Exam Summary

Exam ID: 3483671   Student ID: 207380528

Course ID: 202200036762100622236760000   Course name: מתקדמת סטטיסטיקה

Question Number	Description	Comments	Max Grade	Question Final Grade
1			5.00	5.00
2		If the expectation is zero it does not imply that the value is zero for each realization	5.00	2.00
3			5.00	5.00
4			5.00	5.00
5		The middle item is incorrect	5.00	4.00
6		You should use properties of the CV error we`ve seen in class.	5.00	4.50
7			5.00	5.00
8			5.00	5.00
10	Part II		20.00	20.00
11	Part II		20.00	19.50
12	Part II		20.00	12.00

Final Exam Grade : 87.00

The checked exam is in the next pages

***** Pay attention, there are sticky note and voice on the exam, for suited best watching, please open the file with acrobat reader *****

Student Number
106

Reichman University

Notebook No.:	1
of 1 notebooks	

Before beginning the exam fill in all of the following details in clear print and read the instructions carefully:

Date of Exam: 13.6.22

Course Name: סטטיסטיקה מתקדמת

Instructor's Name: אילן קיננים

Study Track: ML & DS

ID Number

2 0 7 3 8 0 5 2 8

302010861725
207380528 6

Please note: Do not write outside the lined area (stay within the margins).
Answers must be written with a pen with blue or black ink.
Answers must be written only on the right hand side of the exam notebook.
Pages must not be torn out of the exam notebooks.

1. Students must provide the information requested on the back of the exam notebooks as soon as they recieve them. Exams are anonymous. Students must not write any identifying details (other than their ID number and the notebook number) on their test forms or exam notebooks.
2. Students must follow the proctor's instructions. Students may not leave the exam room without the proctor's permission. Students must raise their hands to make a request or ask a question.
3. All students who enter the exam room and receive an exam (test forms) are considered as having taken the exam on the date. Should they decide not to take the exam, they will not be permitted to leave the room until 30 minutes have elapsed from the start of the exam and until they have returned the test forms and the exam notebooks to the proctor.
4. It is strictly forbidden to have any supplementary material in your possession, in or outside the classroom, except for the material allowed by the course instructor. Possession of supplementary material is considered a fraud, and may result in a disciplinary action, including expulsion. Study materials cannot be disposed of in the trash cans near or around the classrooms, including those in the restrooms.
5. All cell phones/smart phones/smart watches must be turned off and placed in the student's bag in the front of the classroom. Students who are found with telephones/devices in their possession against the instructions mentioned above, even if they did not use the telephone/device, **their exam will be disqualified on the spot,** according to the IDC regulations. Holding a telephone/smart watch or operating one during an exam may lead to, among other things, suspension from studies.
6. Students must write clearly and neatly with a pen with blue or black ink (as noted above).

Good Luck!

Exam Grade ....

Instructor's Signature ....

Reichman
University

ID: 207380528

Notebook No: 166

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:	

Please notice:
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).

5
(1)

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 0
Explain: Z has Full rank ⇒ Z is regular ⇒ H is P pro
⇒ $H = H^2$ ⇒ $H(I-H) = H - H^2 = H - H = 0$ ⇒ $tr(0) = 0$.

2. (5 points) We fit a linear model using ordinary least squares regression and obtain the fitted response $\hat{\epsilon}$. It is possible that
$$\hat{\epsilon} = \begin{pmatrix}-1 & -1 & 1 & 1 & 1\end{pmatrix}^{\mathsf{T}}.$$
(True/[✓ False])
Explain: False, $E(\hat{\epsilon})$ has to be 0, and it is $\frac{3}{5} \neq 0$.

2
(2)

$E(\hat{\epsilon}) \neq 0$, we can fit a better line.

What is |X| ⇒ abs

5
(3)

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

Explain: $y \sim N(0,1)$, $|X| \sim \chi_1^2$ $\Rightarrow \frac{y}{|X|} \sim t_{n-1}$

details in notebook.

