סטטיסטיקה
למנהל עסקים א'

ד"ר שרבל שוקייר

דוקטורנטית סיון ריף

ד"ר שרבל שוקייר

דוקטורנטית סיון ריף

הקריה האקדמית אונו, רח' צה"ל 104, קרית אונו

הקריה האקדמית אונו
הפקולטה למנהל עסקים
לתואר בוגר
שם הקורס :סטטיסטיקה למנהל עסקים א'
סוג הקורס: חובה
דרישות קדם: אין

מטרת הקורס

רכישת ידע לגבי תיאור גרפי של נתונים ,ארגון נתונים בטבלאות וסיכום הנתונים על ידי מדדים, בשילוב עם תורת ההסתברות כהכנה להסקה סטטיסטית הנלמדת בקורס ההמשך סטטיסטיקה ב.' בנוסף נלמד בקורס בדיקת קשר ליניארי בין משתנים כמותיים למטרת ניבוי על פי עקרון הריבועים הפחותים.

רשימת קריאה חובה

הרצאות הקורס לא יינתנו בצמוד לחומר הלימוד .עם זאת מומלצים ארבעה ספרים המכסים את החומר הנלמד:

1. רונית איזנבך ,**סטטיסטיקה ללא סטטיסטיקאים** ,הוצאת אקדמון
2. **מבוא לסטטיסטיקה למדעי החברה** ,האוניברסיטה הפתוחה יחידות5 , 0992.-0
3. **מבוא לסטטיסטיקה למדעי החברה** ,האוניברסיטה הפתוחה יחידות8 , 0995.-6
4. שולה ישראלית ,**סטטיסטיקה הלכה למעשה** ,הוצאת לוגיק0997
5. שרון אביטל ,**פתרונות למבחנים בסטטיסטיקה** ,הוצאת אקדמון 3113
6. המבורגר ,ר .,מרדכי ,ח .,**מבוא לסטטיסטיקה א('ללא סטטיסטיקאים)** פתרונות למבחנים ,הוצאת אקדמון. 3112

חובות הקורס

קורס השלמה. חובת מעבר של כל הבחנים והחובות הנדרשים בכל שלבי הקורס.

להלן הפרקים עבורם קיימת דרישת מעבר של בחנים:

פרק 1- סטטיסטיקה תיאורית

פרק 2- הסתברות

פרק 3 – אמידת קשר בין משתנים ורגרסיה לינארית

יתר הפרקים הינם בגדר קריאת רשות.

<u>הנושאים שילמדו בקורס</u>

פרק 1	**סטטיסטיקה תיאורית**	
נושא 1	מבוא- השיטה הסטטיסטית, מדגם ואוכלוסייה, סוגי משתנים	
נושא 2	טבלאות- שכיחות, שכחיות יחסית, שכיחות מצטברת.	
נושא 3	גרפים-דיאגרמת מקלות, היסטוגרמה, עוגה שכיחות יחסית מצטברת ופוליגון	
נושא 4	מדדים לנטיה מרכזית (שכיח, חציון, ממוצע)	
נושא 5	מדדי מיקום (אחוזונים)	
נושא 6	מדדי פיזור (תחום, שונות וסטית תקן,תחום בין רבעוני).	
נושא 7	סוגי התפלגויות	
נושא 8	גבול מדומה וגבול אמיתי	
נושא 9	שינוי לינארי	
נושא 10	ציון תקן	
פרק 2	**הסתברות**	
נושא 1	מושגים בתורת הקבוצות	
נושא 2	פעולות בין מאורעות ודיאגרמת וואן- חיתוך ואיחוד מאורעות	
נושא 3	תצורות שונות של דיאגרמת וואן- קבוצות חלקיות, זרות ולא זרות	
נושא 4	כללי דה מורגן	
נושא 5	כתיב קבוצות ודיאגרמת וואן לשתי קבוצות	
נושא 6	כתיב קבוצות ודיאגרמת וואן לשלוש קבוצות	
נושא 7	יסודות ההסתברות- מושג ההסתברות ותכונות ההסתברות	
נושא 8	ההסתברות של איחוד מאורעות זרים ומאורעות לא זרים	
נושא 9	הסתברות מותנית	
נושא 10	מאורעות תלויים ובלתי תלויים	
נושא 11	עצי החלטות	
פרק 3	**אמידת קשר בין משתנים ורגרסיה לינארית**	
נושא 1	הצגה וניתוח של קשר בין משתנים, דיאגרמת פיזור	
נושא 2	מקדם המתאם של פירסון ושונות משותפת	
נושא 3	מודל הרגרסיה הלינארי והשימוש בו לניבוי	
פרק 4	**משתנים מקריים בדידים**	**קריאת רשות**
נושא 1	משתנה מקרי בינומי	
נושא 2	התפלגות פואסונית	
פרק 5	**התפלגות נורמלית**	**קריאת רשות**
נושא 1	מבוא- מדד פרמטר ומדד סטטיטי ותכונות ההתפלגות הנורמלית	
נושא 2	ערכי Z ושימוש בלוח ההתפלגות הנורמלית הסטדנרטית	
נושא 3	חישוב הסתברויות ההתפלגות הנורמלית	
נושא 4	חישוב ערכים לפי אחוז התצפיות בהתפלגות הנורמלית	
פרק 6	**יישומים באקסל**	**קריאת רשות**

<u>השיטה הסטטיסטית</u>

בעולם של היום אנחנו חשופים לכמויות עצומות של מידע. נושא חשוב ביותר הוא כיצד להשתמש במידע ולהפיק ממנו תועלת. סטטיסטיקה עוזרת לנו לארגן ולהציג את הנתונים בצורה שימושית, לנתח את הנתונים ולהסיק מסקנות. ידע בסטטיסטיקה חשוב בכל תחומי עיסוק השונים כגון: רפואה, שיווק, מנהל עסקים, פסיכולוגיה ועוד.

נבדיל בין שני חלקים של השיטה הסטטיסטית:

1. **סטטיסטיקה תיאורית-** שיטות שעוסקות בסיכום ותיאור המידע בצורה שימושית יותר. לדוגמא: בניית לוחות סטטיסטיים, הצגה גרפית וחישוב מדדים סטטיסטיים שונים שיעזרו לנו להבין את התמונה בצורה טובה יותר.

2. **סטטיסטיקה היסקית-** שיטות שעוזרות לנו להסיק מסקנות לגבי כלל האוכלוסייה על בסיס מידע ותוצאות של מדגם מתוך האוכלוסייה.

<u>הגדרות בסיסיות</u>

<u>אוכלוסיית המחקר</u>

את מי אנחנו חוקרים? אוסף של פרטים אותם אנחנו מעוניינים לחקור. דוגמא: סטודנטים בישראל, מסעדות ועוד.

