Exam Summary

Exam ID: 3483671   Student ID: 204502926

Course ID: 2022000367621006222367600000   Course name: סטטיסטיקה מתקדמת

Question Number	Description	Comments	Max Grade	Question Final Grade
1			5.00	5.00
2		We asked to assume an intercept term.	5.00	2.00
3			5.00	5.00
4			5.00	5.00
5			5.00	5.00
6		H_{ii}	5.00	3.00
7			5.00	5.00
8			5.00	5.00
10	Part II		20.00	19.00
11	Part II		20.00	16.00
12	Part II		20.00	12.00

Final Exam Grade : 82.00

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

114

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: Advanced Statistics for DS

Instructor's Name: Dr. Alon Kipnis

Study Track: MSc MLDS

ID Number

204502926

302010861726
204502926 8

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: 204502926

Notebook No: 114

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).

~~then~~ trace

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 (zero)! $H = \mathrm{diag}(H_{11}, \dots, H_{nn})$

   Explain: $I-H$ is a PPM. The eigenvalues of a PPM are all either 0 or 1. ~~invertible, H has non-zero eigenvalues~~ → $H(I-H) = HI - H^2 = H - H = 0$

2. (5 points) We fit a linear model using ordinary (ols) least squares regression and obtain the fitted ~~response~~ residuals $\hat{\epsilon}$. It is possible that

$$\hat{\epsilon} = \begin{pmatrix} -1 & -1 & 1 & 1 & 1 \end{pmatrix}^{\top}.$$

   (**True**/**False**) $y = (-2 \; \dots$

   Explain: These are the residual values (error) vector. Each entry indicates $y_i - \hat{y}_i$ for every $y_i \in y$ $(n=5)$

   $\hat{\epsilon}_i = y_i - \hat{\beta}^T z_i = y_i - \hat{y}_i$

2

(2)

$$H = \begin{bmatrix} H_{11} & & & \\ & H_{22} & & \\ & & H_{33} & \\ & & & \ddots H_{nn} \end{bmatrix}$$

$$H(I-H) = \begin{bmatrix} H_{11} - (H_{11})^2 & \\ & H_{22} - (H_{22})^2 \end{bmatrix} \quad \Leftarrow H = H^2$$

$$I - H = \begin{bmatrix} 1-H_{11} & & & \\ & 1-H_{22} & & \\ & & 1-H_{33} & \\ & & & \ddots 1-H_{nn} \end{bmatrix}$$

$|x| = \sqrt{x^2}$

5
(3)

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

Explain: Y is standard Normal, $|X| = \sqrt{X^2} = \sqrt{\frac{X^2}{1}} \overset{\sim}{\Rightarrow} chi^2_1$

X, $|X|$ are independent. ~~DoF~~ $k=1$ DoF.

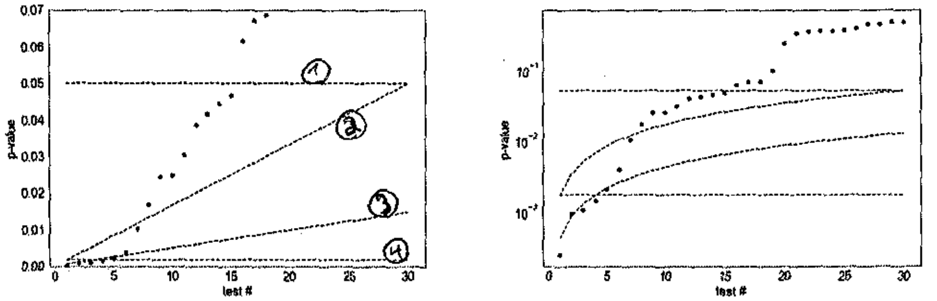
$l_i = \frac{\alpha \cdot i}{m} = \frac{0.05}{10} \cdot$

~~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)

Explain: BH is less strict than Bonf. We set a line $l_i = \alpha \cdot \frac{i}{m}$ and reject the hyp.s with $P(i) \leq P(i^*)$ s.t. $i^* = \max\{i, P(i) \leq l_i\}$ for $i \in [1, 10]$ → i^* can be 2 and $P_1 = 0.006 < \frac{0.05 \cdot 2}{10} = 0.01$, and $P(2) = 0.007$ for example, all other P's may be larger.

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)	$y = 0.05/30$

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

5
(5)

- 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 1 - 7. all under (2)
- 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 1 - 4. all under (4)
- 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 1 - 5. all under (3) (general dependency)

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

$$SS_{RES} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \|\hat{\varepsilon}\|_2^2$$

3 (6)

~~True~~ False $\boxed{\text{False}}$ LOOCV

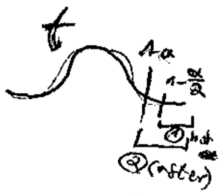
6. (5 points) The cross-validation ~~(CV)~~ residuals sum-of-squares is never smaller than the residuals sum-of-squares. ~~(True/False)~~

Explain: $CV = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(1-H_{ii})^2}$, $\frac{SS_{RES}}{CV}$ so when the denominator < 1 ⇒ $|H_{ii}|$ ... than it is smaller, ~~but it cannot be ... larger than~~

$$V = \sum_{i=1}^{n}(y_i - \hat{y}_i^{(i)})^2 = \sum_i \frac{\hat{\varepsilon}_i^2}{(1-H_{ii})^2}$$

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:** a one-sided is less recommended to begin with, but anyway, $t^{1-\alpha}$ has more area to the right of it (see figure) so the probability to get a value that exceeds it is larger.



