VU University	Statistical Data Analysis
Faculty of Sciences	29 May 2015

SOLUTION

Question 1 [6 points]

a. [2 points] The data x ranges between 0 and 6, with median approximately 2.7. The IQR ≈ 3. The shape is rather symmetric and uniform. There are no extremes.

b. [2 points] The uniform location scale family has the best fit, since that QQ-plot shows the most straight line.

c. [2 points] The range of x is roughly $[0, 6]$. Hence, compare to a standard uniform distribution on $[0, 1]$, we have $a = 0$,$b = 6$. This corresponds to the line visible in the QQ-plot $y = 0 + 6x$.

Question 2 [4 points]

a. [2 points] The empirical distribution functions of x and y have different slope, because the scale is different. The empicirical distribution function of y is the steepest. They cross in the middle of the plot, since the location (median) is approximately equal.

b. [2 points] No. A chi-square goodness of fit test is for testing a simple null hypothesis, that consists of one distirbution, based on one sample. The question is about a two sample problem.

Question 3 [8 points]

a. The steps to find the bootstrap estimates for the variances of T and S are:

1. i. Generate B times a sample $X_1^*, \ldots X_n^*$ by resampling with replacement from the initial sample $X_1, \ldots X_n$.
   ii. Generate B times a sample $X_1^*, \ldots X_n^*$ from the exponential distribution with parameter $\hat{\lambda}$, with $\hat{\lambda}$ the estimated parameter based on $X_1, \ldots, X_n$, e.g. $\hat{\lambda} = 1/\overline{X}$
2. Generate for each X^*-sample $T^* = T(X_1^*, \ldots X_n^*)$ and $S^* = S(X_1^*, \ldots X_n^*)$. This yields the bootstrap values $T_1^*, \ldots T_B^*$ and $S_1^*, \ldots S_B^*$.
3. Estimate the variances of T and S respectively by sample variance of the bootstrap values $T_1^*, \ldots T_B^*$ and $S_1^*, \ldots S_B^*$.

b. [2 points] The two errors are:

- estimate P by $\tilde{P}$,
- estimate the distribution of T resp. S by the empirical distribution of T^*'s resp. S^*'s

The first error is different in nature: under (i.) $\tilde{P} = \hat{P}_n$ a discrete distribution, under (ii.) $\tilde{P} = exp(\hat{\lambda})$ a continuous, parametric distribution.

Question 4 [6 points]

a. [2 points] The asymptotic variance is lowest for $\alpha = 0.5$, i.e. for the median. For this trimmed mean the asymptotic variance reaches the Cramer Rao lower bound. Hence it is an optimal estimator amongst the unbiased location estimators.

b. [2 points] Incorrect. It should be linear in the parameters β_j, but may contain any function of explanatory variables.

c. [2 points] This is a nonsense statement. The sign test is for one sampe, whereas Kendall's rank correlation test is for two paired samples.

Question 5 [6 points]

a. [2 points] The raw data does not follow a normal distribution. A normal distribution for the log transformed data of the unseeded data is also not plausible. THe fourth root transformed data shows two pretty linear QQ-plots in the bottom right corner of figure 3. Therefore, a t-test applied to these last two datasets seems ok. Hence, only the last p-value (0.0124) is reliable.

b. [2 points] The Wilcoxon two-sample test is based on ranks. The log transformation as well as the fourth root transformation is a monotone transformation. Hence the ranks are not changed by these transformation. Therefore, the value of the test statistic is the same in all three cases, as is the p-value.

c. [2 points]

- Kolmogorov-Smirnov two sample test. Appliciable for any difference in shape, requires no assumptions, expect for no ties, which is indeed the case here.
- median test, for testing whether underlying distributions are equal, without assumptions
- permutation test with a sensible test statistic. No assumptions. The permutation test permutes the labels seeded and unseeded. The test statistic should express difference in the two samples, e.g. mean(seeded)-mean(unseeded).

Question 6 [6 points]

a. [3 points] In practice one will draw two samples: one from accidents and one on the street. Hence, we have one sample per row (model B in syllabus). The first sample has size $N_{1\cdot} = 1124$, and has multinomial distribution with parameters $(1124, p_{11}, p_{12}, p_{13}, p_{14}, p_{15})$ with $p_{11} + p_{12} + p_{13} + p_{14} + p_{15} = 1$. The second sample has size $N_{2\cdot} = 429$, and has multinomial distribution with parameters $(429, p_{21}, p_{22}, p_{23}, p_{24}, p_{25})$ with $p_{21} + p_{22} + p_{23} + p_{24} + p_{25} = 1$. The

null hypothesis of homogeneity amongst the "not injured" and "injured" population is $H_0 : p_{ij} \equiv p_j$ for all i. This H_0 expresses the situation that the distribution amongst colours of helmets is the same for the "not injured" and "injured" population.

b. [1 point] The minimum of EN_{ij}'s equals 10.8, hence all EN_{ij}'s are larger than 5. Therefore, the rule of thumb for applying the chi-square test for contingency tables if fulfilled.

c. [1 point] $\chi^2_{(r-1)(c-1)} = \chi^2_4$.

d. [1 point] By computing the standardized residuals that follow a standard normal distribution under H_0. One can compare these standardized residuals (also called normalized contributions) to quantiles of the standard normal distribution.

Question 7 [9 points]

a. [2 points] The multiple linear regression model is:

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

with

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

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. [3 points] Use the t-test with test statistic $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$, which has the $t_{n-p-1} = t_{22}$-distribution under H_0, when the errors $e_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. The value of T is $-0.16/0.094 = -1.70$. The critical value for this two-sided test at level 0.05 is 2.07, hence, H_0 is not rejected.

c. [2 points] By removing an explanatory variable, the residuals will possibly increase in size (i.e. not decrease). Since $R^2 = 1 - RSS/SSY$ with RSS the sum of squared residuals, R^2 will decrease.

d. [2 points] Using the QQ-plot normality of the residuals can be investigated. Since the residuals are the estimated errors, one can (approximately) check the normality assumption of the errors using this plot. In Figure 4 the QQ-plot is quite straight for $n = 25$, hence the assumption is plausible for these data. If the assumption turns out to be false for a data set, the p-values of the tests are unreliable, since all tests are based on the normality assumption.