Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam I
Faculty of Science	March 31, 2023

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 [9 points]
Figure 1 contains four panels: a histogram of a sample A, a normal QQ-plot of a sample B, a histogram of a sample C, and a scatterplot of paired data (x, y).

a. [3 points] Describe the sample distribution of sample A. (See the top left panel of Figure 1.) Comment on three aspects, e.g., symmetry/skewness, modality, location parameter (approximately), range (approximately), outliers.

b. [2 points] Comment on the heaviness of the tails of the sample B distribution. (See the top right of Figure 1.)

c. [3 points] Sketch a boxplot of sample C and comment on some aspects of your boxplot. (See the bottom left panel of Figure 1 for a histogram of sample C.)

d. [1 points] Comment on the relation between the x- and the y-variables. (See the bottom right panel of Figure 1 for a scatterplot of the paired data (x, y).)

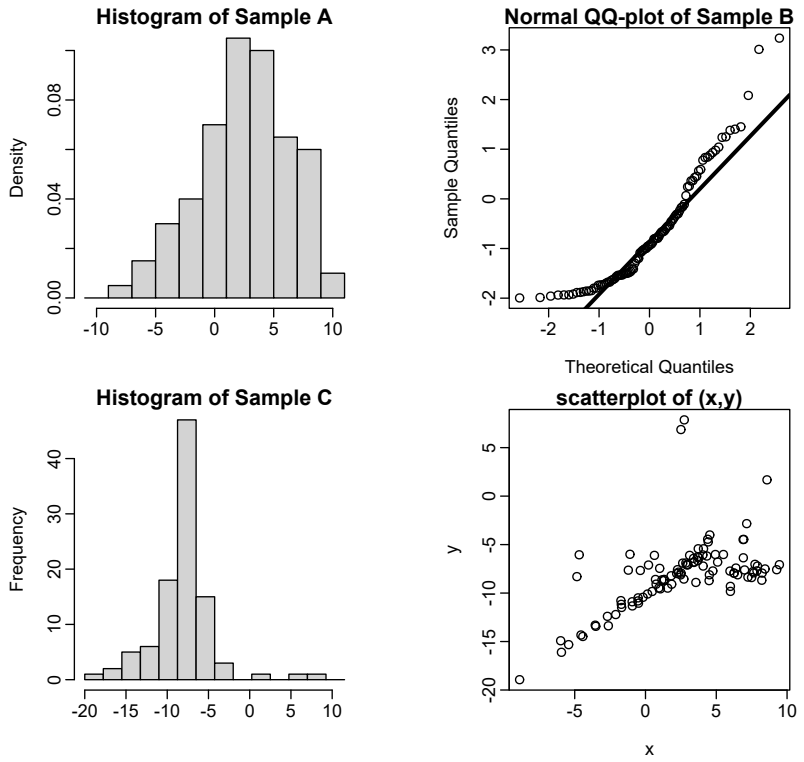

Figure 1: Histogram, normal QQ-plot, and histogram of samples A, B, C, respectively, and a scatterplot of paired data (x, y).

Question 2 [8 points]
In this question, we consider a cumulative distribution function F with continuous density f and with $F(0) = 0$.

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

b. [2 points] Explain how a kernel density estimate based on the uniform kernel differs from a kernel density estimate based on the Gaussian kernel, when the underlying sample is the same for both estimates.

c. [2 points] In the lectures, we have discussed the following choice of bandwidth that is "optimal" in a certain sense:

$$h_{opt} = \Big\{ \int K(x)^2 dx \Big\}^{1/5} \Big\{ \int (f''(t))^2 dt \Big\}^{-1/5} n^{-1/5}.$$

Explain in what sense this bandwidth is optimal; also provide a formula for the specific "error" criterion that is considered here.

d. [1 point] The middle term in the formula provided in c. depends on the unknown density f. Explain what choice(s) can be made for this term when one wishes to use the h_{opt}-formula in practice.

e. [2 points] Explain the symmetrization approach for ensuring that no mass is assigned to $\hat{f}(x)$ for negative values of x; assume for this that a random sample $x_1, \ldots, x_n$ from F is at your disposal.

Question 3 [10 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 chi-squared goodness-of-fit test, we would like to investigate whether F could be equal to a certain chosen distribution F_0, i.e., we consider the simple null hypothesis $H_0 : F = F_0$.

a. [3 points] Describe in detail the test statistic of the chi-squared test and explain what kind of test this is: a left-tailed, a right-tailed, or a two-tailed test.

b. [1 points] In this context of the chi-squared test procedure, describe the rule of thumb for justifying the approximation of the distribution of the test statistic under the null hypothesis by the chi-squared distribution.

c. [3 points] If the rule of thumb in part b. is not satisfied, the parametric bootstrap can be used to approximate the distribution of the test statistic of the chi-squared test under H_0. Describe all steps of the simulation method.

d. [3 points] Explain what kinds of bootstrap errors there are and how they are affected by the choice of bootstrap method (empirical or parametric).