VU University	Statistical Data Analysis, part I
Faculty of Sciences	26 March 2015

Use of a basic calculator is allowed. Graphical calculators and mobile phones are not allowed. This exam consists of 4 questions (27 points).

Please write all answers in English. Grade $= \frac{total+3}{3}$.

GOOD LUCK!

Question 1 [7 points]

a. [2 points] Can the empirical distribution function of a sample be a continuous function instead of a step function? Motivate your answer.

b. [2 points] Describe the difference between a two sample QQ-plot and a two sample scatter plot.

c. [2 points] Is the 10%-trimmed mean expected to be smaller or larger than the median of samples from an exponential distribution? Motivate your answer.

d. [1 point] Sketch the influence function of the sample mean.

Question 2 [7 points]
Consider the data presented in Figure 1 (see page 3). We want to test the null hypothesis that the underlying distribution of this data set is the standard normal distribution, N(0,1).

a. [2 points] Suppose we want to use the chi-square test for goodness of fit. Describe the rule of thumb that the intervals in a chi-square goodness-of-fit test should fulfill in general.

b. [1 point] Why should the rule of thumb in part (a) be fulfilled?

c. [1 point] Do you think the chi-square test for goodness of fit is a good choice for testing the given hypothesis for this data set? Motivate your answer.

d. [1 point] Can we apply the Shapiro-Wilk test to test the given null hypothesis? Motivate your answer.

e. [2 points] Suppose we want to use the Kolmogorov-Smirnov (KS) test. Give the formula of, or describe in words, the test statistic of the KS-test and find its value (approximately) from Figure 1.

Question 3 [6 points]
Consider the data presented in Figure 2 (see page 3). The 10% trimmed mean of this sample equals 3.76 and the 30% trimmed mean equals 3.10. Empirical bootstrap values for the 10% trimmed mean and the 30% trimmed mean of this data set were computed. Histograms of these two sets of bootstrap values are given in Figure 2 (middle and right, in unknown order). Some quantiles of these bootstrap values of both location estimators are:

quantile	0.025	0.05	0.5	0.95	0.975
10% trimmed mean	2.52	2.69	3.72	5.07	5.37
30% trimmed mean	2.23	2.34	3.11	4.38	4.61

    a. [2 points] In the histograms in Figure 2 it is not indicated which histogram shows the bootstrap values of the 10% trimmed mean. Is this the middle plot or the right plot? Motivate your answer. Do not use the numbers in the table in your motivation, but motivate your answer using the histogram of the sample (left plot) **only**.

    b. [3 points] Give the formula for a bootstrap confidence interval and determine the 95% bootstrap confidence intervals for both the 10% and the 30% trimmed mean, using the given numbers.

    c. [1 point] Which estimator for location do you prefer for this data set? Motivate your answer.

Question 4 [7 points]
Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with unknown distribution P. In Figure 3 (see page 4) the histogram, the boxplot and QQ-plots against N(0,1), Exp(1), χ_1^2 and χ_4^2 are shown for this data set. The sample mean equals 1.66, the sample median 0.59, the sample standard deviation is 2.71, and the sample variance equals 7.34.

    a. [1 point] Which of the four location-scale families mentioned above do you think is most appropriate for these data? Motivate your answer.

    b. [2 points] Using the QQ-plot of the location-scale family that you have selected under part (a), determine the location a and scale b approximately. (*Hint: you may use that the expectation and variance belonging to a χ_k^2 distribution equal k and $2k$ respectively.*)

Suppose that the sample mean is used to estimate the location of P. To determine the accuracy of this estimator, its standard deviation is estimated by means of the bootstrap.

    c. [1 point] Which procedure would you prefer for this data set, empirical bootstrap or parametric bootstrap? Motivate your answer.

    d. [2 points] Describe the steps in the scheme of your preferred bootstrap method to find a bootstrap estimate of the standard deviation of T_n.

