Universiteit Leiden
Mathematisch Instituut
Dr. A. Dürre

16.12.2024

Exam

Introduction Mathematical Statistics

Semster I 2024-2025

Family name: First name:

Student number:

Remarks:

- The exam consists of **6** tasks.
- You have **180** minutes to complete the exam.
- You can only use the provided short version of our lecture notes, the handbook of distributions an the R introduction.
- The solution must be documented well with calculations and if necessary references to theorems of our lecture. Giving just the solution is not sufficient.
- Whenever you need a certain quantile or probability try to express it using quantiles respectively the distribution function of standard distributions, like standard normal, t or F distribution.

Task	1	2	3	4	5	6	$\sum$
Points possible	13	11	9	8	5	6	52
Points achieved							

Grade exam: Grade homework: Final grade:

Task 1 (3 + 3 + 1 + 1 + 2 + 3)

In this task (and the whole exam) you can use that for independent and identically distributed random variables $X_1, \ldots, X_n$ with distribution function F :

$$F_{\min(X_1,\ldots,X_n)}(x) = 1 - [1 - F(x)]^n.$$

Let $X_1, \ldots, X_n$ be independent and identically distributed with density $f_a, a \in \mathbb{R}$ given by

$$f_a(x) = \begin{cases} e^{-(x-a)} & \text{if } x \geq a \\ 0 & \text{if } x < a. \end{cases}$$

a) Show that

$$\hat{a} := \min(X_1, \ldots, X_n)$$

is the maximum likelihood estimator of a.

b) Show that the distribution function of $\hat{a}$ equals

$$F_{\hat{a}}(x) = \begin{cases} 1 - e^{n(a-x)} & x \geq a \\ 0 & x < a. \end{cases}$$

We define $T := \hat{a} - a$.

c) Show that

$$F_T(x) = \begin{cases} 1 - e^{-nx} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

d) Show that

$$F_T^{-1}(p) = -\frac{\log(1-p)}{n} \quad \text{for } p \in (0,1).$$

e) Construct a two-sided $1 - \alpha$ confidence interval for a.
[You will not get points for the trivial confidence interval $(-\infty, \infty)$.]

f) Is $\hat{a}$ a (weakly) consistent estimator (for a)?

Task 2 (2+4+2+3)

Let $X_1, \ldots, X_n$ be independent and identically distributed with a discrete density given by

$$f_a(x) = \begin{cases} \frac{1}{3} - a & x = -1 \\ \frac{1}{3} & x = 0 \\ \frac{1}{3} + a & x = 1 \end{cases}$$

for $-\frac{1}{3} \leq a \leq \frac{1}{3}$.

a) Show that there exists a moment estimator defined by

$$\hat{a} = \frac{\overline{X}}{2}.$$

Note that this estimator can produce values smaller than $-\frac{1}{3}$ *or larger than* $\frac{1}{3}$. *We will not worry about this unfavourable behaviour during the exam.*

b) Compute the MSE of $\hat{a}$.

c) Show that

$$\sqrt{n}\sqrt{12}\left(\frac{\hat{a} - a}{\sqrt{2 - 12a^2}}\right) \xrightarrow{\mathcal{D}} N\left(0, 1\right).$$

d) Construct an two-sided asymptotic $1 - \alpha$ confidence interval for a.
[You will not get points for the trivial confidence interval $(-1/3, 1/3)$.*]*

Task 3 (4+3+2)

Let $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be independent and identically distributed random vectors in $\mathbb{R}^2$ following a multivariate normal distribution $N(\mu, \Sigma)$.

a) Show that $\begin{pmatrix} \overline{X} \\ \overline{Y} \end{pmatrix} \sim N(\mu, \frac{1}{n}\Sigma)$.

b) Show that

$$n\left[\begin{pmatrix} \overline{X} \\ \overline{Y} \end{pmatrix} - \mu\right]' \Sigma^{-1} \left[\begin{pmatrix} \overline{X} \\ \overline{Y} \end{pmatrix} - \mu\right] \sim \chi^2_2.$$

Now we want to test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

if Σ is known.

c) Construct a test of level α for this hypothesis.

