Vrije Universiteit Amsterdam	Statistical Data Analysis, Resit exam
Faculty of Sciences	7 July 2017

Use of a basic calculator is allowed. Graphical calculators, laptops, e-readers, mobile phones, smartphones, smartwatches etc. are not allowed. This exam consists of 6 questions on 4 pages (45 points). You have 2 hours 45 minutes to write the exam.

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

GOOD LUCK!

Question 1 [2+2+2=6 points]
In Figure 1 the histogram, boxplot and QQ-plots against the standard Laplace, standard Cauchy, standard normal, and Uniform$[0, 1]$ distributions are shown for a data set `x`.

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 do you think is most appropriate for these data? Explain your answer.

c. Briefly describe how you would approximately determine the location a and scale b using the QQ-plot that you have selected under part b.
You do not have to determine a and b.

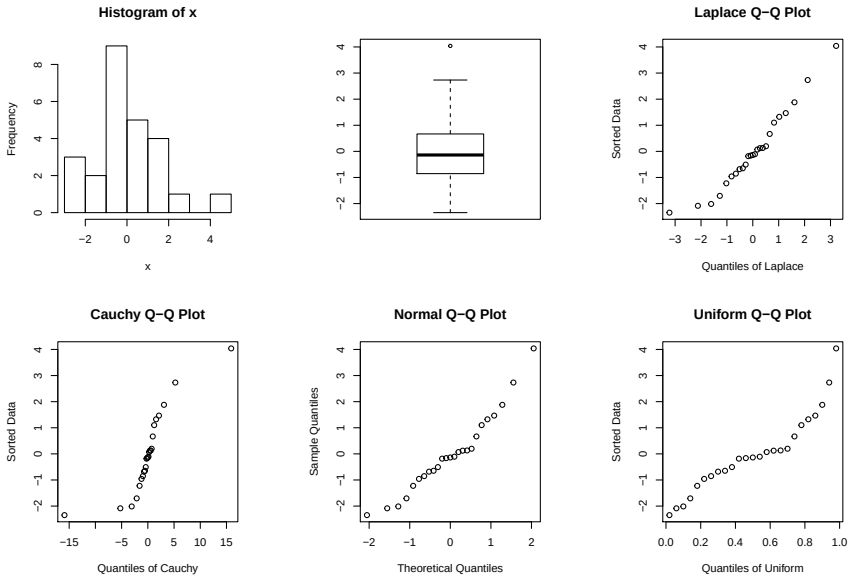

Figure 1: Histogram, boxplot and QQ-plots against indicated distributions of data set `x`.

Question 2 [2+2+2+2=8 points]
Indicate for each of the following statements whether it is correct, incorrect, or nonsensical (i.e., makes no sense). Motivate your answers shortly.

a. For the chi-square goodness-of-fit test it is recommended to choose the intervals such that the number of observed values in each interval is at least 5.

b. Spearman's ρ and Kendall's τ are two measures for the linear correlation in bivariate data.

c. The Shapiro–Wilk test can be used to test the hypothesis that the observations originate from the standard normal distribution.

d. Given a sample x from $\mathcal{N}(2,4)$ and a sample y from $\mathcal{N}(4,4)$, the t-test is more powerful for finding a difference between the underlying distributions than the Wilcoxon two-sample test.

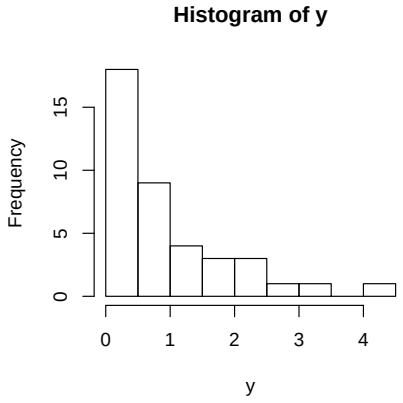

Figure 2: Histogram of data set `y`.

Question 3 [3+2+2=7 points]
Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables with unknown distribution P. Suppose that the α-trimmed mean $T_{n,\alpha}$ is used to estimate the location of P. To determine the accuracy of this estimator, its standard deviation 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 standard deviation of $T_{n,\alpha}$.

b. Consider the data set `y` presented in Figure 2. Empirical bootstrap values for the α_1-trimmed mean and the α_2-trimmed mean of this data set were computed and some quantiles of these bootstrap values of both location estimators are:

quantile	0.025	0.05	0.5	0.95	0.975
α_1-trimmed mean	0.415	0.459	0.654	0.932	0.986
α_2-trimmed mean	0.380	0.407	0.592	0.867	0.947

Indicate whether $\alpha_1 < \alpha_2$ or $\alpha_1 > \alpha_2$. Motivate your answer.

c. The α_2-trimmed sample mean of `y` equals 0.595. Determine the 95% bootstrap confidence interval for the α_2-trimmed mean.

k	p						
	0.025	0.05	0.33	0.5	0.67	0.95	0.975
0	0.738	0.540	0.008	0.000	0.000	0.000	0.000
1	0.965	0.882	0.057	0.003	0.000	0.000	0.000
2	0.997	0.980	0.188	0.019	0.000	0.000	0.000
3	1.000	0.998	0.403	0.073	0.004	0.000	0.000
4	1.000	1.000	0.641	0.194	0.018	0.000	0.000
5	1.000	1.000	0.829	0.387	0.063	0.000	0.000
6	1.000	1.000	0.937	0.613	0.171	0.000	0.000
7	1.000	1.000	0.982	0.806	0.359	0.000	0.000
8	1.000	1.000	0.996	0.927	0.597	0.002	0.000
9	1.000	1.000	1.000	0.981	0.812	0.020	0.003
10	1.000	1.000	1.000	0.997	0.943	0.118	0.035
11	1.000	1.000	1.000	1.000	0.992	0.460	0.262

Table 1: Probabilities $P(X \leq k)$ for binomially distributed random variable X with parameters $n = 12$ and p as given in table, for different values of k.

