains four exercises and comprises five pages (in addition to the present introduction page, 'page 0').

Importantly, the PhD candidates taking the **STK 9011** version of the court, need to work also with one more exercise, namely 3.9 in Hjort & Stoltenberg's PartOne.pdf, using the `smallbabies` dataset on the course website.

It is expected that you give a clear presentation with all necessary explanations, but write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Remember to include all relevant plots and figures. These should preferably be placed inside the text, close to the relevant subquestion.

For a few of the questions setting up an appropriate computer programme might be part of your solution. The code ought to be handed in along with the rest of the written assignment; you might place the code in an appendix.

All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

Application for postponed delivery: If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (email: `studieinfo@math.uio.no`) well before the deadline.

The mandatory assignment in this course must be approved, in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments, along with a 'log on to Canvas', can be found here:

`www.uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html`

Enjoy [imperative pluralis].

Nils Lid Hjort

1. Quartiles & range

For a distribution with continuous and increasing cumulative distribution function F, the lower and upper quartiles are $q_{0.25} = F^{-1}(0.25)$ and $q_{0.75} = F^{-1}(0.75)$. Their empirical cousins $Q_{n,0.25}$ and $Q_{n,0.75}$, for a sample $Y_1, \ldots, Y_n$, are the corresponding 0.25 and 0.75 sample quantiles, which one finds via `quantile(yy,0.25)` and `quantile(yy,0.75)` in `R` (here with `yy` being the data vector).

(a) For F being the uniform distribution on the unit interval, and with $n = 100$, write down the explicit density of $Q_{n,0.75} = U_{(75)}$, order statistic no. 75, along with a few arguments. For a general n, give similarly the density $g_{n,0.75}(u)$ for $Q_{n,0.75} = U_{([0.75\,n])}$. Derive an expression for the density $h_{n,0.75}(z)$ of the transformed variable $Z_{n,0.75} = \sqrt{n}(U_{n,0.75} - 0.75)$. Attempt to show that this density converges to the density of a $\mathrm{N}(0, 0.25 \cdot 0.75)$.

(b) Consider now i.i.d. data $Y_1, \ldots, Y_n$ from the normal $\mathrm{N}(\xi, \sigma^2)$. Show that the lower and upper quartiles for this distribution are $\xi - c\sigma$ and $\xi + c\sigma$, with $c = 0.674$. Use results from Hjort & Stoltenberg, Ch. 2, to show that

$$\sqrt{n}(Q_{n,0.75} - (\xi + c\sigma)) \to_d Z_{0.75} \sim \mathrm{N}(0, \tau_{0.75}^2), \quad \text{with } \tau_{0.75}^2 = 0.25 \cdot 0.75\, \sigma^2/\phi(c)^2,$$

with $\phi(c) = \exp(-\frac{1}{2}c^2)/(2\pi)^{1/2}$ the usual standard normal density.

(c) There is similarly convergence $\sqrt{n}(Q_{n,0.25} - (\xi - c\sigma)) \to_d Z_{0.25} \sim \mathrm{N}(0, \tau_{0.25}^2)$, say. Show that $\tau_{0.25}$ is the same as $\tau_{0.75}$. Then use results from the book-to-be's Exercise 2.20 to show that

$$\begin{pmatrix} \sqrt{n}(Q_{n,0.25} - (\xi - c\sigma)) \\ \sqrt{n}(Q_{n,0.75} - (\xi + c\sigma)) \end{pmatrix} \to_d \begin{pmatrix} Z_{0.25} \\ Z_{0.75} \end{pmatrix},$$

a binormal limit with zero means and covariance matric

$$\Sigma = \frac{\sigma^2}{\phi(c)^2} \begin{pmatrix} 0.25 \cdot 0.75, & 0.25^2 \\ 0.25^2, & 0.25 \cdot 0.75 \end{pmatrix}).$$

(There is a misprint in the PartOne pdf, for this exercise: in (g) the covariance should be $q_j(1 - q_\ell)/\{f(\mu_i)f(\mu_\ell)\}$ for $j < \ell$.)

