Universiteit Leiden
Mathematisch Instituut
Dr. A. Dürre

20.1.2025

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	15	9	8	5	6	56
Points achieved							

Grade exam: ____________ Grade homework: ____________ Final grade: ____________

Task 1 $(3 + 1 + 3 + 4 + 2)$
We define a continuously distributed random variable X by the density

$$f_a(x) = \frac{3}{4a^3}(a^2 - x^2)I[|x| \le a]$$

for $a > 0$. We observe a sample $X_1, \dots, X_n$ of independent random variables distributed as X.

a) Show that

$$E(X^2) = \frac{a^2}{5}.$$

b) Show that

$$\hat{a} := \sqrt{5\overline{X^2}}$$

where $\overline{X^2} := \frac{1}{n}\sum_{i=1}^n X_i^2$ is a moment estimator for a.

In the following you can use that $\mathbb{E}(X^4) = \frac{3}{35}a^4$.

c) Show that

$$\sqrt{n}\left(\overline{X^2} - \frac{a^2}{5}\right) \xrightarrow{\mathfrak{D}} N\left(0, \frac{8}{175}a^4\right).$$

d) Show that

$$\sqrt{n}\,(\hat{a} - a) \xrightarrow{\mathfrak{D}} N\left(0, \frac{2}{7}a^2\right).$$

e) Construct an asymptotic confidence interval of level $1 - \alpha$ for a.

Remark: You will not get points for the trivial interval $(0, \infty)$.

Task 2 $(8 + 3 + 4)$
We define a discrete random variable X by the density

$$f_b(x) = \frac{1-b}{1+b}b^{|x|}$$

for $x \in \mathbb{Z}, b \in (0, 1)$. We observe a sample $X_1, \dots, X_n$ of independent random variables distributed as X.

a) Show that

$$\hat{b} = -\frac{1}{\overline{|X|}} + \sqrt{\frac{1}{(\overline{|X|})^2} + 1}$$

is the maximum likelihood estimator of b where $\overline{|X|} := \frac{1}{n}\sum_{i=1}^n |X_i|$.

Remark: To simplify the exam we ignore the case $\overline{|X|} = 0$ *(the probability of this event goes to 0 anyway).*

b) Show that
$$\mathbb{E}(|X|) = \frac{2b}{1-b^2}.$$
Hint: You can use that a geometrically distributed random variable Y defined by the density:
$$f_Y(x) = (1-p)^x p \quad \text{for } x \in \mathbb{N}_0, p \in (0,1]$$
has expectation:
$$\mathbb{E}(Y) = \frac{1-p}{p}.$$
(The definition is slightly different to the one in the handbook of distributions.)

c) Show that $\hat{b}$ is a weakly consistent estimator for b.

Task 3 (3 + 2 + 4)
Let V, W be two independent random variables with $\mathbb{E}(V^2), \mathbb{E}(W^2) < \infty$.

a) Show that $\mathrm{Var}(V \cdot W) = [\mathbb{E}(V)]^2 \mathrm{Var}(W) + [\mathbb{E}(W)]^2 \mathrm{Var}(V) + \mathrm{Var}(V)\mathrm{Var}(W)$.

Let now $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be a sample of independent and identically distributed random vectors in $\mathbb{R}^2$. We assume that X_1 and Y_1 are continuously distributed with distribution functions F_X respectively F_Y. Furthermore the inverse functions of F_X and F_Y exist and are denoted by F_X^{-1} respectively F_Y^{-1}.

b) Show that $F_X(X_1)$ has a rectangular distribution (continuous uniform distribution) on the interval $[0,1]$.

Now we want to test whether X_1 and Y_1 are independent. We use the following test:
$$\Psi_\alpha(\mathbb{X}, \mathbb{Y}) = \begin{cases} 1 \\ 0 \end{cases} \text{if } \left| \sqrt{n} \frac{\frac{1}{n}\sum_{i=1}^n F_X(X_i)F_Y(Y_i) - \frac{1}{4}}{\sqrt{\frac{7}{144}}} \right| \begin{matrix} > \\ \leq \end{matrix} \Phi^{-1}(1-\alpha/2)$$

c) Show that under the given assumptions Ψ_α is an asymptotic level α test for
$$H_0: \ X_1, Y_1 \text{ are independent} \quad \text{vs.} \quad H_1: \ X_1, Y_1 \text{ are dependent}.$$
Hint: You can use that for a rectangular distribution Z on $[0,1]$*:*
$$\mathbb{E}(Z) = \frac{1}{2} \quad \text{and} \quad \mathrm{Var}(Z) = \frac{1}{12}.$$

