Vrije Universiteit Amsterdam	Statistical Data Analysis, Resit Exam
Faculty of Science	8 July 2022

EXEMPLARY SOLUTIONS

Question 1 [8 points]

a. The data are right-skewed, the data distribution is unimodal (except for a small gap around 13), the mode is about 3 or 4, the extremes are about 0 (min) and 20 (max).

b. The LSF with respect to χ^2_4 or $LN(\cdot, 0.55^2)$ both seem fine. Because the QQ-plots are quite straight for these distributions (except perhaps for some few extremely large points) (and the histogram also seems to suggest one of these right-skewed distributions).

c. We could draw a straight line that somehow fits the QQ-plot appropriately. The scale parameter is the slope and the location parameter is the y-axis intercept.

Or, preferably, with the exact method: We could equate the sample mean and standard deviation with the theoretical mean and standard deviation of the location scale family, and then solve for the location and scale parameters.

d. The sign test is a nonparametric test which has as its only assumption that the median is unique. Based on the available information and in general, it is difficult to decide whether this assumption is satisfied, but it rather safe to assume that it is satisfied. So, yes, the sign test is a safe choice.

Question 2 [10 points]

a. The rule of thumb says that the chi-squared approximation for the distribution of the test statistic under the null hypothesis is only accurate if the expected numbers of observations ($n \cdot p_i$, where $p_i = F_0(a_i) - F_0(a_{i-1})$) within each bin (expected under the null hypothesis) are all at least 5.

b. One test is the one based on the sample correlation (potentially conducted as a permutation test). Advantage: if the data are not corrupted by outliers, this test uses more information than the nonparametric tests, i.e. the test could be more powerful.

Another one is Spearman's rank correlation test. Advantage: it is nonparametric and thus exact under the null hypothesis of independence. Another advantage: ranks are robust, so the test statistic is not sensitive to outliers.

Yet another one is Kendall's rank correlation test. Advantage: it is nonparametric and thus exact under the null hypothesis of independence. Another advantage: this test is also very robust robust against outliers.

c. It can be tested whether the data originate from some normal distribution (mean and standard deviation unspecified).

d. For samples from normal distributions, the t test is more powerful (in terms of ARE) than the Wilcoxon test in detecting a shift in the location parameters.

Question 3 [7 points]

a. We do the following steps a large number of times, e.g. $B = 1000$ or more:

   - Draw a random sample, the bootstrap sample, $Y_1^*, \ldots, Y_{40}^*$ of size 40 from $Y_1, \ldots, Y_{40}$.
   - Recalculate $T_{40,0.2}$ based on the bootstrap sample, call it $T^*_{40,0.2,i}$ for iteration i.

   Take the sample variance of $T^*_{40,0.2,1}, \ldots, T^*_{40,0.2,B}$ as an estimate of the actual variance of $T_{40,0.2}$.

b. The 2-sided 95% confidence interval is given as $[2T_{40,0.2} - T^*_{40,0.2,([0.975 \cdot B])}, 2T_{40,0.2} - T^*_{40,0.2,([0.025 \cdot B])}]$. In our case, $[2 \cdot 0.66 - 0.99, 2 \cdot 0.66 - 0.42] = [0.33, 0.90]$.

c. The data distribution is very right-skewed. That's why the 30% trimmed mean will be smaller than the 20% trimmed mean.

Question 4 [10 points]

a. The added variable plot shows no clear linear correlation pattern between the residuals for the model with age as the response variable and the residuals for the model with width as the response variable. As a conclusion, this gives us no indication that age adds more information on width, next to sex and length, so we don't need to include it in the model.

b. The overall F test tests $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_a : \beta_j \neq 0$ for some $j \in \{1, 2, 3\}$. Based on the information from the R-output, the p-value of that test is 0.000059, which is less than 0.05. As a conclusion, we reject the null hypothesis which means that at least one variable should be included in the model.

c. The multiple R squared of 0.47 means that 47% of the variation in the width values can be explain with the help of age, length, sex.

d. Based on the sample correlation and the right plot in Figure 4, we see a rather strong linear relationship between age and length, at least for the subsample of boys. The variance inflation due to possible collinearity, however, is only marginally bigger than 1 (maximum: 1.26).

   The variance decomposition printed first however indicates that there are 3 (multi-)collinearities, involving the intercept and all three explanatory variables to some extent.

   Yet, after standardizing the explanatory variables, the secondly printed variance decomposition does not reveal these collinearity problems any more.

   As a conclusion, there are apparently no collinearity problems; the variance decompositions indicate that the explanatory variables are not really linearly dependent and the variance inflation factors reveal that there are no bad practical consequences in terms of parameter estimation if all variables are included in the model.

   (On a side note: it seems worthwhile to include an interaction term for sex and at least one of the other explanatory variables.)

Question 5 [10 points]

a. Model: $mult(39, p_{11}, p_{21}, p_{12}, p_{22})$
   $H_0 : p_{jk} = p_{\cdot k} \cdot p_{j\cdot}$ (i.e. independence) vs. $H_0 : p_{jk} \neq p_{\cdot k} \cdot p_{j\cdot}$ for some j, k (i.e. dependence of sex and handedness).

b. $X^2 = \sum_{\text{cell } (j,k)} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}$, where O_{jk} are the observed counts in the cells, and $E_{jk} = 39\hat{p}_{j\cdot} \cdot \hat{p}_{\cdot k} = N_{j\cdot} \cdot N_{\cdot k}/39$ are the expected counts under H_0.

c. Approximate distribution under H_0 if the rule of thumb is satisfied: χ^2 with $(2-1) \cdot (2-1) = 1$ degree of freedom.

d. Test score: $t = 15$. Right p-value: $P(T \geq 15)$. Based on the bootstrap realizations, the p-value is 0.61. This is greater than 0.05, hence we cannot reject the null hypothesis that sex and gender are independent; there is not enough evidence for this.