Task 4 (3+3+2)

Let $Y = X\beta + \epsilon$ where $X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix}$ fulfils $\det(X'X) \neq 0$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with $\epsilon_1 \sim N(0, \sigma^2)$. Further we denote by $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ and $\hat{\sigma}^2$ the ML estimators of β and σ^2.

a) Show that $\hat{\beta}_1 - \hat{\beta}_2 \sim N(\beta_1 - \beta_2, Z)$ where $Z = \sigma^2(a - 2b + c)$ and a, b, c are defined by

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = (X'X)^{-1}.$$

b) Show that

$$\frac{\hat{\beta}_1 - \hat{\beta}_2 - (\beta_1 - \beta_2)}{\sqrt{\frac{n}{n-2}\hat{\sigma}^2(a - 2b + c)}} \sim t_{n-2}.$$

c) We define a test by

$$\Psi(X, Y) = \begin{cases} 1 & \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\frac{n}{n-2}\hat{\sigma}^2(a-2b+c)}} > F^{-1}_{t_{n-2}}(1-\alpha) \\ 0 & \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\frac{n}{n-2}\hat{\sigma}^2(a-2b+c)}} \leq F^{-1}_{t_{n-2}}(1-\alpha) \end{cases}$$

where $F^{-1}_{t_{n-2}}(1-\alpha)$ denotes the $1-\alpha$ quantile of a t-distribution with $n-2$ degrees of freedom. Show that this is a level α test for

$$H_0: \ \beta_1 \leq \beta_2 \text{ vs. } H_1: \ \beta_1 > \beta_2.$$

Task 5 (2 + 1 + 1 + 1)

The following R-function estimates a statistical model based on a data-vector `X` representing a sample $X_1, \dots, X_n$.

```
fu <- function(X) {
A <- function(Y) {
   return(sum(log(dweibull(X,Y[1],Y[2]))))
   }
B <- function(Y) return(-A(Y))
C <- optim(c(1,1),B)$par
D <- matrix(ncol=2,nrow=1000)
for(i in 1:1000) {
   F <- sample(X,replace=TRUE)
   G <- function(Y) {
      return(-sum(log(dweibull(F,Y[1],Y[2]))))
      }
   D[i,] <- optim(c(1,1),G)$par
   }
H <- sort(D[,1])[c(1+floor(1000*0.025),1000-floor(1000*0.025))]
G <- sort(D[,2])[c(1+floor(1000*0.025),1000-floor(1000*0.025))]
return(list(C,H,G))
}
```

a) Which statistical model is assumed?

b) What is represented by A?

c) What values are represented by C?

d) What values are represented by H?

Task 6 (3+3)

We want to check whether there is dependence between the average income of an inhabitant of a municipality and the voting result of the PVV at the last election. We have data from 207 municipalities.

Municipality	Average income in 1000 €	Voting share of PVV
Aalsmeer	31.3	25.1
Aalten	24.4	19.8
Achtkarspelen	22.0	34.5
⋮	⋮	⋮
Zwijndrecht	26.5	29.6
Zwolle	27.3	16.1

The following pages contain some analysis in R. Use the information for the following tasks.

a) Choose a suitable statistical model and justify your choice. Formulate a reasonable null- and alternative hypothesis.

b) Choose a suitable test for your statistical model using a significance level of $\alpha = 0.05$. Can you reject the null-hypothesis? What can you conclude?

