



Bar Ilan University

אוניברסיטת בר אילן

Department of Mathematics

המחלקה למתמטיקה

הנחיות לעבודה מסכמת בסטטיסטיקה וניתוח נתונים, 88624,

סמסטר א', שנה"ל תשפ"ד (2024).



## מידע מנהלי





- ✚ העבודה מורכבת מ-3 חלקים: 1) מודל סטטיסטי; 2) ניתוח פלטים; 3) מבחנים אפרמטריים.
- ✚ העבודה הינה עבודה עצמאית ויחידנית.
- ✚ לאחר הגשת העבודה, תיתכן בחינה בעל-פה, לפי שיקול דעת המרצה.
- ✚ העבודה תוגש בשפה העברית (לבירור בנוגע לאפשרות הגשה בשפה אחרת, ניתן לפנות למרצה).
- ✚ שאלות, הערות או הארות בנושאי העבודה יש להעלות בפורום הייעודי לכך באתר הקורס או בשעות קבלה ייעודיות (לא יהיה מענה פרטני במייל).

# חלק א'

מודל סטטיסטי



בעבודה תתרגלו תוכנה מתקדמת, Python, לניתוח ובדיקה של נתוני הרגלי אכילה.   
מודגש בזאת כי התשובות לשאלות ההנחיה כוללים רכיבי ניתוח והבנה משמעותיים,   
וחלקים אלו הינם עיקר ליבת העבודה. הצגת תדפיסי התוכנה - Python, אין בהם בכדי  
לגלות אלא כישרון טכני בלבד. ניתוח התדפיס והבנתו לעומק, במענה ישיר וברור הינם  
חלק עיקרי ונכבד בעבודה.

## תיאור קובץ הנתונים והנחיות לבניית מדגם



קובץ הנתונים לעבודה היישומית הנו קובץ בשם `onlinefood.csv`. זהו קובץ המכיל מידע שנאסף מפלטפורמת הזמנת מזון באינטרנט לאורך פרק זמן. ניתן לנצל מערך נתונים זה כדי לחקור את הקשר בין גורמים דמוגרפיים/מיקומיים לבין התנהגות הזמנת מזון באינטרנט, לנתח משוואות של לקוחות כדי לשפר את איכות השירות, ובאופן פוטנציאלי לחזות העדפות או התנהגות של לקוחות בהתבסס על מאפיינים דמוגרפיים ומיקומיים.

במסגרת הפרויקט, עלייך לבחור תת-מדגם מתוך הקובץ. מאפייני המדגם והמשתנה התלוי צריכים להיות ייחודיים לכל סטודנט. על בסיס ההנחיות הללו, יש לייצר את בסיס הנתונים לפרויקט, מתוך הקובץ. את הצעדים ליצירת בסיס הנתונים הייעודי יש לכלול בקובץ `Python` נפרד.

בכל השאלות שלהלן, כל בחינות ההשערה הינם ברמת מובהקות מקובלת של 5%.



פרק ראשון : בניית מודל

1) **מודל תיאורטי (כלכלי)** - התשובה על כל הסעיפים בשאלה זו תהיה עד

עמוד 1

- א. מבין המשתנים המופיעים בקובץ, יש לבחור משתנה שלדעתך משפיע באופן משמעותי על משתנה התוצאה (לחלופין, באפשרותכם ליצור משתנה חדש על בסיס המשתנים הקיימים). יש לנסח שאלת מחקר ביחס למשתנה זה. האם זה רלוונטי גם לישראל? (יש לצטט לפי כללי הציטוט המקובלים, בונס 10 נק').
- ב. בנוסף למשתנה המסביר שנבחר בסעיף א', יש לבחור 3 משתני בקרה מתוך המשתנים המופיעים בנתונים (חוץ מהמשתנה בשאלת המחקר). עליך להסביר מדוע כל אחד מהמשתנים חשוב לצורך מענה על שאלת המחקר- כיצד הוא מתקשר למשתנה התלוי וכיצד הוא מתקשר למשתנה בשאלת המחקר. בהתאמה, יש לבחור את הצורה הפונקציונלית בה נכנס כל משתנה למודל. לשם כך ניתן, ואף מומלץ, להמיר את הצורה הפונקציונלית של משתנים קיימים על מנת להתאים למודל- למשל, לוג, יצירת קטיגוריות מקבוצות וכו'.
- חשוב:** מתוך שלושת המשתנים, לפחות אחד צריך להיות רציף, ואחד בינרי (משתנה דמי).

