STK 4011-9011 Autumn 2024
Statistical Inference Theory: the Oblig

This is The Oblig, the mandatory assignment, for STK 4011-9011, Statistical Inference Theory, Autumn 2024. It is made available at the course website Monday October 7, and the submission deadline is Monday October 21, 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 course need to work also with one more exercise: struggle through as much as you manage of Story vii.8, from Hjort & Stoltenberg's PartTwo.pdf, 'Boys, girls, and mathematics scores' – which I constructed after having read a Verdens Gang newspaper story December 2023.

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.

Also, please include half a page or so, at the start of your report, regarding how your work proceeded, what was difficult, the extent to which you have worked independently, which aids you used, how satisfied you are, what you learned, etc.

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. Summing uniforms to compute e

Consider i.i.d. uniform variables $U_1, U_2, \ldots$ on the unit interval. How many of these do we need to observe, in order for their sum to exceed 1?

(a) Show that the densities for $U_1 + U_2$ and $U_1 + U_2 + U_3$ become

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

and

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

Draw these two densities in a diagram and comment briefly on how they look.

(b) Find a formula for the moment-generating function $M_0(t) = \mathrm{E}\exp\{t(U_i - \frac{1}{2})\}$ for $U_i - \frac{1}{2}$. With $\bar{U}_n = n^{-1}\sum_{i=1}^n U_i$, work out also a formula for the moment-generating function $M_n(t)$ for $Z_n = \sqrt{n}(\bar{U}_n - \frac{1}{2})$, and plot it for say $n = 10$. Comment on what you see.

(c) It is perfectly possible but a bit cumbersome to find an explicit formula for the density $f_n(x)$ of $U_1 + \cdots + U_n$, with $0 \le x \le n$, for the general case. On this occasion we are content to study this $f_n(x)$ only for $x \in [0, 1]$. Show, perhaps by induction, that $f_n(x) = x^{n-1}/(n-1)!$ on this startinterval $[0, 1]$. Deduce in particular from this that $F_n(1) = \Pr(U_1 + \cdots + U_n \le 1) = 1/n!$, valid for each n.

(d) Now study N, the first n such that the partial sum $U_1 + \cdots + U_n$ is bigger than 1. Explain that $\Pr(N > n) = \Pr(U_1 + \cdots + U_n \le 1)$. From this, derive a formula for the point probabilities $g(n) = \Pr(N = n)$, and deduce the peculiar property that $\mathrm{E}\,N = e$.

(e) This invites an idiosyncratic way of finding the number e with say three decimal places: set up a computer simulation script that gives you a million N^*, the number of terms needed for a sum of uniforms to exceed 1. Explain that the average of this million N^* really must be close to e. Give this estimate, along with an approximate 99 percent confidence interval. Check if you get Ibsen's and Tolstoy's birth years right.

Лва? –
Лва!

2. Spherical coordinates

Points in the plane may be transformed from Cartesian to polar coordinates, and vice versa; similarly, but with more delicate details, points in the three-dimensional space can be represented in spherical coordinates. Start with a triple (X, Y, Z) of independent standard normals, and from these transform to (R, A, B), via

$$X = R\sin A\cos B, \quad Y = R\sin A\sin B, \quad Z = R\cos A.$$

(a) Show that in fact

$$\begin{aligned} R &= (X^2+Y^2+Z^2)^{1/2}, \\ A &= \arccos\frac{Z}{(X^2+Y^2+Z^2)^{1/2}}, \\ B &= \mathrm{sign}(Y)\arccos\frac{X}{(X^2+Y^2)^{1/2}}, \end{aligned}$$

featuring the arccos function (the inverse of the cosinus) and the $\mathrm{sign}(u)$ function, the latter equal to 1 for $u \ge 0$ and -1 for $u < 0$.

(b) Set up the 3×3 Jacobi matrix $\partial(x,y,z)/\partial(r,a,b)$ of partial derivatives of (x,y,z) with respect to (r,a,b), and show that its determinant becomes $r^2\sin a$. From this demonstrate that the joint density $g(r,a,b)$ can be written as the product $g_1(r)g_2(a)g_3(b)$, with $g_1(r)$ the density proportional to $r^2\exp(-\frac{1}{2}r^2)$, $g_2(a)=\frac{1}{2}\sin a$ over $[0,\pi]$, and g_3 the uniform density $1/(2\pi)$ over $[-\pi,\pi]$. Comment on what this implies.