<u>מדגם</u>

תת קבוצה בתוך האוכלוסייה. לדוגמא: 50 סטודנטים בקריה האקדמית אונו.

<u>משתנה</u>

תכונה המקבלת ערכים שונים באוכלוסיית המחקר.

לדוגמא: באוכלוסיית הסטודנטים בישראל המשתנים יכולים להיות גיל, מגדר, צבע שיער, מצב משפחתי ועוד.

<u>סוגי משתנים</u>

1. **משתנה איכותי (שמי)**

הערכים של המשתנה הם שמות ללא ביטויים מספרים. לדוגמא: מגדר, מצב משפחתי, צבע שיער ועוד.

2. **משתנה כמותי**

הערכים של המשתנה הם במספרים. לדוגמא: גיל, גובה, משקל, מספר נפשות וכו'

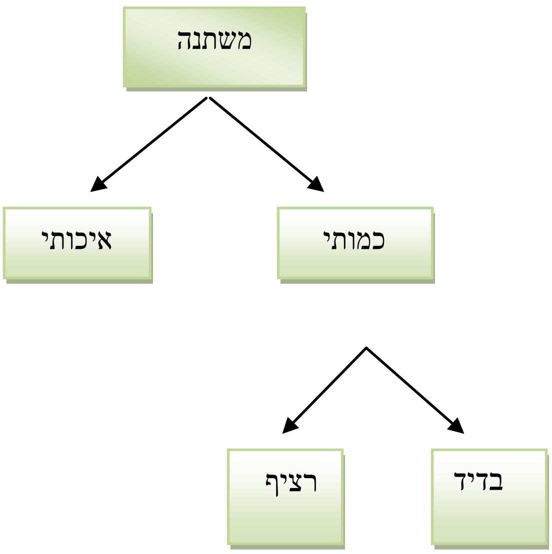
משתנה כמותי ניתן למיין לשני סוגים:

- **משתנה כמותי בדיד**

ערכי המשתנה הם מספרים בודדים, כגון: מס' קורסי החובה, מספר עובדים בחברה.

- **משתנה כמותי רציף**

ערכי המשתנה רציפים. משתנה המקבל אין סוף ערכים אפשריים באותה קטגוריה. לדוגמא: משקל, גובה.



<u>משתנה כמותי בדיד</u>

טבלת שכיחויות למשתנה כמותי בדיד

טבלת שכיחויות הינה טבלה המציגה את השכיחות של הערכים השונים.

<u>דוגמא:</u>

להלן התפלגות הגילאים של חברה עם 25 עובדים:

20,35,20,45,25,35,35,45,50,20,25,25,45,50,35,25,50,45,50,35,35,45,45,50,35

P = F/n שכיחות יחסית מצטברת	p= f/n שכיחות יחסית	F שכיחות מצטברת	f(x) שכיחות	x
3/25	3/25	3	3	20
7/25	4/25	7	4	25
14/25	7/25	14	7	35
20/25	6/25	20	6	45
25/25	5/25	25	5	50
	סה"כ 25/25=1		סה"כ n=25	

<u>הגדרות:</u>

<u>שכיחות</u> - מספר המקרים לכל ערך- f(x) או f

<u>שכיחות מצטברת</u> – מספר המקרים עד (וכולל) אותו ערך. נסמן את השכיחות המצטברת באות F.

<u>גודל המדגם</u>- מספר הערכים במדגם. נסמן באות n.

<u>שכיחות יחסית</u>– השכיחות מובעת באחוזים. כלומר, שכיחות לחלק לגודל המדגם. נסמן באות p.

<u>שכיחות יחסית מצטברת</u>- השכיחות המצטברת מובעת באחוזים. כלומר, השכיחות המצטברת לחלק לגודל המדגם. נסמן באות P.

כללים לבניית טבלת שכיחויות:

א. לסדר את המשתנה מהערך הנמוך לגבוה.

ב. במחלקה האחרונה השכיחות המצטברת שווה לגודל המדגם.

ג. השכיחות היחסית המצטברת במחלקה האחרונה שווה ל- 1.

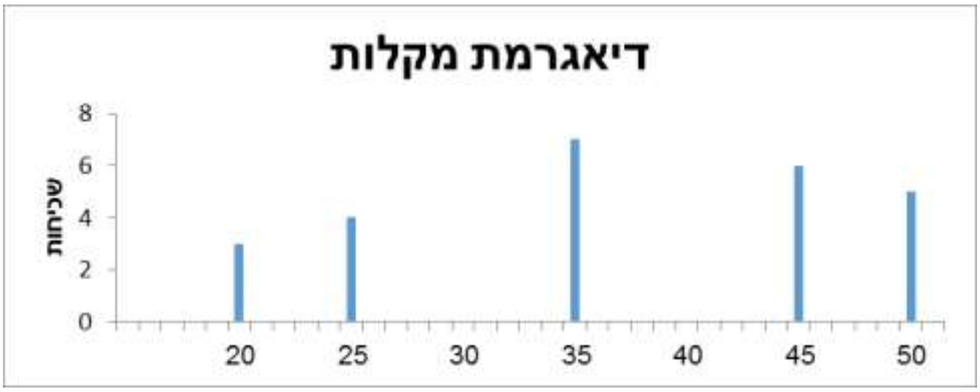
ד. בסטטיסטיקה תיאורית שכיחות יחסית גם שווה להסתברות. ניתן לדעת מה ההסתברות רק מלהסתכל על הטבלה, אין צורך לעשות חישובי הסתברות מסובכים. לדוגמא:

מה הסיכוי בחברה זו לבחירת עובד שהגיל שלו הוא 25? ניתן לראות לפי השכיחות היחסית שהסיכוי הוא 4/25.

הצגה גרפית עבור משתנה כמותי בדיד

הדיאגרמה המתאימה לתיאור משתנה כמותי בדיד הינה דיאגרמת מקלות.

על ציר ה –X נציג את המשתנה ועל ציר ה Y את השכיחות. בדוגמא שלנו נקבל את הדיאגרמה הבאה:



משתנה כמותי רציף

טבלת שכיחויות עבור משתנה כמותי רציף

נשתמש בטבלת שכיחויות במחלקות עם טווח כאשר המשתנה הוא רציף או כאשר הנתונים רבים ויש צורך לפשט את טבלת השכיחויות על מנת שלא תהיה ארוכה מידי.

דוגמא:

בהסתמך על הדוגמא הקודמת יש להציג את הנתונים בטבלת שכיחויות המחולקת לשלוש מחלקות.

כיצד נקבע את רוחב המחלקה?

