Universiteit Leiden
Mathematisch Instituut
Dr. A. Dürre

22.1.2024

Exam

Introduction Mathematical Statistics

Semster I 2023-2024

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 and the handbook of distributions.
- 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	15	8	13	5	9	6	56
Points achieved							

Grade exam: Grade homework: Final grade:

Task 1 (6 + 3 + 3 + 3)
Let $X_1, \ldots, X_n$ be independent and identically distributed with density $f_\lambda, \lambda > 0$ given by

$$f_\lambda(x) = \begin{cases} \frac{2x}{\lambda^2} e^{-(x/\lambda)^2} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

a) Show that

$$\hat{\lambda} := \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

is the maximum likelihood estimator of λ.

In the following you can use that $\mathbb{E}(X_1^2) = \lambda^2$ and $\mathbb{E}(X_1^4) = 2\lambda^4$.

b) Show that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \lambda^2\right) \xrightarrow{\mathcal{D}} N\left(0, \lambda^4\right).$$

c) Show that

$$\sqrt{n}\left(\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2} - \lambda}{\frac{1}{2}\lambda}\right) \xrightarrow{\mathcal{D}} N(0,1).$$

d) Construct an asymptotic two-sided $1-\alpha$ confidence interval for λ.

Task 2 (3 + 5)
Let $X_1, \ldots, X_n$ be independent and identically distributed with discrete density $f_{a,\lambda},\ a \in (0,1), \lambda > 0$ given by

$$f_{a,\lambda}(x) = \begin{cases} (1-a)\frac{\lambda^x}{x!} e^{-\lambda} + a \cdot I[x=0] & \text{if } x \in \mathbb{N} \\ 0 & \text{else} \end{cases}$$

a) Compute a moment estimator for λ if a is known.

b) Compute a moment estimator for $(a, \lambda)'$.

Task 3 (3 + 4 + 4 + 2)
Let $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ be iid random vectors with

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where $\rho \in (-1, 1)$.

a) Show that

$$\begin{pmatrix} X_1 + Y_1 \\ X_1 - Y_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2+2\rho & 0 \\ 0 & 2-2\rho \end{pmatrix} \right).$$

b) Show that

$$\hat{\rho} := \frac{\frac{\sum_{i=1}^n (X_i+Y_i)^2}{\sum_{i=1}^n (X_i-Y_i)^2} - 1}{\frac{\sum_{i=1}^n (X_i+Y_i)^2}{\sum_{i=1}^n (X_i-Y_i)^2} + 1}$$

is a (weakly) consistent estimator for ρ.

c) Show that

$$\frac{\sum_{i=1}^n (X_i + Y_i)^2}{\sum_{i=1}^n (X_i - Y_i)^2} \cdot \frac{1-\rho}{1+\rho} \sim F_{n,n}$$

where $F_{n,n}$ denotes the F distribution with n and n degrees of freedom.

d) Construct a test of level α for

$$H_0: \ \rho = \rho_0 \quad \text{vs.} \quad H_1: \ \rho \neq \rho_0.$$

Task 4 (2 + 1 + 1 + 1)

The following R-function performs a test based on a data vector `x` representing a sample $X_1, \dots, X_n$ from a normal distribution and a number `y`.

```
fu <- function(x,y) {
A <- length(x)
B <- x[1]
for (i in 2:A) B <- B+x[i]
B <- B/A
C <- (x[1]-B)^2
for (i in 2:A) C <- C+(x[i]-B)^2
C <- C/A
D <- qchisq(p=y,df=A-1)
if (A*C < D) return(1)
if (A*C >= D) return(0)
}
```

a) What hypothesis is tested by the function?

b) What value is represented by B?

c) What value is represented by C?

d) What value is represented by D?

