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

EXEMPLARY SOLUTIONS

Note: solutions other than those presented in this document might be acceptable as well.

Question 1

a. A symplot displays the distances of the smaller (sorted) half of the data points to the median against the distances of the greater (sorted) half of the data points to the median.

From this one can usually conclude whether the sample distribution is symmetric (if the plot follows the straight line $y = x$) or skewed (otherwise).

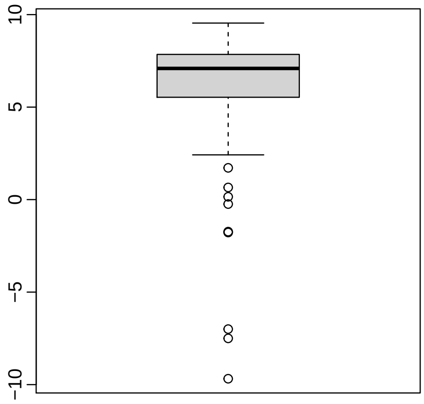
b. Sample A seems to be bimodal (with a small gap between 15-17) or tri-modal. It is right-skewed. The location parameter (i.e. center of the distribution, e.g. sample mean or median) is somewhere around 8-10. The range of the distribution is about 0-20.

c. Sample B distribution has a heavier left tail than the normal distribution (because the lower-left part of the plot is below the straight line).

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

d. (On the right you can see what the true boxplot looks like.)

- The median is in the upper part of the box.
- The lower whisker is longer than the upper whisker.
- There are extremely small outliers (at least one).
- The box is concentrated around the larger values.



Question 2

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

b. For very small bandwidths the KDE can have lots of modes and one would be able to see more clearly where the different sample points are located.

For very large bandwidths the KDE will be very smooth and almost ressemble the kernel function. (Smoothness of course also depends on the kernel...)

c. The kernel determines the shape of the "bumps" around the data points. Many properties of the kernel directly carry over to the KDE, for example, continuity, differentiability etc.

d. First option: The "optimal" bandwidth as the minimizer of the MISE.

For this, a trade-off between the variance and the bias-term must be made.

This optimal consists of a part that depends on the second derivative of the unknown density, a part that depends on the kernel, and the sample size (to a negative power).

—

Second option: The minimizer of the ISE.

The ISE consists of different parts one of which depends on the unknown density; it can be estimated with the help of cross-validation.

For the cross-validation, we leave out one observation at a time and compute the KDE based on all others. Finally, an average is taken.

e. First option: symmetrization i.e. for each data point x_i include $-x_i$ in the sample. Compute the KDE based on the symmetrized sample. Fold the mass that was assigned to negative values of x back to the positive real half-axis.

—

Second option: via transformation i.e. use e.g. the log-transformation to transform the purely positive sample into a general real-valued sample. Compute the KDE based on the transformed sample. Transform the KDE back to the positive real half-axis.

Question 3

a. KS-statistic = maximum distance between the empirical and the model cumulative distribution function, i.e. the cumulative distribution function F_0 under H_0. The test is right-tailed, i.e. we reject for large values of the statistic.

b. 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.

- Based on this drawn sample, re-compute the KS-score.

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

c. For the parametric bootstrap, one first has to estimate the parameters of the parametric class, say $\mathcal{F}$, (and we would assume that $F \in \mathcal{F}$) then draw bootstrap samples from the estimated parametric distribution. For the empirical bootstrap, one draws the bootstrap samples from the empirical distribution of the original sample.

d. The first error results from approximating the unknown distribution of the data with the help of the empirical distribution or an estimated parametric distribution.

The second error comes from approximating the distribution of a statistic by a collection of bootstrap-realizations of that statistic.