(2) סטטיסטיקה תיאורית- עד חצי עמוד של טקסט.

מטרתה של הסטטיסטיקה התיאורית היא כפולה: (1) לקבל תמונת מצב של האוכלוסייה הנבדקת על מאפייניה; (2) לוודא שהמשתנים "מתנהגים יפה": אין מקרי קיצון שישפיעו על תוצאות האמידה; האם יש נתונים חסרים; יש שונות במשתנים המסבירים (מתקיים  $var(x) \neq 0$ )

א. יש להציג סטטיסטיקה תיאורית של משתנה התוצאה ושל המשתנים סטטיסטיקה תיאורית כוללת מאפיינים כמו מספר תצפיות (שאינן חסרות), ממוצע, סטיית תקן, מינימום ומקסימום. בנוסף הציגו באמצעו היסטוגרמה את התפלגות כל המשתנים הרציפים הכלולים במודל. מה ניתן ללמוד מהסטטיסטיקה התיאורית אודות אוכלוסיית המדגם?  
ב. יש להציג בצורה גרפית את הקשר בין אחד מהמשתנים המסבירים הרציפים במודל למשתנה התלוי. מה ניתן ללמוד מהגרף?

(3) מודל. את תוצאות כל הרגרסיות יש להציג בטבלה מרוכזת (דוגמה לטבלה

מרוכזת ניתן למצוא בסוף הקובץ). פרשנות והסברים על תוצאות הרגרסיה יש לכתוב בהיקף של עד עמוד אחד.

- א. יש לנסח מודל הנובע מהמודל התיאורטי של סעיף (1) כמשוואה- הקפידו לדייק בציון רמת התצפית של כל משתנה. (לצורך כך עליך להשתמש בעורך המשוואות ב-word).
- ב. יש לאמוד מודל מצומצם כאשר המשתנה המסביר היחיד הוא המשתנה שנבחר בסעיף א'.
- ג. כעת יש לאמוד את המודל המלא (מסעיף א).
- ד. יש לדון בתוצאות שתי האמידות ולהסביר מה ניתן להסיק מהן.
- ה. האם התוספת של משתני הבקרה שנבחרו מוסיפה לכוח ההסבר של המודל?

# חלק ב'

ניתוח פלטים





1. נתונות שלוש קבוצות של נשים, עבור כל אישה לידה ידוע משקלה.  
 הקבוצות מקודדות באופן הבא: קבוצה 1 – בחורות, 2 – ילדות ו- 3 – מבוגרות. האם יש הבדלים בתוחלות המשקלים בקבוצות השונות?  
 לפי הפלט הנתון ענה/י על השאלות הבאות:
- רשמו את הנחות המודל.
  - מהן השערות המחקר?
  - כמה נשים בכל קבוצה?
  - מה ערכם של הממוצעים וסטיות התקן בכל קבוצה?
  - מה בודק מבחן "Test of Homogeneity of Variances" ומדוע יש לבדוק זאת?
  - האם יש לאין הבדל בין ממוצעי הקבוצות? לפי מה ניתן להסיק זאת? הוכיחו, על פי חישוב את המסקנה.
  - חשבו את  $\eta^2$  ותנו פירוש מילולי לערך שנמצא.
  - הסברו את הצורך במבחן "Post Hoc Test" ומה ניתן להסיק ממבחן זה?

	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
1.00	13	53.5385	2.02548	52.3145	54.7624	50.00	55.00
2.00	30	13.8000	2.44103	12.8885	14.7115	8.00	18.00
3.00	18	55.3333	3.06786	53.8077	56.8589	50.00	65.00
Total	61	34.5246	20.72085	29.2177	39.8314	8.00	65.00

### Test of Homogeneity of Variances

weight

Levene Statistic	df1	df2	Sig.
.093	2	58	.911

### ANOVA

Weight

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	25379.182	2	12689.591	1926.537	.000
Within Groups	382.031	58	6.587		
Total	25761.213	60			

## Post Hoc Tests

### Multiple Comparisons

Weight

LSD

(I) group	(J) group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	39.73846*	.85219	.000	38.0326	41.4443
	3.00	-1.79487	.93413	.060	-3.6647	.0750
2.00	1.00	-39.73846*	.85219	.000	-41.4443	-38.0326
	3.00	-41.53333*	.76517	.000	-43.0650	-40.0017
3.00	1.00	1.79487	.93413	.060	-.0750	3.6647
	2.00	41.53333*	.76517	.000	40.0017	43.0650

\*. The mean difference is significant at the 0.05 level.

