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

EXEMPLARY SOLUTIONS

Question 1 [9 points]

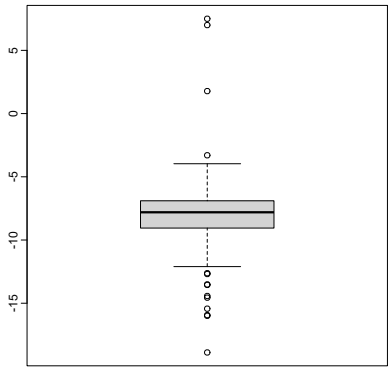
a. Sample A is to be unimodal. It is left-skewed. The location parameter (i.e. center of the distribution, e.g. sample mean or median) is somewhere between 0 and 5. The range of the distribution is about -10 to 11. There are apparently no outliers.

b. Sample B distribution has a lighter left tail than the normal distribution (because the lower-left part of the plot is above the straight line).

Sample B distribution has a heavier right tail than the normal distribution (because the top-right part of the plot is above the straight line.)

c.

On the right you can see what the true boxplot looks like. There are quite some "outliers" in the smaller numbers, and a few in the bigger numbers. The tails are quite heavy, e.g., because the whiskers are longer than the box. Apart from some very small values, the boxplot is rather symmetric.



d. Many data points seem to follow an upwards trend, so there seems to be a positive correlation between x and y.

Question 2 [8 points]

a. $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{t-X_i}{h})$

b. The estimate based on the uniform kernel will have many discontinuities and slopes of 0 between data points, whereas the estimate based on the Gaussian kernel will be very smooth.

c. This optimal bandwidth is the minimizer of the mean integrated squared error (MISE). It is defined as $MISE = \int E[(\hat{f}(t) - f(t))^2]dt$.

d. One could make a parametric assumption to get (hopefully) reasonable proxies for the integral. For example, if a location-scale family w.r.t. a certain distribution is assumed, the only thing left to do is then to involve an estimator of the scale parameter.

e. Given the sample $x_1, \ldots, x_n$, we extend the sample by also including $-x_1, \ldots, -x_n$, and then compute the kernel density estimate of the enlarged sample. Next, the mass the left of $t = 0$ is set to 0 and the mass to the right of $t = 0$ is doubled.

Question 3 [10 points]

a.

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i are the observed numbers of data points in the interval I_i, and $E_i = n \cdot F_0(I_i)$ are the expected numbers of data points in the interval I_i under H_0. $I_1, \ldots, I_k$ are a partition of $(-\infty, \infty)$ into k intervals. It is a right-tailed test.

b. All E_i have to be at least 5.

c. For simulating the distribution of the test statistic, we do the following steps a large number of times (e.g. B=1000 times):

- Generate a data set of size n according to F_0 (all realizations independently).
- Based on this drawn sample, re-compute the χ^2-score.

Use the collection of all B realized scores to approximate the null distribution of the test statistic.

d. The first type of bootstrap error results from the approximation of the unknown distribution P by an estimate $\hat{P}$.

Depending on whether a certain parametric choice is appropriate for describing P, the parametric bootstrap might lead to a smaller first bootstrap error than the empirical bootstrap.

The second type of bootstrap error results from the number of bootstrap samples drawn to approximate the distribution of the target statistic. (This error is in general not affected by the choice of bootstrap method.)