Task 4 (3 + 3 + 2)
We observe a multiple linear regression model like in Definition 6.6. Let $\epsilon_1, \ldots, \epsilon_n$ be iid with $\mathbb{E}(\epsilon_1) = 0$ and $\mathrm{Var}(\epsilon_1) = \sigma^2$. Define then

$$Y = X\beta + \epsilon$$

where

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad X = \begin{pmatrix} x_{11} & \ldots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \ldots & v_{nk} \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Furthermore we assume $\det(X'X) \neq 0$. In contrast to most of the theorems of our lecture we do not assume that ϵ_1 is normally distributed. We define a linear estimator [linear in Y] for β by

$$\tilde{\beta} := [(X'X)^{-1}X' + D]Y = \hat{\beta} + DY$$

where $\hat{\beta}$ is the ML estimator under normality and $D \in \mathbb{R}^{k \times n}$.

a) Show that $\tilde{\beta}$ is unbiased if and only if $DX = 0$.

b) Show that for every unbiased linear estimator $\tilde{\beta}$:

$$\mathrm{Cov}(\tilde{\beta}) = \sigma^2(X'X)^{-1} + \sigma^2 DD'.$$

c) Show that

$$\mathrm{Var}(a'\tilde{\beta}) \geq \mathrm{Var}(a'\hat{\beta})$$

for $a \in \mathbb{R}^k$.
Hint: You can use that $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ *which follows from Theorem 6.8.*

Task 5 (2 + 1 + 1 + 1)
The following R-function estimates a statistical model based on two data-vectors `X` and `Y`.

```
fu <- function(X,Y) {
A <- length(X)
B <- sum(X^2)
C <- sum(X*Y)
D <- C/B
E <- Y-D*X
F <- sum(E^2)/A
G <- sqrt(A/(A-1)*F/B)*qt(0.975,df=A-1)
H <- c(D-G,D+G)
return(H)
}
```

a) Which statistical model is assumed?

b) What is represented by D?

c) What values are represented by E?

d) What values are represented by H?

Task 6 (3 + 3)
We want to investigate whether the running distance of a football team influences the number of created scoring chances. We use a data from the first round (3 games of the group phase) of the Euro 2024. The following table contains data of all 24 teams

Team	Chances	Distance	Team	Chances	Distance
Albania	33	347.48	Netherlands	39	324.36
Austria	31	344.13	Poland	36	338.83
Belgium	46	331.95	Portugal	52	342.80
Croatia	42	356.71	Romania	32	333.86
Czechia	43	351.01	Scotland	16	334.93
Denmark	42	336.47	Serbia	26	340.71
England	28	336.09	Slovakia	37	342.36
France	49	325.61	Slovenia	25	339.58
Georgia	27	338.10	Spain	48	353.15
Germany	57	344.43	Switzerland	30	340.95
Hungary	30	339.26	Türkiye	50	338.28
Italy	33	357.74	Ukraine	42	343.49

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?

