Reichman University

ID: ____________

Notebook No: ________

Student Guidelines	
Course Name: סטטיסטיקה מתקדמת	
Lecturer Name: דר קיפניס אלון	
Exam Date: **13/06/2022**	Term: **1**

Extra Material: No Reference Allowed except	
Time Limit: 3	
Dictionary: **Yes**	
Calculator: **Simple**	
Student Formula Sheet: **Yes**	Number Of Formula Pages Allowed: **2 (דו-צדדי)**
Lecturer Formula Sheet: **No**	
Answer Written on Exam File: **Yes**	Answer Written on Notebook: Yes
Other, Specify:	

[redacted]

Answers must be written only on the right hand side of the exam notebook.
Do not use Marker.

Good Luck!

Final Exam

Advanced Statistics for Data Science

Spring 2022

Instructions

- You have 3 hours to complete the exam.
- The exam contains two parts. Part I contains 8 problems, each has a maximal credit of 5 points. Part II contains 3 questions, each has a maximal credit of 20 points. The maximal number of points in the exam is 100.
- For maximal grade, you should answer *all* problems correctly.
- You may bring to the exam up to two personal two-sided A4 pages containing relevant material.

Part I

For the following problems, either indicate **True** or **False** or fill-in-the-blanks to complete correct statement or answer (whichever applies).

1. (5 points) Let H be the hat matrix for a regression with n observations and p predictors. The underlying design matrix $Z \in \mathbb{R}^{n \times p}$ has full rank. The trace of $H(I - H)$ is __________
   Explain: ____
   ____.

2. (5 points) We fit a linear model using ordinary least squares regression and obtain the fitted response $\hat{\epsilon}$. It is possible that
   $$\hat{\epsilon} = \begin{pmatrix} -1 & -1 & 1 & 1 & 1 \end{pmatrix}^\top .$$
   (True/False)
   Explain: ____
   ____.

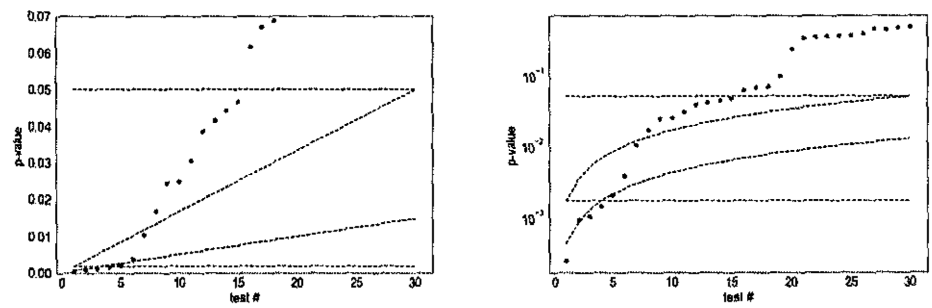
3. (5 points) The random variables X and Y are independent $\mathcal{N}(0,1)$. The distribution of $Y/|X|$ is called ____________ ____________ .

   Explain: ____.

4. (5 points) Suppose we run 10 independent hypotheses tests and obtained P-values $p_{(1)} \leq \ldots \leq p_{(10)}$. If $p_{(1)} = 0.006$ and $p_{(10)} = 0.1$, it is possible that we reject 2 hypotheses after using the Binjamini-Hochberg procedure for controlling the false-discovery rate at level 0.05. **(True/-False)**

   Explain: ____.

5. (5 points) The figures bellow describe sorted P-values obtained from 30 individual hypothesis tests (the only difference between the figures is the scale of the y-axis, which is logarithmic on the right).



We also have the following legend:

curve number	curve description
(1)	$y = 0.05$
(2)	$y = 0.05 \cdot x/30$
(3)	$y = 0.05 \cdot x/(30 \cdot C_{30})$
(4)	$y = 0.05/30$

$(C_m = \sum_{i=1}^{m} i^{-1})$

- The tests selected by Binjamin-Hochberg's (BH) procedure for controlling the false discovery rate (FDR) at level $\alpha = 0.05$ are those whose P-values have ranks ____________ .
- The tests selected by a Bonferroni correction to control the family-wise error rate at level $\alpha = 0.05$ are those whose P-values have ranks ____________ .
- The tests selected by Binjamin-Hochberg's (BH) procedure for controlling the false discovery rate (FDR) at level $\alpha = 0.05$ for any type of dependency among the tests are those whose P-values have ranks ____________ .