רוחב מחלקה = טווח לחלק למספר המחלקות

טווח = הערך המקסימאלי Xmax, פחות הערך המינימאלי Xmin

בדוגמא שלנו:

טווח 30= 50-20

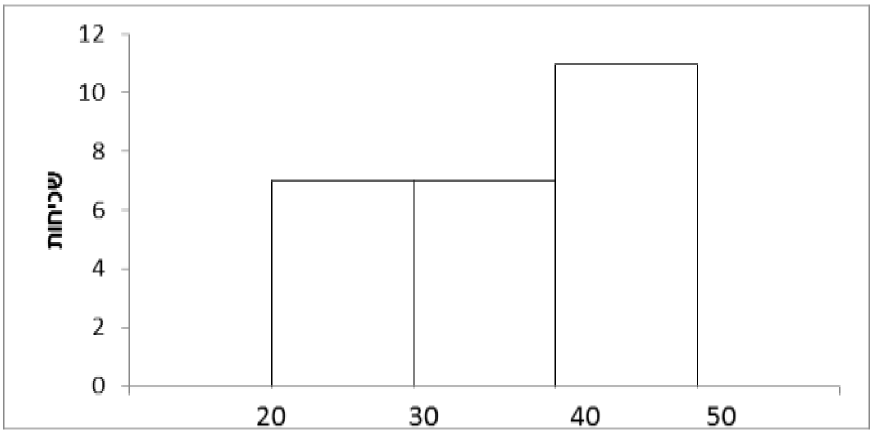
רוחב המחלקה 10= 30/3

x	f(x) שכיחות	F שכיחות מצטברת	p= f/n שכיחות יחסית	P = F/n שכיחות יחסית מצטברת
20-30	7	7	7/25	7/25
30-40	7	14	7/25	14/25
40-50	11	25	11/25	25/25=1
	סה"כ n=25		סה"כ 25/25=1	

הערה: לכל ערך חייבת להיות קטגוריה, ולא יכול להיות ערך עם יותר משתי קטגוריות. לכן, בקורס זה נגדיר כל מחלקה עד וכולל את אותו מספר. (כלומר לדוגמא מחלקה 30-40 היא בעצם 30.0001 – ועד 40 כולל).

הצגה גרפית עבור משתנה כמותי רציף

היסטוגרמה (דיאגרמת מלבנים) הינה הדיאגרמה המתאימה לתיאור משתנה כמותי רציף.



בניית היסטוגרמה כאשר רוחב המחלקות שונה

כשבונים היסטוגרמה ורוחב המחלקות שווה ניתן לבחור בין שכיחות רגילה לצפיפות. אם רוחב המחלקות אינו שווה, ניתן לבנות היסטוגרמה לפי צפיפות בלבד.

צפיפות הינה השכיחות ליחידה. נסמן צפיפות ב 'f.

כאשר:

C = רוחב מחלקה

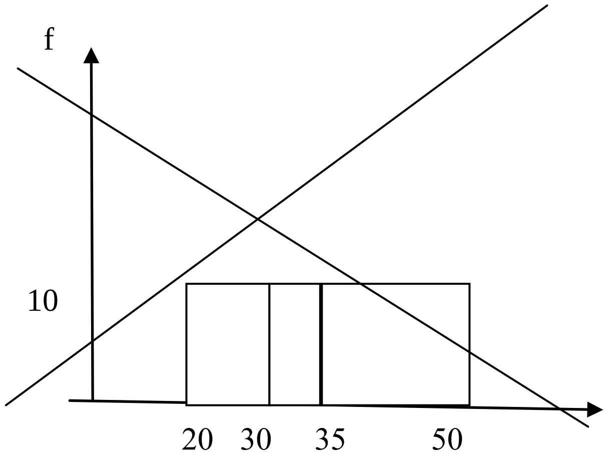
$f' = f/c$ (צפיפות= השכיחות ליחידה אחת)

דוגמא:

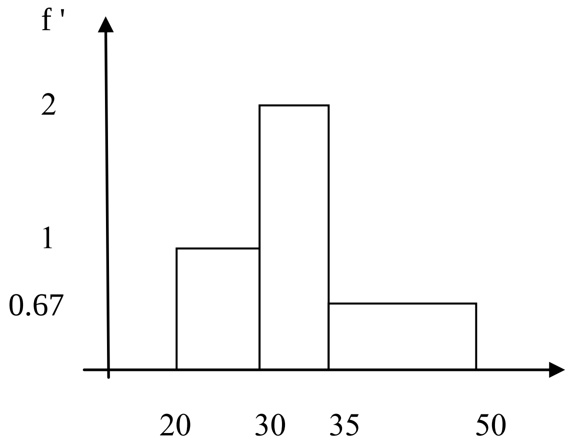
x	f	f'=f/c
20-30	10	10/10=1
30-35	10	10/5=2
35-50	10	10/15=0.67

n=30

הגרף מטה אינו נכון ומטעה.



לכן כאשר בונים היסטוגרמה ורוחב המחלקות הוא שונה אז יש לבנות את ההיסטוגרמה לפי צפיפות:



משתנה איכותי

טבלת שכיחויות עבור משתנה איכותי

לפי סקר של 1,500,000 צופים, להלן נתונים לגבי תוכנית הטלויזיה האהובה עליהם.

נתונים (באלפים):

