Vrije Universiteit Amsterdam	Statistical Data Analysis, Resit exam
Faculty of Sciences	5 July 2019

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 6 pages (45 points).
In the Appendix R output for simple linear regression models can be found.
You have 2 hours 45 minutes to write the exam.

Please write all answers in English. Motivate your answers. Grade $= \frac{total+5}{5}$**.**

GOOD LUCK!

Question 1 [10 points]
Indicate for each of the following statements whether it is correct or incorrect.
Irrespective of “correct” or “incorrect”, motivate your answers shortly!

a. [2 points] The 10% trimmed mean ignores the 10% most extreme observations.

b. [2 points] Maximum likelihood estimators are special cases of M-estimators.

c. [2 points] The χ^2-goodness-of-fit test based on 5 bins is reliable for any sample size.

d. [2 points] In some situations, the signed rank test is applicable when the t-test is not.

e. [2 points] Fisher’s exact test uses that the test statistic has a Poisson distribution under the null hypothesis.

Question 2 [8.5 points]

A histogram of a dataset `x` and several QQ-plots are displayed in Figure 1.

a. [1 point] Based on the plots in Figure 1, which location-scale family do you think is the most appropriate for these data? Motivate your answer.

b. [2 points] With regard to the location-scale family that you chose in part a., determine the location and scale parameters with respect to the corresponding reference distribution given in the table on the right. Use that the sample mean, standard deviation, and variance are $\bar{\mathtt{x}} = 77.0, \hat{\sigma} = 83.9$, and $\hat{\sigma}^2 = 7046$, respectively.

distribution	expectation	variance
standard normal	0	1
uniform(0,1)	0.5	$\frac{1}{12}$
standard lognormal	1.65	4.67
χ^2_4	4	8
χ^2_8	8	16

c. [3 points] We are interested in the statistic $T(z) = median(z) - \bar{z}$ which can be used as a measure of skewness of a dataset z. Here, $median(z)$ is the sample median of z. Propose a suitable bootstrap method to approximate the distribution of T and briefly describe all steps that are made in that method.

d. [1 point] The statistic T applied to our dataset `x` gave a value of $T(\mathtt{x}) = -29.86$. Explain whether this indicates that `x` is rather left-skew, right-skew or symmetric?

e. [1.5 points] Discuss and conclude whether it is safe to apply the sign test to `x` without bootstrapping.

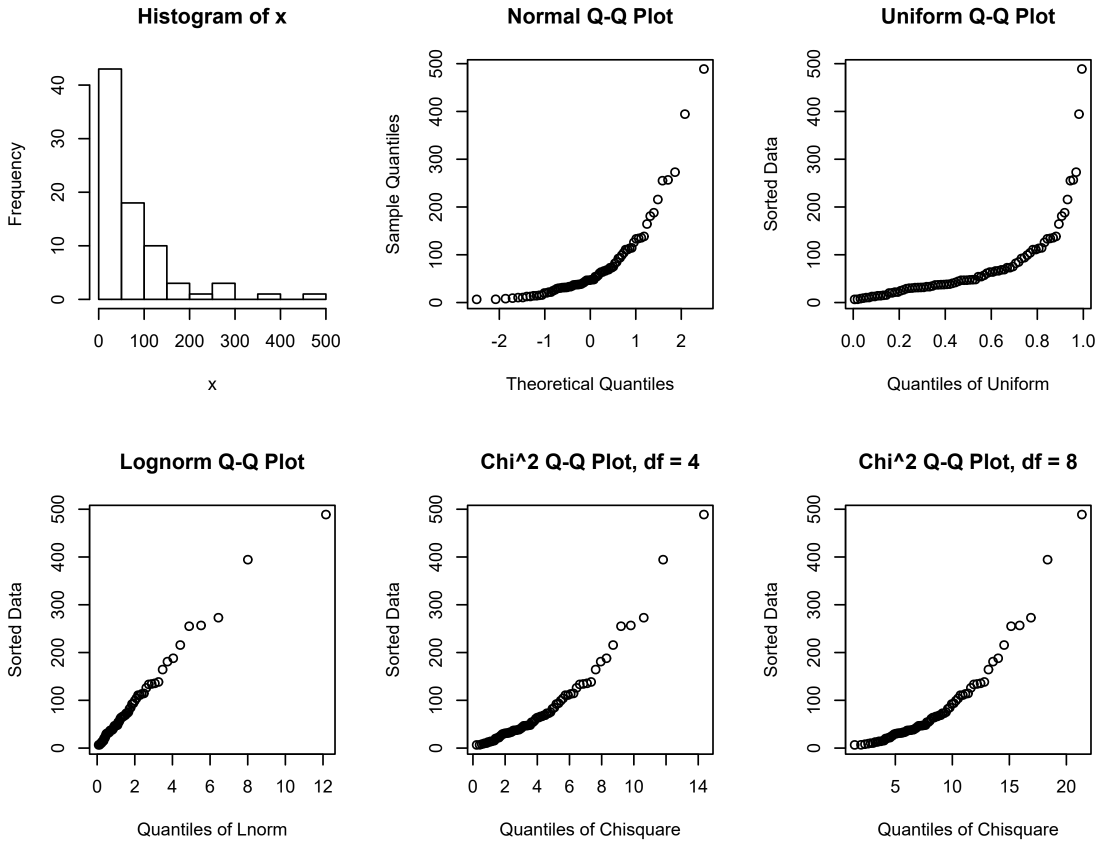

Figure 1: Histogram and QQ-plots against indicated distributions for a sample `x` of size $n = 70$.