5
(4)

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)

independent

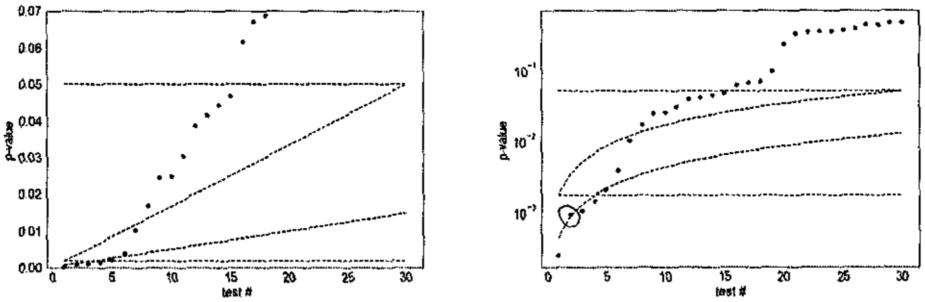
Explain: $\alpha = 0.05$, $m = 10$, BH $= \frac{q}{m} \cdot i = 0.005 \cdot i$, $1 \leq i \leq 10$

For example, $p2 = 0.0061 < 0.005 \cdot 2 = 0.01$, $p3 = 0.0062 < 0.005 \cdot 3 = 0.015$

and all other pi's are $pi = 0.1$. $0.0062 \geq 0.0061 \geq 0.006$

4
(5)

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 $7 + 1 = 8$.
- 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 $4 + 1 = 5$.

[redacted]he 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 $4 + 1 = 5$.

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

Just to make sure! I understand that the rank = number of points under the corresponding curve +1. In case of missing points due to eye.

~~residuals sum of squares = $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$~~ ~~not divided by N~~

4.5
(6)

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

Explain: True. when we use CV, we have less data to fit on. hence, our model is less accurate, if we check it on the not fitted data. (If we do this for all the slices)

5
(7)

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:** explaining in notebook. In high-level: $[t^{1-\frac{\alpha}{2}}] > [t^{1-\alpha}] \Rightarrow$ easier to reject.

5
(8)

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	R^2	variables included	R^2	variables included
0	$\emptyset$	.85	$\{1,2,3,4,5\}$	1	$\emptyset$
.3	$\{2\}$	.81	$\{1,2,3,4\}$	.65	$\{2\}$
.5	$\{2,3\}$	.79	$\{2,3,4\}$	.6	$\{2,3\}$
.6	$\{2,3,5\}$	.78	$\{2,3\}$	.5	$\{2,3,4\}$
.62	$\{2,3,5,4\}$	.785	$\{2\}$	.3	$\{2,3,4,5\}$

Explain: backward selection starts with all features. The only one complies is the middle. hence: answer: middle

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.

Question 7

If for example all β_j's are positive but not significantly $\neq 0$, it could be that using a one-sided t-test (in which we reject if $t > [t_{n-k}]^{1-\alpha}$ instead of if $t > [t_{n-k}^{1-\frac{\alpha}{2}}]$) we'll reject β_j's that we didn't reject with a two-sided test, since $[t_{n-k}^{1-\frac{\alpha}{2}}] > [t_{n-k}^{1-\alpha}]$ which means it's easier to reject.

Instructor's notes:

part 2

1. $H_0: \mu_1 = \mu_2 + 10 \Rightarrow \mu_1 - \mu_2 = 10$

$H_1: \mu_1 \neq \mu_2 + 10 \Rightarrow \mu_1 - \mu_2 \neq 10$

It is a two-sample t-test with 2 balanced groups.