"בשלנים ללא הפסקה" 1,350

"ראשון בתקשורת" 45

"תא 22" 15

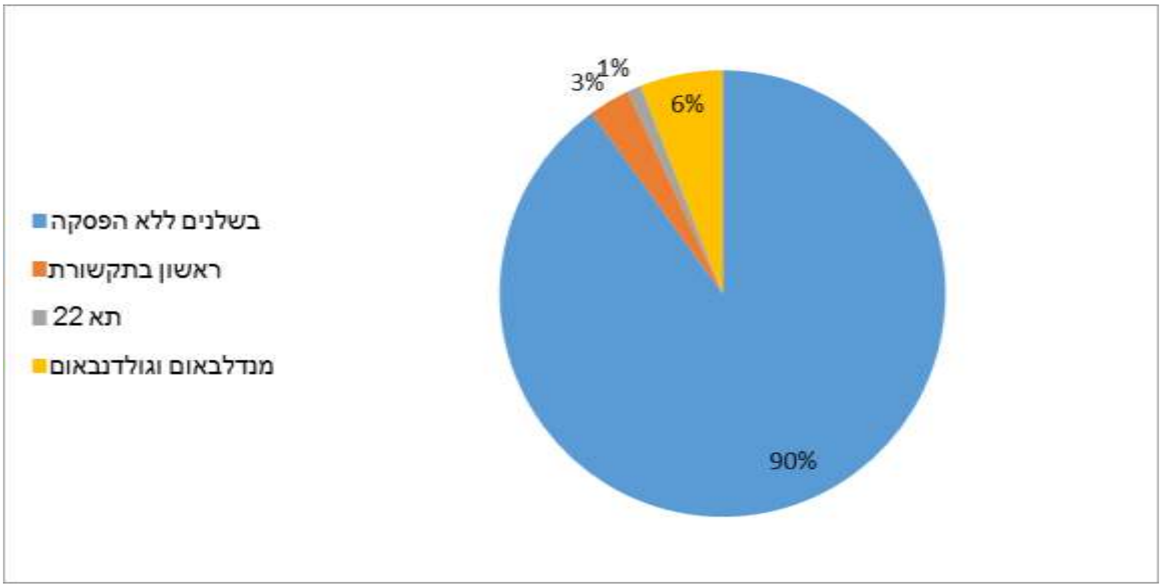
"מנדלבאום וגולדנבאום" 90

נסדר את הנתונים בטבלת שכיחויות כדלקמן (נתונים באלפים):

p= f/n שכיחות יחסית	f(x) שכיחות	x
(1350/1500)=90%	1350	"בשלנים ללא הפסקה"
(45/1500)=3%	45	"ראשון בתקשורת"
(15/1500)=1%	15	"תא 22"
(90/1500)=6%	90	"מנדלבאום וגולדנבאום"
סה"כ 100%	סה"כ n=1500	

הצגה גרפית עבור משתנה איכותי

הדיאגרמה המתאימה עבור משתנה שמי או איכותי הינה דיאגרמת עוגה (מעגל).



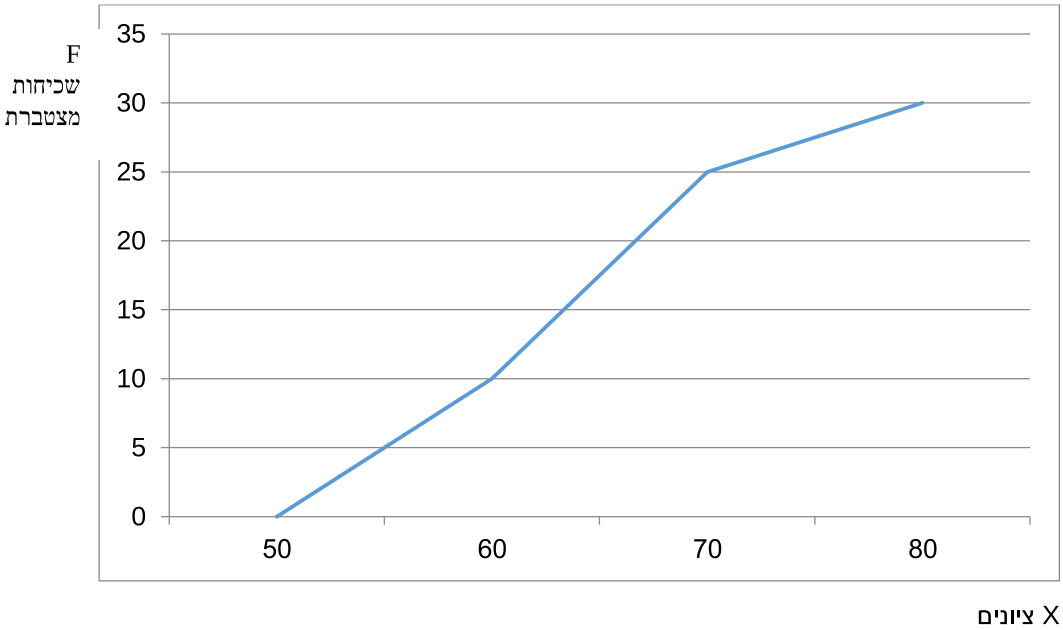
משתנה כמותי רציף- דיאגרמות נוספות

1. גרף שכיחות מצטברת

דיאגרמה זו מציגה את השכיחות המצטברת באופן גרפי.

F	f	X (ציונים)
10	10	50-60
25	15	60-70
30	5	70-80

סה"כ n=30



על ידי גרף זה ניתן לענות על שאלות כגון:

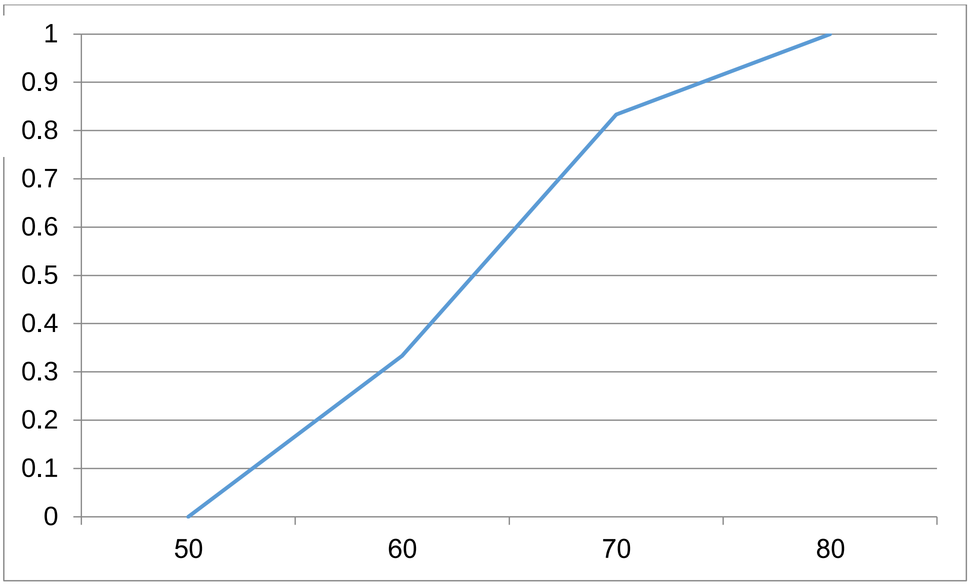
- מה מספר הסטודנטים שציון שלהם בין 60 ל 80? (תשובה 10-30=20).
- מה מספר הסטודנטים שקיבלו ציון נמוך או שווה ל 70 (תשובה 25).

2. <u>גרף שכיחות יחסית מצטברת (באחוזים)</u>

זהה לגרף שכיחות מצטברת רק שהתוצאות באחוזים.

X	f	F	P	P
50-60	10	10	10/30=33.33%	33.33%
60-70	15	25	15/30=50%	83.33%
70-80	5	30	5/30=16.67%	100%

P
שכיחות
יחסית
מצטברת



X ציונים

מהו אחוז הסטודנטים שהציון שלהם קטן או שווה ל 70? תשובה : 0.8333

מהו אחוז הסטודנטים שקיבלו ציון בין 80 ל 60 ? 1-0.3333=0.66667

<u>הערות</u>

- שימו לב שלא מדובר בקו ישר. הגרף מראה מגמה. הגרף מתחיל בחיתוך עם ציר ה- X ומראה איך הסתברות מתחילה מאפס ועד 100%.
- ישנן שיטות נוספות להציג את השכיחות היחסית המצטברת באופן גרפי, למשל על ידי דיאגרמה בצורת מלבנים (מדרגות).

3. <u>גרף פוליגון (מצולע שכיחויות)</u>

מצולע שכיחויות נותן לנו קו שמפשט את ההיסטוגרמה על ידי חיבור מרכזי המלבנים. הפוליגון מאפשר לנו לראות באופן ברור יותר את צורת ההתפלגות והוא גם נוח כאשר משווים בין כמה התפלגויות. כדי להשלים את גרף הפוליגון נחבר את אמצעי הקבוצות הקיצוניות לאמצעי קבוצות "דמיוניות" בקצוות של הגרף.

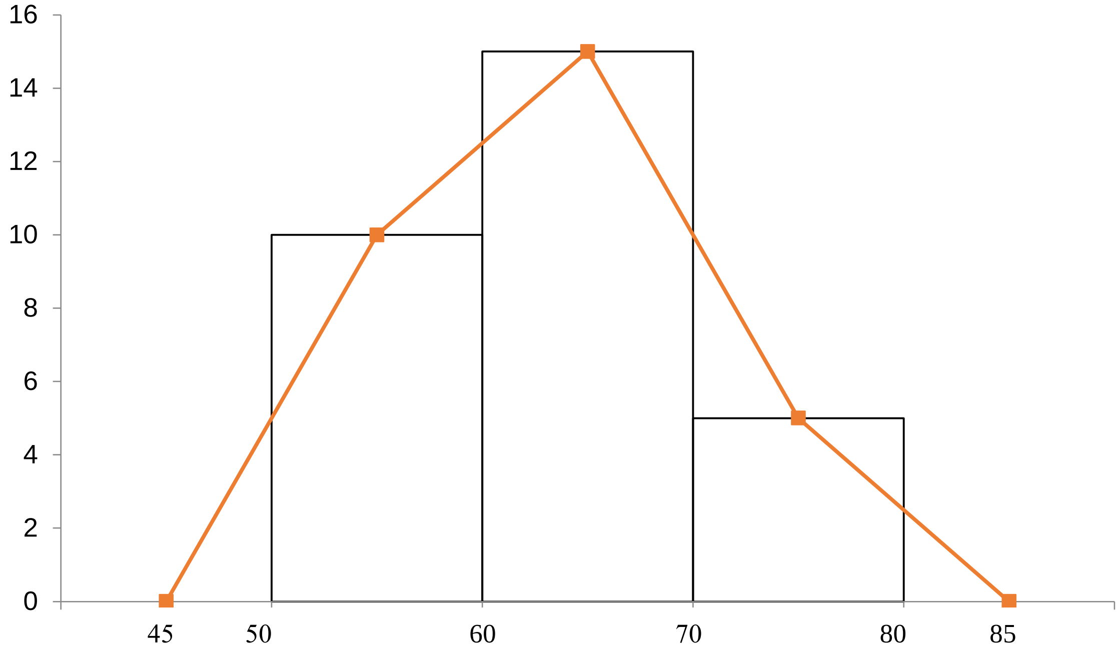
אופן בניית גרף פוליגון:

1. לוקחים את רוחב המחלקה הראשונה ונחלק אותו ב-2. ניקח את המספר הראשון ונחסיר ממנו את התוצאה.
2. לוקחים את רוחב המחלקה האחרונה, מחלקים ב- 2 ומוסיפים את התוצאה למספר האחרון.
3. על מנת לצייר פוליגון יש למצוא את אמצע המחלקות ולהעביר ישר בין כל אמצע המחלקה.

<u>דוגמא:</u>

ציונים X	f	F
50-60	10	10
60-70	15	25
70-80	5	30

n=30


16
14
12
10
8
6
4
2
0
45
50
60
70
80
85

שאלות לתרגול (1)

שאלה 1

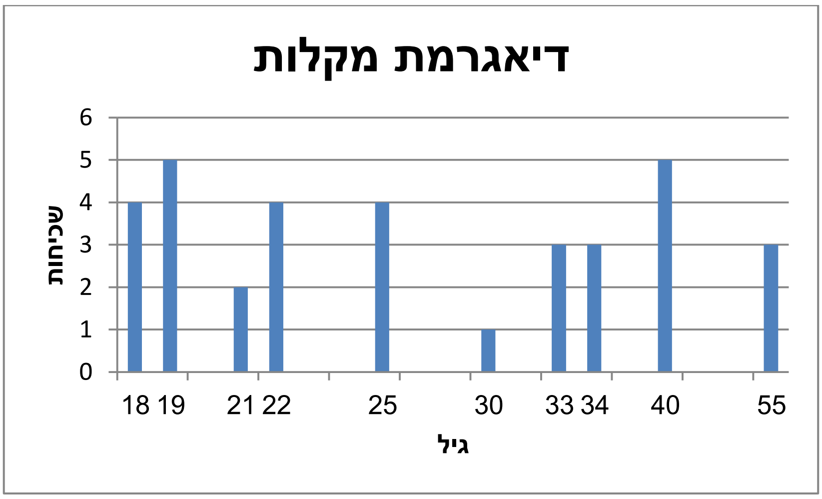
להלן התפלגות גילאים של חברת הייטק בתל אביב המעסיקה 34 עובדים:

22, 25, 30, 33, 55, 40, 18, 19, 21, 34, 33, 34, 40, 22, 55, 18, 19,
34, 33, 40, 19, 18, 22, 25, 21, 19, 40, 55, 19, 18, 25, 22, 25, 40

בנו טבלת שכיחויות מלאה והציגו את הנתונים באמצעות דיאגרמת מקלות.

