VU university	Statistical Data Analysis
Faculty of Sciences	29 May 2012

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!

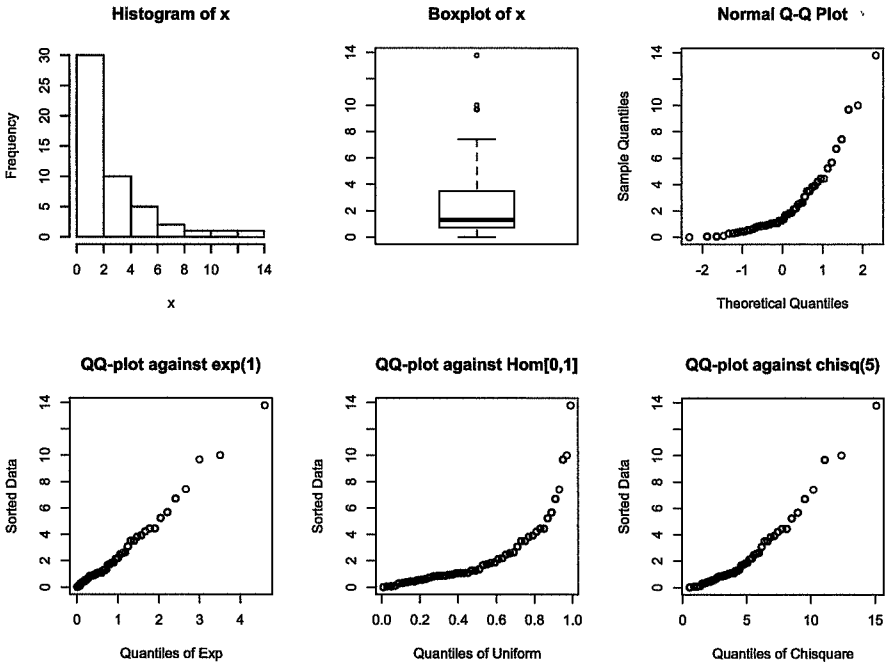

Figure 1: Histogram, boxplot and QQ-plots against the normal, exponential, uniform and χ^2_5 distributions of a data set.

Question 1 [6 points]

In Figure 1 a histogram, boxplot and several QQ-plots of a data set x 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] Suppose we want to test whether the underlying distribution is equal to the exponential distribution with parameter 1, using a χ^2 goodness-of-fit test. Indicate how we should select the intervals in that case. Can we find these intervals from the histogram in Figure 1? (You are not asked to give *numerical* intervals.)

Question 2 [6 points]
Are the following statements correct/sensible? Motivate your answer by a short argument, or a sketch.

a. [2 points] The influence function of the median is bounded.

b. [2 points] In the context of linear regression: variance inflation factors (VIF_i's) are more informative for detecting outliers than hat values (h_{ii}'s).

c. [2 points] Given a sample x from Unif[0,8] and a sample y from $N(4,4)$, the median test has higher power for finding a difference between the two underlying distributions than the Kolmogorov-Smirnov test.

Question 3 [8 points]
Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with unknown distribution P. We assume that P is an exponential distribution with unknown parameter λ. Suppose that $T_n(X_1, \ldots, X_n) = X_{(n)} - X_{(1)}$ (the difference between the last and first order statistics) is used to estimate the scale of P. To determine the accuracy of this estimator, its standard deviation is estimated by means of the bootstrap.

a. [4 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 in a parametric bootstrap procedure.

b. [2 points] Describe shortly which two errors are (necessarily) made in an empirical bootstrap procedure.

c. [2 points] Which of the two errors in part (b) can be made arbitrarily small? What do you have to change in the procedure under (a) to make this error smaller?

Question 4 [7 points]
We investigate the null hypothesis that handedness and sex are independent by considering the following data:

	right-handed	left-handed
men	150	13
women	63	0

The p-value of Fisher's (two-sided) test is 0.02, whereas the p-value of the χ^2 test for contingency tables is 0.05.

a. [3 points] Is the p-value of Fisher's test reliable? Is the p-value of the χ^2 test for contingency tables reliable? Motivate your answers.

b. [4 points] This null hypothesis can also be tested using a permutation test. Describe the steps of such a permutation test. (Several answers are possible — give one appropriate option.)

