STK 4900-9900 Spring 2024
Statistical Inference Theory: Oblig One

This is Oblig One, the first of two mandatory assignments for STK 4900-9900, Statistical Methods and Applications, Spring 2024. It is made available at the course website Saturday March 23, and the submission deadline is Monday April 8, 15:46, *via the Canvas system*. Reports may be written in nynorsk, bokmål, riksmål, English, or Latin, should preferably be text-processed (for instance with TeX, LaTeX, `word`), and must be submitted as a single pdf file. The submission must contain your name, the course, and assignment number.

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

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 two obligs 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:

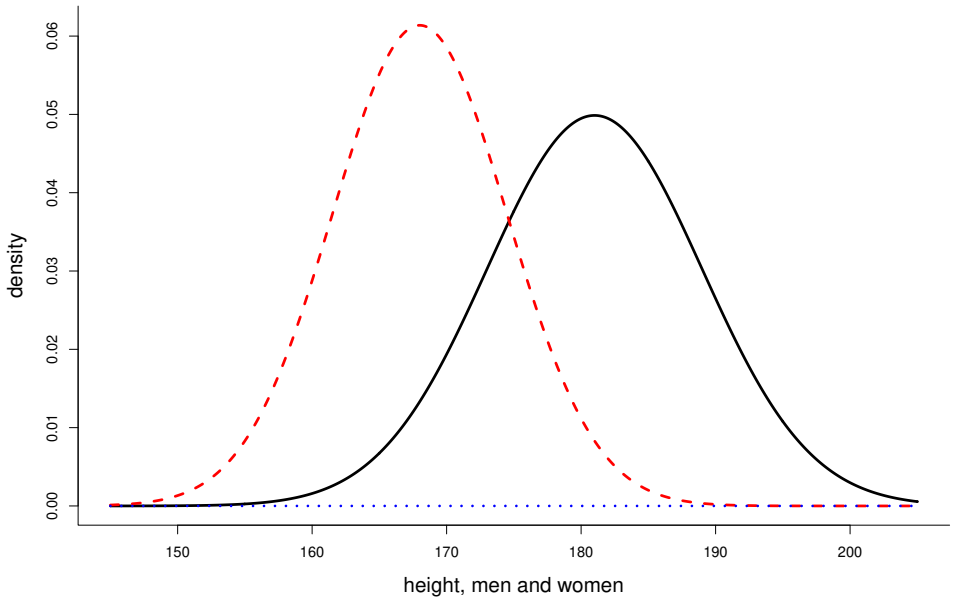
`www.uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html`

Enjoy [imperative pluralis].

Nils Lid Hjort

1. Men & women (and their heights)

If you aren't in over your head, how do you know how tall you are? We visit an island where the heights of men and women above the age of twenty are normally distributed. For the island in question, we take this to mean that with heights X_m and X_w randomly sampled from the men and women populations, then $X_m \sim \mathrm{N}(\mu_m, \sigma_m^2)$ for men and $X_w \sim \mathrm{N}(\mu_w, \sigma_w^2)$ for women, with $(\mu_m, \sigma_m) = (181.0, 8.0)$ and $(\mu_w, \sigma_w) = (168.0, 6.5)$ (the scale being in cm, of course). These two normal densities are shown in the figure below.



Densities $f_m(x_m)$ for men and $f_w(x_w)$ for women, both normal.

(a) Out of one million men, about how many are taller than 200 cm?

(b) Show that 95 percent of women have heights in the interval $[155.26, 180.74]$. What is the chance that the first ten women you meet all have heights in this interval?

(c) You meet a random pair, one man and one woman. What is the chance that the woman is taller than the man?

(d) I have met $n = 25$ people, all of the same gender, gotten their heights $x_1, \ldots, x_n$, and inform you that their average is $\bar{x} = 172.0$ cm. Test the null hypothesis that they are women. Explain your testing procedure, and compute the associated p-value.

2. Mothers & babies

Data from Berlin have shown a significant correlation between the increase in the stork population around the city and the increase in deliveries outside city hospitals (check 'New evidence for the theory of the stork', 2004, T. Höfer et al., *Paediatric and*

Perinatal Epidemiology). We shall not investigate those themes here, but concentrate on *the birthweights* and some potentially influential factors. Access the dataset `mothers-babies.txt`, for $n = 189$ US born babies and their mothers, available at the course website. It can be read into `R` as either