פתרון:

x	f(x) שכיחות	F שכיחות מצטברת	p= f/n שכיחות יחסית	P = F/n שכיחות יחסית מצטברת
18	4	4	4/34	4/34
19	5	9	5/34	9/34
21	2	11	2/34	11/34
22	4	15	4/34	15/34
25	4	19	4/34	19/34
30	1	20	1/34	20/34
33	3	23	3/34	23/34
34	3	26	3/34	26/34
40	5	31	5/34	31/34
55	3	34	3/34	34/34
	סה"כ n=34		סה"כ 34/34=1	



<u>שאלה 2</u>

להלן הגילאים של 42 תושבים מיישוב בצפון הארץ:

50, 37, 48, 44, 53, 66, 70, 69, 63, 67, 57, 32, 41, 86,

37, 68, 67, 73, 37, 60, 64, 61, 62, 69, 54, 75, 68, 57,

67, 49, 71, 60, 72, 71, 38, 48, 90, 76, 88, 20, 64, 89

חלק את הנתונים לחמש מחלקות שוות רוחב, בנה טבלה שכיחויות מלאה והצג את הנתונים באמצעות היסטוגרמה.

<u>פתרון:</u>

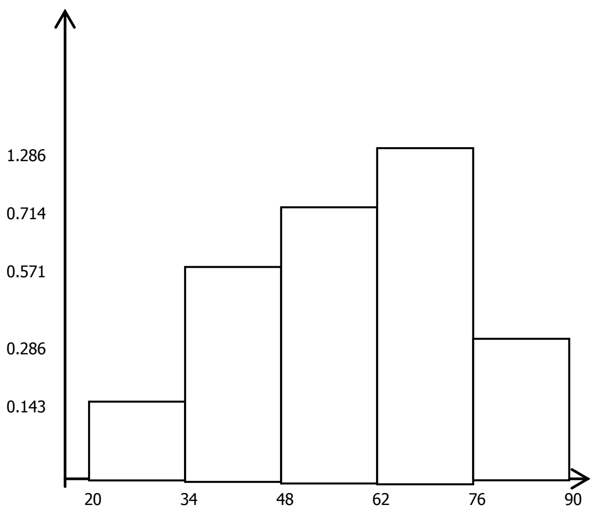
טווח = 70=90-20

רוחב מחלקה = 14=70/5

x	f(x) שכיחות	F שכיחות מצטברת	p= f/n שכיחות יחסית	P = F/n שכיחות יחסית מצטברת	ƒ' צפיפות ƒ/c
20-34	2	2	$^{2}/_{42}$	$^{2}/_{42}$	0.143
34-48	8	10	$^{8}/_{42}$	$^{10}/_{42}$	0.571
48-62	10	20	$^{10}/_{42}$	$^{20}/_{42}$	0.714
62-76	18	38	$^{18}/_{42}$	$^{38}/_{42}$	1.286
76-90	4	42	$^{4}/_{42}$	$^{42}/_{42}$	0.286

n = 42

נבנה היסטוגרמה לפי צפיפות (במקרה זה ניתן גם להציג שכיחות כי רוחב המחלקות שווה)



שאלה 3

בחברה חקלאית מדדו את מספר השתילים בכל חלקה. להלן התוצאות ממדגם מקרי של 60 חלקות:

11, 33, 52, 61, 22, 27, 16, 38, 49, 13, 20, 27, 37, 30, 16,
35, 49, 27, 29, 28, 15, 18, 30, 65, 7, 11, 32, 25, 19, 23,
37, 25, 24, 26, 30, 21, 20, 30, 32, 10, 8, 20, 22, 59, 12,
33, 67, 9, 36, 50, 57, 19, 37, 48, 26, 9, 66, 53, 28, 29

א. חלק את הנתונים ל- 6 מחלקות שוות רוחב.
ב. בנה טבלת שכיחות מלאה.
ג. שרטט היסטוגרמה, פוליגון שכיחויות, וגרף שכיחות יחסית מצטברת.

פתרון:

