Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam I
Faculty of Sciences	30 March 2017

Use of a basic calculator is allowed. Graphical calculators and mobile phones are not allowed. This exam consists of 4 questions on 2 pages (27 points).

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

You have 120 minutes to write the exam.

GOOD LUCK!

Question 1 [8 points]
Indicate for each of the following statements whether it is correct, incorrect, or nonsensical (i.e., makes no sense). Motivate your answers shortly.

a. [2 points] The standard bootstrap confidence interval with confidence level $(1-2\alpha)$ is given by $\left[T^*_{([\alpha B])}, T^*_{([(1-\alpha)B])}\right]$.

b. [2 points] The (non-bootstrap version of) Kolmogorov–Smirnov test can be used to test a simple hypothesis about any fixed probability distribution.

c. [2 points] The α-trimmed mean for $\alpha > 0$ is B-robust.

d. [2 points] A straight line in a two sample QQ-plot of two data sets clearly indicates that they originate from the same distribution.

Question 2 [7 points]
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with unknown distribution P.

a. [2 points] The Shapiro–Wilk test statistic is

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

where $a_1, a_2, \ldots, a_n$ are certain constants. State the null hypothesis, in terms of the unknown distribution P, that can be tested with this test statistic.

b. [1 point] What is the quantity in the denominator of the test statistic W?

c. [1 point] What are the possible values of the test statistic W?

d. [3 points] The distribution of the Shapiro–Wilk test statistic is available in many statistical packages, including *R*. Nevertheless, suppose that the distribution of W is not available and we have to use a bootstrap test. Describe the steps that are made in a bootstrap test for the given null hypothesis using W as test statistic.

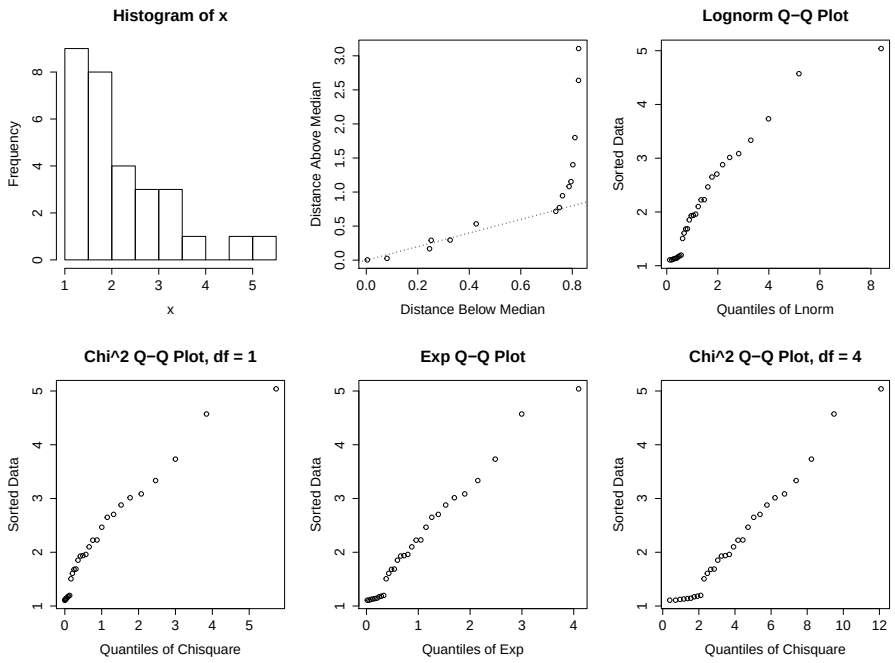

Figure 1: Histogram, symplot and QQ-plots against indicated distributions of a sample `x`.

Question 3 [5 points]
In Figure 1 the histogram, symplot and QQ-plots against the standard lognormal, χ^2_1, standard exponential, and χ^2_4 distributions are shown for a data set `x`. The following tables contain theoretical values of expectation and variance of the reference distributions, and some sample characteristics

distribution	expectation	variance
standard lognormal	1.649	4.671
χ^2_1	1	2
standard exponential	1	1
χ^2_4	4	8

	`x`
sample mean	2.150
sample standard deviation	1.041
sample variance	1.083

.

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.

c. [2 points] The sample median is one of the two numbers: 2.212, or 1.933. Indicate which number it is and motivate your answer.

Question 4 [7 points]
Let $X_1, X_2, \ldots, X_n$ be a sample from the Poisson distribution with rate parameter $\lambda > 0$. Suppose that the sample variance $T_n(X_1, \ldots, X_n) = S^2_X$ is used to estimate λ. To determine the accuracy of this estimator, its standard deviation is estimated by means of the bootstrap.

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

b. [2 points] Describe shortly which two errors are necessarily made in this bootstrap procedure. Can one of these errors be made arbitrarily small? If yes, what do you have to change in the procedure in part a to make this error smaller?

c. [2 points] In this situation the parametric bootstrap can also be used. Describe briefly what has to be changed in the procedure in part a.