(d) Now consider the so-called interquartile range, $\mathrm{iq}_n = Q_{n,0.75} - Q_{n,0.25}$. Explain from the above that

$$\sqrt{n}(\mathrm{iq}_n - 2c\sigma) \to_d \mathrm{N}(0, \kappa^2), \quad \text{with } \kappa = \tfrac{1}{2}\sigma/\phi(c).$$

(e) The standard estimator for σ in such a normal sample is $\widehat{\sigma}$, the empirical standard deviation. From results in Hjort & Stoltenberg, we have $\sqrt{n}(\widehat{\sigma} - \sigma) \to_d \mathrm{N}(0, \frac{1}{2}\sigma^2)$. Now argue that the above considerations invite another estimator for σ, namely $\sigma^* = \mathrm{iq}_n/(2c) = \mathrm{iq}_n/1.349$. Find the limit variance for $\sqrt{n}(\sigma^* - \sigma)$. How much does σ^* lose, to $\widehat{\sigma}$, if the normal model holds? Are there arguments in support of sometimes using σ^*, in spite of the fact that it is less precise under normality?

(f) Now attempt to generalise that above, which is about $F^{-1}(0.75) - F^{-1}(0.25)$, to $F^{-1}(1-q) - F^{-1}(q)$, for other suitable $q \in (0,1)$. Form a good estimator of σ, say σ_q^*, by scaling $Q_{n,1-q} - Q_{n,q}$. Which, of all such symmetric quantile difference type estimators for σ, has the smallest limiting variance?

2. Estimation in the geometric model

"If that dice has a 'one' face up, I thought, I'm going downstairs to rape Arlene" (says the main character in Luke Rhineheart's *Dice Man* to himself). But what it the dice are rigged?

(a) Let Y be the number of independent dice throws required to have 'one' the first time. With $p = \Pr(\text{'one'})$, explain that the point-mass probabilities are

$$f(y,p) = \Pr(Y = y) = (1-p)^{y-1}p \quad \text{for } y = 1, 2, \ldots.$$

Show that $\mathrm{E}\,Y = 1/p$ and $\mathrm{Var}\,Y = (1-p)/p^2$.

b) Find an expression for $\Pr(Y \geq y_0 + y \,|\, Y \geq y_0)$, and give an interpretation of this.

(c) Suppose you repeatedly carry out this simple experiment many times, leading to counts $Y_1, \ldots, Y_n$, each with the $f(y,p)$ probabilities. With $\bar{Y}_n$ the sample average, use the central limit theorem to find the approximate distribution of $\sqrt{n}(\bar{Y}_n - 1/p)$.

(d) Show that the moment estimator for p is $\widehat{p} = 1/\bar{Y}_n$. Find the limit distribution of $\sqrt{n}(\widehat{p} - p)$.

(e) I did this (with my admittedly artificial die), getting

`1 2 7 1 20 3 7 17 6 8 2 13 8 9 17 1 26 2 13 4 1 3 9 5 24`

for 25 experiments. Find a 90 percent confidence interval for p, and test the null hypothesis that my die is fair, i.e. has $\Pr(\text{'one'}) = 1/6$.

(f) Changing gears a little, suppose you throw a fair coin 5 times, getting your little sequence of 'kron' and 'mynt'. How many times do you expect to need to do this 5-in-a-row operation, until you finally have kron-kron-kron-kron-kron? (Jo Røislien, tv person and professor of statistics, once proclaimed for his millions of viewers, in one of the *Siffer* episodes, 'I will now throw my coin ten times and have the same outcome in each'. And proceeded, rather impressively, to do so. He looked tired, though – since he had done it a very high number of times, and only showed us that final successful clip.)

3. Transformers (and back-transformers)

The true transformation takes place within, my local psychologist claims – in the present exercise we shall transform from one thing to another but then perhaps transform partly back the other way again. We start with (X, Y) being independent standard normals, so that their joint density may be written

$$f_0(x,y) = \frac{1}{2\pi} \exp\{-\tfrac{1}{2}(x^2 + y^2)\} \quad \text{for } x, y \text{ on the real line.}$$