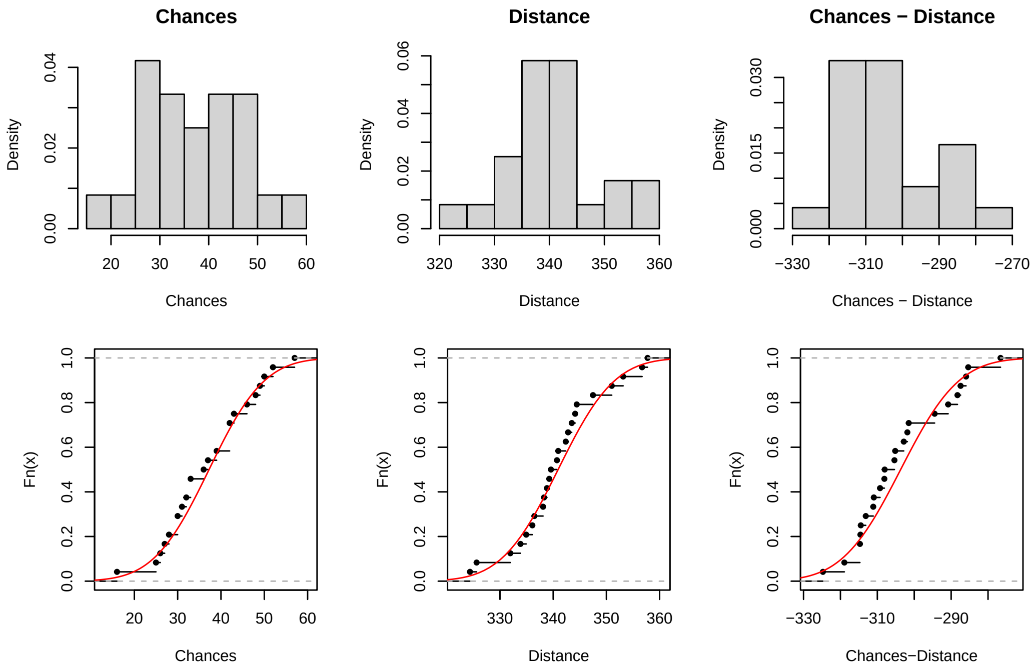

Figure 1: Histograms of the variables: Chances (left), Distance (middle), Chances-Distance (right) in the first row and empirical distribution functions with fitted normal distributions in red in the second row.

```
> t.test(Chances,Distance)

Welch Two Sample t-test

data:  Chances and Distance
t = -113.73, df = 44.473, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -309.0582 -298.2985
sample estimates:
mean of x mean of y
  37.2500  340.9283

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

> wilcox.test(Chances,Distance)

Wilcoxon rank sum test with continuity correction

data:  Chances and Distance
```

```
W = 0, p-value = 3.046e-09
alternative hypothesis: true location shift is not equal to 0

Warning: cannot compute exact p-value because of ties

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

> t.test(Chances-Distance)

One Sample t-test

data:  Chances - Distance
t = -119.41, df = 23, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -308.9391 -298.4176
sample estimates:
mean of x
-303.6783

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

> wilcox.test(Chances-Distance)
Wilcoxon signed rank test with continuity correction

data:  Chances - Distance
V = 0, p-value = 1.94e-05
alternative hypothesis: true location is not equal to 0

Warning: cannot compute exact p-value because of ties

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

> asympcor.test(Chances,Distance)

Asymptotic correlation test

data:  Chances and Distance
t = 0.4634654 p-value = 0.6430
alternative hypothesis: true correlation is not equal to 0
```

```
sample estimates:
       cor
0.09460448
```

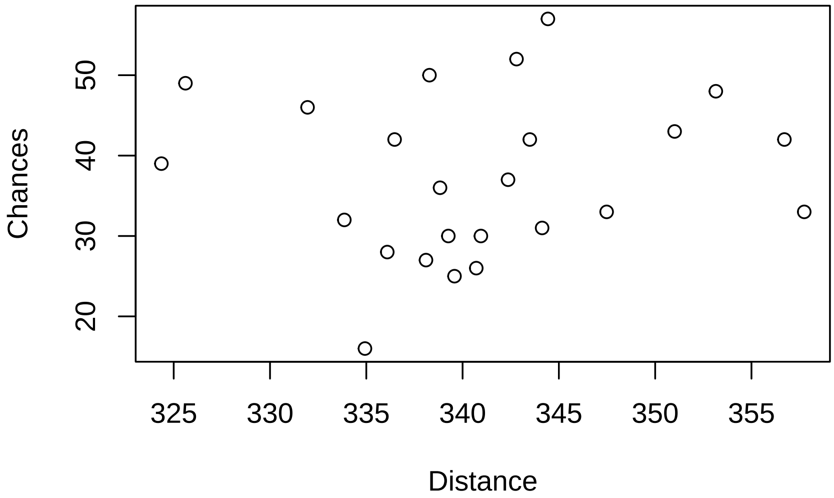

Figure 2: Scatter plot of the variables Distance and Chances

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

> chisq.test(Chances,Distance)

Pearson's Chi-squared test

data:  Chances and Distance
X-squared = 456, df = 437, p-value = 0.2559

Warning:
In chisq.test(Chances,Distance) : Chisquared approximation may be incorrect

