Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam II
Faculty of Science	2 July 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 [7.5 points]

a. [1.5 points] This is correct.

Motivation: The Epanechnikov kernel minimises (an upper bound of) the MISE (mean integrated squared error) of $\hat{f}$. As such, it can generally be considered a good choice for a kernel

b. [1.5 points] This is incorrect.

Motivation: Even the distribution of the sample mean cannot always be well approximated via the bootstrap; c.f. example 5.6 in the lecture notes. In general, both the distribution of the data and the specific statistic being bootstrapped influence the performance of the bootstrap.

c. [1.5 points] This is correct.

Motivation: If a permutation procedure is possible, then it provides exact control of type I error (albeit at perhaps not every possible significance level.)

d. [1.5 points] This is correct.

Motivation: There is nothing preventing us from applying the χ-square goodness-of-fit test to test if a sample comes from a χ-square distribution with a pre-specified degrees of freedom.

e. [1.5 points] This is incorrect.

Motivation: A point can be a leverage point without being an influence point. This happens, for instance, if the response associated with the leverage point is *in line* with the regression line fitted without the use of that leverage point.

Question 2 [7 points]

a. [2 points] Examples of justifications:

i) The empirical distribution function is an unbiased (non-parametric) estimator of the true cumulative distribution function.

ii) The integrated kernel density estimator is a smooth function which often might seem more realistic than a step-function.

b. [2 points] There is no relation: the normal kernel is appropriate irrespectively of the distribution of the data. The normal kernel can be used for non-normal data, and non-normal kernels can be used for normal data.

c. [3 points] Repeat these steps a large number (e.g. $B = 1000$) of times:
(denote iteration index by i)

– Randomly draw $X_1^*, \ldots, X_n^*$ with replacement from $X_1, \ldots, X_n$.
– Calculate the kernel density estimate $f_i^*(t)$ based on $X_1^*, \ldots, X_n^*$.

Derive the empirical $\alpha/2$- and the $(1-\alpha/2)$-quantiles of $f_1^*(t), \ldots, f_B^*(t)$, denote them $q^*(\alpha/2)$ and $q^*(1-\alpha/2)$.

Create the confidence interval $[2 \cdot \hat{f}(t) - q^*(1-\alpha/2), 2 \cdot \hat{f}(t) - q^*(\alpha/2)]$, where $\hat{f}(t)$ is the kernel density estimator at t based on $X_1, \ldots, X_n$.

Question 3 [11 points]

a. [4 points] The test statistic is $T = \sum_{i=1}^{12} 1(X_i > 20)$ which, when $m = 20$, has a binomial distribution with $n = 12$ and $p = 0.5$.

b. [4 points] From the data, we see that T takes the value $t = 10$, and no observations are equal to 20. We compute the one-sided p-value, because H_1 is one-sided. Assuming $T \sim \text{Bin}(12, 0.5)$,

$$p = \mathbb{P}(T \geq 10) = 1 - \mathbb{P}(T \leq 9) = 1 - 0.981 = 0.019 < 0.05.$$

Hence, we reject the null hypothesis at significance level $\alpha = 5\%$: there is enough evidence to state that the claim being made by the company is incorrect (at the above stated significance level).

c. [2 points] For the Wilcoxon signed rank test the underlying data distribution is assumed to be symmetric. Symmetry can be checked either with a histogram, or a symplot.

d. [1 points] Let T be the test statistic of a test that is *distribution free*. Then the distribution of T under H_0 does not depend on specifically which distribution under H_0 is the true distribution.

Question 4 [9.5 points]

a. [1.5 points] Model B is appropriate where we compare the distribution of the two rows.

The hypotheses are H_0 : *the probability of winning is the same online or in person* and H_1 : *the probability of winning is not the same online or in person.* You can also formulate the hypotheses in terms of the $p_{i,j}$ as in the syllabus but we omit this in the solutions for brevity.

b. [4 points] i) the test tells us to reject the null hypothesis if the value of the test statistic (0.11 in our case) exceeds the 0.95-quantile of the χ-square distribution with $(2-1)(2-1) = 1$ degree of freedom. Since $0.11 < \chi^2_{1;0.95} = 3.84$, we do not reject the null hypothesis at significance level 5%.

ii) To check the rule of thumb, one of the expected frequencies under the null is $8 \cdot 9/23 \approx 3.13 < 5$ this is enough to conclude that it is not true that at least 80% of the cells have expectation greater than 5. We should therefore not place much trust in this test.

iii) The conclusion is that (at significance level 5%) we cannot conclude from the data that you perform differently in person than you do online.

c. [2 points] The standardised residuals are the cell entries minus their expected value under the null hypothesis, then divided by an estimator of its standard deviation , which makes comparisons with the standard normal quantiles reasonable. We have that $\pm 0.78 \in [-1.96, 1.96] = [q_{0.025}, q_{0.975}]$, thus no cell entry can be considered extreme which is in line with not rejecting H_0 in b.

d. [2 points] i) Fisher's exact test is applicable to 2×2 contingency tables so it is applicable here. ii) The test statistic in Fisher's exact test takes the value 4 here, as this is the top-left entry of the table. And finally, iii) for 2×2 contingency tables, Fisher's exact test is preferred since it is exact (as its name suggests) and thus does not require any rule of thumb to be satisfied. This is particularly relevant here as the rule of thumb for the χ-square test is not satisfied.

Question 5 [10 points]

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

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

with

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

The assumption on the errors is that $e_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. with $\sigma^2 > 0$ the unknown error variance.

b. [4 point]

- **linearity** from scatter plots (Y, X_j) for $j = 1, \ldots, p$ and added variable plots;
- **independence of errors** from context, scatter plots;
- **normality of errors** from Normal QQ-plot of residuals $R_Y(X)$;
- **constant error variance** from scatter plots $(\hat{Y}, R_Y(X))$ or $(X_j, R_Y(X))$ for $j = 1, \ldots, p$.

c. [2 point] Multicollinearity occurs when there is (nearly) linear dependence of explanatory variables, or more formally, among columns of the design matrix X. It may cause problems because it can contribute to increase the variances of the $\hat{\beta}_j$'s corresponding to columns of X involved in a collinearity. As a result, estimates of the respective coefficients can be unreliable.

d. [2 points] The observation with 'weight of incoming packages' approximately equal to 15kg is a potential leverage point since it is an outlier in the explanatory variable. In order to confirm that, the corresponding diagonal entry in the hat matrix can be computed and if it is close to 1, then the corresponding residual will be small (so the fit will be pulled towards a perfect fit for that observation regardless of the value of the response variable).

To study the effect of the point as an influence point, Cook's distance for that point can be computed. If Cook's distance is larger than 1, it is considered an influence point.