VU University	Statistical Data Analysis
Faculty of Sciences	28 May 2014

SOLUTION

PART 1

Question 1 [6 points]

a. [2 points] Location: minimum ≈ 0, maximum ≈ 8.8, median ≈ 1.4, IQR ≈ 2.2, data are skewed to the right, and the sample contains two high extremes (around 7.0 and 8.8).

b. [2 points] The QQ-plot against the exponential distribution shows the most straight line, and therefore that location-scale family is most appropriate.

c. [2 point] Fitting a line by hand in the QQ-plot yields approximately $y = 0 + 2x$, so location $a = 0$ and scale $b = 2$. Since we chose the exponential distribution this corresponds to the Exp(1/2) distribution.

Question 2 [5 points]

a. [1 point] The given expression is correct.

b. [2 points] Under the null hypothesis the test statistic has approximately a χ^2-distribution with $k-1$ degrees of freedom. This approximation is reliable if $np_i \geq 5$ for $i = 1, \ldots, k$. In other words: the expected number of observations under the null hypothesis has to be at least 5 in all intervals.

c. [2 point] One could you the one sample Kolmogorov–Smirnov test, which is also a goodness-of-fit test for a simple null hypothesis.

Question 3 [7 points]

a. [3 points] Given a sample $Z_1, \ldots Z_n \sim P$ 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 $Z_1^*, \ldots Z_n^*$ by resampling with replacement from the initial sample $Z_1, \ldots Z_n$.

- Generate for each Z^*-sample $T^* = T_{n,\alpha}(Z_1^*, \dots Z_n^*)$. 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:

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

b. [2 points] 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 α_2-trimmed mean is lower than the α_1-trimmed mean it follows that $\alpha_1 < \alpha_2$.

c. [2 points] The formula for the bootstrap confidence interval is $[2T - T^*_{\lfloor(1-\alpha_1)B\rfloor}, 2T - T^*_{\lfloor\alpha_1 B\rfloor}]$, which equals $[2 * 0.7 - 1.16, 2 * 0.7 - 0.45] = [0.24, 0.95]$.

PART 2

Question 4 [6 points]

a. [2 points] Incorrect. The distribution of the KS test statistic need not be normal if the data is from a normal distribution. In general, one uses bootstrap methods to mimic the distribution of a statistic because it is unknown (and certainly not normal).

b. [2 points] Correct. Hat values only show that an observation is potentially influential, whereas Cook's distances show the real influence.

c. [2 points] Correct. The a.r.e. of the Wilcoxon signed rank test to the sign test is 3/4 which is lower than 1. So the Wilcoxon signed rank test needs more data to obtain the same power as the sign test. In other words: the sign test has higher power.

Question 5 [5 points]

a. [4 points]

- sign test: not suitable since the sign test is for testing the underlying median of ONE sample.
- Wilcoxon two sample (rank sum) test : suitable.
- Kendall's rank correlation test: not suitable, since a rank correlation test is testing dependence in paired samples.
- Kolmogorov–Smirnov two sample test: suitable

b. [1 point] Since the Wilcoxon rank sum test is best suited for shift alternatives, it is expected not to have much power in this situation, because the biggest difference between x and y is in scale. The Kolmogorov–Smirnov test is suited for any differences between underlying distributions and is expected to have higher power in this situation.

Question 6 [7 points]

a. [3 points] This concerns one multinomial sample of size $n = 20$. The underlying distribution is multinomial$(20, p_{11}, p_{12}, p_{21}, p_{22})$ with $p_{11} + p_{12} + p_{21} + p_{22} = 1$. The null hypothesis of independence between handedness and having voted is $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$ with $p_{i\cdot} = pi1 + p_{i2}$ and $p_{\cdot j} = p1j + p_{2j}$..

b. [3 point] Fisher's exact test uses N_{11} as test statistic which has under the given H_0 a hypergeometric distribution with $n = 20$, $l = 7$, $m = 2$. Since the observed count in cel (1,1) equals 0 the left p-value equals

$$P(X = 0) = \frac{\binom{2}{0}\binom{18}{7}}{\binom{20}{7}} = 0.41.$$

This one-sided p-value is bigger than any realistic significance level, so H_0 is not rejected. Alternative, one may use $n = 20$, $l = 2$, $m = 7$ and

$$P(X = 0) = \frac{\binom{7}{0}\binom{13}{2}}{\binom{20}{2}} = 0.41.$$

c. [1 point] 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. Here we have $\mathrm{E}N_{11} = 2 \times 7/20 = 14/20 < 1$ so the rule of thumb does not hold. The chi-square test is therefore not applicable for these data.

Question 7 [9 points]

a. [3 points] The multiple linear regression model is:

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

with

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

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]

- **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 $\hat{e}$ against the normal distribution
- **constant error variance** from scatter plots $(Y, \hat{e})$, $(\hat{Y}, \hat{e})$

c. [3 points] The main problem to expect is collinearity. Apart from that there may be some influence points because of outliers in the `Armed.Forces` values. Remedies for collinearity: scale the design matrix, compute variation inflation factors, condition indices and variance decompositions, and based on this information take a selection of the collinear explanatory variables. Remedies for the influence points: compute hat values and Cook's distances.