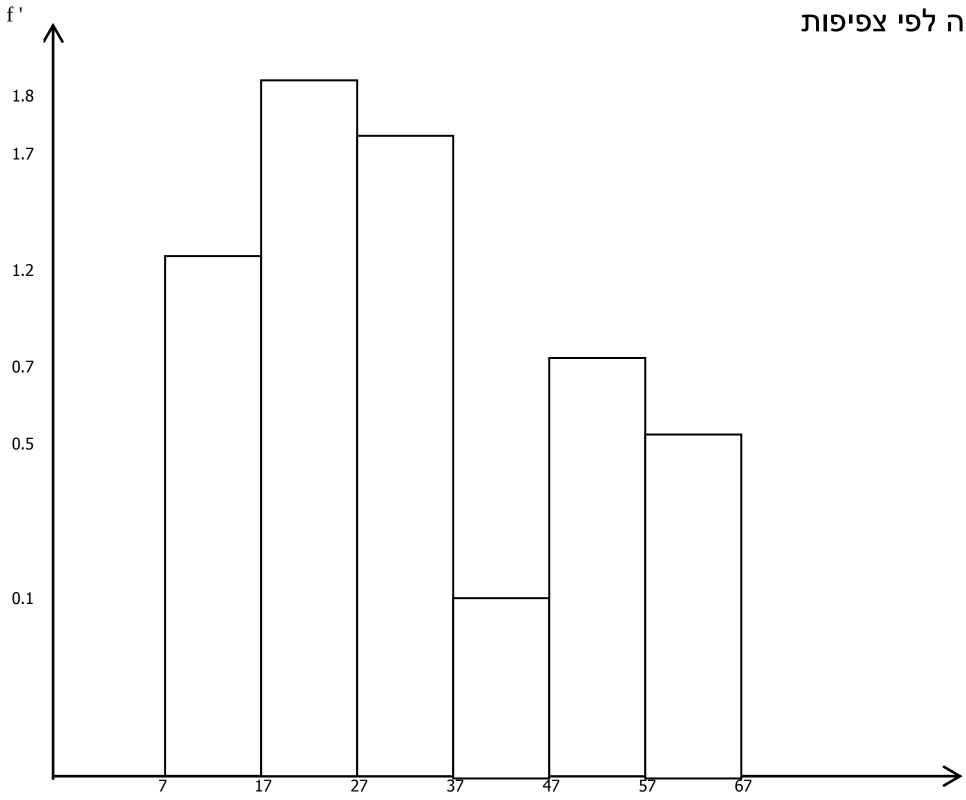
טווח= 60=67-7

רוחב מחלקה=10=60/6

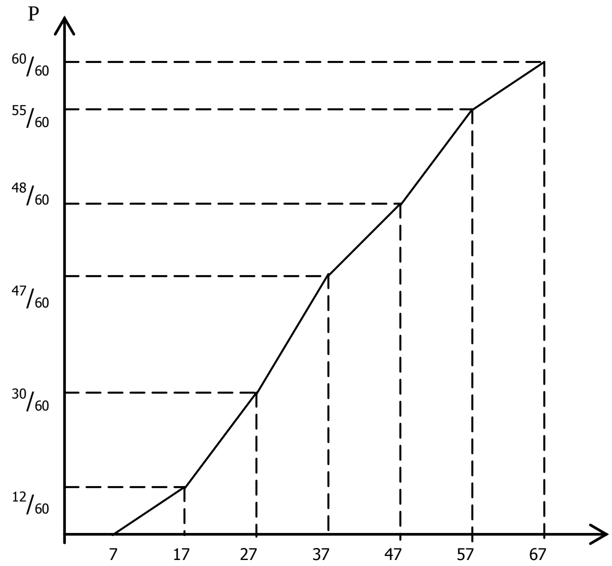
x	f(x) שכיחות	F שכיחות מצטברת	p= f/n שכיחות יחסית	P = F/n שכיחות יחסית מצטברת	ƒ' צפיפות ƒ/c
7-17	12	12	$^{12}/_{60}$	$^{12}/_{60}$	1.2
17-27	18	30	$^{18}/_{60}$	$^{30}/_{60}$	1.8
27-37	17	47	$^{17}/_{60}$	$^{47}/_{60}$	1.7
37-47	1	48	$^{1}/_{60}$	$^{48}/_{60}$	0.1
47-57	7	55	$^{7}/_{60}$	$^{55}/_{60}$	0.7
57-67	5	60	$^{5}/_{60}$	$^{60}/_{60}$	0.5

n = 60

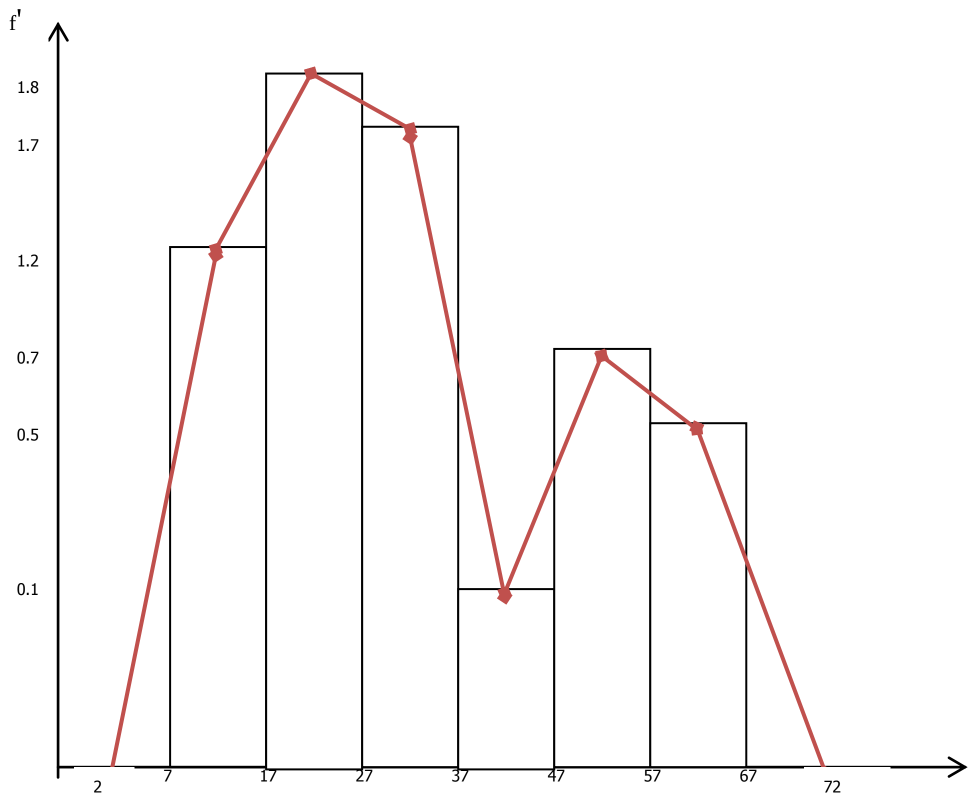
נבנה היסטוגרמה לפי צפיפות



גרף שכיחות יחסית מצטברת



גרף פוליגון



<u>שאלה 4</u>

מתוך מדגם של 5,000 גברים נמצא כי : 1,795 הינם רווקים, 2,685 נשואים , 290 אלמנים ו- 230 גרושים.

מתוך מדגם של 5,000 נשים נמצא כי: 925 רווקות, 3,505 נשואות, 320 אלמנות ו- 250 גרושות.

א. הציגו בטבלה את אחוז הגברים ואחוז הנשים (שכיחות יחסית) עבור כל מצב משפחתי

ב. הציגו את הנתונים בשתי דיאגרמות עוגה.

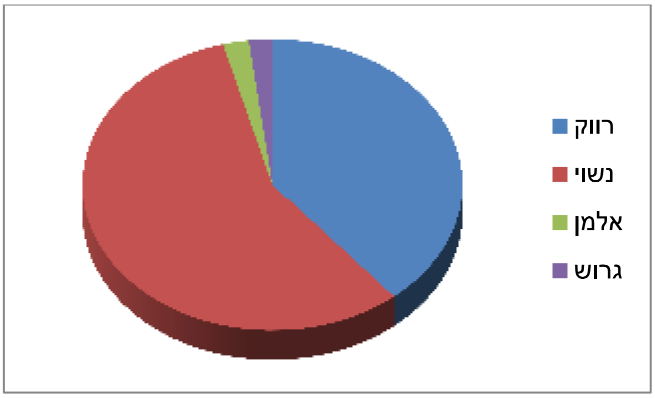
<u>פתרון:</u>

א.

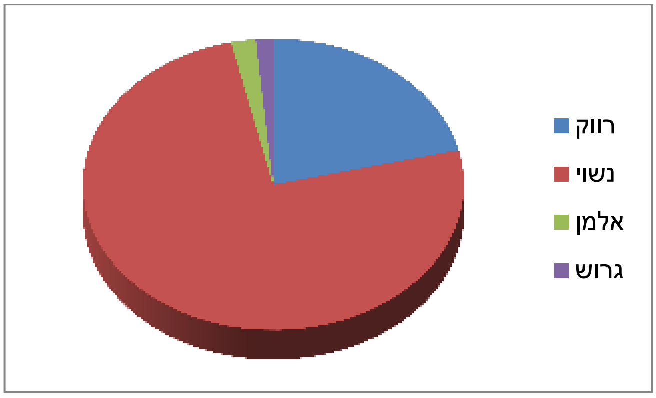
מצב משפחתי	אחוז גברים	אחוז נשים
רווק	35.9	18.5
נשוי	53.7	70.1
אלמן	5.8	6.4
גרוש	4.6	5

ב. שתי דיאגראמות עוגה

גברים



נשים



הקדמה: סימן הסיגמא $\sum$

לפני שנתחיל להציג שיטות שונות בסטטיסטיקה התיאורית, נקדים ונסביר מהו סימן ה- $\sum$. הסיבה לכך היא שניתקל בהמשך בסימן זה בקורס, כיוון שנהוג להשתמש בסיגמא בנוסחאות רבות כסימן קיצור לסכום של איברים.

