Test Mathematical Statistics (202001348)
Friday 23 October 2020 (09:00-12:00)
Instructors: Julio Backhoff and Johannes Schmidt-Hieber
Module Coordinator: Johannes Schmidt-Hieber

A formula sheet is added. A regular scientific calculator is allowed, a programmable calculator ("GR") is not. No tables of probability distributions are needed for the test.

PART A: BASIC CONCEPTS

a) (1 Point) What is the sample median?

b) (1 Point) Describe two tools from descriptive statistics to visualize data.

c) (1 Point) Provide the unbiased estimator of the variance.

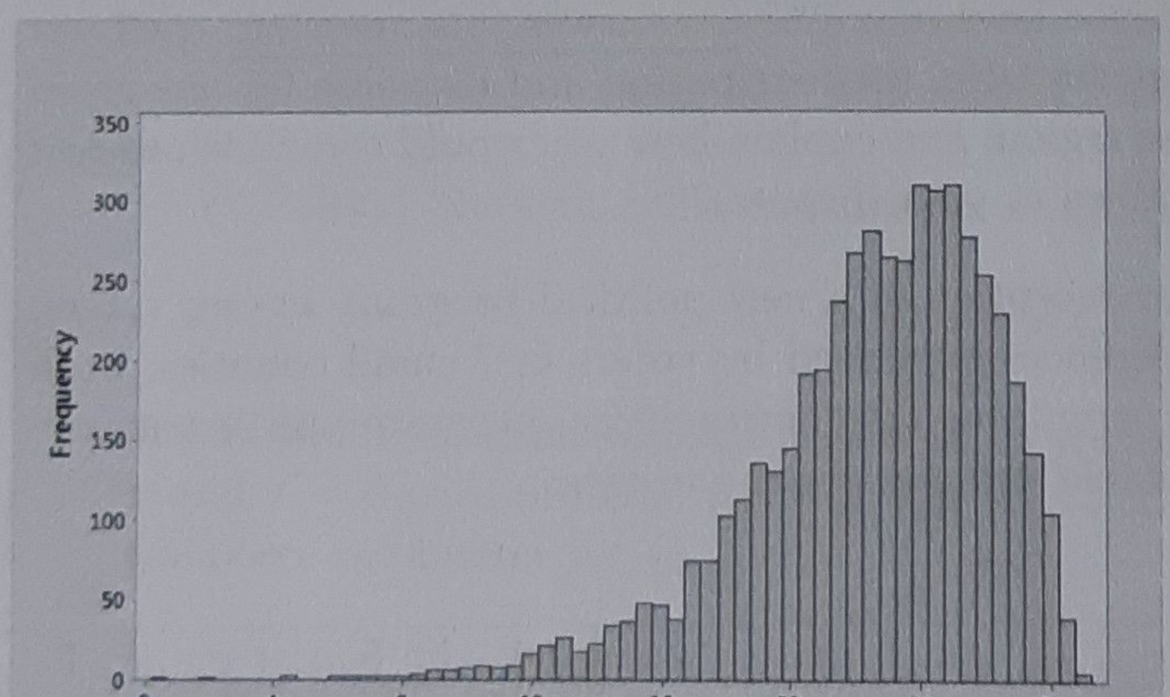
d) (1 Point) Define the Type I error of a hypothesis test. In a hypothesis test, do we aim to minimize this error or keep it upper-bounded?

e) (1 Point) Explain, via examples, what is the difference between a "two samples" and a "paired samples" problem.

f) (1 Point) How do you expect the coefficient of determination R^2 to change, if more explanatory variables are added in a linear regression model?

PART B: VISUALIZATION OF DATA

1. (1 Point) Consider the histogram below. Do you expect the mean to be greater, smaller or roughly equal to the median. Justify.



2. (1 Point) Sketch a scatter plot in which the correlation without an outlier is positive, but with the outlier it becomes negative. Indicate in your plot which point is the outlier.

3. (2 Points) Provide the boxplot (including outliers) associated to the 10 data points:

6	4	9	1	6	5	3	15	4	6

PART C: THEORY

1. Let $\mathcal{X}$ be a subset of $\mathbb{R}$, and consider

$$f_\phi(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\} \ (x \in \mathcal{X}), \qquad (*)$$

which is a collection of density functions (or p.m.f. in the discrete case) indexed by a parameter $\phi \in \Phi \subset \mathbb{R}$. Such a collection is called a "1-parameter exponential family".

a) (2 Points) Fix $n \in \mathbb{N}$ and consider the binomial distribution with parameters (n, p), as p ranges in $(0, 1)$. Show that this gives rise to a 1-parameter exponential family.

b) (2 Points) Let $X_1, \ldots, X_n$ be iid samples, with $X_1 \sim f_\phi$ and where f_ϕ is as in $(*)$. Find the joint density (or joint p.m.f. in the discrete case) of $X_1, \ldots, X_n$ and show that $\tau(x_1, \ldots, x_n) := \sum_{i=1}^n T(x_i)$ is a sufficient statistic for ϕ.

