VU University Statistical Data Analysis

Faculty of Sciences 3 July 2014

Use of a basic calculator is allowed. Graphical calculators are not allowed. Please write all answers in English.
The exam consists of 6 questions (45 points). Grade = $\frac{total+5}{5}$.

GOOD LUCK!

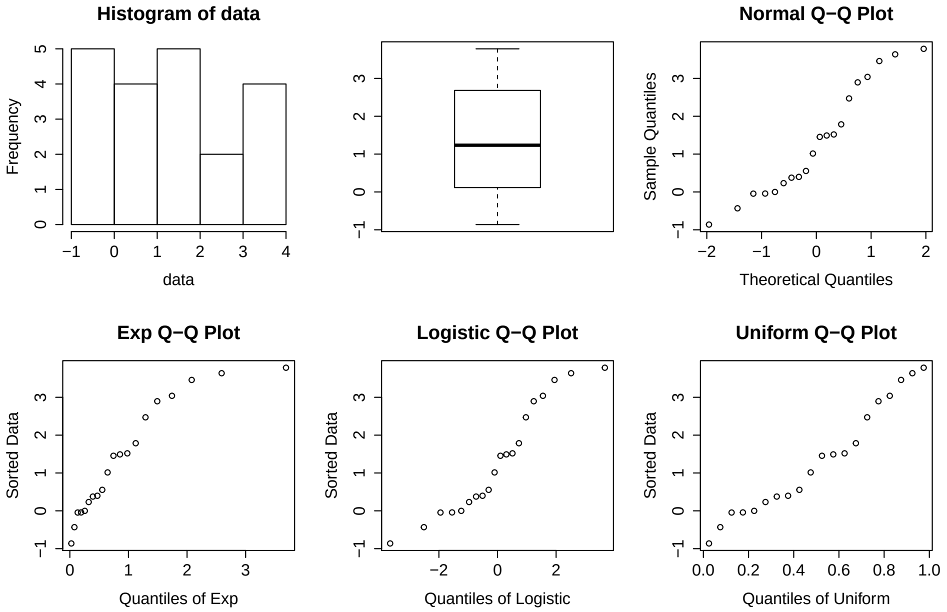

Figure 1: Histogram, boxplot and QQ-plots of a data set against the standard normal, standard exponential, standard logistic and standard uniform distributions.

Question 1 [8 points]

In Figure 1 a histogram, boxplot and several QQ-plots of a data set with sample size 20 are presented.

a. [3 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 that are mentioned above 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.

d. [1 point] Give a sketch of the QQ-plot of this data set against the standard Cauchy distribution.

Question 2 [7 points]
Are the following statements correct? Motivate your answer by a short argument or sketch.

a. [2 points] For the chi-square test for goodness-of-fit it is recommended to choose the intervals such that the number of observed values in each interval is at least 5.

b. [1 point] The influence function of the median is bounded.

c. [2 points] A two sample permutation test is in fact a bootstrap test.

d. [2 points] To test whether an explanatory variable should be included in a multiple linear regression model with uncorrelated, normally distributed measurement errors a t-test can be used.

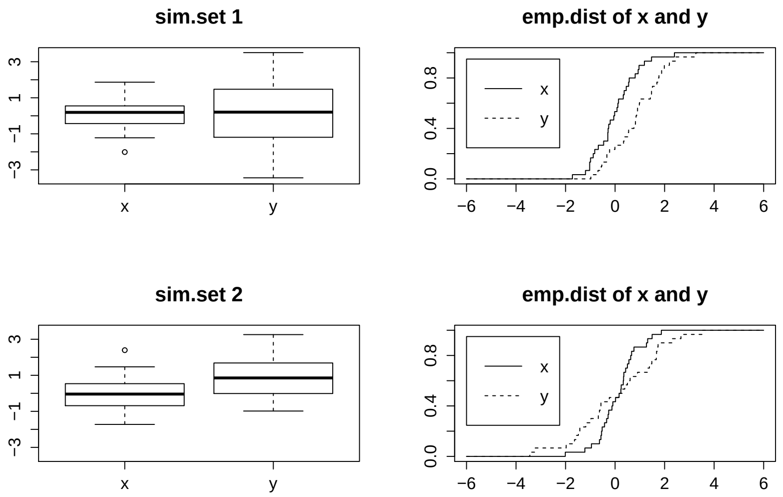

Figure 2: Boxplots and empirical distribution functions of two data sets x and y.

Question 3 [6 points]
Consider the data presented in Figure 2, consisting of four plots. Two plots show boxplots of data sets x and y, and two plots show empirical distribution functions of data sets x and y. For this figure we generated twice a simulation set of two data sets (x and y). The top left boxplot corresponds to simulation set 1, the bottom left boxplot corresponds to simulation set 2. The empirical distribution function plots at the right may or may not have been swapped.

a. [2 points] Which empirical distribution plot corresponds to simulation set 1? Motivate your answer clearly.

b. [2 points] Suppose you want to investigate a possible difference in the underlying distributions of samples x and y. You may use for one simulation set the Kolmogorov-Smirnov two sample test and for the other the Wilcoxon rank sum test (i.e. the Wilcoxon two sample test). In which simulation set would you apply which test in order to have the maximum power within the given contraints? Motivate your answer clearly.

c. [2 points] Suppose you are allowed to use any test for the hypothesis in part (b). Which test would you use in simulation set 1, and in simulation set 2? Motivate your answer.

Question 4 [10 points]
Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables with unknown distribution P. Suppose we assume that P is a parametric distribution P_θ. Suppose that the location of P is estimated by $T_n(X_1, \ldots, X_n) = \overline{X}$. To determine the accuracy of this estimator, its standard deviation is estimated by means of the parametric bootstrap.

a. [3 points] Describe the steps of the parametric bootstrap scheme that you would use to find the bootstrap estimate of the standard deviation of T_n.

b. [2 points] Which two errors are necessarily made in this bootstrap procedure? Indicate for both errors whether they can be made arbitrarily small. Motivate your answer.

c. [2 points] Suppose we substitute the procedure in part (a) by an empirical bootstrap procedure. What would change in your answer to part (b)? Motivate your answer.

d. [1 point] Explain the difference between

   (i) the sample standard deviation of bootstrap values of the sample mean (in R: `sd(bootstrap(data,mean))`)

   (ii) the sample mean of bootstrap values of the sample standard deviations (in R: `mean(bootstrap(data,sd))`).

e. [2 points] Indicate for both (i) and (ii) in part (d) what these quantities estimate. Which one of (i) and (ii) in part (d) is used in parts (a) and (c)?

Question 5 on next page

Question 5 [6 points]
We asked at random 76 inhabitants of Amsterdam whether they have watched any match of the 2014 FIFA World Cup Brazil at Museumplein. Moreover we asked them how many football matches they have watched during the period January 2014 – May 2014. The results we found are presented in the following table:

	Museumplein	not Museumplein	total
0 matches	2	29	31
>0 matches	4	41	45
total	6	70	76

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 watching at Museumplein and having watched any matches between January and May 2014. (You may formulate your hypotheses either in words or in formulas.)

b. [1 point] Check whether the rule of thumb for applying the chi-square test is fulfilled.

c. [2 points] Which test would you use to test the null hypothesis in part (a)? Describe the test statistic and its distribution under the null hypothesis of part (a).

Question 6 [8 points]

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

b. [2 points] For each assumption shortly describe a method to verify the validity of that assumption for a given data set.

c. [3 points] Consider the data on criminality in Figure 3 (see next page). The variables are violent crimes per 100,000 people (`crime`), murders per 1,000,000 (`murder`), the percentage of population with a high school education or above (`pcths`), percentage of population living under poverty line (`poverty`), and percentage of population that are single parents (`single`). It has 51 observations, one per state in the US. We want to investigate a linear regression model for these data with `crime` as response variable.
Indicate for each of the following problems whether you do expect this problem when a linear regression model is fitted to these data (motivate your answer):

   – outliers

   – leverage/influence points

   – collinearity

   Moreover, for each problem that you expect, indicate at least one way you would investigate that problem.

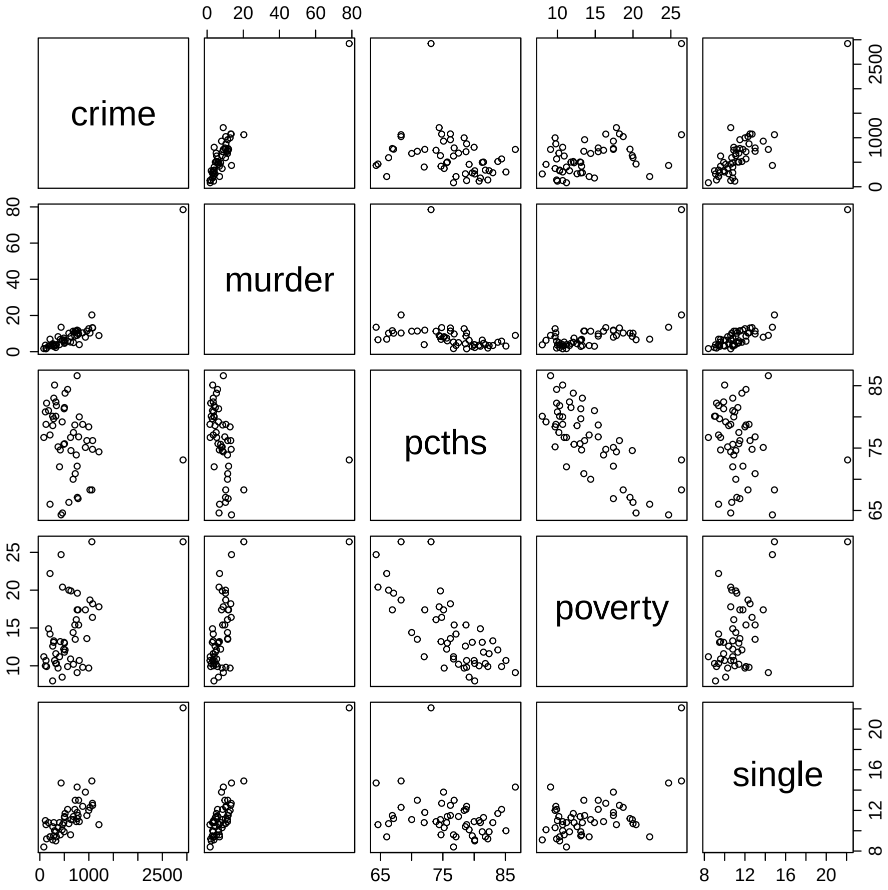

Figure 3: Scatter plots of the data on criminality.

THE END