    e. [1 point] How do you like the sample mean as estimator for location of P for this data set? Would you prefer a different estimator for location? Motivate your answer.
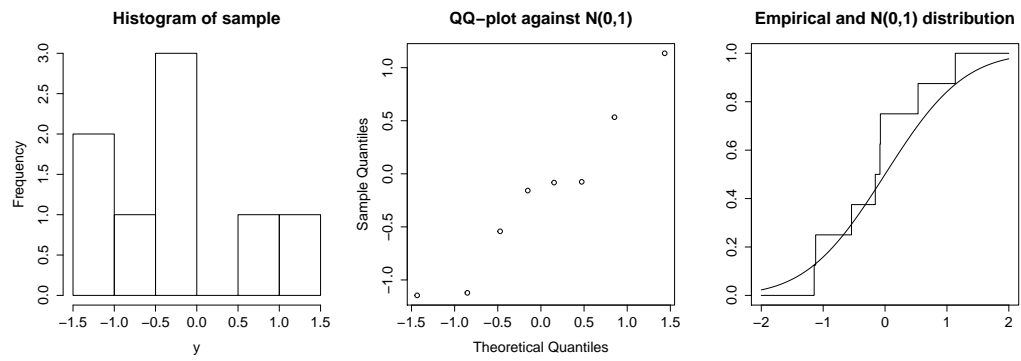
Figure 1: Histogram of a sample (left), QQ-plot against N(0,1) (middle) and empirical distribution function together with the N(0,1) distribution function (right).
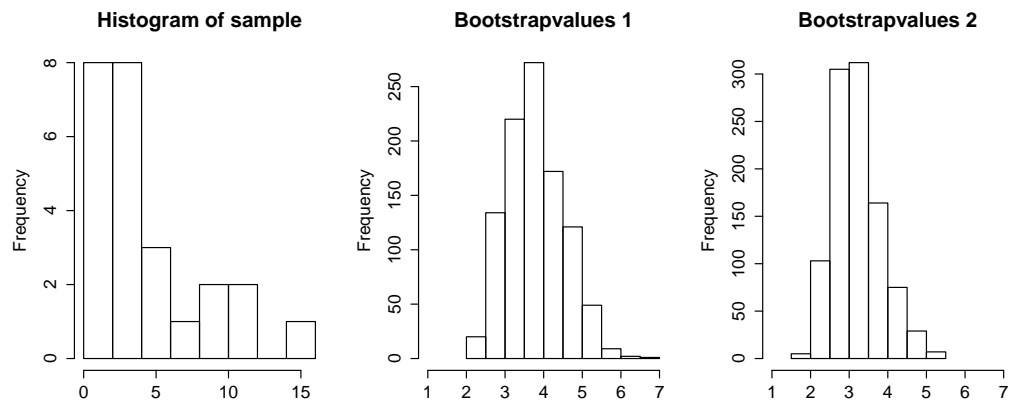

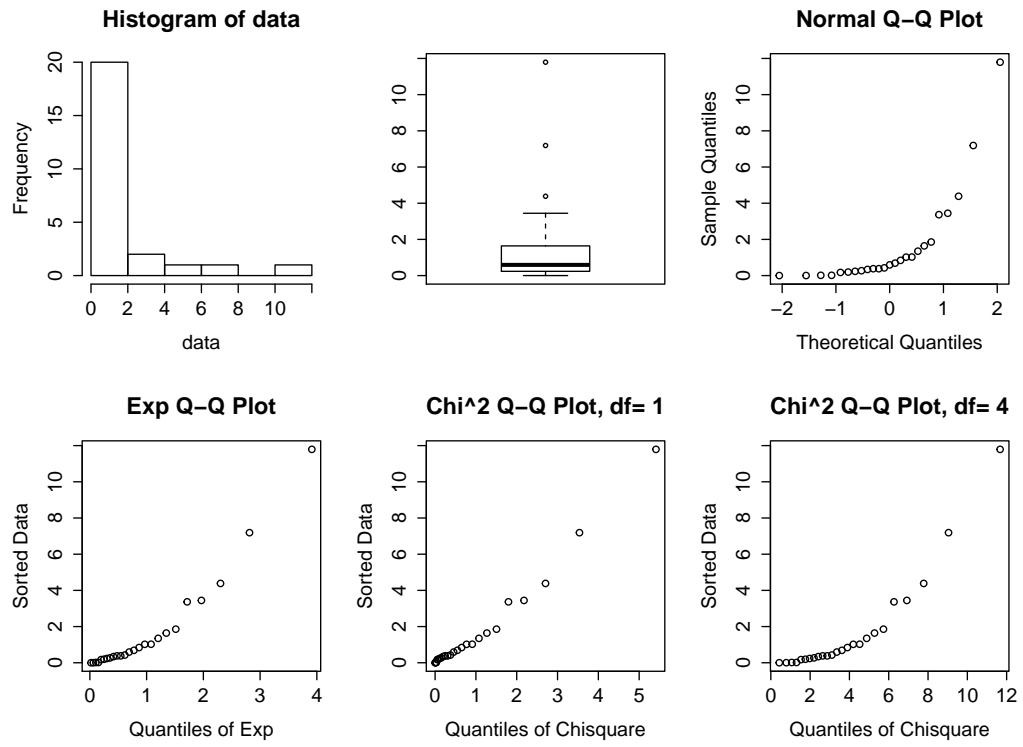Figure 2: Histogram of a sample (left), and bootstrap values of two different trimmed means (middle and right).

Figure 3: Histogram and boxplot of a data set, and QQ-plots against standard normal, standard exponential and χ_1^2 and χ_4^2.

THE END