(c) To illustrate these transformation results, and a test of your simulation and plotting skills, simulate 10^3 triples (X,Y,Z), and from these compute the corresponding 10^3 values of (R,A,B). Give histograms for these three variables, along with plots of their densities.

(d) Explain briefly that if you tweak the distribution of R slightly, and keep the start transformation from (R,A,B) to (X,Y,Z), then you have generalised the normal distribution.

3. Quantiles via sums

Transformation can be a transformative experience. Two well-known distributions from the course material are as follows. First, the $\mathrm{Gam}(a,1)$ distribution has density $\Gamma(a)^{-1}x^{a-1}\exp(-x)$ for x positive, and has mean a and variance a. Second, the $\mathrm{Beta}(a,b)$ has density $\{\Gamma(a+b)/(\Gamma(a)\Gamma(b))\}x^{a-1}(1-x)^{b-1}$ on the unit interval, with mean $\xi=a/(a+b)$ and variance $\xi(1-\xi)/(a+b+1)$.

(a) When $X\sim\mathrm{Gam}(a,1)$, show that its moment-generating function $\mathrm{E}\exp(tX)$ takes the form $1/(1-t)^a$. Use this to show that a sum of independent $\mathrm{Gam}(a_i,1)$ variables, for $i=1,\ldots,k$, is a $\mathrm{Gam}(\sum_{i=1}^k a_i,1)$.

(b) Suppose X and Y are independent with the same $\mathrm{Gam}(a,1)$ distribution. Put up an expression for the joint density $f(x,y)$ for these. Then transform these to $R=X/(X+Y)$ and $Z=X+Y$. Find their joint density $g(r,z)$. Show from this that $R\sim\mathrm{Beta}(a,a)$, and give its mean and variance.

(c) Let $U_1,\ldots,U_n$ be i.i.d. on the unit interval, with n odd, so we can write $n=2m+1$. With $M_n=U_{(m+1)}$ the median, show that $M_n\sim\mathrm{Beta}(m+1,m+1)$. Explain that this leads to the representation

$$M_n=\frac{X_1+\cdots+X_{m+1}}{X_1+\cdots+X_{m+1}+Y_1+\cdots+Y_{m+1}}=\frac{\bar X_{m+1}}{\bar X_{m+1}+\bar Y_{m+1}},$$

in which the X_i and the Y_i are all independent and standard exponentially distributed.

(d) Use the Central Limit Theorem to show that

$$\sqrt{m+1}(\bar{X}_{m+1}-1)\to_d U, \quad \sqrt{m+1}(\bar{Y}_{m+1}-1)\to_d V,$$

where U and V are independent standard normals. Then use the delta method to find the limit distribution for $\sqrt{m+1}(M_n-\frac{1}{2})$.

(e) In one of the many exercises from Ch. 2 in the course you have all cleverly shown that $\sqrt{n}(M_n-\frac{1}{2})\to_d \mathrm{N}(0,1/4)$, by working out an expression for its density and its limit. Now deduce this result from the above.

4. How special are You?

How special are You, gentle course participant, among the other mammals on this planet? The dataset `mammals4011` gives the average body weight and average brain weight for 56 mammals, in kg, from tiny short-tailed shrew (0.005 kg) to the African elephant (6654 kg), and with brain to body ratios ranging from the cow and Brazilian tapir (about 0.08 percent) to You (with a somewhat modest 2.1 percent) up to the thirteen-lined ground squirrel with the impressive 3.9 percent. Intriguingly, the (log-body, log-brain) data pairs follow approximately a binormal distribution, with a relatively high correlation, making it feasible to assess the biological variability and from this the extent to which You might consider yourself special.

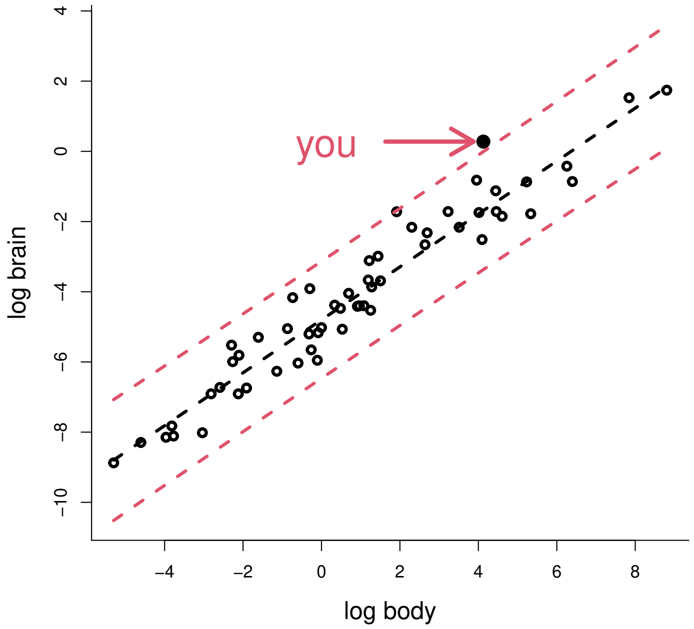
I have compiled the dataset via a fuller version, with supplementary information regarding e.g. sleep characteristics, available at

