Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam II
Faculty of Science	26 May 2021

SOLUTIONS

Below follows a summary of possible answers to the exam questions. These solutions are for your reference only and, as such, you may be expected to provide more complete answers in the exam.

Question 1 [4.5 points]

a. [1.5 points] This is correct.

   Motivation: "The test is non-parametric because it does not specify a parametric family for the (common) X, Y distribution", and "The test is distribution-free because the test statistic under the null does not depend on the distribution of the X, Y." alternatively, "only depends on the ranks of the Y observations within the full X, Y sample but not the distribution of X, Y.")

b. [1.5 points] This is incorrect.

   Motivation: "Continuous variables can be discretised into two categories each to construct the 2×2 contingency table."

c. [1.5 points] This is incorrect. Motivation: a) "Anscombe's quartet is an example of how different datasets can lead to the same $\hat{\beta}_0, \ldots, \hat{\beta}_p, \hat{\sigma}^2$, and $\mathcal{R}^2$," or b) "Even just the $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are not unique if the underlying design matrix is not full rank," or c) other reasonable counter-example.

Question 2 [5.5 points]

a. [1.5 points] If F is strictly increasing, then $m_q = F^{-1}(q)$ and therefore all quantiles are unique, so also the median. If F corresponds to a continuous distribution then the probability that an observation equals 0 is zero and so no observations will equal the proposed median.

b. [2 points] In this case, the three observations are dropped and we use $T = \sum_{i=4}^{20} 1\{X_i > 0\}$ which (conditional on there being 3 observations equal to the median under the null) has a $\text{Bin}(17, 1/2)$ distribution under the null.

c. [2 points] If under the null $m_{0.8} = 0$, then under the null $1\{X_i > 0\} \sim \text{Ber}(0.2)$ and so, since the observations are i.i.d., then T has a $\text{Bin}(20, 0.2)$ distribution under the null.

Question 3 [5 points]

a. [1.5 points] Model A is appropriate.

   The hypotheses are H_0 : *the variables hair colour and eye colour are independent* and H_1 : *the variables hair colour and eye colour not independent.* You can also formulate the hypotheses in terms of the $p_{i,j}$ as in the syllabus.

b. [2 points]

   i) The expectation under the null must be greater than 1 for every cell and greater than 5 for at least 80% of the cells (equivalently, 12 cells or more);

   ii) These can be obtained from the totals.

$a = 64 - 29 - 14 - 16 = 108 - 68 - 20 - 15 = 5$,

$b = 220 - 68 - 119 - 26 = 7$,

$c = 93 - 15 - 54 - 14 = 10$,

$d = 592 - 108 - 286 - 127 = 71$.

iii) to compute the expectations under the null, only the marginals are needed so we only need to know d.

c. [1 point] If $N_{i,j}$ is the (i,j) entry of the contingency table, $N_{i,\cdot}$ and $N_{\cdot,j}$ the respective row- and column totals, and n the overall total, then

$$n\hat{p}_{i,j} = n\frac{N_{i,\cdot} \times N_{\cdot,j}}{n^2} = \frac{N_{i,\cdot} \times N_{\cdot,j}}{n}.$$

d. [2 points] Under the null hypothesis, X^2 is approximately distributed like a chi-square distribution with $(r-1)(c-1) = 3 \times 3 = 9$ degrees of freedom. The test that rejects H_0 when $X^2 > \chi^2_{9,0.95} = 16.92$ is a test of level $\alpha = 5\%$, so since $138.29 > 16.92$ we reject the null hypothesis at this significance level. We therefore conclude that for this underlying population, hair- and eye-colour are not independent.

e. [1.5 points] i) A downside of the bootstrap test is that it is computationally costly compared to the chi-square test. ii) This discrepancy may happen by chance but it might also be an indication that the chi-square approximation of X^2 is not a good; the bootstrap test always has (approximately) the correct significance level and is preferred.

Question 4 [5 points]

a. [3 points] The test is an F-test which, in this particular case, reduces to a t-test. For each $j = 0, \dots, p$, we test $H_0 : \beta_j = 0, \beta_k$ arbitrary versus $H_1 : \beta_j \neq 0, \beta_k$ arbitrary, where $k = 0, \dots, p$, $k \neq j$. In words: we compare the model with- and without each of the variables (and intercept.)

The step-down procedure would then remove the variable whose respective test has the highest p-value, except if the p-value is below the significance level, and we don't remove the intercept.

In this particular case, the step-down procedure prescribes that we remove the variable **critical** since it has the highest p-value above 0.05.

b. [1 point] Since at level 5% in all 5 models **complaints** is significant and the other variable being introduced into the model is not significant we stop the procedure and keep the model with the intercept and **complaints**.

c. [2 point] The coefficient being positive means that a higher value of **complaints** leads on average to higher **rating** and, specifically, an increase of one unit in **complaints** leads, on average, to an increase of 0.75461 units in **rating**.

The student can combine both answers into one for full points.

d. [2 points] The QQ-plot allows us to verify if the normality assumption on the noise is reasonable, and the fitted versus residual plots allows us to detect deviations from linearity and/or equal variance of the noise.

Subjectively, in this particular case, both plots look consistent with the model being appropriate.

e. [1 point] We flag a point as being an influence point if the respective Cook's distance is above 1. In this particular case there are, therefore, no influence points.