Vrije Universiteit Amsterdam	Statistical Data Analysis, part I and part II
Faculty of Sciences	25 May 2016

SOLUTION

Question 1 [6 points]

a. [2 points] The underlying distribution is right-skewed and probably ranges approximately from 2 to 6. The median lies around 2.5. There are no extreme values.

b. [2 points] The location-scale family of exponential distributions seems most appropriate for this data, since the points in the QQ-plot of the data x against the standard exponential distribution follow approximately a straight line.

c. [2 points] Write $X = a + bU$, with $U \sim Exp(1)$ and X the underlying distribution of the data x. Since U has a positive density for positive values and x only has datapoints larger than or equal to 2, we see the $a = 2$. The points in the QQ-plot approximately follow the line $x = 2 + u$, so $b = 1$.

Question 2 [5 points]

a. [2 points] The Shapiro-Wilk test. This test tests for $H_0 : F \in \{\mathcal{N}(\mu, \sigma^2) : \mu \in R, \sigma^2 > 0\}$, a composite H_0, with which we can test for normality. The Kolmogorov-Smirnov test and the Chi-square-test are both tests with a simple H_0. So they are not suitable here.

b. [3 points] All tests are applicable for the hypothesis $H_0 : F = G$. But since the medians of both samples are approximately equal, the median test will not be able to distinguish differences in the underlying distributions. Because both distributions are more or less symmetric around their identical median, the Wilcoxon rank sum test will have the same problem, and will not see the difference. The Kolmogorov-Smirnov two-sample test is best here, since it can detect any difference between two underlying distributions.

Question 3 [7 points]

a. [3 points] Given a sample $X_1, \dots X_n \sim \text{Exp}(\lambda)$ the parametric bootstrap estimate of the standard deviation of T_n =median$(X_1, \dots, X_n)$ is found by estimating the distribution Q_P of T_n by the following two steps

   (i) Estimate P by $P_{\hat{\lambda}}$, the exponential distribution with parameter $\hat{\lambda}$, where $\hat{\lambda}$ is the estimated parameter based on $X_1, \dots X_n$, e.g. $\hat{\lambda} = 1/\bar{X}$, and, hence, Q_P by $Q_{P_{\hat{\lambda}}}$.

   (ii) Estimate $Q_{P_{\hat{\lambda}}}$ 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 exponential distribution with parameter $\hat{\lambda}$, where $\hat{\lambda}$ is the estimated parameter based on $X_1, \dots X_n$, e.g. $\hat{\lambda} = 1/\bar{X}$.

   (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 standard deviation of T_n by the sample standard deviation of the bootstrap values $T_1^*, \dots T_B^*$.

b. [2 points] The two errors are estimating P by $P_{\hat{\lambda}}$ and estimating by $Q_{P_{\hat{\lambda}}}$ by the empirical distribution of a sample $T_1^*, \dots T_B^*$.

c. [2 points] Bootstrap $(1-\alpha)$-confidence interval for the population median, based on T_n, looks as follows:

$$[2T - T^*_{([(1-\frac{\alpha}{2})B])}, 2T - T^*_{([\frac{\alpha}{2}B])}],$$

   where T =median$(X_1, \dots, X_n)$, $T_1^*, \dots T_B^*$ are the bootstrap values of T and $T^*_{(i)}$ is the i-th order statistic of T_i^*'s.

Question 4 [6 points]

a. [2 points] Mean (0-trimmed mean) is optimal for normal samples, since $\alpha = 0$ reaches the asymptotic Cramér-Rao lower bound, which means that the corresponding location estimator has the smallest asymptotic variance.

b. [2 points] Incorrect, since overall F-test considers $H_0 : \beta_1 = \ldots = \beta_p = 0$ against $H_1 : \beta_j \neq 0$ for some $j, 1 \leq j \leq p$. This means that at least one of the β_j's will differ from 0, but not necessarily all β_j's.

c. [2 points] Nonsensical ("apples and oranges"). The Wilcoxon signed rank test considers a one-sample problem and the Wilcoxon rank sum test is two-sample test.

Question 5 [6 points]

a. [2 points]

- The test statistic for the sign test is

$$T = \sum_{i=1}^{n} 1_{X_i > 4}$$

  where $n = 15$ and 1 is indicator function.
- The distribution of the test statistic under the assumption $m = 4$ is binomial with parameters $n = 15$ and $p = \frac{1}{2}$.

b. [3 points] From 15 given numbers, 6 of them are larger than 4: 4.47, 4.86, 5.09, 7.94, 9.05, 10.69. Therefore, the observed value of the test statistic is $t_{\text{obs}} = 6$.

