Vrije Universiteit Amsterdam	Statistical Data Analysis, Full Exam
Faculty of Science	27 May 2020

Other than an electronic device to display the exam and another one to stream yourself via Zoom, no other electronic device is allowed.
This exam consists of 6 questions on 3 pages (36 points).

Please write all answers in English. Grade = $\frac{total+4}{4}$.

You have 165 minutes to write the exam.

GOOD LUCK!

Question 1 [6 points]
Indicate for each statement whether it is correct or not. Motivate/explain your answers shortly;
also if you choose "correct", show with your explanation that you know why it is correct.
Note: just "correct" or "false" without further explanation will result in at most 0.5 point for each question.

a. [2 points] A normal QQ-plot could indicate that one should choose a right-skewed distribution as a model for the data distribution.

b. [2 points] If a conducted Kolmogorov-Smirnov test for the standard normal distribution at significance level $\alpha = 5\%$ results in a p-value of $p = 0.11$, it has been significantly shown that the data originate from the standard normal distribution.

c. [2 points] The center of a bootstrap confidence interval for a parameter θ is in general equal to an estimator $\hat{\theta}$ of θ.

Question 2 [6 points] Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with twice continuously differentiable density function f.

a. [1 point] Explain in your own words what the idea behind kernel density estimation for f is.

b. [2 points] Explain the idea of the cross-validation approach for selecting a bandwidth of a kernel density estimator. Also explain how the out-of-sample procedure works in this context.

Now suppose that $\hat{f}(t; X_1, ..., X_n)$ denotes a reasonable estimator for $f(t)$, $t \in \mathbb{R}$.

c. [3 points] Describe all steps for the construction of a bootstrap-based two-sided level-$(1-\alpha)$-confidence interval for $f(t)$, where $\alpha \in (0, 1)$.

Question 3 [5 points]
Let $X_1, X_2, \ldots, X_n$ ($n \geq 2$) be independent random variables that follow an unknown distribution P.
A goodness-of-fit test for "normality" could be based on the test statistic

$$T_n(X_1, \ldots, X_n) = \frac{|\bar{X}_n - \text{med}(X_1, \ldots, X_n)|}{S_n},$$

where $\bar{X}_n$ denotes the empirical mean, $\text{med}(X_1, \ldots, X_n)$ the empirical median, and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ the empirical variance of $X_1, \ldots, X_n$. The idea of the test statistic is that it tends to take large values if P is left- or right-skewed. On the other hand, for symmetric distributions, such as normal distributions, the test statistic will tend to take small values.

a. [2 points] Describe all steps for testing for "normality" based on T_n, including a formulation of the null and alternative hypotheses and a description of how to obtain the critical or p-value and how to reach a test conclusion.
Note: assume for the moment that the distribution Q_P of T_n under the null hypothesis is known.

b. [3 points] Now suppose that the distribution Q_P of T_n under the null hypothesis is *unknown*.
Describe a method to approximate Q_P and explain all steps that are done in this method.
Justify your choice of method and avoid unnecessary parameter estimation whenever possible.
Hints: 1. You should choose a method that minimizes the approximation error(s).
2. The test statistic T_n is independent of location and scale parameters.

Question 4 [5 points]
Independence in paired numerical data can be analyzed with the help of Spearman's rank correlation coefficient.

a. [1 point] Explain one advantage of Spearman's rank correlation coefficient over Pearson's correlation coefficient, i.e. the sample correlation divided by the product of both sample standard deviations.

b. [3 points] Describe all steps you do when you conduct the independence test based on Spearman's rank correlation coefficient *as a permutation test.*

c. [1 point] Explain the advantage of the permutation test from part b. over the test that uses Spearman's rank correlation coefficient in combination with a normal approximation.

Question 5 [7 points]
In a study about sport preferences, 100 randomly selected men and 100 randomly selected women have been asked which of the following sports they prefer to do: archery, boxing, cycling. The data is summarized in the following table:

sex / sport	archery	boxing	cycling	total
female	17	8	25	50
male	5	15	30	50
total	22	23	55	100

a. [3 points] Motivate which model for categorical data is appropriate for this dataset and state the null and alternative hypotheses that can be tested for the chosen model.

b. [3 points] Do a χ^2-test at level $\alpha = 5\%$ to test the hypotheses from part a. Use that the value of the test statistic is $\chi^2 \approx 9.13$ and that the critical value of the test statistic, that would be used by default in R, is 5.99.
Note: do not forget the check all required assumptions.

c. [1 point] Which specific quantile of which specific distribution is the critical value given in part b.?

Question 6 [7 points]
By means of a linear regression analysis, we analyze the steam amount used by an engine with the help of $n = 25$ independent measurements. The model with $p = 2$ explanatory variables, *fatty acid* and *temperature*, in addition to the intercept and measurement errors, seems fine, and the diagnostic plots seem fine too. It remains to continue with additional model diagnostics.

a. [2 points] Mention two reasons why the normality assumption for the measurement errors is important.

b. [2 points] Observation number 7 has a hat-value, i.e. the corresponding diagonal entry h_{77} of the hat matrix $H = X(X^TX)^{-1}X^T$, that exceeds $2 \cdot (p+1)/n = 0.24$. Explain the implications of this and discuss under which circumstance this would cause a problem.
Note: X denotes the design matrix of full rank.

c. [3 points] The variance inflation factors for the explanatory variables were both 1.000004 and the variance decomposition resulted in the following condition indices and variance decomposition proportions:

condition indices	intercept	fatty acid	temperature
1.000	1	0.000	0.000
4.995	0	0.499	0.499
5.005	0	0.501	0.501

Explain which specific presence or absence of problems can be analyzed with the help of the variance inflation factors, the condition indices, and the variance decomposition proportions. Discuss whether the present numbers indicate possible issues with the final linear regression model; motivate your answer.