Answers to Statistical Data Analysis, part I

27 March 2014

Question 1

a. Incorrect. If no parametric model applies, parametric bootstrap is not feasible, and empirical boostrapping is better. However, there are also situations where the parametric bootstrap works better than the empirical (cf. Assignment 5.2.)

b. Correct. The influence function is bounded (see Figure 5.2b in the syllabus).

c. Incorrect. In the separate plots the dependence between X_i and Y_i within pairs is lost, whereas that is visible in the scatter plot. (NB. There are other possible motivations for this answer).

d. Correct. An S-shape shows that the relative distances between quantiles in the tails of F_0 are bigger than in the distribution of the data. Hence, there is more mass in the tails of F_0.

Question 2

a. Yes, that is plausible since the QQ-plot shows a straight line.

b. Nothing, since we do not have a QQ-plot of x versus $N(0,1)$. (NB. Considering the spread in the x-coordinates of the elements of the QQ-plot it seems that x has a skewed distribution, so normality is unlikely.)

c. i) This test is suitable, since $H_0 : F = N(0.5, 1)$ is a simple hypothesis.

ii) This test is also suitable, for the same reason as in i).

iii) This test is not suitable, since it is for a composite hypothesis $H_0 : F \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ instead of the required simple hypothesis. In other words, it tests whether the sample comes from *a* normal distribution, and not from *the* normal distribution with mean 0.5 and variance 1.

d. This means that the distribution of T under H_0 does not depend on which distribution $F_0 \in \mathcal{F}_0$ is the true underlying distribution of the data.

Question 3

a. The data are skewed to the right. Therefore, the mean exceeds the median, and so will bootstrap values of the corresponding estimators do. So, estimator1 is the median, estimator2 is the mean.

b. The bootstrap confidence interval is given by $\left(2T - T^*_{n,[(1-\alpha)B]}, 2T - T^*_{n,[\alpha B]}\right)$, so the length of the interval is $T^*_{n,[(1-\alpha)B]} - T^*_{n,[\alpha B]}$. Therefore, using $\alpha = 0.025$:
CI for the mean: $[1.49, 2.78]$, length: 1.29;
CI for the median: $[0.68, 1.50]$, length: 0.82.

c. The median is preferred, as the corresponding confidence interval is shorter, hence it is more accurate (less variance).

Question 4

a. There are two possible answers:

(1) Estimate P by $\hat{P}_n$, the empirical distribution of the sample $X_1, \ldots, X_n$. And hence, estimate Q_P (distribution of T_n) by $Q_{\hat{P}_n}$.
(2) Estimate $Q_{\hat{P}_n}$ by the empirical distribution of a sample $T^*_1, \ldots, T^*_B$ from $Q_{\hat{P}_n}$.
(3) Estimate the standard deviation of T_n by the sample standard deviation of $T^*_1, \ldots, T^*_B$.

or

(1) Generate B times a sample $X^*_1, \ldots, X^*_n$ from the empirical distribution $\hat{P}_n$ of the sample $X_1, \ldots, X_n$.
(2) Compute $T^* = T_n(X^*_1, \ldots, X^*_n)$ for each of the samples generated in (1).
(3) Estimate the standard deviation of T_n by the sample standard deviation of $T^*_1, \ldots, T^*_B$.

b. Error 1: Estimate P by $\hat{P}_n$, which yields an error.

Error 2: Estimate $Q_{\hat{P}_n}$ by the empirical distribution of $T^*_1, \ldots, T^*_B$. If B is larger, this error decreases (but it is still present).

c. Error 1: No such error, since we simulate X^*-samples according to H_0.

Error 2: Yes, this error is still present, since we approximate the true distribution of the test statistic T by the empirical distribution of T^*'s. Again, increasing B will decrease the error (but will not remove it).