VU University	Statistical Data Analysis
Faculty of Sciences	26 March 2015

SOLUTION

Question 1 [7 points]

a. [2 points] No. The empirical distribution of a sample is always a discrete distribution, and hence the distribution function will be a discontinuous (step) function instead of a continuous function.

b. [2 points] A two sample QQ-plot is a scatter plot of matching order statistics of two samples, whereas a two sample scatter plot can only be made of a bivariate sample $(x_1, y_1), \ldots, (x_n, y_n)$ and plots the points (x_i, y_i) for $i = 1, \ldots, n$.

c. [2 points] Since the exponential distribution is skewed to the right, trimming will lower the mean. The more trimming, the lower the value. Hence the 10%-trimmed mean is expected to be larger than the median, whcih is the 50%-trimmed mean.

d. [1 point] See syllabus, figure 5.3, top left plot (proportional to the identity function)

Question 2 [7 points]

a. [2 points] The intervals should be choses such that the number of observations expected under H_0 is at least 5 in each interval.

b. [1 point] If not fulfilled, the chi-square approximation of the distribution of the test statistic under th H_0 is not reliable.

c. [1 point] No. Since the sample size is only 8, we can only choose 1 interval, and the distribution of the test statistics would be χ^2_0, which is not possible

d. [1 point] No. The Shapiro-Wilk test is for testing the composite null hypothesis of normality, whereas here we want to test a simple null hypothesis, existing of only one normal distribution.

e. [2 points]

$$D_n = \sup_{-\infty<x<\infty} |\hat{F}_n(x) - F_0(x)|$$

where $\hat{F}_n$ is the empirical distribution function and F_0 is the hypothesized distribution function. In words: D_n is maximum vertical distance between $\hat{F}_n$ and F_0. In this case $F_0(x) = \Phi(x)$. The value of D_n for this found at $x \approx 0$, and equals approximately $D_n \approx 0.26$.

Question 3 [6 points]

a. [2 points] The histogram of the sample is skewed to the right, so the 10% trimmed is expected to be higher than the 30%-trimmed mean. Moreover, the 30%-trimmed mean is expected to be more robust, i.e. have smaller variance. Therefore the middle plot corresponds to the 10%-trimmed mean, since this histogram shows higher values and more spread than the right plot, which corresponds to the 30%-trimmed mean.

b. [3 points] The $(1-2\alpha)$ bootstrap confidence interval for a statistic T_n based on bootstrap values $T_1^*, \ldots, T_B^*$ equals

$$\left[2T_n - T^*_{[(1-\alpha)B]}, 2T_n - T^*_{[\alpha B]}\right]$$

For the 10%-trimmed mean the 95% interval is:
$[2 \times 3.76 - 5.37, 2 \times 3.76 - 2.52] = [2.15, 5.00]$.
For the 30%-trimmed mean the 95% interval is:
$[2 \times 3.10 - 4.61, 2 \times 3.10 - 2.23] = [1.59, 3.97]$

c. [1 point] The interval for the 30%-trimmed mean is smaller, which means that that estimator is more accurate, and hence preferred.

Question 4 [7 points]

a. [1 point] Because the QQ-plot against the χ^2_2 family shows the most straight line, that familiy is preferred.

b. [2 points] If X has the standard χ^2_2 distribution, we have $\mathrm{E}X = 2$ and $\mathrm{Var}X = 4$. For $Y = a + bX$ we get $\mathrm{E}Y = a + 2b$ and $\mathrm{Var}Y = 4b^2$. Matching this with the given sample values 1.66 (sample mean) and 7.34 (sample variance) we solve $1.66 = a + 2b$ and $7.34 = 4b^2$, yielding $a =$ and $b = 1.35$ and $a = -1.05$.

c. [1 point] I would prefer the empirical bootstrap because that is generally safer if we do not know anything about the background of the data.

d. [2 points] Given a sample $X_1, \ldots X_n \sim P$ the empirical bootstrap estimate of the **distribution** Q_P of $T_{n,\alpha}$ is found as follows:

- Estimate P by $\hat{P}_n$ the empirical distribution of the sample, and, hence, Q_P by $Q_{\hat{P}}$
- Estimate $Q_{\hat{P}}$ by the empirical distribution of a sample $T_1^*, \ldots T_B^*$ from it

In computational steps this scheme equals:

- Generate B times a sample $X_1^*, \ldots X_n^*$ by resampling with replacement from the initial sample $X_1, \ldots X_n$.
- Generate for each X^*-sample $T^* = T_{n,\alpha}(X_1^*, \ldots X_n^*)$. This yields the bootstrap values $T_1^*, \ldots T_B^*$.

The bootstrap estimate of the **standard deviation** of $T_{n,\alpha}$ is found in both schemes by the last step:

- Estimate the sd of $T_{n,\alpha}$ by the sd of the bootstrap values $T_1^*, \ldots T_B^*$.

e. [1 point] Because the sample is rather skewed to the right, the mean is not the best choice, since it is influenced a lot by the high values in the sample. Better to use a trimmed mean, e.g. 20%-trimmed mean.