2. חברה רוצה לבדוק האם כל הסניפים של החברה עומדים בדרישות "זמן משלוחים מהיר" באופן זהה. לשם כך דגם באופן מקרי, חוקר החברה, חמישה סניפים בגדלים שונים ומדד את זמן המשלוח מרגע בקשת הלקוח. החוקר קבל את הפלט הבא:
- א. רשמו את הנחות המודל.
- ב. מהן השערות המחקר?
- ג. עבור כמה משלוחים בדק החוקר את זמן המשלוח בכל סניף?
- ד. מה ערכם של הממוצעים וסטיות התקן בכל סניף?
- ה. האם יש לאין הבדל בין ממוצעי הקבוצות? לפי מה ניתן להסיק זאת? הוכיחו, על פי חישוב את המסקנה.
- ו. חשבו את  $eta^2$  ותנו פירוש מילולי לערך שנמצא
- ז. מה ניתן להסיק ממבחן "Post Hoc Test" ?
- ח. מהו מבחן טוקי? הסברו את הצורך במבחן. האם החישוב זהה למבחן ?LSD

	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
1.00	5	606.4000	206.13903	350.4446	862.3554	354.00	854.00
2.00	5	240.4000	32.29241	200.3037	280.4963	200.00	278.00
3.00	4	385.0000	104.52751	218.6734	551.3266	235.00	478.00
4.00	3	850.3333	8.14453	830.1012	870.5655	841.00	856.00
5.00	5	632.8000	231.59814	345.2329	920.3671	254.00	851.00
Total	22	522.2273	250.10301	411.3378	633.1167	200.00	856.00

### Test of Homogeneity of Variances

time\_trans

Levene Statistic	df1	df2	Sig.
2.718	4	17	.065

### ANOVA

זמן המעבר עבור 5 הקבוצות

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	891975.997	4	222993.999	8.992	.000
Within Groups	421605.867	17	24800.345		
Total	1313581.864	21			

## Post Hoc Tests

Tukey HSD	1.00	2.00	366.00000*	.014	62.9696	669.0304
		3.00	221.40000	.266	-100.0123	542.8123
		4.00	-243.93333	.256	-593.8428	105.9761
		5.00	-26.40000	.999	-329.4304	276.6304
	2.00	1.00	-366.00000*	.014	-669.0304	-62.9696
		3.00	-144.60000	.654	-466.0123	176.8123
		4.00	-609.93333*	.000	-959.8428	-260.0239
		5.00	-392.40000*	.008	-695.4304	-89.3696
	3.00	1.00	-221.40000	.266	-542.8123	100.0123
		2.00	144.60000	.654	-176.8123	466.0123
		4.00	-465.33333*	.009	-831.2774	-99.3893
		5.00	-247.80000	.179	-569.2123	73.6123
	4.00	1.00	243.93333	.256	-105.9761	593.8428
		2.00	609.93333*	.000	260.0239	959.8428
		3.00	465.33333*	.009	99.3893	831.2774
		5.00	217.53333	.358	-132.3761	567.4428
	5.00	1.00	26.40000	.999	-276.6304	329.4304
		2.00	392.40000*	.008	89.3696	695.4304
		3.00	247.80000	.179	-73.6123	569.2123
		4.00	-217.53333	.358	-567.4428	132.3761
LSD	1.00	2.00	366.00000*	.002	155.8626	576.1374

	3.00	221.40000	.051	-1.4844	444.2844
--	------	-----------	------	---------	----------

	4.00	-243.93333*	.049	-486.5791	-1.2876
	5.00	-26.40000	.794	-236.5374	183.7374
2.00	1.00	-366.00000*	.002	-576.1374	-155.8626
	3.00	-144.60000	.189	-367.4844	78.2844
	4.00	-609.93333*	.000	-852.5791	-367.2876
	5.00	-392.40000*	.001	-602.5374	-182.2626
3.00	1.00	-221.40000	.051	-444.2844	1.4844
	2.00	144.60000	.189	-78.2844	367.4844
	4.00	-465.33333*	.001	-719.0984	-211.5683
	5.00	-247.80000*	.031	-470.6844	-24.9156
4.00	1.00	243.93333*	.049	1.2876	486.5791
	2.00	609.93333*	.000	367.2876	852.5791
	3.00	465.33333*	.001	211.5683	719.0984
	5.00	217.53333	.076	-25.1124	460.1791
5.00	1.00	26.40000	.794	-183.7374	236.5374