$P_r(|t| > t^{1-\alpha}) > P_r(|t| > t^{1-\frac{\alpha}{2}})$

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
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: The (middle) one. You start with all predictors and remove 1 at every step until no statistically significant improvement (using F test extra sum of squares for example)
("drop") (the least sig.) or max SS_{RES}

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:

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

$$t = \frac{\mu_1 - (\mu_2 + 10)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad s^2 = \sum_{j=1}^{n_1}(y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \mu_2 + 10)^2$$

- 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 $y_1, \ldots, y_n$. We are given some $\mu_0 \in \mathbb{R}$ and would like to test the hypothesis

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

(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

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

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	$\vdots$	$\vdots$
Series2	$\vdots$	$\vdots$
Series2	$\vdots$	$\vdots$
Series1	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
Series2	$\vdots$	$\vdots$

(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.

University

Do not write anything in the margins

Part II

1. (i) We perform a two-sample t-test

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

$(N = n_1 + n_2)$

$$t = \frac{\mu_1 - (\mu_2 + 10)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{N-2}, \quad s^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \mu_2)^2}{n_1 + n_2 - 2}$$

Our test rejects H_0 at significance level α if $|t| > t_{N-2}^{1-\frac{\alpha}{2}}$

-1 (10)

(ii) Same but $H_0: \mu_1 = 10\mu_2$ 0.5

19 (10)

DF=2n-2

$$t = \frac{\mu_1 - 10\mu_2}{s\sqrt{\frac{1}{n_1} + 10^2\frac{1}{n_2}}} \sim t_{N-2}, \quad s^2 = \text{same as above}$$

2. (i) We set H_0 as the hyp. that $Y_i \sim N(\mu_0, \sigma^2)$.

For each Y_i we can check if it comes from a normal distribution $N(\mu_0, \sigma^2)$.

16 (11)

by $t = \frac{Y_i - \mu_0}{s\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu_0)^2$

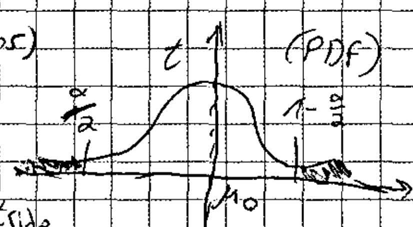
$\boxed{H_0: \mu = \mu_0}$

~~after we sample n points from a normal dist $N(\mu_0, \sigma^2)$~~

(ii) Then, we'll be able to reject / fail to reject H_0 for each Y_i and see if it comes from such distribution.

* reject if $|t| > t_{n-1}^{0.975}$ $(\alpha = 0.05)$

This is the CDF of t which gives us the area outside of $\frac{\alpha}{2}$ to $1-\frac{\alpha}{2}$ in blue



Instructor's notes:

2. (iii) The factors affecting my ability to detect $\mu_1 \neq \mu_0$ may be the sample size (too ~~small~~ small) so less accurate predictions.

Also, outliers may affect - by leading us to false assumptions (Type I (FP) or type II (FN) errors).

$$\frac{|\mu_1 - \mu_2|}{\sigma}$$

you did not mentione the "signal strength"

3. (i) We take the average score per series, based on all its unique contest ID (simply average score from all contests). denote μ_1 or μ_2

Now, we perform a two-sample t-test between ~~every~~ the averge scores (μ_i) ~~pair of~~ of series_i & series_j to check $H_0: \mu_i = \mu_j$

$$t = \frac{\mu_1 - \mu_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{N-2}, \quad N = n_1 + n_2$$

where n_i = number of scores for series i (a.k.a. number of contests it participated in)

$$s^2 = \frac{1}{N-2} \sum_{k=1,2} \sum_{i=1}^{n_k} (y_{ki} - \mu_k)^2$$

Our test rejects H_0 at significance level $\alpha = 0.05$ if $|t| > t_{N-2}^{0.975\ (\to 1-\frac{\alpha}{2})}$. if we reject H_0 we may assume that one series is better than the other! we can then check $H_0: \mu_1 > \mu_2$ and $H_0: \mu_1 < \mu_2$

(ii) In this case the process is similar, but the n_i (sample size) for each series may change and affect the numerical result of the t-statistic and therefore our final ~~result~~ conclusion (reject/fail to reject H_0) (may change the denominators of t and s)

-8 (12)

Instructor's notes:

12 (12)

Paired ttest

Instructor's notes:

Instructor's notes:

Instructor's notes:

Instructor's notes: