VU university	Statistical Data Analysis
Faculty of Sciences	30 June 2016

Use of a basic calculator is allowed. Graphical calculators are not allowed. This exam consists of 7 questions (45 points).

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

GOOD LUCK!

Question 1 [7 points]

Figure 1 shows a histogram and a boxplot for two data sets x and y.

a. [2 points] (This part of the question only considers data set x.) Describe briefly what the graphical summaries tell you about the underlying distribution of data set x. Consider (at least) the aspects location, scale, shape and extreme values.

b. [2 points] Decide for each of the two data sets whether the median will be larger, smaller, or approximately equal to the mean? Motivate your answer.

c. [3 points] Which tests would be appropriate to test the null hypothesis that the underlying distributions of x and y are equal? Give at least two tests and briefly explain why these tests are appropriate.

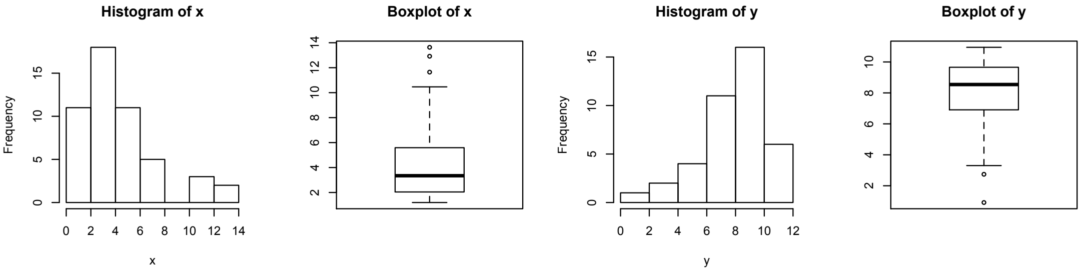

Figure 1: Histograms and boxplots of data sets x and y.

Question 2 [6 points]
In Figure 2 a histogram, boxplot and several QQ-plots of a data set are presented.

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

b. [2 points] Using the QQ-plot you have selected under part (a) determine the location a and scale b approximately.

c. [2 points] What is in general the influence of the sample size on QQ-plots? How confident are you about the conclusion made in (a)?

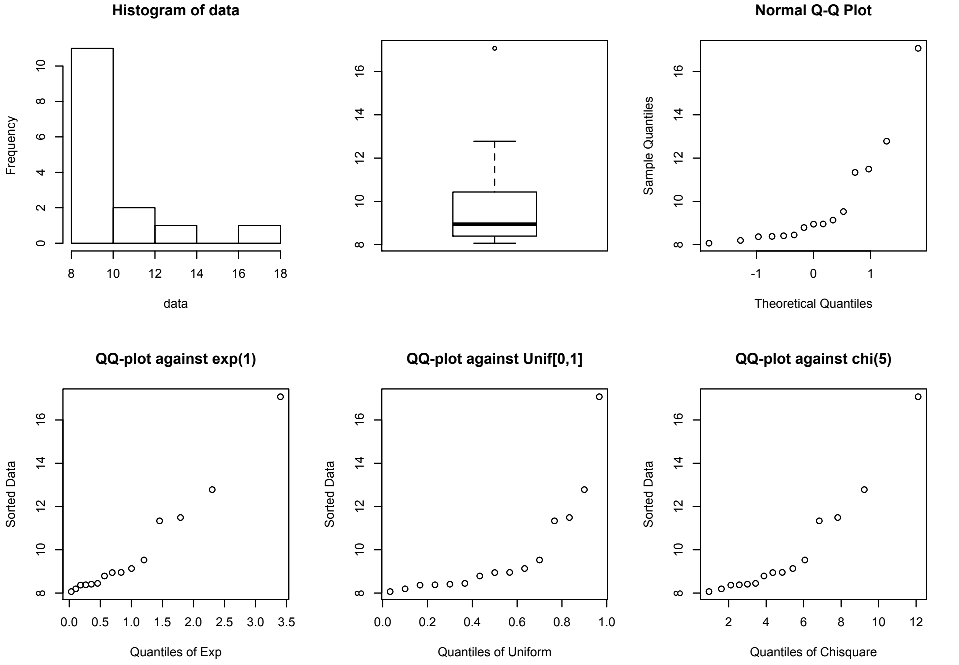

Figure 2: Histogram, boxplot and QQ-plots against the normal, exponential, uniform and χ^2_5 distributions of a data set.

Question 3 [6 points]
Are the following statements correct, incorrect or nonsensical (i.e. makes no sense)? Choose one of the three options for each statement. Motivate your answer by a short argument or sketch.

a. [2 points] The influence function of the sample mean is unbounded.

b. [2 points] In the context of linear regression: Cook's distances are more informative for detecting potential points than condition indices.

c. [2 points] The chisquare goodness-of-fit test always has higher power than the Kolmogorov-Smirnov goodness-of-fit test for testing a simple null hypothesis ($H_0 : F = F_0$).

	p						
k	0.025	0.05	0.33	0.5	0.67	0.95	0.975
0	0.684	0.463	0.002	0.000	0.000	0.000	0.000
1	0.947	0.829	0.021	0.000	0.000	0.000	0.000
2	0.994	0.964	0.083	0.004	0.000	0.000	0.000
3	1.000	0.995	0.217	0.018	0.000	0.000	0.000
4	1.000	0.999	0.415	0.059	0.002	0.000	0.000
5	1.000	1.000	0.629	0.151	0.008	0.000	0.000
6	1.000	1.000	0.805	0.304	0.029	0.000	0.000
7	1.000	1.000	0.916	0.500	0.084	0.000	0.000
8	1.000	1.000	0.971	0.696	0.195	0.000	0.000
9	1.000	1.000	0.992	0.849	0.371	0.000	0.000
10	1.000	1.000	0.998	0.941	0.585	0.001	0.000
11	1.000	1.000	1.000	0.982	0.783	0.005	0.000
12	1.000	1.000	1.000	0.996	0.917	0.036	0.006
13	1.000	1.000	1.000	1.000	0.979	0.171	0.053
14	1.000	1.000	1.000	1.000	0.998	0.537	0.316
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 1: Probabilities $P(X \leq k)$ for binomially distributed random variable X with parameters $n = 15$ and p as given in table, for different values of k.

Question 4 [6 points]

The data presented in Figure 2 contains the following 15 numbers: 8.07, 8.19, 8.37, 8.38, 8.41, 8.45, 8.79, 8.95, 8.95, 9.13, 9.53, 11.34, 11.49, 12.78, 17.07. We conjecture that the median m of the underlying distribution is smaller than 10, and want to test $H_0 : m \geq 10$ against $H_1 : m < 10$, using the sign test.

a. [2 points] Formulate the test statistic for the sign test, and give its distribution under the assumption $m = 10$.

b. [2 points] Perform the sign test at significance level $\alpha = 0.05$ using Table 1. Give the p-value and the conclusion.

c. [1 point] Is it appropriate to use the t-test in stead of the sign test for these data? Motivate your answer.

d. [1 point] Is it appropriate to use the Wilcoxon signed rank test in stead of the sign test for these data? Motivate your answer.

Question 5 [7 points]
Using a questionnaire we want to investigate the relation between political preference and ethnic background. We ask 178 people to fill out our questionnaire. We find the following data

	liberal	social democrates	green	total
western Europe	19	19	7	45
northern Africa	2	68	10	80
Suriname	19	27	7	53
total	40	48	24	178

a. [3 points] Specify a suitable model and state the corresponding null and alternative hypothesis for investigating whether there is a relationship between ethnic background and political preference using a chi-square test. (You may give your answer in formulas or in words.)

b. [2 points] State the rule of thumb for applying the chi-square test and check whether it is fulfilled.

c. [1 point] Suppose that the null hypothesis in part (a) is rejected. Shortly describe a method to investigate in what way the data differ from what is expected under the null hypothesis.

d. [1 point] In case the rule of thumb in part (b) is not fulfilled, which method would you use in order to test the null hypothesis of part (a)?

Question 6 [6 points]
Consider the data in Figure 3 about precipitation values of seeded clouds. As estimators for spread we computed the sample MAD and the sample standard deviation. To assess the accuracy of these estimators, we determined 90 % bootstrap confidence intervals for both. We found the following two intervals: [425, 976] and [122, 333].

a. [2 points] Give the formula for a 90 % bootstrap confidence interval for an estimator T based on a data sample $X_1, \ldots, X_n$. Explain your notation.

b. [2 points] Which interval is for the MAD? Motivate your answer.

c. [2 points] Which estimator for spread is more appropriate for these data? Motivate your answer.

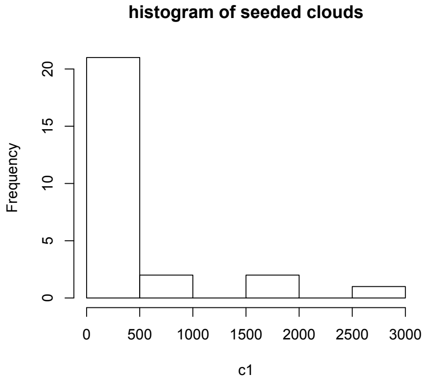

Figure 3: Histogram of precipitation values of the seeded clouds.

Question 7 [7 points]

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

b. [2 points] For each assumption, shortly describe a method to verify that assumption.

c. [2 points] Consider the data shown in Figure 4. The response variable is the number of violent crimes per 100,000 people (`crime`) and the available explanatory variables are population fraction of individuals that are single parents (`single`) and population fraction of individuals living under the poverty line (`poverty`). There are 51 observations.
What problem(s) do you expect when the full model is fitted to these data? Indicate at least one way you would investigate this/these problem(s).

THE END

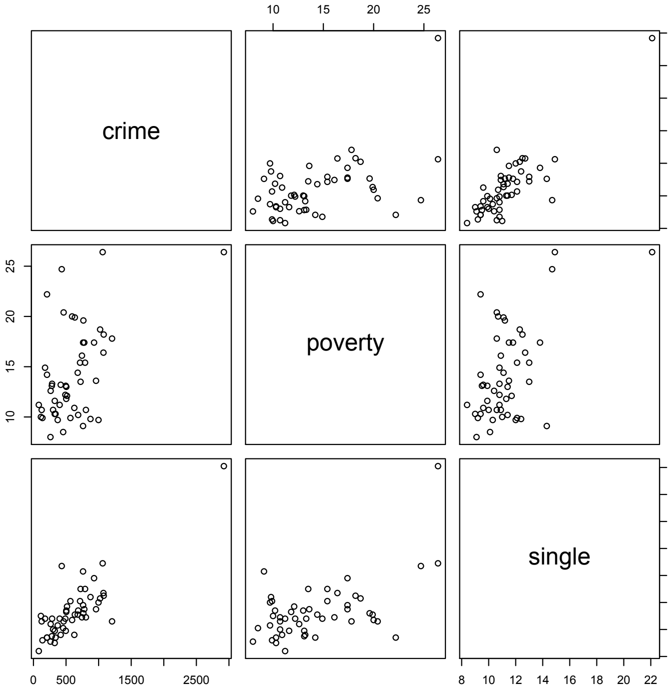

Figure 4: Scatter plots of the crime data.