Vrije Universiteit Amsterdam	Statistical Data Analysis, Resit Exam
Faculty of Sciences	July 8, 2022

Use of a basic calculator is allowed. Graphical calculators, laptops, e-readers, mobile phones, smartphones, smartwatches etc. are not allowed. This exam consists of 5 questions on 5 pages (45 points). You have 2 hours and 45 minutes to write the exam.

Please write all answers in English. Grade = $\frac{total+5}{5}$**.** **GOOD LUCK!**

Question 1 [2+2+2+2=8 points]
In Figure 1 the histogram, boxplot and QQ-plots against the standard normal, standard Laplace, chisquared (with 4 degreees of freedom), lognormal (with `sdlog` parameter 0.55) distributions are shown for a data set `A`.

a. Describe briefly what these graphical summaries tell you about the underlying distribution of the data set. Consider at least two aspects, e.g., shape, extreme values, etc.

b. Which of the four location-scale families indicated by the QQ-plots do you think is most appropriate for these data? Explain your answer.

c. Briefly describe how you would determine the location a and scale b for an appropriate location-scale family by using either the QQ-plot or the sample mean and the sample standard deviation. *Note: You do not actually need to determine a and b.*

d. Answer the following questions, and motivate your answers:
i) Is the sign test nonparametric? ii) Is it safe to apply the sign test to `A`?

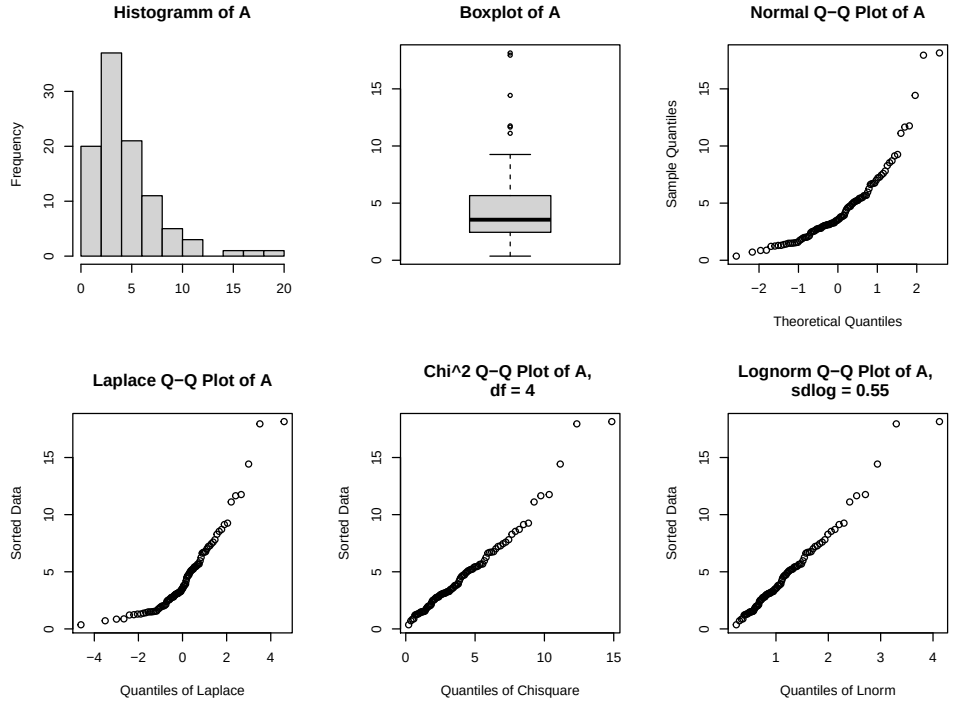

Figure 1: Histogram, boxplot and QQ-plots against indicated distributions of data set `A`.

Question 2 [3+3+2+2=10 points] Elaborate on and answer the following questions.

a. Which condition should be met so that the chi-square goodness-of-fit test can be considered reliable in terms of the type-I-error control? (Also provide a formula for the condition.)

b. Mention two testing procedures that can be used to test for independence of paired data, and give at least one reason why one might have an advantage over the other.

c. Which statement(s) about normal distributions can be tested with the help of the Shapiro–Wilk test?

d. For a sample B from $\mathcal{N}(\mu, \sigma^2)$ and an independent sample C from $\mathcal{N}(\nu, \sigma^2)$ for some unknown $\sigma^2 > 0$, explain which of the t-test and the Wilcoxon test is more powerful for finding a difference between the location parameters, i.e. testing $H_0 : \mu = \nu$ against $H_a : \mu \neq \nu$.