Question 3 [10 points]

a. [1 point] For independent, real-valued random variables $X_1, \dots, X_n$, $n \in \mathbb{N}$, with continuous cumulative distribution function F, mean $\mu \in \mathbb{R}$, and variance $\sigma^2 > 0$, state the null and alternative hypotheses that can be tested with the classical Kolmorogov-Smirnov test.

b. [1.5 points] The one-sample Kolmogorov-Smirnov test is an example of a nonparametric test. Describe in your own words in general what a nonparametric test is and what the benefits of these tests are.

c. [3 points] The two-sample Kolmogorov-Smirnov test can be applied to two independent sets of random variables $X_1, \dots, X_n$ with distribution F and $Y_1, \dots, Y_m$ with distribution G. It uses the test statistic $D_{n,m} = \sup_{-\infty \leq z \leq \infty} |\hat{F}_n(z) - \hat{G}_m(z)|$, where $\hat{F}_n$ and $\hat{G}_m$ are the empirical distribution functions of $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$, respectively. This test can be conducted as a permutation test. Describe all steps that are made in the permutation procedure to find critical values for this test.

d. [2 points] If the data come in pairs, i.e. in case of independent and identically distributed pairs of random variables $(X_1, Y_1), \dots, (X_n, Y_n)$, Spearman's correlation test is applicable. State the null and alternative hypotheses that can be tested with Spearman's test and describe which information of the data is used by Spearman's test.

e. [1 point] Discuss the advantage of Spearman's test over Pearson's correlation test which uses the (Pearson) sample correlation of $(X_1, Y_1), \dots, (X_n, Y_n)$ as a test statistic.

f. [1.5 points] Spearman's correlation test uses less information from the data than Pearson's correlation test. This is why it is possible that Spearman's test is less efficient. Briefly explain in your own words and in terms of sample sizes what it means if the asymptotic relative efficiency of Spearman's test with respect to Pearson's test equals 0.5.

Question 4 [7 points]

Suppose that, before the 2019 European elections, we have asked randomly chosen EU-citizens of different countries about their voting preferences for the elections: 48 from Germany, 37 from the Netherlands, and 57 from France. The aim of the survey was to analyze differences in the voting preferences among these countries. The following table shows the results for the following groups of political parties in detail: EPP, S&P, RE, and Green-EFA; preferences for other parties are displayed as "others":

group / country	Germany	Netherlands	France	total
EPP	15	6	6	27
S&D	8	9	4	21
ALDE	4	8	16	28
Greens-EFA	11	5	10	26
others	10	9	21	40
total	48	37	57	142

We wish to do the analysis with the help of a chi-square test which uses the test statistic $X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij}-n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$. Throughout this question we choose the significance level $\alpha = 5\%$.

a. [1.5 points] Formulate a suitable model (II-A, II-B, or II-C) of multinomial distribution(s) and state the corresponding null and alternative hypotheses for the intended investigation. *You may formulate your hypotheses in words or in formulas.*

b. [2 points] Under the null hypothesis, the test statistic approximately follows a chi-square-distribution if a certain rule of thumb is satisfied. Explain what this rule of thumb is, and check whether it is satisfied for the present data.

c. [2 points] Irrespective of your conclusion in part b., test the hypotheses from part a. with the help of the chi-square test. Use that the value of the test statistic is $X^2 \approx 19.72$ for the present data, and use a suitable critical value from Table 1. What is your conclusion?

	γ						
k	0.025	0.050	0.330	0.500	0.670	0.950	0.975
1	0.00	0.00	0.18	0.45	0.95	3.84	5.02
2	0.05	0.10	0.80	1.39	2.22	5.99	7.38
3	0.22	0.35	1.55	2.37	3.43	7.81	9.35
4	0.48	0.71	2.36	3.36	4.61	9.49	11.14
5	0.83	1.15	3.19	4.35	5.76	11.07	12.83
6	1.24	1.64	4.05	5.35	6.90	12.59	14.45
7	1.69	2.17	4.92	6.35	8.03	14.07	16.01
8	2.18	2.73	5.80	7.34	9.15	15.51	17.53
9	2.70	3.33	6.68	8.34	10.26	16.92	19.02

Table 1: γ-quantiles of χ_k^2-distribution for indicated values of γ and k.

d. [1.5 points] In cases where the chi-square approximation for the distribution of X^2 under the null hypothesis is not appropriate, a certain bootstrap for contingency tables can be applied. Explain the main idea behind this bootstrap approach.