The test is left sided. Indeed, under H_0 the test statistic T will have larger values which implies that H_0 is rejected for small values of T. Therefore, the p value is

$$P(T \leq t_{\text{obs}}) = P(T \leq 6) = 0.304,$$

which is obtained from Table 1, for $k = 6$ and $p = 0.5$.

c. [1 point] No. The Kolmogorov-Smirnov test is used to test the simple null hypotheses that the underlying distribution F is equal to a given distribution F_0. The sign test here is used to test the composite null hypothesis that the underlying distribution F belongs to a family of probability distributions with median $m \geq 4$.

Question 6 [7 points]

a. [3 points] We have three samples (of persons given drugs A, B, C respectively), each of size 50. For $i = 1, 2, 3$, the counts follow a multinomial distribution with parameters $(150, p_{i1}, p_{i2}, p_{i3}, p_{i4})$ with $p_{i1} + p_{i2} + p_{i3} + p_{i4} = 1$. The null hypothesis of homogeneity amongst the 3 drug populations can be stated as

$$H_0 : p_{ij} \equiv p_j \text{ for all } i.$$

This H_0 expresses that the incidence of types of allergic reactions is the same, regardless of whether drug A, B or C is assigned.

b. [1 point] Rule of thumb: $\mathrm{E}\, N_{ij} \geq 1$ for all i, j and $\mathrm{E}\, N_{ij} \geq 5$ for at least 80% of the cells. In this case, the smallest expected counts are in the first column. Under the null, $\mathrm{E}\, N_{i1} = 150 \frac{16}{150} \frac{50}{150} \approx 5.33 > 5$, so the rule of thumb is fulfilled.

c. [1 point] Under the null hypothesis, the chi-square test statistic follows the $\chi^2_{(r-1)(c-1)} = \chi^2_6$ distribution, for which the critical value equals 12.59 (the 0.95 quantile according to Table 2). The critical region is therefore $[12.59, \infty)$. The observed value (3.06) is not in this region. Hence, we do not reject the null hypothesis of homogeneity.

d. [1 point] Decreases. Let X^2 be the chi-square test statistic for the original data and let Y^2 be the test statistic for the data multiplied by 10. We have,

$$Y^2 = \sum_{i,j} \frac{(10N_{ij} - \mathrm{E}\, 10N_{ij})^2}{\mathrm{E}\, 10N_{ij}} = \sum_{i,j} \frac{100(N_{ij} - \mathrm{E}\, N_{ij})^2}{10\, \mathrm{E}\, N_{ij}} = 10X^2.$$

Since $X^2 > 0$ we have that $Y^2 > X^2$. Because the distribution is χ^2 the (right-sided) p-value decreases. (*Note that we have seen this in one of the assignments*)

e. [1 point] Stays the same. The number of degrees of freedom equals $(r-1)(c-1)$ and neither r (number of rows) nor c (number of columns) would change.

Question 7 [8 points]

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

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

with
Y_i : i-th observed response value
$\beta_0, \ldots, \beta_p$: unknown parameters
$x_{i1}, \ldots, x_{ip}$: measured explanatory variables for the 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. [2 points] Use the t-test with test statistic

$$T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)},$$

which has the $t_{n-p-1} = t_{13}$-distribution under H_0, when the errors $e_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. The value of T is $\frac{0.0044}{0.0012} \approx 3.67$. The critical value for this two-sided test at level 0.05 is 2.16, hence H_0 is rejected.

c. [2 points] By removing an explanatory variable, the residuals will possibly increase in size (i.e. not decrease). Since $R^2 = 1 - \frac{RSS}{SSY}$ with RSS the sum of squared residuals, R^2 will decrease. It will not decrease a lot though, because from b) we know that the variable birth weight does not add significantly.

d. [2 points]

- Normal QQ-plot of the residuals: can be used to investigate the normality of the residuals. Since the residuals are the estimated errors, one can (approximately) check the normality assumption of the errors.
- Scatter plot of the residuals against Y: can be used to check the constant error variance assumption.
- Scatter plot of Y against X: can be used to check linearity of the relation between Y and X