Vrije Universiteit Amsterdam	Statistical Data Analysis, part I
Faculty of Sciences	24 March 2016

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 [8 points]
Indicate for each of the following statements whether it is correct (or makes sense at all) or not. Motivate your answers shortly.

a. [2 points] The empirical bootstrap is always a better choice than the parametric bootstrap.

b. [2 points] Dots in a QQ-plot always follow a non-decreasing curve.

c. [2 points] A bivariate scatter plot shows more information than a plot of the empirical distribution function of a sample.

d. [2 points] M-estimators are robust estimators.

Question 2 [7 points]
In Figure 1 (see page 3) the histogram, symplot and QQ-plots against the lognormal, standard exponential, χ^2_3 and χ^2_{10} distributions are shown for a data set `x`. The following values of the sample are computed:

	sample `x`	log(sample `x`)
sample mean	3.89	0.91
sample sd	3.79	1.08
sample variance	14.34	1.172

a. [1 point] Which of the four location-scale families in Figure 1 do you think is the 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^2_k *distribution equal* k *and* $2k$ *respectively.*)

c. [2 points] Sketch the shape of the QQ-plot of sample `x` against the Uniform[0,1]-distribution. Sketch clearly whether that is approximately a straight, convex or concave curve. Motivate the shape of your sketch.

d. [2 points] The 0, 10%, 20%, 30%, 40% and 50% trimmed means of sample `x` are, in arbitrary order, equal to 3.89, 3.22, 2.89, 3.01, 2.94, 3.02. Indicate which number is the 40% trimmed mean, and motivate your answer.

Question 3 [7 points]
Suppose we are given a sample $X_1, \ldots, X_n$ from the uniform distribution on $[0, \theta]$ with $\theta > 0$ unknown. We are interested in the variance of the statistic $T_n = \max(X_1, \ldots, X_n)$ and will use the parametric bootstrap for that.

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

b. [3 points] Which two errors are necessarily made in this bootstrap procedure? One of the two errors can be avoided in this specific situation. Which error is that? Motivate your answer clearly!

c. [1 point] Is the variance of bootstrap values of the sample maximum the same as the maximum of bootstrap values of the sample variance? I.e. in R-code is `var(bootstrap(sample, max, B=1000))` equal to `max(bootstrap((sample, var, B=1000))` ? Motivate your answer.

Question 4 [5 points]
Consider the data presented in Figure 2 (see page 3). The figure on the left is a histogram, and the right panel shows the empirical distribution function $\hat{F}_n$ for a sample $Y_1, \ldots, Y_{20}$ from an unknown distribution F, together with the distribution function F_0 of the uniform distribution on [0,5] (dashed line). We want to test the null hypothesis $H_0 : F = F_0$.

a. [2 points] Describe in words or formulas the Kolmogorov-Smirnov test statistic for the stated null hypothesis and determine its observed value (roughly) from the figure.

b. [3 points] The χ^2 goodness-of-fit test is based on the test statistic

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

which has approximately a χ^2_{k-1}-distribution under H_0. Explain the notation k, N_i, n and p_i. Describe the rule of thumb that needs to be satisfied for the χ^2-approximation to be reliable.

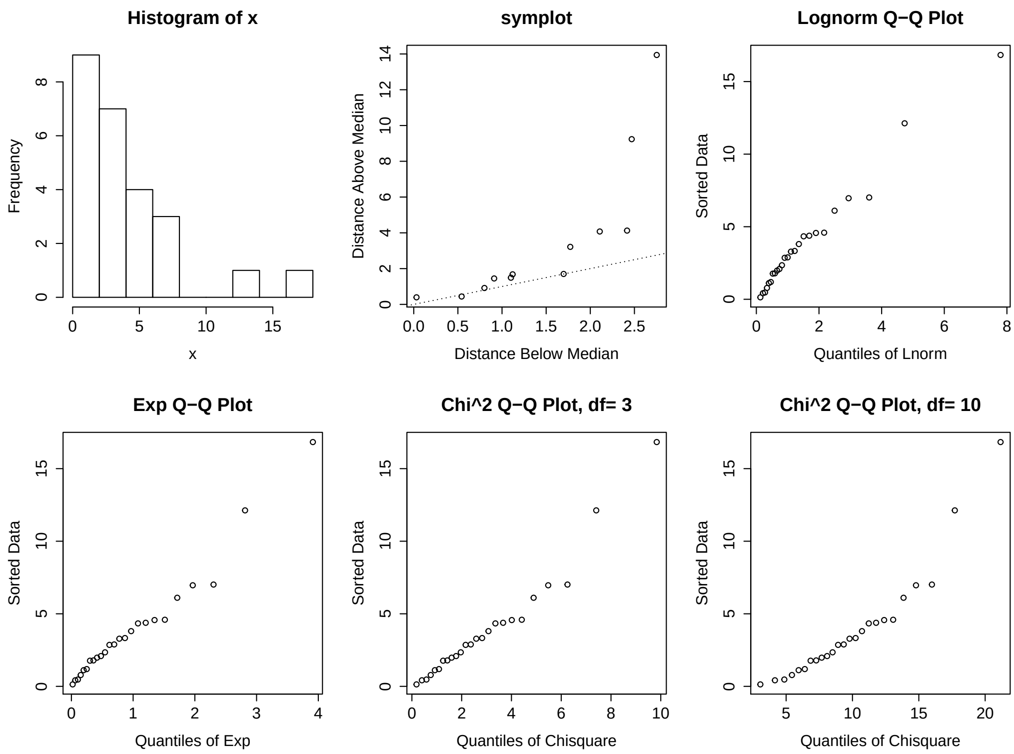

Figure 1: Histogram, symplot and QQ-plots against indicated distributions of a sample `x`.

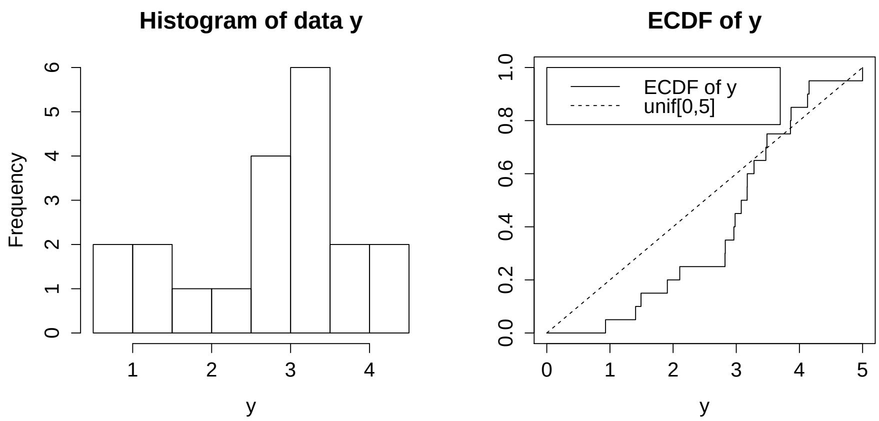

Figure 2: Histogram (left) and empirical cumulative distribution function (ECDF) of a sample `y` together with the uniform[0,5] distribution function (right).