(the rank of a P-value p is said to be k is there are $k-1$ P-values that are smaller than p)

6. (5 points) The cross-validation (CV) residuals sum-of-squares is never smaller than the residuals sum-of-squares. **(True/False)**

   Explain: ____

   ____.

7. (5 points) We fit a linear model with $p = 5$ predictors using least squares and obtain coefficients $\hat{\beta}_j$ for $j = 1, \ldots, 5$. We conduct a t-test for each one of the coefficients to check whether they are different than zero – we obtain that only 2 out of the 5 tests are significant in the sense that the absolute value of their t statistics exceed the $1-\alpha/2$ quantile of the t distribution, where $\alpha \in (0,1)$ is some significant level. Is it possible that all coefficients will turn out to have significant t-test P-values if we replace each test by a one-sided t-test test that rejects only when the coefficient is significantly *larger* than zero? **(True/False) Explain:** ____

   ____.

8. (5 points) We examine a linear model with 5 predictors. Below are three tables, each potentially describing a path of a model/variable selection procedure for our model. Which of the following paths may correspond to a *backward* step-wise selection procedure?

R^2	variables included
0	$\emptyset$
.3	$\{2\}$
.5	$\{2,3\}$
.6	$\{2,3,5\}$
.62	$\{2,3,5,4\}$

R^2	variables included
.85	$\{1,2,3,4,5\}$
.81	$\{1,2,3,4\}$
.79	$\{2,3,4\}$
.78	$\{2,3\}$
.785	$\{2\}$

R^2	variables included
1	$\emptyset$
.65	$\{2\}$
.6	$\{2,3\}$
.5	$\{2,3,4\}$
.3	$\{2,3,4,5\}$

   Explain: ____

   ____.

Part II

The questions below may have multiple sections. You should write your response on a separate piece of paper.

1. (20 points) We consider a balanced 2-group model:

$$y_{1j} = \mu_1 + \epsilon_{1j}, \qquad y_{2j} = \mu_2 + \epsilon_{2j}, \qquad j = 1, \ldots, n$$

   (it is called *balanced* because $n_1 = n_2 = n$). The standard assumption $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $j = 1, 2$, applies. We have the null hypothesis:

$$H_0 : \mu_1 = \mu_2 + 10$$

- Design a level-α test against H_0: Describe the test statistic and explain for what values of this statistic you decide to reject H_0 and why (you can use the quantile function of any of the distributions we have seen in class).
- Repeat the previous item for testing

$$H_0' : \mu_1 = 10\mu_2$$

2. (20 points) We observe $y_1, \dots, y_n$. We are given some $\mu_0 \in \mathbb{R}$ and would like to test the hypothesis

$$H_0 \; : \; y_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2), \qquad i = 1, \dots, n.$$

   (i) Propose a test for H_0.

   (ii) Express the test's P-value in terms of the quantile function of one of the distributions we have seen in class.

   (iii) Suppose that in reality

$$y_i \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2), \qquad i = 1, \dots, n.$$

   Explain what factors affecting your ability to detect $\mu_1 \neq \mu_0$ and how they affect.

3. We would like to compare the quality of two wine series based on a dataset containing scores of many participating wines in many contests. Each series is rated only once in each contest it participated. For each competing wine we record the following variables: series name, contest id, and score. The table below provides a general description of how the data may look like.

series name	contests id	score
Series1	⋮	⋮
Series2	⋮	⋮
Series2	⋮	⋮
Series1	⋮	⋮
⋮	⋮	⋮
Series2	⋮	⋮

   (i) Describe a process to decide which series is better. Write out the form of the t statistic for testing this hypothesis. State the null distribution of the t statistic and give conditions under which we reject H_0. Introduce and define the notation you need. We can assume that the measurements are independent normally distributed random variables and that they all have the same variance.

   (ii) Suppose that we know that both series have competed in each contest in the dataset. Would that change your process? If yes, explain the new process.