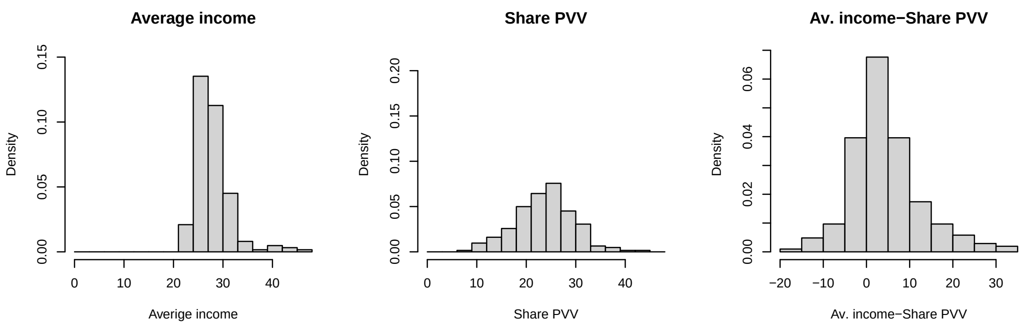

Figure 1: Histograms of the variables: average income (left), voting share PVV (middle), average income - voting share PVV (right).

```
> t.test(Income,PVV)

Welch Two Sample t-test

data:  Income and PVV
t = 9.02, df = 343.16, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.413664 5.317608
sample estimates:
mean of x mean of y
 28.00338  23.63775

#################################################################

> wilcox.test(Income,PVV)

Wilcoxon rank sum test with continuity correction

data:  Income and PVV
W = 31835, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

#################################################################

> t.test(Income-PVV)

One Sample t-test

data:  Income - PVV
t = 7.6069, df = 206, p-value = 9.833e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.234163 5.497109
sample estimates:
mean of x
 4.365636

#################################################################
```

```
> wilcox.test(Income-PVV)

Wilcoxon signed rank test with continuity correction

data:  Income - PVV
V = 16919, p-value = 9.838e-13
alternative hypothesis: true location is not equal to 0

##################################################################

> asympcor.test(Income,PVV)

Asymptotic correlation test

data:  Income and PVV
t = 6.5337 p-value = 6.42e-11
alternative hypothesis: true correlation is not equal to 0
sample estimates:
       cor
-0.4541117
```

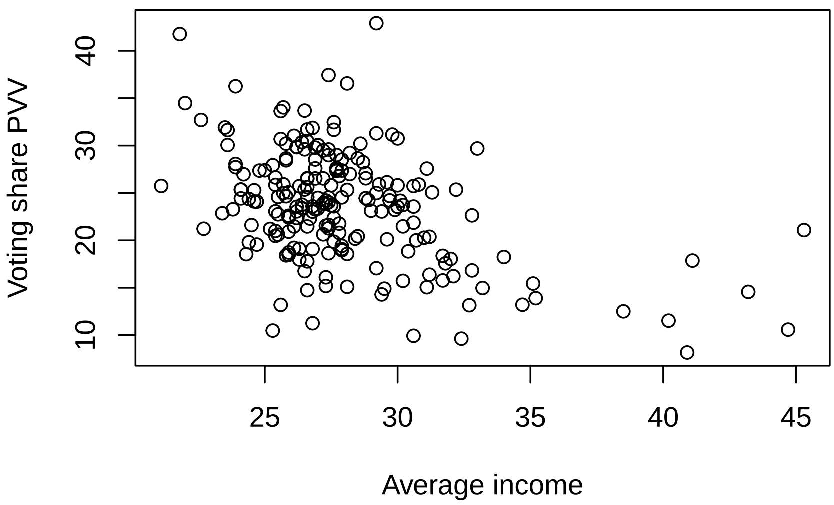

Figure 2: Scatter plot of the variables average income and voting share PVV

```
################################################################

> chisq.test(Income,PVV)

Pearson's Chi-squared test
```

```
data:  Income and PVV
X-squared = 19251, df = 19158, p-value = 0.3164

Warning:
In chisq.test(Income, PVV) : Chisquared approximation may be incorrect

######################################################################

> summary(lm(PVV~Income))

Call:
lm(formula = PVV ~ Income)

Residuals:
    Min      1Q  Median      3Q     Max
-15.145  -3.348  -0.301   3.647  20.154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.2310     2.8457  15.543  < 2e-16 ***
Income       -0.7354     0.1008  -7.298 6.31e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.292 on 205 degrees of freedom
Multiple R-squared:  0.2062,Adjusted R-squared:  0.2023
F-statistic: 53.26 on 1 and 205 DF,  p-value: 6.31e-12
```