Task 5 (5 + 4)
We consider the following linear model:

$$Y_i = x_i\beta + \epsilon_i,\ i = 1, \dots, n$$

where $\epsilon_1, \dots, \epsilon_n$ are independent and $\epsilon_i \sim N(0, \sigma_i^2)$ for $i = 1, \dots, n$ and $\sigma_1^2, \dots, \sigma_n^2 > 0$ known. Furthermore we assume that $(x_1, \dots, x_n)' \neq (0, \dots, 0)'$.

a) Show that the maximum likelihood estimator of β equals

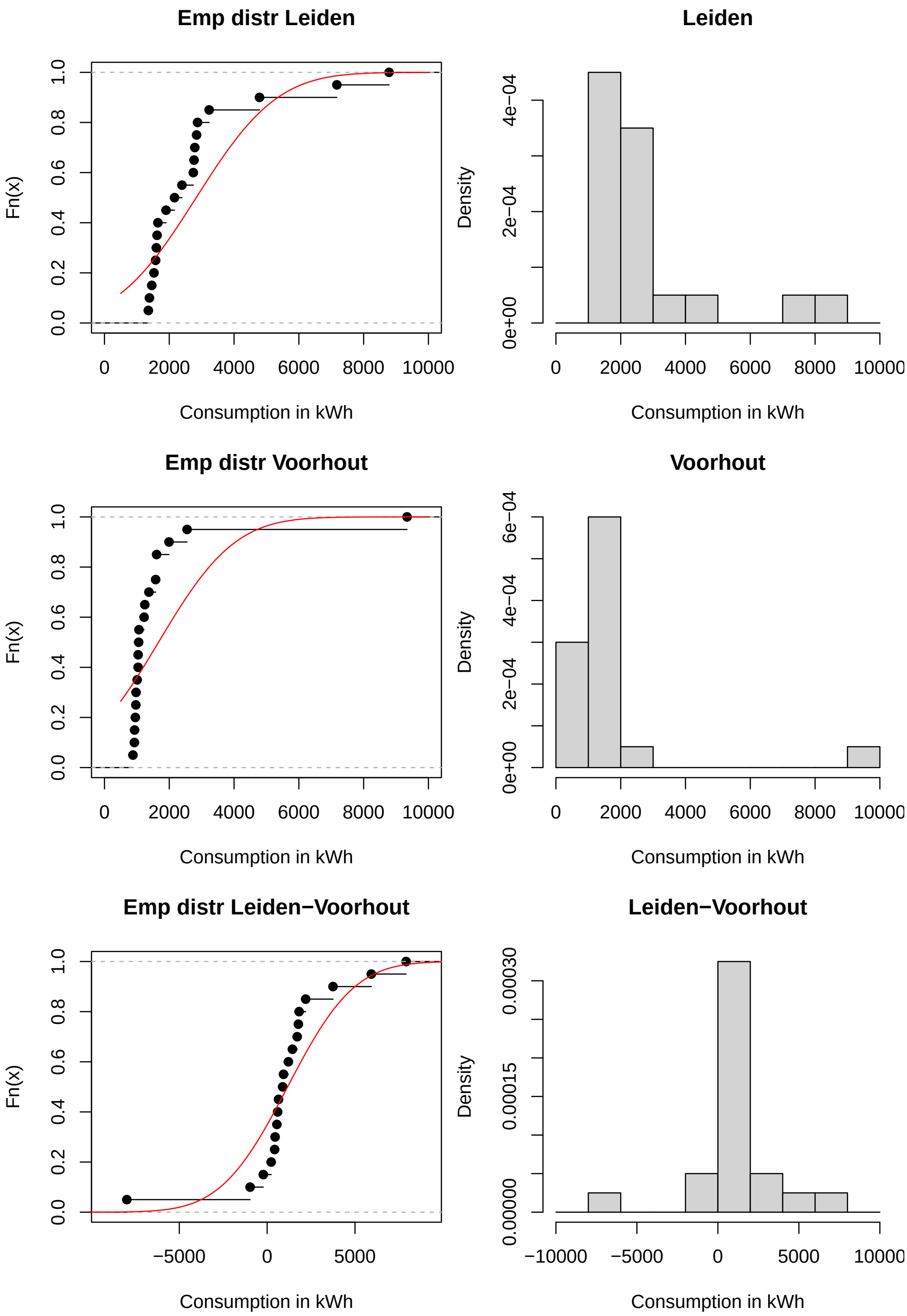
$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i/\sigma_i^2}{\sum_{i=1}^n x_i^2/\sigma_i^2}.$$

b) Derive the distribution of $\hat{\beta}$.