`github.com/tidyverse/ggplot2/blob/main/data-raw/msleep.csv`

from which one may also identify the 56 mammals in question; You happen to be data pair no. 25. With the `R` software, the data may be written into your computer using

```
mammals <- matrix(scan("mammals4011", skip=7),byrow=T,ncol=3)
id <- mammals[ ,1]
xx0 <- mammals[ ,2]
yy0 <- mammals[ ,3]
xx <- log(xx0)
yy <- log(yy0)
xxnew <- xx[-25]
yynew <- yy[-25]
```

(a) Plot first (body, brain) and then the more statistically informative (log-body, log-brain). For the latter, compute the correlation 0.965, and use the method outlined in Exercise 2.48 to give an approximate 90 percent confidence interval. Discuss briefly why correlation on this log-log scale might be a more meaningful measure of association than correlation on the original body-brain scale.



Plot of log body-weight, log brain-weight, both in log-kg, for 56 mammals, including You, plotted at $(\log 62.00, \log 1.32)$. *The regression line and 99 percent prediction band are computed based on having You pushed out of the data, i.e. carried out based on the other 55 mammals.*

(b) To have a fair assessment of how special You are, we carry out modelling and analysis on the revised dataset where You have gently removed yourself. For this edited dataset with $n = 55$ pairs (x_i, y_i) of (log-body, log-brain) measurements, carry out linear regression for y on x. Parts of the mathematics below become simpler writing the model as

$$y_i = a + bx_i^* + \varepsilon_i = a + b(x_i - \bar{x}) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with $\bar{x} = (1/n)\sum_{i=1}^n x_i$. Give point estimates of b and σ, along with 95 percent confidence intervals for these. Relevant quantities may of course be computed from scratch, so to speak, but it is convenient to use e.g. R software, with

```
xxstar <- xxnew - mean(xxnew)
hello <- glm(yy ~ xxstar, family=gaussian)
ahat <- hello$coef[1]
bhat <- hello$coef[2]
resid <- hello$res
```

followed by $Q_0 = \sum_{i=1}^n \text{res}_i^2$ and $\widehat{\sigma} = \{Q_0/(n-2)\}^{1/2}$, etc. Writing `summary(hello)` gives a certain default output, from which a few central lines should be like these:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.14171    0.08331  -49.71   <2e-16 ***
xxstar       0.75393    0.02637   28.59   <2e-16 ***
```

(c) Suppose you encounter a new species, while exploring some African wilderness, with weight 111 kg. Estimate its brain size.

(d) For a mammal with weight x_0, and hence x factor $x = \log x_0$, consider the associated log-brain weight Y, which under model conditions may be expressed as $a+b(x-\bar{x})+\varepsilon$, with $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Using results from the book's Ch. 2, show that

$$Y - \widehat{a} - \widehat{b}(x - \bar{x}) \sim \mathrm{N}(0, \sigma^2 g(x)), \quad \text{with } v(x) = 1 + 1/n + (x - \bar{x})^2/M_n,$$

with $M_n = \sum_{i=1}^n (x_i - \bar{x})^2$. Show from this that

$$\frac{Y - \widehat{a} - \widehat{b}(x - \bar{x})}{\widehat{\sigma}\, v(x)^{1/2}} \sim t_{n-2}.$$

Then use this to show that we for each x have

$$\Pr\{Y \in \widehat{a} + \widehat{b}(x - \bar{x}) \pm t_0 \widehat{\sigma}\, v(x)^{1/2}\} = 0.99,$$

with t_0 the 0.995 quantile of the t_{n-2}. Construct a version of the plot given in the figure above.

(e) So, how special are You? Give a brief discussion, perhaps also pointing to yet other questions which could be raised.