20 (10)

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 10}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$n = n_1 = n_2$

$n_1 + n_2 = 2 \cdot n$

$$s^2 = \frac{1}{2 \cdot n - 2} \cdot \left(\sum (y_{1i} - \bar{y}_1)^2 + \sum (y_{2i} - \bar{y}_2)^2 \right)$$

our test rejects H_0 if $|t| > t_{2n-2}^{1-\frac{\alpha}{2}}$

Now for

$H_0': \mu_1 = 10\mu_2 \Rightarrow \mu_1 - 10\mu_2 = 0$

$H_1': \mu_1 \neq 10\mu_2 \Rightarrow \mu_1 - 10\mu_2 \neq 0$

we'll define $y_{3j} = 10\mu_2 + \varepsilon_{3j}$

Instructor's notes:

We'll perform a two-sample t-test on y_1, y_3, but now they are NOT iid. because $Var(y_3) = Var(10\, y_2) = 100 \cdot Var(y_2) = 100 \cdot \sigma^2$. However, $n_1 = n_3$, so, we can still use a t-test (althoun perhaps Walchs t-test is better) $Var(\varepsilon_{3j}) = 100\sigma^2$.

$$t = \frac{\bar{y}_1 - \overset{=\bar{y}_3}{10\bar{y}_2} - 0}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n-2}$$

$$s^2 = \frac{1}{2n-2} \cdot \left(\Sigma(y_{1i} - \bar{y}_1)^2 + \Sigma(10 \cdot y_{2i} - \bar{y}_2)^2 \right)$$

Part 1 Question 3 $\quad x \sim N(0,1), \; y \sim N(0,1)$

We know that $|x| = \sqrt{x^2} \Rightarrow |x| \sim \chi^2$

✓ We also know that the dist is:

$t = \frac{z}{\sqrt{\frac{x}{q}}} \sim t_{n-1}$ where $z \sim N(0,1)$ $\quad x \sim \chi^2_v$

$\frac{y}{|x|} \sim t_{n-1}$

We have $\quad$ because $y \sim N(0,1)$ and $|x| \sim \chi^2$ with 1 dof.

Instructor's notes:

part 2 Question 2

19.5
(11)

H0: ~~$y_i \sim N(\mu_0, \sigma^2)$~~ $\mu = \mu_0$

H1: $\mu \neq \mu_0$

(i) I'll use a one-sample t-test: ✓

$$t = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

where

$$s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$$

(ii) we'll reject if $|t| > t_{n-1}^{1-\frac{\alpha}{2}}$ ✓

(iii) n - The more samples, the bigger ✓ the chance we'll reject it

α - the lower α, it will become harder to reject ($t_{n-1}^{1-\frac{\alpha}{2}}$ will be bigger)

✓ (the effect size): $\left|\frac{\mu_0 - \mu_1}{s}\right|$ sigma , the bigger it is (in abs value) the easier it is will be to reject (it'll increase the t statistic).

Instructor's notes:

part 2 question 3

(i) i'll use two-sample t-test ✓ for this test,

in which we'll check all scores from each of the 2 series. and average them.

~~the t-test~~ score = μ

$H0: \mu 1 = \mu 2 \Rightarrow \mu 1 - \mu 2 = 0$

$H1: \mu 1 \neq \mu 2 \Rightarrow \mu 1 - \mu 2 \neq 0$

$$t = \frac{\overline{score\_1} - \overline{score\_2} - 0}{s\sqrt{\frac{1}{n1} + \frac{1}{n2}}} \sim t_{n1+n2-2}$$

$$s^2 = \frac{1}{n1+n2-2} \cdot \left(\sum (score\_1_i - \overline{score\_1})^2 + \sum (score\_2_i - \overline{score2})^2 \right)$$

We reject if $|t| > [t_{n-2}^{1-\frac{a}{2}}]$ at some significance level a.

12
(12)

(ii) NO. this would just mean that n1 = n2 and will probably result in a more accurate test.

Instructor's notes:

-8
(12)

Paired ttest

You described a provate case of section 1...

Instructor's notes: