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

SOLUTIONS

Question 1

a. The top-right of the QQ-plot relates to the right tails of distributions.

"If the points in the top-right of a QQ-plot lie above the best-fitting straight line, then the (sample) distribution whose quantiles are given on the vertical axis seems to have a heavier right tail compared to a (hypothetical) distribution whose quantiles are given on the horizontal axis."

b. The SW-test is a test about a specific LSF, not about a big class such as that of symmetric distributions.

"The Shapiro-Wilk test tests whether the true distribution underlying a sample is normal."

c. The properties of the kernel density estimator strongly depend on the properties of the chosen kernel function.

"Kernel density estimates are smooth if and only if the chosen kernel function is smooth."

d. The estimated member of a parametric family is used in a step prior to the approximation of a statistic's distribution.

"The parametric bootstrap uses an estimated member of a parametric family and uses samples of it to approximate the distribution of a statistic of interest."

Question 2

a. Quantile functions of two members of a LSF are connected in a linear relationship.
Hence, not the shape of the QQ-plot changes but only the location and scale on the vertical axis.

b. The Kolmogorov-Smirnov test statistic compares the empirical distribution function to the hypothetical one *at each point* and then takes the *maximum*,
whereas the Chi-squared test compares the masses of the empirical distribution with those of the hypothetical one *over a bunch of intervals* and then takes the *sum* (of weighted squares).

c. h_{opt} is optimal in the MISE-sense.

The MISE is $MISE(\hat{f}) = \int (\hat{f}(x) - f_0(x))^2 dx$, where f_0 is the true density.

d. CV is used for estimating the unknown integral $\int \hat{f}(t) f(t) dt$.
$\int \hat{f}(t) f(t) dt = E(\hat{f}(Y) | X_1, \ldots, X_n)$, where Y and $X_1, \ldots, X_n$ are i.i.d.

Question 3

a. A correct answer would be based on a combination of the following:

The sample seems to be quite right-skewed (histogram, KDE, and symplot). One should thus not use a symmetric distribution (e.g. normal) as a model.

Among the right-skewed distributions the lognormal distribution seems to show the best fit to a straight line in the QQ-plot in both the lower-left and the upper-right part (except for two very large points).

The $\chi^2_1 0$-distribution shows a quite good fit for the left tails but not for the right tails.

And the $Exp(1)$-distribution shows a good fit for the right tails but not for the left tails.

This is why we choose the lognormal distribution as a model, i.e. it seems reasonable to assume that the data originates from a distribution in the location-scale family of shifted and scaled $logN(1, 0.65)$-distributions.

b. For $logN(1, 0.65)$-distribution.

First possibility: Graphically it seems that the y-axis intercept of the straight line is very close to 0, hence $a \approx 0$. The scale parameter is equal to the slope which is about $b \approx 16/10 = 1.6$. (Such an argument is only heuristic, so such an answer would not get you full points.)

Second possibility: Based on computations; $\hat{\sigma} = b\sqrt{5.928} \Leftrightarrow b \approx 3.818/2.435 \approx 1.568$.

Consequently, $\bar{x} = a + b3.358 \Leftrightarrow 5.581 - 1.568 \cdot 3.358 \approx 0.3157$.

c. The data distribution is right skewed as apparent from basically all of the plots above.

For right-skewed distributions the (sample) mean is usually to the right of the (sample) median.

Hence, the sample median must be 4.734.

Question 4

a. for a large number of times (e.g. $B = 1000$),

generate a sample from the empirical distribution $x_1^*, \ldots, x_n^* \overset{i.i.d.}{\sim} \hat{P}_n$, i.e. independent drawing with replacement from the original sample,

and recalculate the value of the statistic: $T_i^* = T_n(x_1^*, \ldots, x_n^*)$.

Take the empirical $\alpha/2$ and $1 - \alpha/2$-quantiles of the empirical distribution of $T_1^*, \ldots, T_B^*$, i.e. $T^*_{(\lfloor \alpha B/2 \rfloor)}$ and $T^*_{(\lceil B - \alpha B/2 \rceil)}$.

The confidence interval is obtained as $[2T_n - T^*_{(\lceil B-\alpha B/2 \rceil)}, 2T_n - T^*_{(\lfloor \alpha B/2 \rfloor)}]$.

b. First error: approximation of the true underlying distribution function P by some estimated distribution, $\hat{P}_n$.

Second error: approximate the estimated distribution of T_n, say $Q_{\hat{P}_n,n}$ by the empirical distribution function of the bootstrapped statistics, say, $\hat{Q}_{\hat{P}_n,n}$.