Task 6 (3 + 3)
We want to check whether the electricity consumption of households in rural areas and cities are different. Therefore we randomly choose 20 postcodes in Leiden and Voorhout and look at the yearly average electricity consumption of all electricity-consumers of that postcode in 2023 (`https://www.liander.nl/partners/datadiensten/open-data/data`).

Leiden		Voorhout	
Street	Consumption	Street	Consumption
Bachstraat 2324GW	1624	Baarslaan 2215XJ	975
Blauwe Tramstraat 2311ZZ	2765	Beukenrode 2215JE	1055
Boshuizerkade 2321TX	1901	Bizetstraat 2215SK	968
Busken Huetplein 2321SJ	2843	Boekenburglaan 2215AE	1065
Etta Palmstraat 2331MG	1579	Buxushaag 2216AD	2550
Haverstraat 2311NN	1601	Dirck Verhagenstraat	1372
Irenestraat 2316RJ	1354	Herenstraat 2215KJ	9340
Kraanbaan 2321PN	2788	Jacoba van Beierenweg 2215KV	1580
Maarsmansteeg 2311EE	8786	Julianalaan 2215HD	879
Menno ter Braakstraat 2321WJ	1649	Kruidenschans 2215BS	1222
Mirakelsteeg 2312WG	2162	Leeuwerikenlaan 2215NS	1608
Mozartstraat 2324XW	3231	Meidoornrode 2215LG	1038
Musschenbroekstraat 2316AW	4786	Mozartlaan 2215LV	1039
Paramaribohof 2315VX	1528	Munthof 2215VE	930
Parkzicht 2317RG	1388	Nachtegalenlaan 2215NR	1608
Sara Knipscheerstraat 2331SL	2873	Prins Bernhardstraat 2215AV	1992
Sophiastraat 2316PP	2742	Raadhuisplein 2215MA	924
Stadhuisplein 2311EJ	7174	Rembrandtlaan 2215CJ	1245
Wielmakersteeg 2311JZ	2392	Waterlelieweg 2215GR	953
Zijloeverpad 2315MK	1460	Waterlelieweg 2215GP	1009

The following plots show the empirical distribution functions (with fitted normal distribution in red) and histograms of the consumption in Leiden, Voorhout and row-wise differences of Leiden and Voorhout.



a) Choose a suitable statistical model and formulate a reasonable null- and alternative hypothesis.

You can find here the results of two sample tests for paired and independent samples.

```
> t.test(Leiden,Voorhout)

Welch Two Sample t-test

data:  Leiden and Voorhout
t = 1.9241, df = 37.868, p-value = 0.06188
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -60.7773 2388.1773
sample estimates:
mean of x mean of y
   2831.3    1667.6

> wilcox.test(Leiden,Voorhout)

Wilcoxon rank sum test with continuity correction

data:  Leiden and Voorhout
W = 342, p-value = 0.0001293
alternative hypothesis: true location shift is not equal to 0

> t.test(Leiden-Voorhout)

One Sample t-test

data:  Leiden - Voorhout
t = 1.7383, df = 19, p-value = 0.09834
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -237.4561 2564.8561
sample estimates:
mean of x
   1163.7

> wilcox.test(Leiden-Voorhout)

Wilcoxon signed rank exact test

data:  Leiden - Voorhout
V = 179, p-value = 0.004221
alternative hypothesis: true location is not equal to 0
```

b) Choose a test which is suitable for your statistical model using a significance level of $\alpha = 0.05$. What can you conclude?