Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam I
Faculty of Science	26 March 2021

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]
Explain for each of the following statements what is wrong and re-state them in a corrected way.

a. [2 points] If the points in the top-right of a QQ-plot lie above the best-fitting straight line, then both distributions that are compared in the QQ-plot have similarly heavy left tails.

b. [2 points] The Shapiro-Wilk test tests whether the true distribution underlying a sample is symmetric.

c. [2 points] Kernel density estimates are always smooth functions.

d. [2 points] The parametric bootstrap approximates the distribution of a statistic of interest by a suitable member of a parametric family.

Question 2 [9 points]
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with unknown cumulative distribution function F with density f.

a. [2 points] Explain why the general shape of a QQ-plot of the points $X_1, X_2, \ldots, X_n$ does not change if they underwent a linear transformation, i.e. $\tilde{X}_i = a + bX_i$, $i = 1, \ldots, n$, for some $a \in \mathbb{R}, b > 0$.

b. [3 points] Explain in your own words the conceptual differences in the test statistics that are used in the Kolmogorov-Smirnov and the Chi-squared goodness-of-fit tests for a simple null hypothesis.

c. [2 points] Explain in what sense the bandwidth $h_{opt} = (\int K^2(x)dx)^{1/5}(\int (f'')^2)^{-1/5}n^{-1/5}$ is optimal for kernel density estimation, and state a formula for the criterion that is optimized.

d. [2 points] Explain what cross-validation is used for in the context of kernel density estimation. Also give a representation of $\int \hat{f}(t)f(t)dt$ as a conditional expectation.
Reminder: $ISE(\hat{f}) = \int \hat{f}(t)^2dt - 2\int \hat{f}(t)f(t)dt + \int f(t)^2dt$*, where* $\hat{f}$ *is a density estimate.*

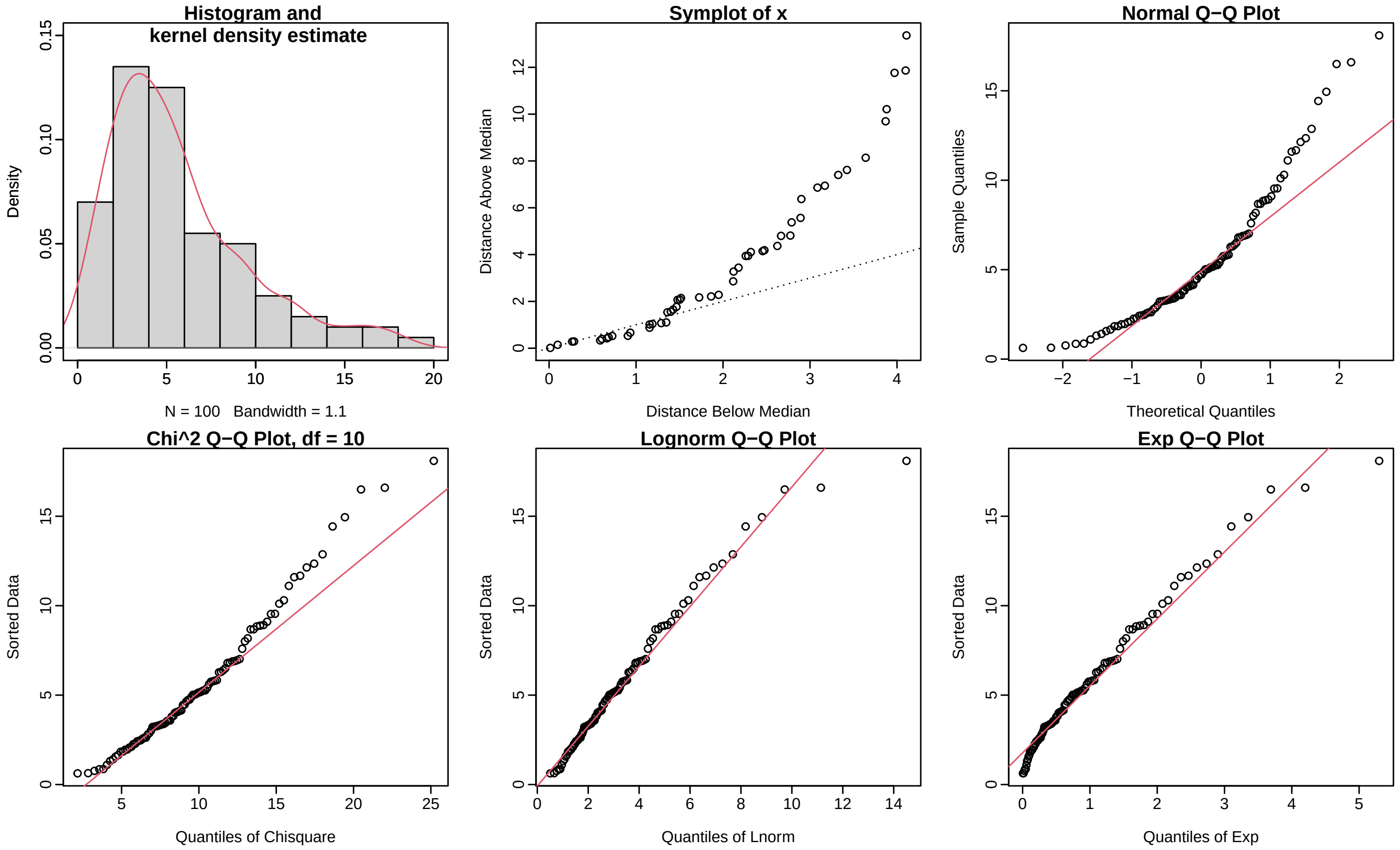

Figure 1: Histogram, kernel density estimate, symplot, and QQ-plots against indicated distributions of a sample x of size $n = 100$.

Question 3 [5 points]

In Figure 1 the histogram, symplot and QQ-plots with respect to the standard normal, χ^2_{10}, a specific lognormal, and standard exponential distributions are shown for a data set x. It has a sample skewness of 0.64. The table to the right contains theoretical values of expectation and variance of the reference distributions.

distribution	expectation	variance
standard normal	0	1
χ^2_{10}	10	20
$logN(1, 0.65)$	3.358	5.928
$Exp(1)$	1	1

a. [1 point] Based on the plots in Figure 1, decide and motivate which location-scale family is in your opinion most appropriate for these data.

b. [2 points] With respect to the location-scale family you chose in part a., determine the parameters of location a and scale b. Use that the sample mean, standard deviation, and variance are about $\bar{x} = 5.581$, $\hat{\sigma} = 3.818$, and $\hat{\sigma}^2 = 14.576$, respectively.

c. [2 points] The sample median is one of the following numbers: 4.734, 5.891, or 6.904. Indicate which one it is, and motivate your answer.

Question 4 [5 points]

Let $X_1, X_2, \ldots, X_n$ $(n \geq 2)$ be independent random variables that follow an unknown distribution P. We would like to investigate whether P is a skewed distribution. To this end, we choose the sample skewness statistic T_n with distribution $Q_{P,n}$. But instead of just the point estimate we want to use an interval estimate, i.e. a $(1 - \alpha)$-level confidence interval for the population skewness θ_P.

a. [3.5 points] Explain all steps that need to be made in order to find a two-sided level $(1 - \alpha)$ confidence interval for θ_P based on the empirical bootstrap, and state a formula for it.

b. [1.5 points] Describe shortly which two approximation errors are typically involved when a bootstrap procedure is applied.