Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam I Solution
Faculty of Sciences	28 March 2019

Question 1 (With *exemplary* motivations.)

a. Incorrect. There are asymmetric distributions with a skewness of zero.

b. Incorrect. It can be used to test any simple null hypothesis about the data distribution.

c. Correct. For asymptotically normal estimators, we have $A(F) = \int IF(y, F)^2 dF(y)$. Typically (not always!), more robust estimators hence have a smaller variance. In general, one has to make a trade-off between robustness and efficiency.

d. Correct. There are two kinds of bootstrap errors and at least the second error is always made: when simulating the distribution of a statistic, only finitley many iterations can be done by a computer, hence there is an approximation error.

Question 2

a. Shapiro-Wilk tests the composite 0-hypothesis $H_0 : X_1 \sim N(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0$. Kolmogorov-Smirnov tests the simple 0-hypothesis $H_0 : F = F_0$ for a fixed c.d.f. F_0.

b. No, it does not depend on the specific choice of F_0 because, as we have seen in the lecture, it is a nonparametric statistic under the null hypothesis if F_0 is continuous.

c. It makes more sense to generate samples from F_0 because this avoids the first (of the two) bootstrap error.

d. We repeat the following two steps a large number (e.g. $B = 1000$) of times:

   generate n numbers from the distribution F_0 to get a bootstrap sample $x_1^*, \dots, x_n^*$.

   calculate D_n based on $x_1^*, \dots, x_n^*$. Call the resulting values $D_1^*, \dots, D_B^*$.

   these $D_1^*, \dots, D_B^*$ can be used to find the p-value or the critical value of the test.

Question 3

a. The data distribution looks reasonably symmetric. The best straight line seems to be the one in the QQ-plot against the Laplace distribution. Also, the histogram indicates that the tails are heavier than those of a normal distribution. This is why the location-scale family with respect to the Laplace distribution seems to be the best choice.

b. If X follows the standard Laplace distribution and $Y = a + bX$, then $\mathrm{E}Y = a + b\mathrm{E}X = a$ (note that $EX = 0$ for Laplace distributions) and $\operatorname{Var} Y = b^2 \operatorname{Var} X = 2b^2$. Our aim is to equate these theoretical values to the sample values.

   Thus, we have to solve the equations

$$a = 7.2$$
$$\sqrt{2}b = 6.6 \ \Rightarrow\ b \approx 4.67.$$

c. The data distribution is a bit right skew as apparent from the histogram and the sample skewness. This is why the trimmed mean should have a value less than the sample mean, i.e. that the trimmed mean equals $t_{50,0.1} = 6.816 < 7.2$ seems to be most reasonable.

Question 4

a. Given a sample $X_1, \dots X_n$ from the lognormal distribution $\log N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$, any application of the parametric bootstrap requires in a first step that the data distribution, i.e. these unknown parameters μ and σ, is estimated. Hence, estimators $\hat{\mu}$ and $\hat{\sigma}$ of those values are required. The remaining steps:

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

   generate a sample $x_1^*, \dots, x_n^* \overset{i.i.d.}{\sim} \log N(\hat{\mu}, \hat{\sigma}^2)$

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

   Use the sample standard deviation of the resulting values $T_1^*, \dots, T_B^*$ to approximate the true standard deviation of T_n.

b. Another possible method to estimate the standard deviation of T_n is the empirical bootstrap but here we prefer the parametric bootstrap because it makes use of the *additional* information that the sample indeed comes from a lognormal distribution. This is why I expect the parametric bootstrap to work (at least a bit) better.

c. Let's say that we are interested in a statistic T_n. The two errors are estimating the data distribution P by some $\tilde{P}_n$ and estimating the approximated distribution of T_n, i.e. $Q_{\tilde{P}_n}$, by the empirical distribution of a bootstrap sample $T_1^*, \dots T_B^*$, i.e. by a $\tilde{Q}_{\tilde{P}_n}$.