Question 5 [9.5 points]

We wish to analyse the dataset `poverty` which was collected from the U.N.E.S.C.O. in 1990 and which contains data about $n = 27$ African states. The dataset contains the following measurements for each state:

`lifeexpF` (the life expectancy at birth for females),

`livebirth` (the live birth rate per 1,000 of the population),

`infantdeath` (the number of deaths of under 1 year old children per 1,000 of the population),

`GNP` (the state's Gross National Product per capita in U.S. dollars).

For these data we assume a multiple linear regression model where the response variable is `lifeexpF`, and the other variables are the possible explanatory variables.

Throughout this question, we choose the significance level $\alpha = 5\%$.

a. [2.5 points] Do the first step in the step up procedure to decide which explanatory variable should be added to the empty model. Write down the resulting linear model, including the model assumptions.
Note: the Appendix below Figure 2 contains the R output of all simple linear models.

b. [1 point] The variance inflation numbers for the parameter estimates $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ in the full model are 1.457, 1.467, and 1.263, respectively. Based on these numbers, discuss whether there is a collinearity problem with the model.

c. [1 point] The condition indices in the full model, i.e.

$$\texttt{lifeexpF} = \beta_0 + \beta_1\texttt{livebirth} + \beta_2\texttt{infantdeath} + \beta_3\texttt{GNP} + \texttt{error},$$

and the variance decomposition proportions of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ are given below. Use these to identify all collinearities, i.e. for each collinearity mention the group of the involved explanatory variables.
(*Note: the columns of the design matrix have been rescaled to the same length.*)

```
conditionindices intercept livebirth infantdeath   GNP
           1.000     0.001     0.001       0.005 0.022
           2.362     0.000     0.001       0.010 0.679
           9.388     0.075     0.037       0.919 0.182
          23.793     0.924     0.962       0.067 0.118
```

d. [2 points] Explain which model assumptions and/or questions regarding the variable selection can be checked with plots like the ones in Figure 2 (on the next page), i.e. with an added variable plot and with a scatterplot of a variable *included in* the model against the residuals.

e. [2 points] Based on the characteristics of the plots in Figure 2, which concrete conclusions do you draw about the linear regression model `lifeexpF =` $\beta_0 + \beta_1$ `livebirth +` β_3 `GNP + error`?

f. [1 point] Briefly describe how it can be tested whether the first observation in the dataset has an outlying value in the response variable.

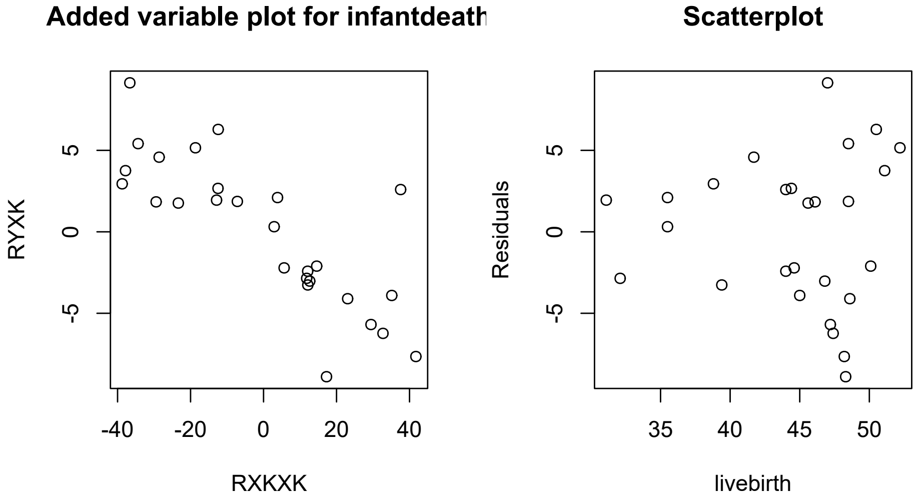

Figure 2: Added variable plot for the variable `infantdeath` (left) and plot of the `livebirth` values against the residuals (right) in the linear model `lifeexpF` = $\beta_0 + \beta_1$ `livebirth` + β_3 `GNP + error`.

Appendix

```
summary(lm(lifeexpF ~ livebirth))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  91.7687     8.1097  11.316 2.50e-11
livebirth    -0.8451     0.1807  -4.676 8.62e-05
---
Residual standard error: 5.24 on 25 degrees of freedom
Multiple R-squared:  0.4666,Adjusted R-squared:  0.4453


summary(lm(lifeexpF ~ infantdeath))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.84803    2.45758  30.049  < 2e-16
infantdeath -0.19750    0.02358  -8.374 1.01e-08
---
Residual standard error: 3.678 on 25 degrees of freedom
Multiple R-squared:  0.7372,Adjusted R-squared:  0.7267


summary(lm(lifeexpF ~ GNP))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.978382   1.432023  35.599   <2e-16
GNP          0.003707   0.001040   3.564   0.0015
---
Residual standard error: 5.842 on 25 degrees of freedom
Multiple R-squared:  0.3369,Adjusted R-squared:  0.3104
```