Vrije Universiteit Amsterdam	Statistical Data Analysis, resit exam
Faculty of Sciences	5 July 2019

SOLUTIONS

Question 1 [2+2+2+2+2=10 points]

a. Incorrect. The most extreme 20% are ignored (10% upper, 10% lower values).

b. Correct. The ρ-function has to be chosen as the (parametric) density, i.e. $\rho = f$. Then the maximizer of the expression $\prod_{i=1}^{n} \rho(X_i - \theta)$ maximizes the likelihood. It is hence the maximum likelihood estimator.

c. Incorrect; this depends on the choice of the bins and the sample size. Every bin should have at least 5 expected observations under the null hypothesis, then the χ^2-goodness-of-fit test can be used (that's the rule of thumb).

d. Correct. The requirement for the signed rank test is that the data are symmetrically distributed.

e. Incorrect. The distribution of the test statistic is a hypergeometric one (with parameters given by the marginal counts).

Question 2 [1+2+3+1+1.5=8.5 points]

a. The location-scale family with respect to the standard lognormal distribution seems to be the best model family for the data; the line in the corresponding QQ-plot is quite straight (and the histogram ressembles the lognormal density).

b. Denote by Z a random variable which has a standard lognormal distribution.

Solving the formulae $77 = \bar{x} = a + bE(Z) = a + 1.65b$ and $7046 = b^2 Var(Z) = 4.67b^2$ for a and b gives $b = \sqrt{7046/4.67} \approx 38.84$ and $a = 77 - 38.84 \cdot 1.65 \approx 12.91$.

c. The empirical bootstrap estimate of the distribution Q_P of T 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 $Z_1^*, \dots Z_n^*$ by resampling with replacement from the initial sample $z_1, \dots z_n$.

(II) Generate for each Z^*-sample $T^* = T(Z_1^*, \dots Z_n^*)$. This yields the bootstrap values $T_1^*, \dots T_B^*$.

(III) Estimate the distribution of T by the empirical distribution of the bootstrap values $T_1^*, \dots T_B^*$.

d. This negative value of T means that `x` is right-skew because its negativity implies that the mean of `x` is (much) greater than the median of `x`. (And such a thing is a characteristic of a right-skewed sample.)

e. The sign test is a quite robust test which in our case means that it does not assume anything about the symmetry or skewness of the data. Hence, it is safe to apply the sign test to our right-skew dataset `x`.

Question 3 [1+1.5+3+2+1+1.5=10 points]

a. H_0: $F = F_0$ vs. H_a: $F \neq F_0$; for a pre-specified cumulative distribution function F_0;

b. A nonparametric test does not depend on parametric assumptions about the data distribution.

This is the case if it uses a test statistic the distribution of which does not depend on the specific parametric family under the null hypothesis.

c. (I) Randomly permute B times the observations of both samples to obtain (each time) permuted samples $X_1^\pi, \ldots, X_n^\pi$ and $Y_1^\pi, \ldots, Y_m^\pi$.

(II) Generate for each such obtained pair of samples $D(X_1^\pi, \ldots, X_n^\pi; Y_1^\pi, \ldots, Y_m^\pi)$. This yields the permutation values $D_1^\pi, \ldots D_B^\pi$.

(III) Estimate the distribution of D by the empirical distribution of the permutation values $D_1^\pi, \ldots D_B^\pi$.

d. H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$; here ρ is the population correlation coefficient,

Spearman's test only uses the ranks of the X-values and the ranks of the Y-values.

e. Spearman's test has the advantage that it is much more robust than Pearson's test because Pearson's test statistic is very sensitive to outliers, whereas Spearman's is not.

f. If are(Spearman, Pearson) = 0.5, this means that, in comparison to Pearson's test, Spearman's test requires twice the sample size to obtain the same power as Pearson's test.

Question 4 [1.5+2+2+1.5=7 points]

a. Model II-C: 3 independent columns of categorical counts because we are interested in the differences in the probabilities among countries.

For each column j: $(N_{1j}, N_{2j}, N_{3j}, N_{4j}, N_{5j}) \sim Mult(N_{\cdot j}, p_{1j}, p_{2j}, p_{3j}, p_{4j}, p_{5j})$.

$H_0 : p_{i1} = p_{i2} = p_{i3}$ for all $i = 1, 2, 3, 4, 5$. against $H_a : p_{i1} \neq p_{i2}$ or $p_{i2} \neq p_{i3}$ or $p_{i3} \neq p_{i1}$ for some $i = 1, 2, 3, 4, 5$.

OR

H_0 : the citizens of each country have the same probabilies to vote for the specific political parties.

against H_1 : the citizens of each country have different probabilies to vote for the specific political parties.

b. Rule of thumb: all expected numbers under the null hypothesis have to be at least 1, and at least 80% of them have to be at least 5.

In the present situation: check smallest expected frequencies: $\frac{37 \cdot 21}{142} \approx 5.47 \geq 5 > 1$.

Hence, all of the expected frequencies are at least 5 and in particular greater than 1, which is why both conditions of the rule of thumb are met.

c. Under H_0, we use the following χ^2-approximation: $X^2 \sim \chi^2_8$, where the distribution has $(5-1) \cdot (3-1) = 8$ degrees of freedom.

The correct quantile from Table 1 is 15.51 (right-tailed test).

We see that $X^2 = 19.72 > 15.51$.

Hence, we reject the null hypothesis and conclude that there is a significant heterogeneity among the voting tendencies of the countries Germany, Netherlands, France.

d. The bootstrap for contingency tables uses the following idea: All counts are redistributed to the table in a way such that the marginal counts are the same as for the original table and the distribution to the cells is according to the null hypothesis (of independence / homogeneity).

Question 5 [2.5+1+1+2+2=9.5 points]

a. The highest R^2 value is achieved with the independent variable "infantdeath" ($R^2 = 0.7372$). It is also a significant variable ($p = 1.01e^{-08} < 0.05$), hence we add it to the empty model.

   The resulting model is

$$\texttt{lifeexpF} = \beta_0 + \beta_2 \texttt{infantdeath} + \texttt{error}$$

   with $\texttt{error} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

b. The variance inflation numbers are still rather small. Hence, these do not point to a multicollinearity problem.

c. All the condition indices are below 30. Hence, there is no collinearity apparent from these numbers.

d. Added variable plot: it can be used to check whether an additional variable has an influence on the dependent variable after having corrected for all the other variables.

   Scatterplot of a variable included in the model vs. residuals: this can be used to test the constant error variance assumption; constant with respect to the variable included in the model.

e. The added variable plot shows a strong negative linear relationship between the residuals. Hence, infantdeath should be added to the linear regression model in addition to the other variables.

   The scatterplot does not show a real trend.
   Hence, it seems relatively save to assume that the constant error variance assumption is met.

f. One can fit a new linear model in which just the first observation receives a non-zero dummy explanatory variable.

   Then it is tested (t-test) whether the parameter that belongs to this dummy variable is different from zero.