VU University	Statistical Data Analysis
Faculty of Sciences	28 May 2014

Use of a basic calculator is allowed. Graphical calculators are not allowed. Please write all answers in English.

The **complete exam** consists of 7 questions (45 points). Grade = $\frac{total+5}{5}$.

The **exam on part 2** consists of 4 questions (27 points). Grade = $\frac{total+3}{3}$.

GOOD LUCK!

PART 1

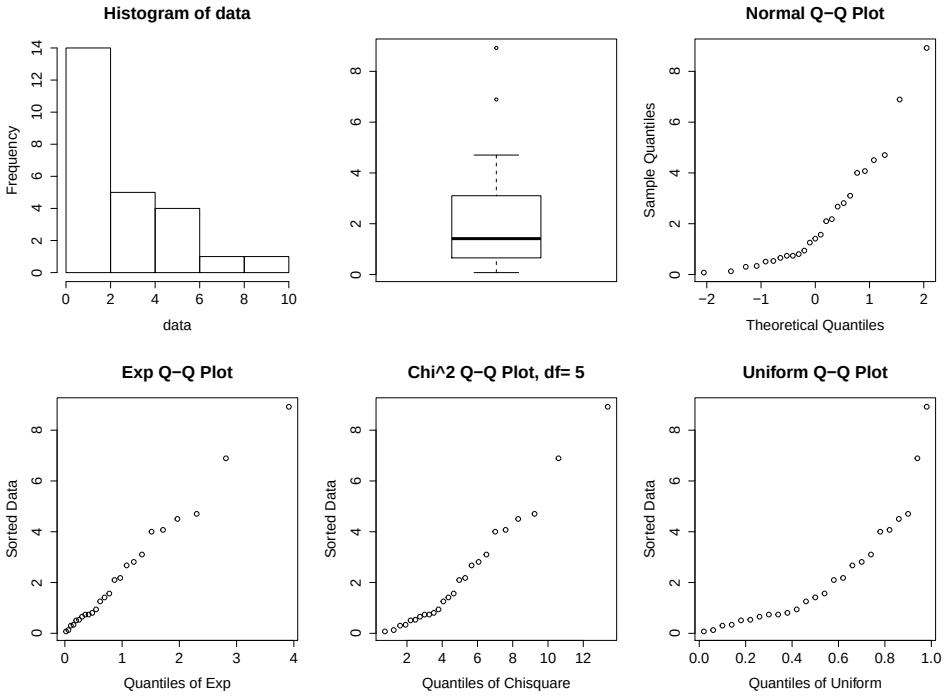

Figure 1: Histogram, boxplot and QQ-plots of a data set against the standard normal, standard exponential, χ^2_5 and standard uniform distributions.

Question 1 [6 points]
In Figure 1 a histogram, boxplot and several QQ-plots of a data set with sample size 25 are presented.

a. [2 points] Describe briefly what these graphical summaries tell you about the underlying distribution of the data set. Consider at least the aspects location, scale, shape and extreme values.

b. [2 points] Which of the four location-scale families do you think is most appropriate for these data? Explain your answer.

c. [2 point] Using the QQ-plot of the location-scale family that you have selected under part (a) determine the location a and scale b approximately.

Question 2 [5 points]
Suppose we want to test the null hypothesis that the data in Figure 1 are from the exponential distribution with parameter 1/2, using the chi-square goodness of fit test.

a. [1 point] Is the following expression for the test statistic correct?

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

where $I_1, \ldots, I_k$ are k disjoint intervals, N_i is the number of observations in interval I_i, $p_i = P\{I_i\}$ is the probability under H_0 that an observations lies in I_i and n is the total number of observations. If this expression is not correct, indicate where the error is.

b. [2 points] Give the degrees of freedom of the approximate χ^2-distribution of the test statistic under the null hypothesis and explain how one has to choose the intervals $I_1, \ldots, I_k$ for this approximate distribution to be reliable.

c. [2 point] Give an alternative test to the mentioned chi-square goodness of fit test for testing this null hypothesis.

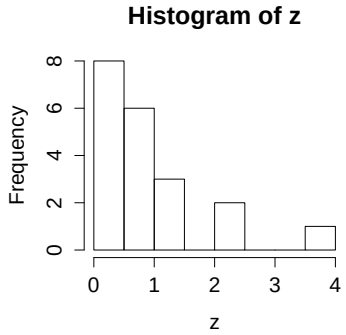

Figure 2: Histogram of data set z

Question 3 [7 points]
Let $Z_1, \ldots, Z_n$ 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. [3 points] 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. [2 points] Consider the data 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	0.45	0.48	0.71	1.07	1.16
α_2-trimmed mean	0.37	0.41	0.60	0.86	0.93

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

c. [2 points] The α_1-trimmed sample mean of z equals 0.70. Determine the 95% bootstrap confidence interval for the α_1-trimmed mean.

Part 2 starts on the next page

PART 2

Question 4 [6 points]
Are the following statements correct? Motivate your answer by a short argument.

a. [2 points] A normal QQ-plot of bootstrap values of the Kolmogorov-Smirnov test statistic is helpful for testing normality of the underlying distribution.

b. [2 points] Cook's distances are more informative for detecting influence points than hat values.

c. [2 points] The sign test has higher power than the Wilcoxon signed rank test for samples from a Laplace distribution (see Figure 3).

	t	s	w
t	1		
s	$\frac{2}{\pi}$	1	
w	$\frac{3}{\pi}$	$\frac{3}{2}$	1

N(0,1)

	t	s	w
t	1		
s	$\frac{\pi^2}{12}$	1	
w	$\frac{\pi^2}{9}$	$\frac{4}{3}$	1

logistic

	t	s	w
t	1		
s	$\frac{1}{3}$	1	
w	1	3	1

uniform

	t	s	w
t	1		
s	2	1	
w	$\frac{3}{2}$	$\frac{3}{4}$	1

Laplace

Figure 3: Asymptotic relative efficiencies (row-variable with respect to column-variable) of t-test (t), sign test (s) and Wilcoxon signed rank test (w) for shift alternatives of different underlying distributions (bottom line).

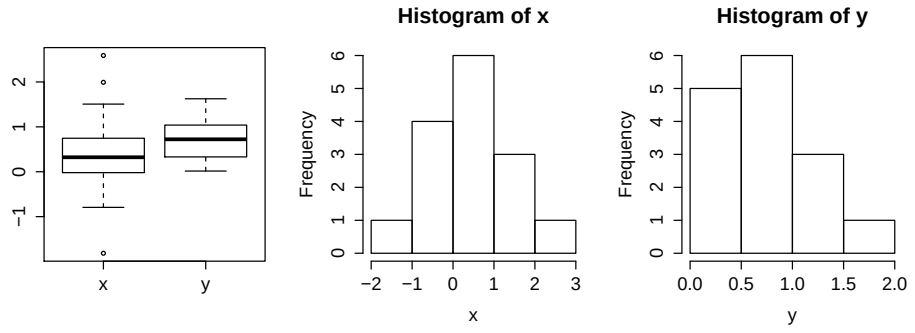

Figure 4: Boxplot and histograms of two data sets x and y.

Question 5 [5 points]
In Figure 4 the boxplots and histograms of two data sets x and y are shown. Suppose we want to test the null hypothesis that the underlying distributions of both data sets are equal.

a. [4 points] Indicate for each test below whether the test is suitable for testing this hypothesis. Motivate your answer clearly in case you think a test is not suitable.

   – sign test

- Wilcoxon two sample (rank sum) test
- Kendall's rank correlation test
- Kolmogorov–Smirnov two sample test

b. [1 point] Which test under part (a) do you expect to be most powerful for detecting a difference between the underlying distributions in the case of the data sets presented in Figure 4? Motivate your answer.

Question 6 [7 points]
We asked 20 students whether they voted for the European Parliament Elections on May 22nd, last week. Moreover we asked whether they were right or left handed. The results we found are presented in the following table:

	voted	not voted	total
left	0	2	2
right	7	11	18
total	7	13	20

a. [3 points] Formulate a suitable model of multinomial distribution(s) and state the corresponding null and alternative hypotheses for investigating whether there is a relationship between handedness and having voted. (You may formulate your hypotheses either in words or in formulas.)

b. [3 point] Perform Fisher's exact test to test the null hypothesis of part (a). You may use that 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. [1 point] Check whether the rule of thumb for applying the chi-square test is fulfilled. Can one use the chi-square test for these data?

Question 7 is on the next page

Question 7 [9 points]

a. [3 points] Formulate the general multiple linear regression model and its assumptions.

b. [3 points] For each assumption shortly describe a method to verify the validity of that assumption for a given data set.

c. [3 points] Consider Longley's economic regression data shown in Figure 5.
For these data we assume a linear regression model where the response variable is `GNP`, the Gross National Product. The available explanatory variables are `Unemployed` (the number of unemployed), `Armed.Forces` (the number of people in the armed forces), `Population` (the noninstitutionalized population $\geq$ 14 years of age), `Year` (the year (time)) and `Employed` (the number of people employed). There are 16 observations.
What problem(s) do you expect when the full model is fitted to these data? For each problem indicate at least one way you would investigate that problem.

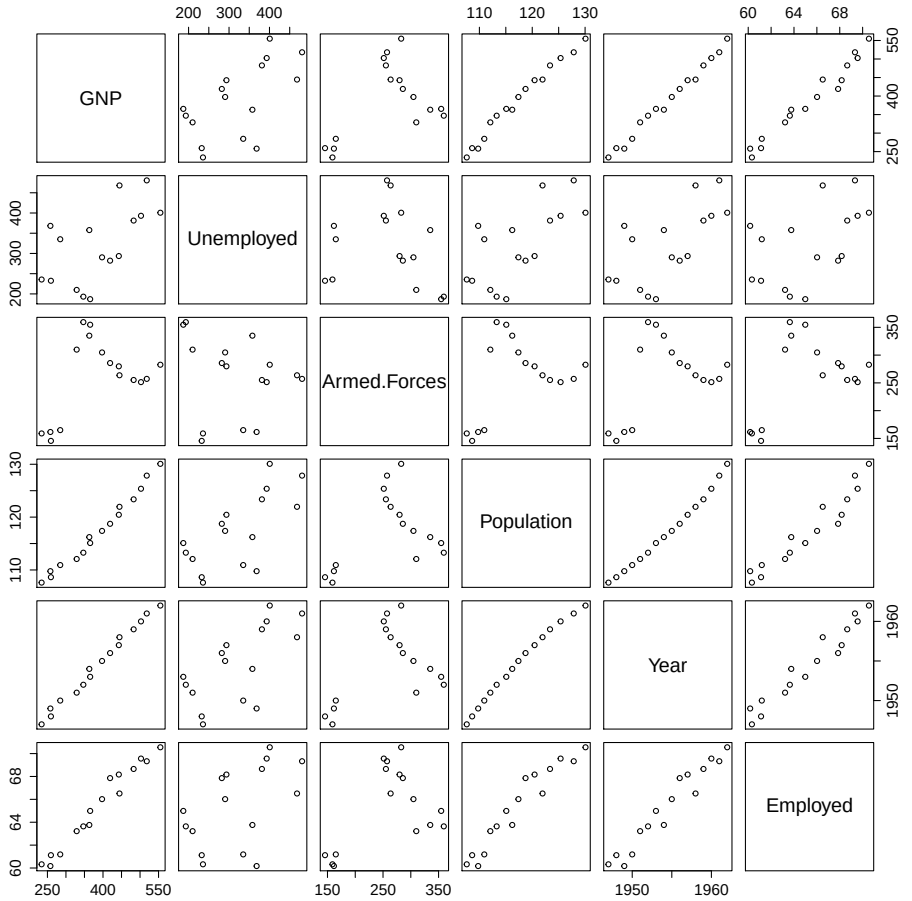

Figure 5: Scatter plots of Longley's data.

THE END