2. For marketing purposes a supermarket chain chose, some years ago, 4 equally large categories of costumers, based on their weekly expenses in the supermarket. Management wants to check whether these categories are still equally large (all 25%) nowadays.

For a group of $n = 205$ customers the marketing department found the following counts in the categories:

Category	1	2	3	4
Observed count	43	56	59	47

[3 Points] Is there sufficient evidence to conclude that the probabilities of belonging to a category have changed, at a significance level $\alpha = 5\%$? To answer this question: spell out the hypothesis test, describe the test statistic, its distribution and its value for the given data, show the shape of the rejection region, and explain how you would conclude the test if you had software or probability tables at your disposal.

3. A political party wants to assess the reception of a new political program among voters. For this reason it considers the preferences expressed by voters in 7 small counties, both in the previous election and in the current one (the new political program was announced between these elections). The next table summarizes the results:

	1	2	3	4	5	6	7
Votes obtained in previous election (x)	3419	4135	4979	3752	6222	4047	3720
Votes obtained in current election (y)	4340	5269	6061	4011	5749	4814	3642

Define $z = y - x$ the difference of votes between the current and the past election in a small county. It can be verified that the sample means of x, y and z are respectively 4324.9, 4840.9 and 516. Similarly the sample standard deviations are respectively 970.8, 901.3 and 622.7.

a) (2 Points) Propose a suitable hypothesis test to check whether the expected increase in voter preferences is positive, assuming that the number of votes is approximately Normal distributed. Your answer should be as detailed as possible.

b) (2 Points) Propose a suitable hypothesis test to check whether the expected increase in voter preferences is positive, but this time do not assume that the number of votes is approximately Normal distributed. Do so providing an expression for the p-value of the test and assuming $\alpha = 5\%$. Your answer should be as detailed as possible.

4. 1000 male volunteers participate in a test where a certain medicine against hair loss is examined. The volunteers are split randomly into two groups of size 500. One group gets the medicine while the other group does not. After an appropriate amount of time, it is determined that 100 out of the 500 treated individuals have significant hair loss, whereas 140 out of the 500 untreated individuals show significant hair loss.

a) (2 Points) Determine a 99%-confidence interval for the difference in the proportions of treated and untreated individuals experiencing significant hair loss. For this, use that $P(|Z| \geq 2.58) = 1\%$ if $Z \sim \mathcal{N}(0,1)$.

b) (2 Points) Is the medicine against hair loss effective? Formulate and solve an appropriate hypothesis test in as much detail as possible. For this, use that $P(Z \leq -2.33) = 1\%$ if $Z \sim \mathcal{N}(0,1)$.

c) (2 Points) Summarize the data of the problem in a 2 x 2 cross table. Propose, and solve in as much detail as possible, an appropriate χ^2-hypothesis test to determine if the medicine has any impact (positive or negative). What particular kind of χ^2-test is this?

5. Recall that the multiple regression model, in matrix-vector notation, is $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. Under the assumption that $X^\top X$ is invertible, we have $\beta^{MLE} = (X^\top X)^{-1} X^\top Y$ is the MLE. The residuals are given by $Y - X\beta^{MLE}$.

a) (2 Points) Show that the residuals are also given by the formula

$$(I - X(X^\top X)^{-1} X^\top)\epsilon.$$

b) (2 Points) Show that the distribution of the residuals is

$$\mathcal{N}(0, I - X(X^\top X)^{-1} X^\top).$$

6. Recall that a random variable is said to be $Bern(p)$-distributed if $Y = 1$ with probability p and $Y = 0$ with probability $1 - p$. Let $X_1, \ldots, X_n$ be iid and $Bern(p)$-distributed for an unknown parameter $p \in [0, 1]$.

a) (1 Point) Show that the likelihood function of $X_1, \ldots, X_n$ is

$$L(p) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}.$$

b) (3 Points) Find the maximum likelihood estimator of p. Why is it unique, and why is it a true maximum (and not, say, a stationary point or a local maximum)? You will have to distinguish the cases when $\sum_{i=1}^n X_i = 0$, $\sum_{i=1}^n X_i = n$, and $\sum_{i=1}^n X_i \in \{1, \ldots, n-1\}$.