(a) We then transform to so-called polar coordinates,

$$X = R\cos\theta \quad \text{and} \quad Y = R\sin\theta,$$

with $\theta \in [-\pi, \pi]$. Show that (R, θ) has the density

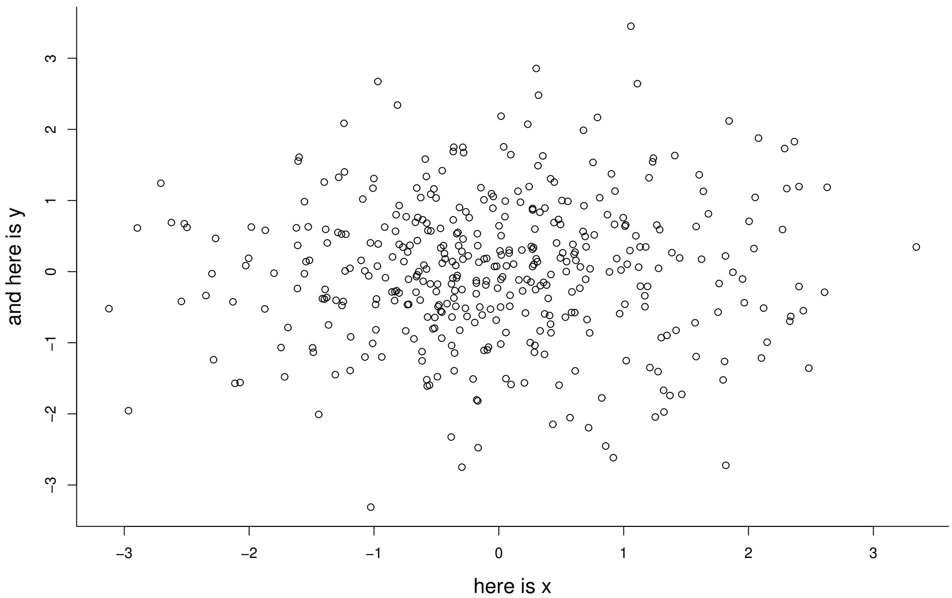
$$g_0(r,\theta) = h_0(r)\frac{1}{2\pi} = r\exp(-\tfrac{1}{2}r^2)\frac{1}{2\pi} \quad \text{for } r > 0 \text{ and } \theta \in [-\pi, \pi].$$

– The density $h_0(r) = r\exp(-\frac{1}{2}r^2)$ reached above is sometimes called the Raleigh distribution (after Lord Raleigh, who won the Nobel in physics in 1904), though a separate name might not really be needed in that it is simply a $\chi_2 = (\chi_2^2)^{1/2}$ – the density of the square-root of a chi-squared with two degrees of freedom. But the representation above opens the door for generalising the binormal distribution we started out with, by inventing a more general density than $h_0(r)$.

(b) Suppose the random radius R has density $h(r)$, rather than the $h_0(r)$, and keep θ uniform on $[-\pi, \pi]$, independent of R. Show that $(X, Y) = (R\cos\theta, R\sin\theta)$ then must have density

$$f(x,y) = h(\sqrt{x^2+y^2})\frac{1}{\sqrt{x^2+y^2}}\frac{1}{2\pi}.$$

You may find a need for the mathematical fact that the derivative of $A(u) = \arctan u$ is $A'(u) = 1/(1+u^2)$.



Simulated pairs (x_i, y_i) from the generalised binormal density, where I used $h(r) = \frac{1}{2}\gamma r^{\gamma-1}\exp(-\frac{1}{2}r^\gamma)$ rather than $h_0(r) = r\exp(-\frac{1}{2}r^2)$, for a secret value of γ.

(c) Consider one such generalisation, namely

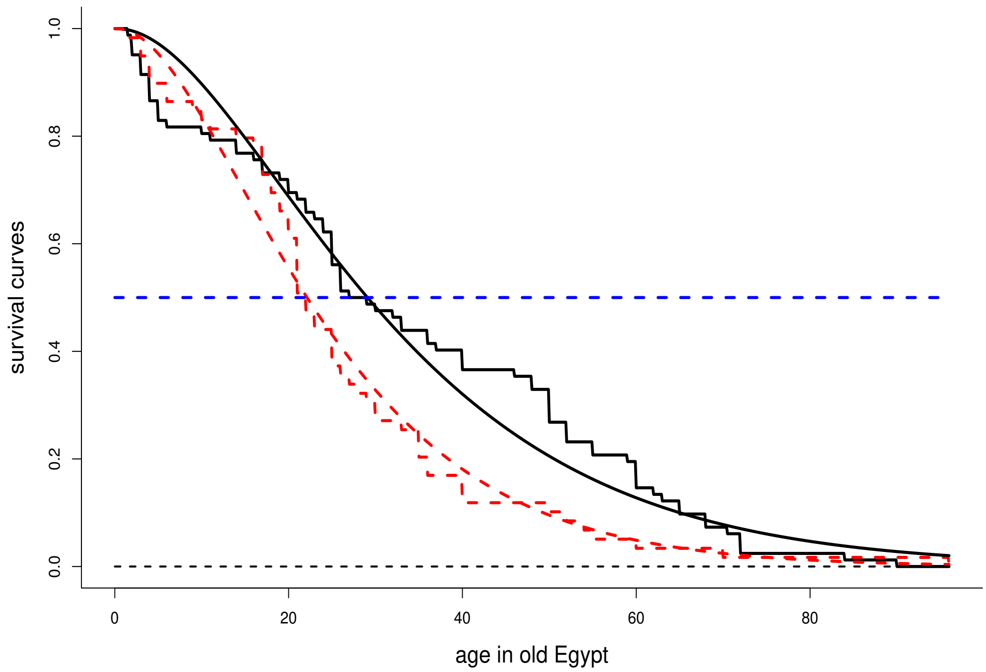
$$h(r) = \tfrac{1}{2}\gamma r^{\gamma-1}\exp(-\tfrac{1}{2}r^{\gamma}) \quad \text{for } r > 0,$$