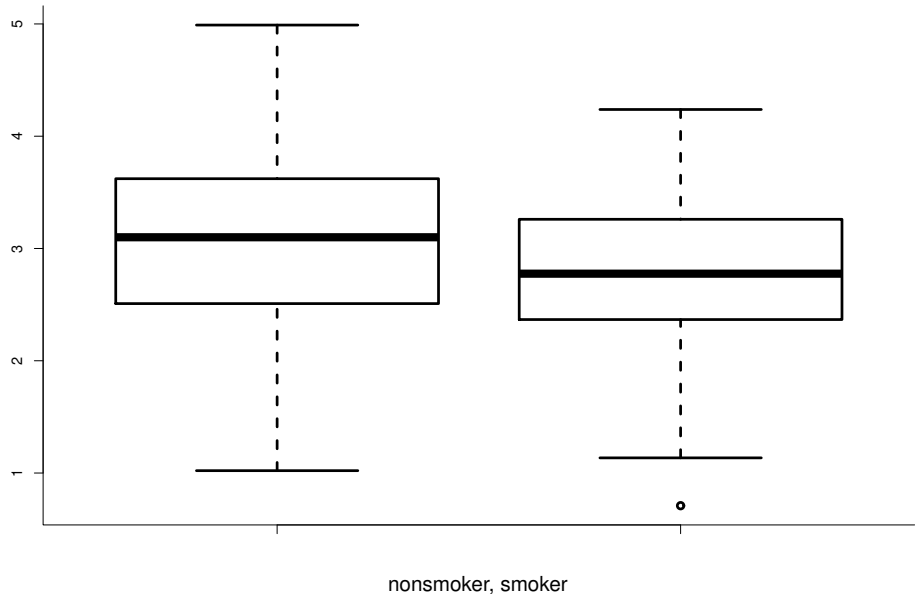
`babies = matrix(scan("mothersbabies.txt",skip=7),byrow=T,ncol=7)`

if you have put the file inside your own system, or with the longer url

`https://www.uio.no/studier/emner/matnat/math/STK4900/v24/mothersbabies.txt`

to get it directly from the website. This might then be followed by a version of these `R` lines:

```
y = babies[ ,1] # birthweight, in kg
x1 = babies[ ,2] # mother's weight prior to pregnancy
x2 = babies[ ,3] # age
x3 = babies[ ,4] # indicator for smoking (1 for yes, 0 for no)
x4 = babies[ ,5] # indicator for ethnic 1, black
x5 = babies[ ,6] # indicator for ethnic 2, neither white nor black
x6 = babies[ ,7] # indicator for ethnic 0, white
```



Boxplots for birthweights, for nonsmoking and smoking mothers.

(a) First ignore the other covariates, and work with the smoking and nonsmoking groups. Make boxplots, as in the figure, and comment briefly on what you learn from these. You may use these lines to sort the data into these two groups:

```
ysmoke = y[x3 == 1]
ynosmoke = y[x3 == 0]
```

(b) Then carry out a t-test for the hypothesis that there is no difference in the two mean parameters. Give also a 95 percent confidence interval for the mean difference $d = \mu_{\text{nonsmoke}} - \mu_{\text{smoke}}$. If you find time, do a test for the assumption that the variances for the two groups are the same. Comment on further underlying assumptions.

(c) Then concentrate on the three ethnic groups, creating datasets

```
yethnic0 = y[x6 == 1]
```

etc. Give a figure with boxplots for the three groups. Carry out an anova (analysis of variance) F test for the null hypothesis that there are no differences in birthweight mean parameters for the three groups. Formulate a conclusion, and again comment on the underlying assumptions for the analysis. – I suppose there are two ok options for carrying out the anova work in `R`: (i) use the `aov` command, but this requires some preliminary bureaucratic work, organising the data in the needed fashion; (ii) ignore the `aov`, but compute the required ingredients using the formulae from the lectures slides.

(d) Before turning to regressions, give a plot of age vs. weight. Compute the empirical correlation between these two, and find also a 95 percent confidence interval for this correlation. Comment on what you find.

(e) Fit the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

perhaps via `ok123 = lm(y ~ x1 + x2 + x3)` followed by `summary(ok123)`. Give 95 percent confidence intervals for $\beta_1, \beta_2, \beta_3$, and explain what you learn from this. Also comment briefly on the R-squared number from this regression.

(f) Then fit the fuller linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \varepsilon_i' \quad \text{for } i = 1, \ldots, n.$$

Comment on what is learned from this, and on whether any of the previous findings need to be modified.

(g) It appears from analyses above that smoking is a significant factor for the birthweight, in the greater US population. Now check whether this statistical association is about the same in the three ethnic groups. Do this by running linear regressions, of the type $y_i = \beta_0 + \beta_1 x_{i,3} + \varepsilon_i''$, inside each group. Comment on what you learn from this.

(h) Say hello to Mrs. Jones, who is fruktsommelig. She is white, age 25, weighing 60 kg before pregnancy, and is a smoker. Her cousin Mrs. Smith is also white, of the same age and weight, but she has never gudsigforbyde smoked in her life. Predict the weights of the two second cousins to come.