Question 3 [3+2+2=7 points]
Let $Y_1, \ldots, Y_{40}$ be independent and identically distributed random variables with unknown distribution P. The 20%-trimmed mean $T_{40,0.2}$ is used to estimate the location of P. To determine the accuracy of this estimator, its variance is estimated by means of the empirical bootstrap.

a. Describe the steps of the empirical bootstrap scheme that you would use to find the bootstrap estimate of the variance of $T_{40,0.2}$.

b. Consider the data set `D` presented in Figure 2. Empirical bootstrap values for the 20%-trimmed mean of this data set were computed and some quantiles of these bootstrap values of this location estimator are:

quantile	0.025	0.05	0.5	0.95	0.975
20%-trimmed mean	0.42	0.46	0.64	0.92	0.99

The 20%-trimmed sample mean of `D` equals 0.66. Determine the two-sided bootstrap confidence interval with confidence level 95% for the 20%-trimmed mean.

c. Explain whether the 30%-trimmed sample mean is smaller than, approximately equal to, or greater than the 20%-trimmed sample mean, for the data set `D` presented in Figure 2.

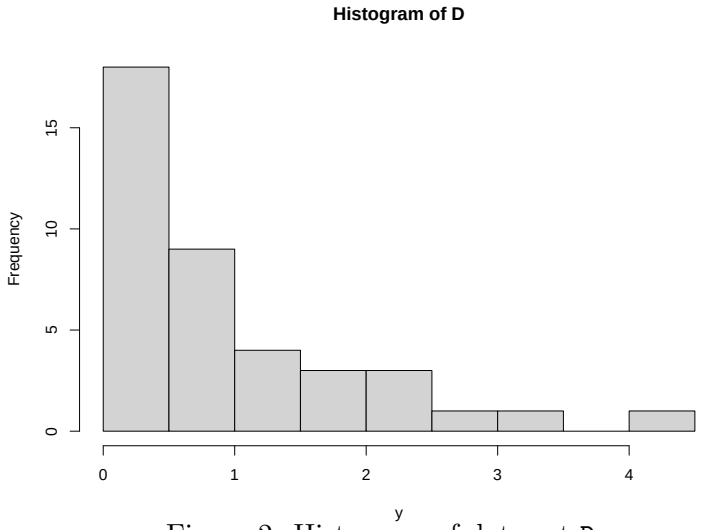

Figure 2: Histogram of data set `D`.

Question 4 [2+2+1+5=10 points]

Children of an elementary school class (4th grade) in the U.S. state Michigan participated in a study about whether shoe manufacturers should make girls' shoes as wide as boys' shoes (with the same length). Several measurements of the 19 girls and 20 boys were taken; in addition to (the explanatory variable) `sex` (boy=0, girl=1), we will focus on:

- `width`, i.e. the widest part (cm) of the longest of the two feet, as the response variable;
- `length`, i.e. the length of the longest of the two feet, as a possible explanatory variable;
- `age` (years), as a potential explanatory variable.

Figure 3 below displays an added variable plot for `age` in the linear regression model that only includes `sex` and `length` as explanatory variables, next to the intercept.

a. Based on Figure 3, discuss whether `age` should also be included in the linear model.

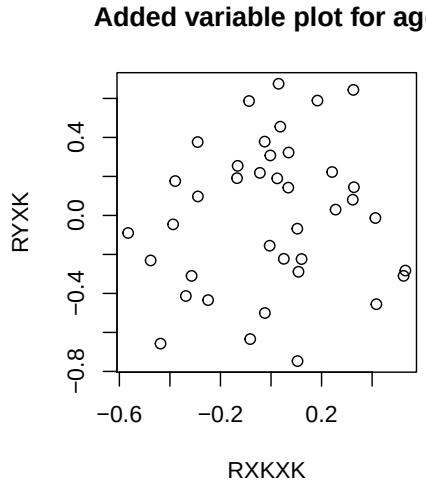

Figure 3: Added variable plot for `age` in the model including the intercept, `sex`, and `length`, i.e. the residuals of the linear model for `age` regressed on the intercept, `sex`, and `length` (x-axis) against the residuals of the linear model for `width` regressed on the intercept, `sex`, and `length` (y-axis)

For arbitrary reasons, we will from now on focus on the full model

$$\texttt{width} = \beta_0 + \beta_1 \cdot \texttt{age} + \beta_2 \cdot \texttt{length} + \beta_3 \cdot \texttt{sex} + error, \quad error \sim N(0, \sigma^2) \tag{1}$$

The following is part of the summary output for the linear model in R (with rounded numbers):

```
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.04       2.07    1.47     0.15
age          0.14       0.23    0.61     0.55
length       0.21       0.05    3.91   <0.001
sex         -0.23       0.13   -1.73     0.09
---
Residual standard error: 0.39 on 35 degrees of freedom
Multiple R-squared:  0.47,Adjusted R-squared:  0.42
F-statistic: 10.15 on 3 and 35 DF,  p-value: 5.9e-05
```

(Question 4 with the parts b., c., d. continues on the next page.)

b. Conduct an overall test at level $\alpha = 5\%$ to check whether at least one explanatory variable should be included in the model. (Do not forget to state the hypotheses to be tested.)

c. Explain the interpretation of the `Multiple R-squared` part of the above R output.