2.00	392.40000*	.001	182.2626	602.5374
3.00	247.80000*	.031	24.9156	470.6844
4.00	-217.53333	.076	-460.1791	25.1124

## Homogeneous Subsets

זמן מעבר

סניפים	N	Subset for alpha = 0.05		
		1	2	3
Tukey B <sup>a,b</sup> 2.00	5	240.4000		
3.00	4	385.0000	385.0000	
1.00	5		606.4000	606.4000
5.00	5		632.8000	632.8000
4.00	3			850.3333

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.225.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.



# חלק ג'

מבחנים אפרמטריים



1. על מנת לבדוק את ההשפעה של ההורמון מלטונין כטיפול בנדודי שינה, חוקר גייס 10 נבדקים. כל נבדק ישן 2 לילות במעבדה. באחד הלילות קיבל placebo ובלילה השני melatonin (double-blind – לצורך השאלה חשוב להבין שלא היה סדר קבוע של לקיחת התרופות). לאחר כל לילה התבקשו הנבדקים לסמן בסקלה שבין 1 ל-10 את איכות השינה שלהם. מה תהיה מסקנת החוקר ברמת בטחון של 95% לגבי יעילותו של מלטונין כטיפול לבעיית שינה?
2. מורה לנהיגה מעוניין לבדוק האם יש הבדל בין גברים לנשים בכמות הטסטים עד שמקבלים רישיון נהיגה, לצורך כך הוא דגם 8 גברים ונשים ולהלן התוצאות, בדקו את ההשערה ברמות מובהקות של 5%.

גברים	3	2	4	1	3	5	6	2
נשים	2	4	5	6	2	5	1	

3. לפני עימות רדיופוני שבא לקבוע מי מבין שני מועמדים פופולרי יותר בקרב בני הנוער, נדרשו 300 בני נוער להכריע בין שניהם. לאחר העימות נשאלו אותם בני נוער בשנית, התוצאות:

	אחרי			
	מועמד 2	מועמד 1		
לפני	B18	A102	מועמד 1	120
	D157	C23	מועמד 2	180
	175	125		300

האם שדרן רדיו יכול לקבוע בר"מ 0.05 שקיים הבדל בין שני המועמדים במספר הקולות שסחפו לטובתם?



■ **הגשת העבודה תיעשה בשני קבצים באתר הקורס ב-Moodle:** קובץ העבודה בפורמט

PDF<sup>1</sup> וקובץ ה-Python (חלק א') שיאפשר שחזור של התוצאות. **אין צורך** להגיש עותק מודפס של העבודה ולא תתקבלנה הגשות בדוא"ל!

■ **קובץ העבודה ב-PDF צריך לכלול פתרון מלא, ענייני ומפורט של כל הסעיפים מכל החלקים, לפי סדר הופעתם בהנחיות. הפתרונות צריכים להיות מוצגים באופן הבא:** תשובות מילוליות לכל החלקים בפרק; טבלאות וגרפים בעמוד נפרד עוקב; תדפיס הקוד הרלוונטי, אם רלוונטי לשאלה (צילום מסך). בפתרון של כל סעיף יש לפרט את כל הפעולות שבוצעו, הערכים שחושבו והמבחנים שנערכו עד להגעה לתוצאה הסופית. מודגש בזאת כי התשובות לשאלות ההנחיה כוללים רכיבי ניתוח והבנה משמעותיים, וחלקים אלו הינם עיקר ליבת העבודה. כמו כן, יש לוודא כי הקשר בין כל נתון ששימש אתכם, לבין התדפיס הרלוונטי מה-Python, יהיה ברור לחלוטין.

א. את הרגרסיות **אין להציג בפלט** אלא באמצעות טבלה מסודרת כפי שמוצג בנספח. שמות המשתנים התלויים והבלתי תלויים צריכים להיות ברורים מתוך קריאת הטבלה (בלי קיצורים או שמות קוד).

