Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam I
Faculty of Sciences	30 March 2017

SOLUTION

Question 1 [8 points]

a. [2 points] Incorrect. The correct formula is $[T - Z^*_{([(1-\alpha)B])}, T - Z^*_{([\alpha B])}]$ or, equivalently, $[2T - T^*_{([(1-\alpha)B])}, 2T - T^*_{([\alpha B])}]$.

b. [2 points] Incorrect. The non-bootstrap version of Kolmogorov–Smirnov test can only be used to test a simple hypothesis about a fixed *continuous* probability distribution.

c. [2 points] Correct. See Figure 5.2.b) in the syllabus.

d. [2 points] Incorrect. This is true only if the line is $y = x$.

Question 2 [7 points]

a. [2 points] The Shapiro–Wilk test is meant for testing normality. The null hypothesis is $H_0\colon P \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$.

b. [1 point] It is a multiple of the sample variance, $(n-1)S_X^2$.

c. [1 point] The possible values of W are $(0, 1]$.

d. [3 points] We need to simulate the distribution of W under H_0 for this situation. Since the test is nonparametric (the distribution of W is the same for all samples of size n from any normal distribution), we can simulate samples from any fixed normal distribution, e.g., the standard normal distribution. We proceed as follows:

   - Generate B times a sample $X_1^*, \ldots, X_n^*$ from $N(0, 1)$.
   - Compute $W(X_1^*, \ldots, X_n^*)$ for each of the B bootstrap samples: $W_1^*, \ldots, W_n^*$.
   - Compute the bootstrap p-value: $p = \#(W_i^* : W_i^* < W)/B$ with $W = W(X_1, \ldots, X_n)$, the value of the test statistic for the given sample.

Question 3 [5 points]

a. [1 point] The best straight line seems to be the one in the QQ-plot against the standard exponential distribution. Therefore this is the best choice. (χ_4^2 is also quite good, if properly motivated).

b. [2 points] If X follows the standard exponential distribution, and $Y = a + bX$ then $\mathrm{E}Y = a + b\mathrm{E}X = a + b$ and $\operatorname{Var} Y = b^2 \operatorname{Var} X = b^2$. Equating these theoretical values to the sample values yields:

$$a + b = 2.150$$
$$b^2 = 1.083.$$

   Solving this yields $b = 1.041$ and $a = 1.109$. Similar values can be obtained by finding the intercept and the slope of the best straight line in the QQ-plot. (In case of χ_4^2 the values are $a = 0.678$ and $b = 0.368$).

c. [2 points] Because the sample is skewed to the right, the median is smaller than the mean. Therefore, the sample median is 1.933.

Question 4 [7 points]

a. [3 points] Given a sample $X_1, \ldots X_n$ from the Poisson distribution with rate λ the empirical bootstrap estimate of the standard deviation of $T_n = S_X^2$ is found by estimating the distribution Q_P of T_n by the following two steps

(i) Estimate P by $\hat{P}_n$, the empirical distribution of the sample $X_1, \ldots, X_n$, and, hence, Q_P by $Q_{\hat{P}_n}$.

(ii) Estimate $Q_{\hat{P}_n}$ by the empirical distribution of a sample $T_1^*, \ldots T_B^*$ from it.

In computational steps this scheme equals:

(I) Generate B times a sample $X_1^*, \ldots X_n^*$ from the empirical distribution of the sample $X_1, \ldots, X_n$.

(II) Generate for each X^*-sample $T^* = T_n(X_1^*, \ldots X_n^*)$. This yields the bootstrap values $T_1^*, \ldots T_B^*$.

The bootstrap estimate of the standard deviation of T_n is found in both schemes by the last step:

(iii) Estimate the standard deviation of T_n by the sample standard deviation of the bootstrap values $T_1^*, \ldots T_B^*$.

b. [2 points] The two errors are estimating P by $\hat{P}_n$ and estimating $Q_{\hat{P}_n}$ by the empirical distribution of a sample $T_1^*, \ldots T_B^*$. The second error can be made arbitrarily small by increasing the value of B.

c. [2 points] Instead of using the empirical distribution $\hat{P}_n$ of the original sample $X_1, \ldots, X_n$, we use the distribution $P_{\hat{\lambda}}$.