where $\gamma > 0$ is such an extra parameter. We let this be the new distribution for the random radius R, and again keep the random angle θ distributed independently and uniform on $[-\pi, \pi]$. In the figure above I have simulated $n = 400$ pairs from the associated distribution $f(x, y)$, with a certain value of γ known so far only to me. Explain how you can assess whether the data are from the simple binormal density, i.e. from the $f_0(x, y)$ of (a), or not.

4. Lifelengths in Roman Era Egypt, 2100 years ago

Intriguingly, archeologists have been able to learn the ages at death of 141 mummified individuals living in Roman Era Egypt, some 2100 years ago, see Spiegelberg (1901). These lifelengths, varying from 1 to 96 years, for 82 men and 59 women, were discussed and analysed by Karl Pearson in the very first volume of Biometrika, see Pearson (1902). We treat them here as a random sample of lifelengths from the upper social class of Roman Era Egypt, during a period of relative societal stability; more details are in Hjort & Stoltenberg (2024, Story xxxiii).

Despite Pearson's not unreasonable comment that "in dealing with [these data] I have not ventured to separate the men from the women mortality, the numbers are far too insignificant" we shall work with parametric modelling for their survival.



Survival curves $S(t) = \Pr(T \geq t)$ for old Egypt, for men (black curves) and women (red curves); the ragged curves are the direct empirical ones, and the smooth curves are based on Gamma distributions. The horizontal 0.50 line is there to help us read off the medians.

(a) Read the data into your computers, via the `egypt-data` file on the course website. Writing n_m and n_w for the sample sizes, and $T_{m,i}$ and $T_{w,i}$ for the lifelengths, consider the empirical distribution functions

$$\widehat{F}_m(t) = (1/n_m)\sum_{i=1}^{n_m} I(T_{m,i} \le t), \quad \widehat{F}_w(t) = (1/n_w)\sum_{i=1}^{n_w} I(T_{w,i} \le t).$$

Show that these are unbiased for the underlying cumulative distribution functions F_m and F_w. Compute them, along with their nonparametric survival curves $\widehat{S}_m = 1 - \widehat{F}_n$, $\widehat{S}_w = 1 - \widehat{F}_w$, and display these in a diagram.

(b) Suppose in general terms that $T_1, \ldots, T_n$ are i.i.d. from the Gamma(a, b) distribution, with density $\{b^a/\Gamma(a)\}t^{a-1}\exp(-bt)$ for $t \ge 0$. Show that the mean and variance are a/b and a/b^2. Explain that the moment estimators are the solutions to the two equations

$$\bar{Y} = a/b \quad \text{and} \quad V = a/b^2,$$

in which $\bar{Y}$ and V are the sample mean and empirical variance of the data. Solve these, to find

$$\widehat{a} = \bar{Y}^2/V \quad \text{and} \quad \widehat{b} = \bar{Y}/V.$$

Show that these are consistent, i.e. converging to the corect values when sample size grows (if the model holds).

(c) Compute these parameter estimates, for the men and women of Old Egypt, and use these to construct a version of the figure above.

(d) Consider $d = \mu_m - \mu_w$, the difference between median life times for men and women. Compute first the direct nonparametric estimate d^*, the difference between the empirical medians. Use methods as for the Quantile Story in Hjort & Stoltenberg, on Oslo girls and boys, to construct a confidence interval for d.

(e) Then we attempt to estimate the d difference parametrically – when it works, it will often be more precise than the nonparametric procedure. Compute therefore the estimate

$$\widehat{d} = \mu_m(\widehat{a}_m, \widehat{b}_m) - \mu_w(\widehat{a}_w, \widehat{b}_w),$$

using the estimated Gamma distributions. Comment on d^* and $\widehat{d}$.

(f) It's a bit laborious, and even if you do not succeed in following your ideas and arguments to the finish, try to explain how you can compute an estimated variance for $\widehat{d}$, and how this may be used to find a parametrically based confidence interval for the median difference.