Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam I
Faculty of Science	1 April 2022

Use of a basic calculator is allowed. Graphical calculators and mobile phones are not allowed. This exam consists of 3 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 [10 points]
Figure 1 contains three panels: a histogram of a sample A, a normal QQ-plot of a sample B, and a histogram of a sample C.

a. [2 points] Explain generally what information symplots provide and what one can conclude from this for the plotted sample. (Note: no symplot is depicted.)

b. [3 points] Describe the sample distribution of sample A. (See the left panel of Figure 1.) Comment on: symmetry/skewness, modality, location parameter (approximately), range (approximately).

c. [2 points] Comment on the heaviness of the tails of the sample B distribution. (See the middle panel of Figure 1.)

d. [3 points] Sketch a boxplot of sample C and comment on at least three features of the boxplot. (See the right panel of Figure 1.)

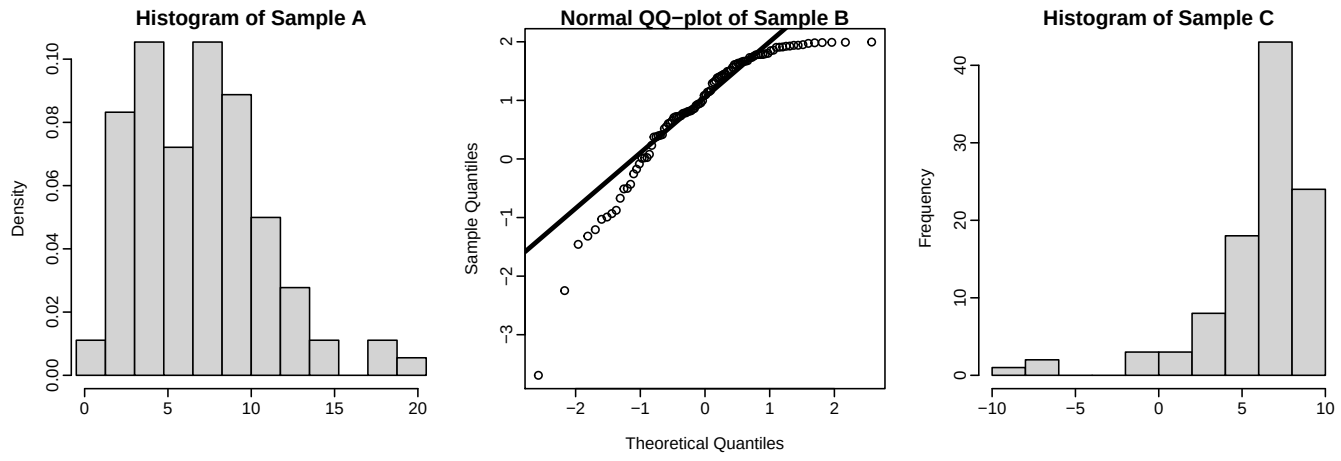

Figure 1: Histogram, normal QQ-plot, and histogram of samples A, B, C, respectively.

Question 2 [8 points]
Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with unknown continuous cumulative distribution function F with density f.

a. [1 point] Given a kernel K and a bandwidth $h > 0$, give the formula for the corresponding kernel density estimator $\hat{f}$.

b. [1.5 points] Explain the influence of the bandwidth on kernel density estimators. In particular, how does the shape of the estimate change when a small or a large bandwidth is chosen?

c. [1.5 points] Explain the influence of the kernel on kernel density estimators. In particular, how do the properties of the kernel influence the properties of the estimate?

d. [2 points] In the lectures we have discussed two approaches for making objective choices of the bandwidth. Choose one of them, describe roughly how it can be computed, give a motivation for the practicality of that approach.

e. [2 points] In the lectures we have discussed different approaches for making sure that the kernel density estimate assigns no mass to negative values of x if all $X_1, \ldots, X_n$ are positive. Two of these approaches were deemed reasonable; choose one thereof and summarize it.

Question 3 [9 points]
Let $X_1, X_2, \ldots, X_n$ ($n \geq 50$) be independent random variables that follow an unknown continuous cumulative distribution function F. Based on the Kolmogorov-Smirnov goodness-of-fit test, we would like to investigate whether F could be equal to a certain distribution F_0, i.e. we consider the simple null hypothesis $H_0 : F = F_0$.

a. [2 points] Describe the test statistic of the Kolmogorov-Smirnov test and explain what kind of test the Kolmogorov-Smirnov goodness-of-fit test is: a left-tailed, a right-tailed, or a two-tailed test.

b. [3 points] Suppose that the distribution of the Kolmogorov-Smirnov test is not implemented in your favorite software package. Describe the steps for approximating its distribution under the null hypothesis $H_0 : F = F_0$ by means of simulation.

c. [2 points] Describe the conceptual difference between the empirical and the parametric bootstrap.

d. [2 points] Explain the two types of errors that are usually involved when using a bootstrap procedure.