Vrije Universiteit Amsterdam	Statistical Data Analysis, Resit exam
Faculty of Sciences	7 July 2017

SOLUTION

Question 1 [2+2+2=6 points]

a. The underlying distribution seems symmetric, with minimum approximately equal to -2 and maximum approximately equal to 4. The median is approximately 0, and only the maximum exceeds 1.5 of the interquartile range (is plotted beyond the whiskers of the boxplot).

b. The QQ-plot against the standard Laplace shows the most straight line, and therefore this location-scale family is the most appropriate. (The normal location-scale family is also OK, but there should be a reasonable motivation why it is chosen instead of the Laplace family).

c. The location a and scale b can be approximated by the intercept and the slope, respectively, of the straight line in the chosen QQ-plot.

Question 2 [2+2+2+2=8 points]

a. Incorrect. The *expected* (under H_0) number of observed values should be at least 5.

b. Incorrect. They measure the *rank* correlation in bivariate data.

c. Incorrect. The Shapiro–Wilk test is used to test the claim that the data originate from *a* normal distribution, without specifying its parameters (the mean and standard deviation).

d. Correct. The t-test is more powerful than Wilcoxon test in case the underlying distributions are truly normal.

Question 3 [3+2+2=7 points]

a. Given a sample $Y_1, \dots Y_n \sim P$ the empirical bootstrap estimate of the **distribution** Q_P of $T_{n,\alpha}$ is found as follows:

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

(ii) Estimate $Q_{\hat{P}}$ 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 $Y_1^*, \dots Y_n^*$ by resampling with replacement from the initial sample $Y_1, \dots Y_n$.

(II) Generate for each Y^*-sample $T^* = T_{n,\alpha}(Y_1^*, \dots Y_n^*)$. This yields the bootstrap values $T_1^*, \dots T_B^*$.

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

(iii) Estimate the standard deviation of $T_{n,\alpha}$ by the standard deviation of the bootstrap values $T_1^*, \dots T_B^*$.

b. Since the data are skewed to the right, the mean (0% trim) is larger than the median (50% trim). Since the data is monotonically skewed, it follows that $T_{n,\alpha}$ decreases when α increases. Since the quantiles corresponding to the α_2-trimmed mean are lower than the corresponding quantiles for the α_1-trimmed mean it follows that $\alpha_1 < \alpha_2$.

c. The formula for the $(1-2\alpha)$ bootstrap confidence interval is $[2T - T^*_{\lfloor(1-\alpha)B\rfloor}, 2T - T^*_{\lfloor\alpha B\rfloor}]$, which equals $[2 \cdot 0.595 - 0.947, 2 \cdot 0.595 - 0.380] = [0.243, 0.810]$.

Question 4 [3+3+2=8 points]

a. The test statistic is $T = \#(X_i > 15)$ which has the binomial distribution with $n = 12$ and $p = 0.5$ if $m = 15.00$.

b. From the data we find: $T = t = 10$, and no observations are equal to 15.00. We compute the one-sided p-value, because H_1 is one-sided:

$$p_r = P_{p=0.5}(T \geq 10) = 1 - P_{p=0.5}(T \leq 9) = 1 - 0.981 = 0.019 < 0.05.$$

Hence, we reject the null hypothesis; there is enough evidence to state that the slogan is incorrect.

c. For the Wilcoxon signed rank test the underlying distribution is assumed to be symmetric. Symmetry can be checked either with a histogram, or a symplot.

d. Let T be the test statistic of a test that is *nonparametric*. Then the distribution of T under H_0 does not depend on specifically which distribution under H_0 is the true distribution.

Question 5 [2+2+2+1=7 points]

a. We have one sample of size $n = 20$. The underlying distribution is multinomial (4-nomial) with parameters 20, $p_{11}, p_{12}, p_{21}, p_{22}$ where

$$\sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij} = 1.$$

For that model the hypothesis of the independence of the row and column variables is tested (that is whether the severity of injury is independent of wearing seat belts). The hypotheses can be stated as

$$\begin{aligned} H_0 &: p_{ij} = p_{i\cdot}p_{\cdot j}, \qquad \text{for all } i = 1, 2,\ j = 1, 2, \\ H_1 &: p_{ij} \neq p_{i\cdot}p_{\cdot j}, \qquad \text{for some } i = 1, 2,\ j = 1, 2, \end{aligned}$$

where

$$p_{i\cdot} = \sum_{j=1}^{2} p_{ij}, \quad i = 1, 2, \qquad p_{\cdot j} = \sum_{i=1}^{2} p_{ij}, \quad j = 1, 2.$$

b. Fisher's exact test uses N_{11} as test statistic which has under the given H_0 a hypergeometric distribution with $n = 20$, $l = 6$, $m = 4$. Since the observed count in cell (1,1) equals 0 the left p-value equals

$$P(X = 0) = \frac{\binom{4}{0}\binom{16}{6}}{\binom{20}{6}} \approx 0.21.$$

This one-sided p-value is bigger than given significance level, so H_0 is not rejected. Alternatively, one may use $n = 20$, $l = 4$, $m = 6$ and

$$P(X = 0) = \frac{\binom{6}{0}\binom{14}{4}}{\binom{20}{4}} \approx 0.21.$$

c. The rule of thumb is: $\mathrm{E}N_{ij} > 1$ for all i, j and $\mathrm{E}N_{ij} > 5$ for at least 80% of the cells. In case of a 2×2 table $\mathrm{E}N_{ij} > 5$ for all i, j (since in this case there are four cells, and three are only 75% of the cells). Here we have $\mathrm{E}N_{11} = \frac{4 \cdot 6}{20} = \frac{24}{20} = 1.2$ so the rule of thumb does not hold. The chi-square test statistic is only approximately chi-square distributed under H_0. The approximation is reliable when the rule of thumb is satisfied.

d. In this case a bootstrap test could be used.

Question 6 [3+2+2+2=9 points]

a. The multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i$$

with

$$\begin{aligned} Y_i &- i^{th}\text{observed response value} \\ \beta_0, \ldots, \beta_p &- \text{unknown parameters} \\ x_{i1}, \ldots, x_{ip} &- \text{measured explanatory variables for } i^{th} \text{ observation} \\ e_i &- \text{error in } i^{th} \text{ observation} \end{aligned}$$

The assumption on the errors is $e_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. with $\sigma^2 > 0$ the unknown error variance.

b.
- **linearity** from scatter plots (Y, X_j) for $j = 1, \ldots, p$ and added variable plots
- **independence of errors** from context
- **normality of errors** from QQ-plot of residuals $R_Y(X)$ against the normal distribution
- **constant error variance** from scatter plots $(\hat{Y}, R_Y(X))$ or $(X_j, R_Y(X))$ for $j = 1, \ldots, p$

c. Multicollinearity is (nearly) linear dependence of explanatory variables, or more formally, amongst columns of the design matrix X. It may cause problems because it can raise the variances of the $\hat{\beta}_j$'s corresponding to columns of X involved in a collinearity. As a result, these estimates can be unreliable.

d. The observation with 'size of incoming data' approximately equal to 15 is a potential point since it is an outlier in the explanatory variable. In order to confirm that the corresponding value can be computed and if it is close to 1, the corresponding residual will be small (so the fit will be pulled towards a perfect fit for that observation regardless of the value of the response variable). To study the effect of a leverage point the Cook's distance for that point can be computed. If the Cook's distance is larger than 1, it is considered an influence point.