Test Mathematical Statistics (202001348)

Friday 06 November 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) Explain in words what a confidence interval is.

b) (1 Point) Consider the following situation: we first compute sample mean and sample median from data, then a new data point arrives, and finally we recompute the sample mean and sample median taking the new data set (which now includes the new point) into consideration. Do you expect the sample mean, or the sample median, to have changed more during the recomputation? Illustrate with an example.

c) (1 Point) Define the Type II error of a hypothesis test.

d) (1 Point) Assume that T is a sufficient statistic for a parameter ϕ. For which values of a constant $c \in \mathbb{R}$ is cT guaranteed to be a sufficient statistic for ϕ?

e) (1 Point) In linear regression, what is the difference between a response and an explanatory variable?

f) (1 Point) Why is the sign test on the median sometimes preferred (over competing approaches)?

PART B: VISUALIZATION OF DATA

1. (2 Point) A coin is tossed three times. We record a zero (0) whenever heads is observed and a one (1) whenever tails is observed, so for example HHT becomes 001. Plot the four different boxplots that could be obtained from the data of this experiment.

2. (2 Point) How do you expect the histogram and the boxplot would look like in a situation where we have many independent samples generated according to a uniform distribution on the interval $[0, 1]$?

PART C: THEORY

1. Suppose that X is a single observation from an experiment with density function given by

$$f(x) = \rho x^{\rho - 1} \text{ if } x \in (0, 1),$$

and $f(x) = 0$ if $x \notin (0, 1)$. We want to test the null hypothesis $\rho = 3$ against the alternative hypothesis $\rho = 2$.

a) (2 Points) What is the likelihood ratio test (i.e. the Neyman Pearson test) for this problem? Describe the form of the rejection region.

b) (2 Points) Consider the significance level $\alpha = 5\%$ and suppose that the observation is $X = 0.7$. Conclude the hypothesis test.

c) (Bonus of 1 Point) If the observation is $X = 0.7$, what is the p-value of the test?

2. Until 2013, a certain country held presidential elections according to the "first register voluntarily, then vote obligatorily" system. The turnout (number of voters) during the six latest elections that took place using this system were:

$$7.055.128,\ 7.178.727,\ 6.942.041,\ 6.959.413,\ 6.977.544,\ 6.958.972$$

In 2013 the same country changed to a "first register obligatorily, then vote voluntarily" system. The turnout in the four elections that have taken place under this new system are:

$$6.699.011,\ 5.697.751,\ 6.703.327,\ 7.032.878$$

a) (1 Points) Assuming that the turnout of elections are independent, normally distributed random variables with a known standard deviation of 500.000, propose a hypothesis test to check whether the turnout decreased under the new system. Provide your reasoning.

b) (2 Points) Complete your hypothesis test with a siginifance level of $\alpha = 5\%$. You may use that $P(Z \leq 1.645) \approx 95\%$ if Z is approximately $\mathcal{N}(0, 1)$-distributed.

3. (Continuation of Question 2) In fact the country under consideration has always used a 2-round election procedure. This means that in every election year (1999,2005,2009,2013,2017) there is a first round where voters choose two candidates, and then a second round where voters choose one of the two candidates to become the next president. The data can be summarized as follows:

	1999	2005	2009	2013	2017
First Round	7.055.128	6.942.041	6.977.544	6.699.011	6.703.327
Second Round	7.178.727	6.959.413	6.958.972	5.697.751	7.032.878
	123599	17.372	-18572	-1001260	329551

a) (2 Points) Propose a suitable hypothesis test to check whether there is an increase in voter turnout during the second rounds (compared to the first rounds), assuming that the number of votes is approximately Normal distributed with and unknown variance. Your answer should be as detailed as possible.

b) (2 Points) Propose a suitable hypothesis test to check whether there is an increase in voter turnout during the second rounds, this time not assuming that the number of votes is approximately Normal distributed. Do so providing an expression for the p-value of the test. Your answer should be as detailed as possible.

4. In a certain region, it was known during pre-corona times, that roughly equal numbers of people considered themselves "rather satisfied", "more or less satisfied" and "not satisfied" with their lives. To check the impact of the corona pandemic, a survey on 100 indivuduals was then carried out, showing that 25% considered themselves "rather satisfied", 35% "more or less satisfied", and 40% "not satisfied". Your task is to propose and solve a hypothesis test, in as much detail as possible, on whether there has been a change in perception concerning life satisfaction in this region. (3 Points)

5. Consider the following modified simple regression model:

$$Y = \beta_1 x + \epsilon.$$

This means that if $\{(x_i, Y_i)\}_{i=1}^n$ are our data points, then $Y_i = \beta_1 x_i + \epsilon_i$ under this model, with the ϵ_i independent and $\mathcal{N}(0, \sigma^2)$-distributed, and we assume that σ is known. Notice that unlike the usual simple regression model, here there is no β_0 term (equivalently, β_0 is set to zero by definition).

a) (2 Points) Prove that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

is the least square estimator of β_1. (If it is easier for you, you may use that the least square and the maximum likelihood estimators coincide in this case.)

b) (2 Points) Find the distribution of $\hat{\beta}_1$.

~~c)~~ (Bonus of 1 Point) What is the general form of a confidence interval for β_1?

6. Let Y be Binomial distributed with parameters (n, p), i.e. $Y \sim \mathit{Binomial}(n, p)$, where $n \in \mathbb{N}$ is fixed, and $p \in (0, 1)$.

~~a) (1~~ Point) Find an unbiased estimator of p.

~~b) (2~~ Points) Show that there is no unbiased estimator of $1/p$.