We now wish to investigate possible collinearity problems. Figure 4 displays two scatterplots (separately for girls and boys) of the `age` and `length` variables. The corresponding sample correlations within the subsamples of girls and boys are 0.12 and 0.55, respectively. In addition, we obtained the following (rounded) variance inflation factors (VIFs),

```
varianceinflation(cbind(age, length, sex))
[1]  1.18  1.26  1.11
```

and the following (rounded) variance decomposition,

```
vardecomposition(cbind(age, length, sex))
     conditionindices 0    1    2    3
[1,]             1.00 0 0.00 0.00 0.00
[2,]            44.91 0 0.07 0.27 0.48
[3,]            58.54 0 0.24 0.68 0.45
[4,]           826.21 1 0.69 0.05 0.07
```

After centering the variables `age` and `length` at their respective means and rescaling the length of these vectors to 1, we obtained the following variance decomposition,

```
     conditionindices    0    1   2    3
[1,]             1.00 0.25 0.00 0.0 0.05
[2,]             2.57 0.59 0.00 0.0 0.82
[3,]             6.52 0.04 0.34 0.3 0.11
[4,]             9.38 0.13 0.66 0.7 0.02
```

d. Discuss whether there are any collinearity problems in model (1) by using all available information. Motivate your findings and arguments with particular care.

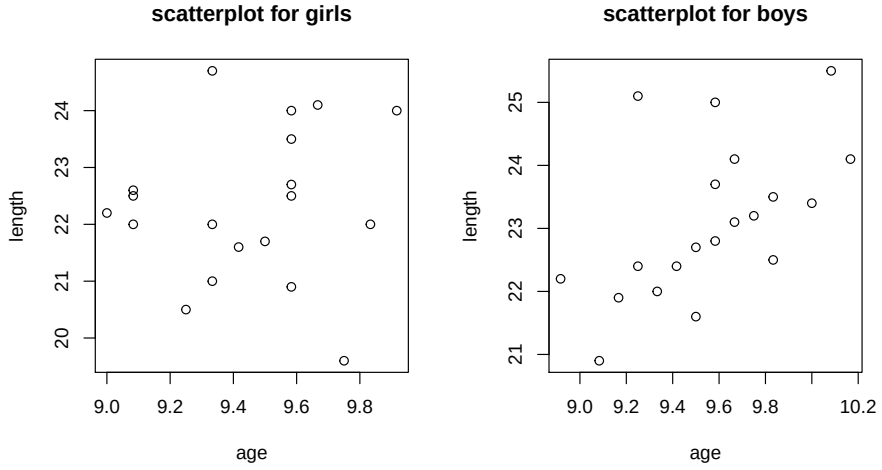

Figure 4: Scatterplots of the `age` against `length` measurements for girls (left) and boys (right).

Question 5 [2+2+2+4=10 points]
For the study described in Question 4 above, also the handedness of the elementary school pupils were taken. Based on this dataset, we could thus also test whether gender and handedness are independent. The following table summarizes the relevant categorical data:

Handedness \ Sex	Boy	Girl	total
Left	5	3	8
Right	15	16	31
total	20	19	39

Table 1: Handedness counts among boys and girls

a. Formulate a suitable model of multinomial distribution(s) and state the corresponding null and alternative hypotheses, H_0 and H_a, respectively, for investigating whether there is a relationship between gender and handedness.
You may formulate your hypotheses either in words or in formulas.

b. State the formula of the chi-squared test statistic that could be used to test the hypotheses from part a.; describe all symbols you are using for this.

c. Describe the approximate distribution of the chi-squared test statistic under the null hypothesis (assuming that a certain rule of thumb is satisfied).
You don't need to state the rule of thumb.

Instead of the chi-squared test, one could also use the second largest cell entry of Table 1 to test the hypotheses from part a.

d. Use the second largest cell entry as the test statistic T of a right-tailed test to test H_0 against H_a at level $\alpha = 5\%$. For this, use the following $B = 100$ bootstrap realizations of the test statistic which have been obtained by means of a valid bootstrap procedure for contingency tables:

```
 [1] 12 13 13 13 13 13 13 13 13 13 13 13 13 13 13 14 14 14 14 14
[21] 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 15
[41] 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
[61] 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
[81] 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
```

(R output of $B = 100$ bootstrap realizations of the statistic T for the present dataset.)