Question 4 [2+2+2+2=8 points]
A photo shop uses the slogan: *Majority of our customers have their photos printed within 15 minutes!* In other words, the shop claims that the median printing time is less than or equal to 15 minutes. Since some customers were complaining that they always had to wait longer than that, the manager of the photo shop decided to test $H_0 : m \leq 15$ versus $H_1 : m > 15$. He randomly selected 12 print tasks and obtained the following measurements (in minutes, sorted): 14.12, 14.13, 15.03, 15.23, 15.38, 16.23, 16.91, 17.71, 20.07, 20.27, 27.00, 30.05. Next, the problem was investigated by performing the sign test on these data.

a. Give the formula for the test statistic and its distribution under the assumption $m = 15$.

b. Perform the sign test at significance level $\alpha = 0.05$ using Table 1. Give the p-value and the final conclusion.

c. The sign test requires a relatively weak assumption: the true underlying distribution of the observations has a unique median. Another test that could be used to test the hypotheses of the manager is the Wilcoxon signed rank test. What is the assumption in this case? Briefly explain how you would investigate whether this assumption is satisfied or not.

d. The sign test and the Wilcoxon signed rank tests are *distribution free* or, in other words, *nonparametric*. Explain what this means.

Question 5 [2+2+2+1=7 points]
We asked 20 people involved recently in car accidents how severely they were injured in the accident. Additionally, we asked whether they were wearing seat belts at the time of accident or not. The results we found are presented in the following table:

	light or no injury	severe injury	total
no seat belts	0	4	4
seat belts	6	10	16
total	6	14	20

a. Formulate a suitable model of multinomial distribution(s) and state the corresponding null and alternative hypotheses for investigating whether there is a relationship between severity of injury and wearing seat belts.
You may formulate your hypotheses either in words or in formulas.

b. Perform Fisher's exact test at significance level $\alpha = 0.05$ to test the null hypothesis of part a. You may use the probability mass function of the hypergeometric distribution: let X be the number of white balls in a pick of l balls from an urn containing n balls of which m balls are white, then

$$P(X = k) = \frac{\binom{m}{k}\binom{n-m}{l-k}}{\binom{n}{l}}.$$

c. Suppose that the chi-square test is to be used, instead of Fisher's exact test. Check whether the rule of thumb for applying the chi-square test is fulfilled. Why is that check important? Motivate your answer.

d. Suppose the rule of thumb in part c. is not fulfilled, which method could be applied if one still wanted to use the chi-square test statistic?
You do not have to present the method in detail.

Question 6 [3+2+2+2=9 points]

a. Formulate the general multiple linear regression model including its assumptions.

b. For each assumption, shortly describe a method to verify that assumption.

c. Explain the concept of (multi-)collinearity in multiple linear regression in your own words. Clearly state what collinearity is and why it may cause problems.

d. Consider the data shown in Figure 3. The response variable is the number of requests processed by a computer program per hour, and the explanatory variable is the size of incoming requests (in GB). There are 20 observations. Do you expect any problems when the (simple) linear regression model is fitted to these data? If yes, how would you investigate it/them?

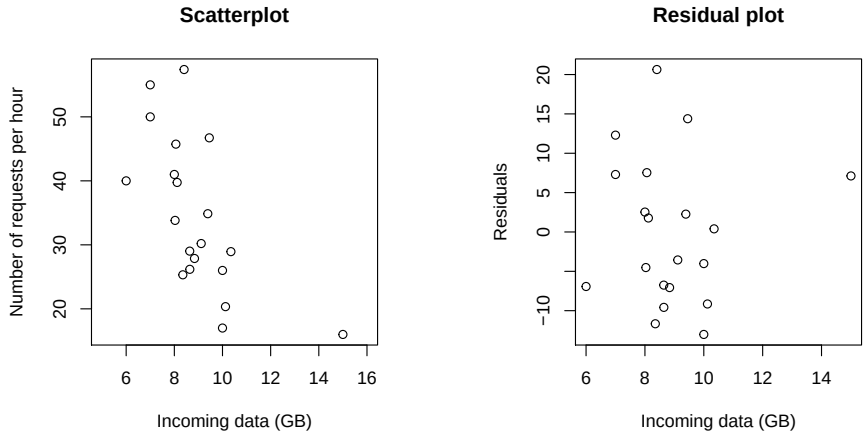

Figure 3: Scatter plot of 'requests per hour' against 'size of incoming data', scatter plot of the residuals against 'size of incoming data'.