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

> summary(lm(Chances~Distance))

Call:
lm(formula = Chances ~ Distance)

Residuals:
    Min      1Q  Median      3Q     Max
-20.566  -7.108  -0.712   6.283  19.350
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6523    87.3022  -0.019    0.985
Distance      0.1141     0.2560   0.446    0.660

Residual standard error: 10.25 on 22 degrees of freedom
Multiple R-squared:  0.00895,Adjusted R-squared:  -0.0361
F-statistic: 0.1987 on 1 and 22 DF,  p-value: 0.6601
```

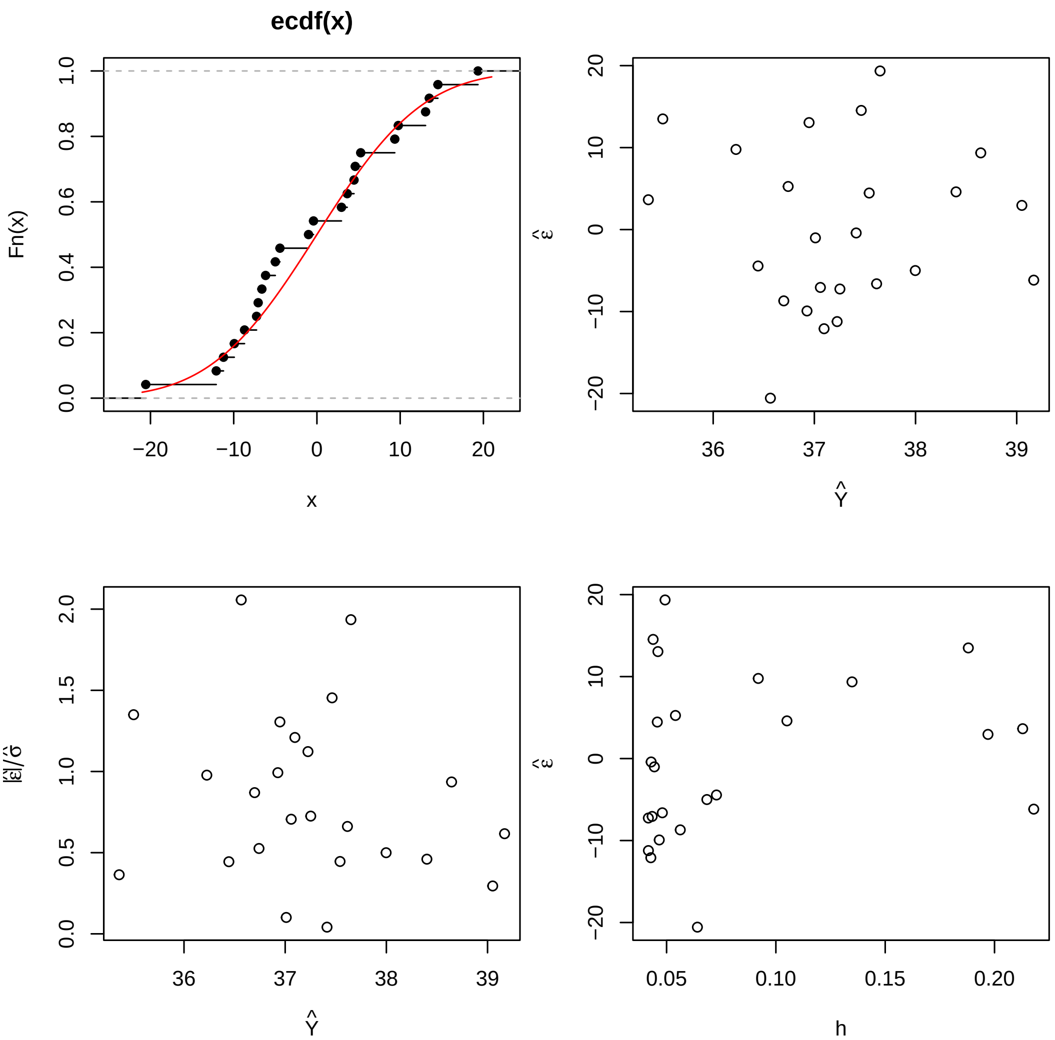

Figure 3: Model assessment for the linear model.