Vrije Universiteit Amsterdam	Statistical Data Analysis, part I
Faculty of Sciences	24 March 2016

SOLUTION

Question 1 [8 points]

a. [2 points] Incorrect. In some cases the parametric bootstrap works better, e.g. in the case of estimating the variance of the statistic $T_n = \max(X_1, \ldots, X_n)$ with $(X_1, \ldots, X_n)$ a uniform sample.

b. [2 points] Correct. Both the x coordinates and y coordinates are non decreasing. The x coordinates are α-quantiles for increasing α and the y coordinates are order statistics of the sample.

c. [2 points] This is nonsense. A bivariate scatter plot is for a bivariate sample, wherease a plot of the empirical distribution function is for a univariate sample.

d. [2 points] Incorrect. The sample mean is also a M-estimator, but not robust.

Question 2 [7 points]

a. [1 point] The best straight line seems to be the one in the QQ-plot against χ_3^2. Therefore this is the best choice. (Exponential is also quite good, and not wrong to answer.)

b. [2 points] If X follows the χ_3^2 distribution, and $Y = a + bX$ then $\mathrm{E}Y = a + b\mathrm{E}X = a + 3b$ and $\mathrm{Var}Y = b^2\mathrm{Var}X = 6b^2$. Equating these theoretical values to the sample values yields:

$$\begin{aligned} a + 3b &= 3.89 \\ 6b^2 &= 14.34 \end{aligned}$$

Solving this yields $b = 1.55$ and $a = -0.76$.

c. [2 points] See Figure 1

d. [2 points] Because the sample is skewed to the right, the median is smaller than the mean and the other trimmed means are monotonically in between. The ordered 0, 10%, 20%, 30%, 40% and 50% trimmed means are 3.89, 3.22, 3.02, 3.01, 2.94, 2.89. Hence the 40% trimmed mean equals 2.94.

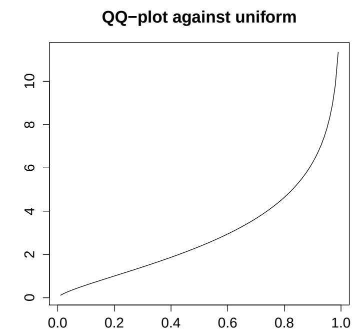

Figure 1: QQ-plots against uniform distribution.

Question 3 [7 points]

a. [3 points] Given a sample $X_1, \dots X_n \sim \text{uniform}[0, \theta]$ the parametric bootstrap estimate of the variance of $T_n = \max(X_1, \dots, X_n)$ is found by estimating the distribution Q_P of T_n by the following two steps

   (i) Estimate P by $P_{\hat{\theta}}$, the uniform distribution on $[0, \hat{\theta}]$, and, hence, Q_P by $Q_{P_{\hat{\theta}}}$.

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

   In computational steps this scheme equals:

   (I) Generate B times a sample $X_1^*, \dots X_n^*$ from the uniform distribution on $[0, \hat{\theta}]$.

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

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

   (iii) Estimate the variance of T_n by the sample variance of the bootstrap values $T_1^*, \dots T_B^*$.

b. [3 points] The two errors are estimating P by $P_{\hat{\theta}}$ and estimating by $Q_{P_{\hat{\theta}}}$ by the empirical distribution of a sample $T_1^*, \dots T_B^*$. The second error can be avoided here, since the variance of the maximum of a uniform sample is known when the underlying uniform distribution of the X_i's is known/estimated.

c. [1 point] No, the variance of bootstrap values of the sample maximum is a bootstrap estimate for the variance of the statistics T_n, whereas the maximum of bootstrap values of the sample variance is some quantile from the bootstrap estimate of the distribution of the sample variance (the latter is in general not very interesting).

Question 4 [5 points]

a. [2 points] The Kolmogorov-Smirnov test statistic equals

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

which is the maximum vertical distance found between the empirical distribution of the sample and the hypothesized distribution function F_0. The value of this statistic from the figure is around 0.3 (around y=2.8).

b. [3 points] k is the number of intervals, N_i is the observed number of observations in interval i, n is the total sample size and p_i is the probability mass in interval i under H_0. The rule of thumb is that $np_i \geq 5$ for each i, so that the χ^2-approximation is reliable.