ב. בעבור מבחנים, יש לרשום בגוף הטקסט את השערות המבחן ואת התוצאות ולצרף **קטעים רלוונטיים מהפלט** אשר אליו מתייחסים בסעיף של השאלה. על קטעי הפלט להיות ערוכים באופן **ברור וקריא**.

#### ■ **פורמט הגשה לקובץ העבודה:**

א. העבודה צריכה לכלול שער הכולל את כל הפרטים הבאים: האוניברסיטה, המחלקה, שם הקורס, מספר הקורס (כולל מספר הקבוצה), שם המרצה, שם פרטי, שם משפחה, מספר תעודת זהות (9 ספרות) ותאריך ההגשה.

ב. פונט David, גודל 12; שורה וחצי רווח בין שורות; יישור לשני הצדדים; שוליים רגילים; יש למספר עמודים

ג. טבלאות בעברית- יישור לימין (עמודת המשתנים מימין)

---

<sup>1</sup> ניתן להמיר קובץ word לקובץ PDF על ידי save as PDF או ע"י הדפסה ל-PDF דרך פונקציית ההדפסה בוורד.

ד. מספרים בטבלאות- עד שלוש ספרות אחרי הנקודה העשרונית.

ה. יש למספר טבלאות ותרשימים ולהפנות אליהם באופן ברור.

ו. בסוף העבודה יש לכתוב **ביבליוגרפיה** מלאה שתכלול פירוט של כל ספר, אתר אינטרנט, מקור מידע או אדם שעזר לכם בביצוע עבודתכם. יש לפרט את מקור המידע, זמן העיון בו ומה תרם לכם מקור זה בביצוע המטלה. חובה להקפיד על רישום מסודר, ברור ואמין של הביבליוגרפיה.

■ **קובץ Python:** קובץ ה-Python צריך להיות מותאם לפתיחה בכל סביבת עבודה של Python. קובץ הפעולה צריך להיות במצב כזה שיאפשר לבדוק לקבל את הפתרון לכל סעיפי הפרויקט בהפעלה בודדת. המשמעות היא שמעבר למבחנים ולרגרסיות, הקובץ צריך לכלול את כל הפקודות שמייצרות נתונים ובוחרות את המדגם הרלוונטי. יש לציין את מספרי השאלות המתאימים ביחס לכל חלק בקובץ.

■ חובה לשמור גיבויים וצילומים של כל החומר המועבר לבדיקה, לכל מקרה של בירור.

■ הוראות הגשה נוספות ייתכן ויפורסמו, סמוך למועד ההגשה - נא להתעדכן בהתאם.

## נספח- טבלאות לדוגמה:

### סטטיסטיקה תיאורית:

לוח 1: אפיונים סטטיסטיים שונים של התלמידים בכיתות מב"ר ובכיתות הרגילות

המשתנה	כלל המדגם	הכיתות הרגילות	כיתות מב"ר
מיצ"ב ח'	54.11 (25.674)	56.39 (25.182)	32.00 (19.031)
שנות הלימוד האם	13.10 (3.092)	13.22 (3.122)	12.03 (2.565)
מספר האחים	2.79 (1.644)	2.78 (1.644)	2.87 (1.643)
שיעור הבנים	47% (0.499)	47% (0.499)	45% (0.498)
הערות: סטיות תקן בסוגריים.			

### תוצאות רגרסיה:

לוח 3

השפעת שיעור היצוא הענפי על השכר תוך בקרה על מאפייני פרט קבועים, סך המשק לעומת התעשייה בלבד, פאנל עובדים לשנים 2008-2015

כל הענפים התעשייה בלבד		
(2)	(1)	
0.126*** (0.00425)	0.389*** (0.00121)	שיעור היצוא (%)
0.128*** (0.000817)	0.153*** (0.000311)	שנות לימוד
0.141*** (0.00168)	0.189*** (0.000731)	גיל
-0.00128*** (2.57e-05)	-0.00194*** (1.15e-05)	גיל בריבוע
כן	כן	FE ברמת העובד
4.136*** (0.0276)	2.654*** (0.0112)	החותך
745,437 0.901	7,276,263 0.765	מספר התצפיות R-squared

הערות: בסוגריים - סטיות התקן. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .  
שיעור היצוא הענפי (נכון לשנת 2006) בענף בו הפרט מועסק (2 ספרות). המקור: קובץ עובד מעבד של הלשכה המרכזית לסטטיסטיקה, ילידי 1985-1975.