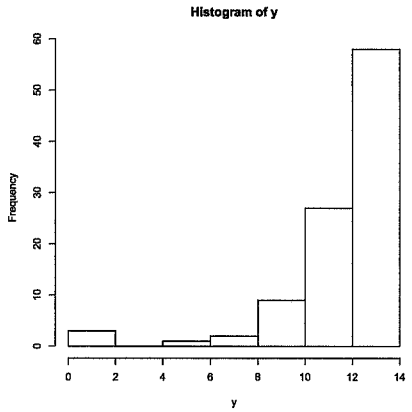

Figure 2: Histogram of a data set

Question 5 [5 points]
Consider the data in Figure 2. The mean of these data is 11.47 and the median is equal to 12.38.

a. [3 points] Empirical bootstrap values for the mean and median were computed and some quantiles of these bootstrap values of both location estimators for these data set are:

quantile	0.025	0.05	0.5	0.95	0.975
bootstrap value mean	11.00	11.07	11.45	11.83	11.90
bootstrap value median	11.94	12.01	12.35	12.61	12.62

Determine 95% bootstrap confidence intervals both for the mean and the median of the underlying distribution.

b. [2 points] Compare the two intervals and explain the difference between the intervals.

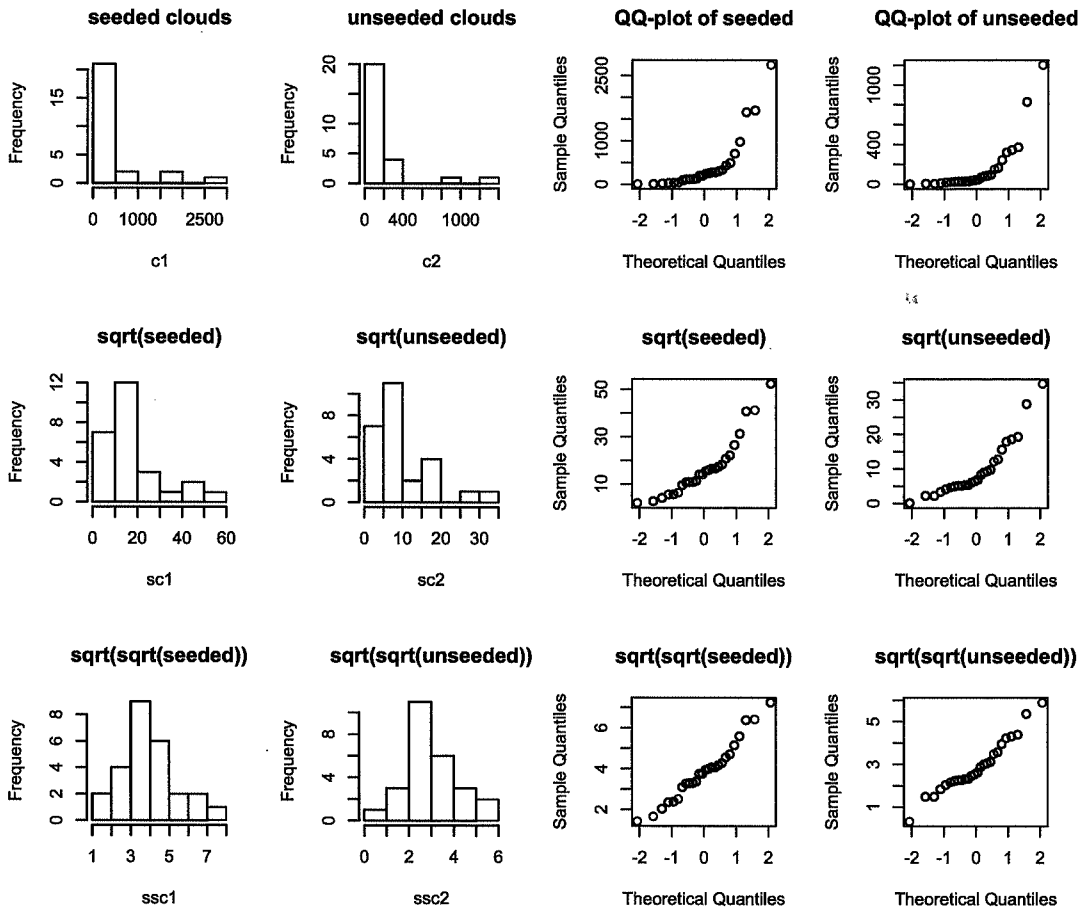

Figure 3: Graphical representations of the clouds data, and its transformations. The QQ-plots are against the $N(0,1)$ distribution.

Question 6 [5 points]
Consider the data on seeded clouds and unseeded clouds in Figure 3. To test the null hypothesis that these two samples come from the same distribution we have performed the two-sample t-test, the Wilcoxon two-sample test on the original data, and on the square root and fourth root of the data. Some of the p-values of these tests are given in the table below.

data	t-test	Wilcoxon test
seeded vs. unseeded	0.05375	0.01383
$\sqrt{\text{seeded}}$ vs. $\sqrt{\text{unseeded}}$	0.01956	p_1
$\sqrt[4]{\text{seeded}}$ vs. $\sqrt[4]{\text{unseeded}}$	0.0124	p_2

a. [2 points] Which p-values in the column under t-test do you trust? Motivate your answer.

b. [2 points] Indicate whether p_1 and p_2 are bigger, equal or smaller than 0.01383. Motivate your answer. (*Hint: consider the form of the test statistic of the Wilcoxon two-sample test.*)

c. [1 point] What is your conclusion about the null hypothesis? Motivate your answer.

Question 7 [8 points]

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

b. [2 points] Describe shortly how the model assumptions can be checked.

c. [3 points] Consider the data shown in Figure 4. The response variable is Gross National Product (`GNP`) and the available explanatory variables are number of unemployed (`unemployed`), number of armed forces (`Armed.Forces`), noninstitutionalized population $\geq$ 14 years of age (`population`), the year (`year`) and number of people employed (`employed`).
What problem(s) do you expect when the full model is fitted to these data? Indicate at least two ways you would investigate this/these problem(s).

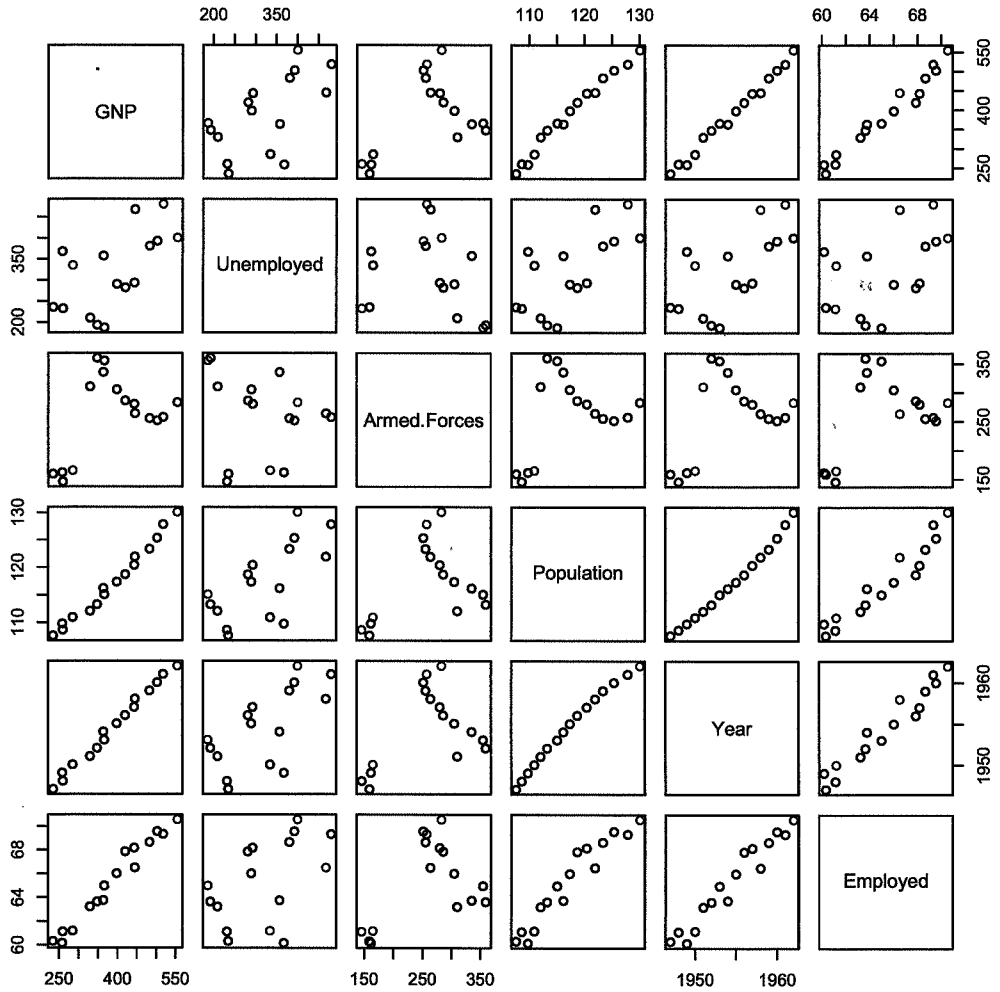

Figure 4: Scatter plots of the GNP data.

THE END