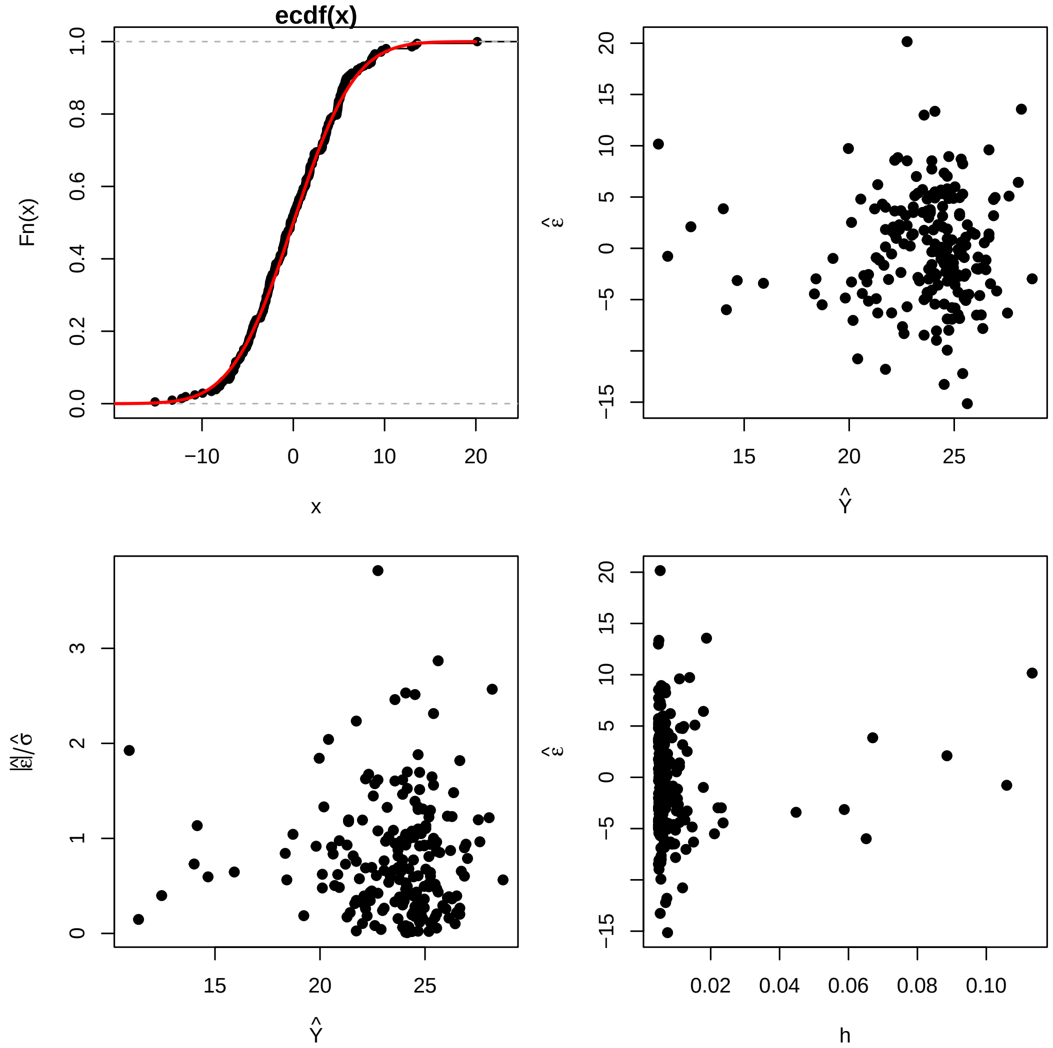

Figure 3: Model assessment for the linear model.