סימן הסיגמא מהווה הוראה לחיבורם של מספר איברים מסודרים:

$$\sum_{i=1}^{n} X_i = x_1 + x_2 + x_3 .... + x_n$$

<u>דוגמאות:</u>

1. $\sum_{i=2}^{4} X_i = x_2 + x_3 + x_4$

2. $\sum_{i=1}^{3} 2X_i = 2x_1 + 2x_2 + 2x_3$

3. $\sum_{i=1}^{3} (2X_i + 5) = 2x_1 + 5 + 2x_2 + 5 + 2x_3 + 5$

<u>דוגמא:</u>

	Xi
1	4
2	1
3	5
4	3
5	2

$$\sum_{i=1}^{5} xi = 4+1+5+3+2=15$$

<u>מדדים סטטיסטיים</u>

לאחר ארגון הנתונים בטבלת שכיחות והצגת הדיאגרמות יש לחשב ערכים מספריים שיעזרו לנו להסיק מסקנות על המשתנה הנחקר.

ניתן למיין את המדדים לשלוש קטגוריות:

1. **<u>מדדי מרכז</u>**
   - שכיח
   - חציון
   - ממוצע

2. **<u>מדדי מיקום</u>**

   אחוזונים (מאונים)

3. **<u>מדדי פיזור</u>**

   א. תחום

   ב. שונות וסטיית תקן

   ג. תחום בין - רבעוני

<u>מדדי מרכז</u>

א. <u>שכיח</u>

המספר הנפוץ ביותר, המספר בעל התדירות הגבוה ביותר.

שכיח נסמן ב **Mo (mode) או** $\tilde{X}$.

<u>חישוב שכיח עבור משתנה כמותי בדיד</u>

כאשר יש לנו טבלת שכיחויות (ציונים לדוגמא) השכיח הוא המספר שמופיע הכי הרבה פעמים. כלומר הציון שיש לו את השכיחות הגבוהה ביותר.

בדוגמא שלפנינו השכיח הוא 65 (עם תדירות 7).

X	f
40	*3*
50	*4*
65	*7*
75	*6*
80	*5*

n=25

יתכן מצב שבו גם יופיעו מספר ערכים עם התדירות הגבוהה ביותר. למשל בדוגמא הבאה ניתן לומר שהשכיחים הם 65 ו 80 (עם תדירות 7).

X	f
40	*3*
50	*4*
65	*7*
75	*6*
80	*7*

n=27

<u>חישוב שכיח עבור משתנה כמותי בדיד</u>

<u>שלבים בחישוב השכיח עבור משתנה כמותי רציף:</u>

1. לבנות עמודה של צפיפות.
2. למצוא את מחלקת השכיח, המחלקה בעלת הצפיפות הגבוהה ביותר.
3. לחשב את השכיח בעזרת הנוסחה הבאה:

$$Mo = L + \left(\frac{f'(i) - f'(\mathrm{i}-1)}{f'(i) - f'(\mathrm{i}-1) + f'(i) - f'(\mathrm{i}+1)} \right) \bullet C$$

$L =$ גבול תחתון של מחלקת השכיח

$f'(i)$ = הצפיפות של מחלקת השכיח

$f'(\mathrm{i}-1)$ = הצפיפות של המחלקה שלפני מחלקת השכיח

$f'(\mathrm{i}+1)$ = הצפיפות של המחלקה שאחרי מחלקת השכיח

C= רוחב מחלקת השכיח

דוגמא:

X (גיל)	f(x) שכיחות	***f'*** צפיפות f/c
20-25	10	2
25-30	25	5
30-40	10	1
40-50	30	3

n = 75

מחלקה (25-30)

$$Mo = L + \left(\frac{f'(i) - f'(\mathrm{i}-1)}{f'(i) - f'(\mathrm{i}-1) + f'(i) - f'(\mathrm{i}+1)} \right) \bullet C = 25 + \left(\frac{5-2}{5-2+5-1} \right) \bullet 5 = 27.14$$

הערה : אם מחלקת השכיח היא המחלקה הראשונה אז המחלקה שלפני תהיה אפס. אם מחלקת השכיח היא הקבוצה האחרונה, המחלקה שאחרי תהיה שווה לאפס.

<u>חישוב שכיח עבור משתנה איכותי</u>

עבור משתנה איכותי השכיח הוא הערך בעל השכיחות הגבוה ביותר.

<u>דוגמא:</u>

לפי סקר של 1,500,000 צופים, להלן נתונים לגבי תוכנית הטלויזיה האהובה עליהם.

<u>נתונים (באלפים):</u>

"בשלנים ללא הפסקה" 1,350

"ראשון בתקשורת" 45

"תא 22" 15

"מנדלבאום וגולדנבאום" 90

x	f(x) שכיחות (באלפים)	p= f/n שכיחות יחסית
"בשלנים ללא הפסקה"	1350	(1350/1500)=90%
"ראשון בתקשורת"	45	(45/1500)=3%
"תא 22"	15	(15/1500)=1%
"מנדלבאום וגולדנבאום"	90	(90/1500)=6%
	סה"כ n=1500	סה"כ 100%

Mo= בשלנים ללא הפסקה

השכיח הוא בשלנים ללא הפסקה עם השכיחות הגבוהה ביותר (תדירות 1,350).

ב. <u>חציון</u>

הערך שמחצית מההתפלגות נמצאת מעליו ומחצית מתחתיו. כלומר, הערך שמחצית מהמקרים קטנים ממנו ומחצית גדולים ממנו או שווים לו.

חציון נסמן ב **ME (median)**.

<u>חישוב חציון עבור משתנה כמותי בדיד</u>

דוגמא: להלן התפלגות הציונים לשלושה סטודנטים בכלכלה

X	f
70	1
80	1
90	1

n=3

בדוגמא שלפנינו 80 הוא החציון (מחצית מהמקרים מעליו ומחצית מתחתיו).

<u>שלבים בחישוב החציון עבור משתנה כמותי בדיד:</u>

1. לבנות עמודה של שכיחות מצטברת (מספר המקרים עד לאותו ערך).
2. לחשב את הערך n/2
3. לבדוק מתי בפעם הראשונה השכיחות המצטברת גדולה או שווה לחצי $F\geq n*1/2$

<u>דוגמא:</u>

x	f	F
50	*2*	*2*
60	*3*	*5*
65	*5*	*10*
70	*8*	*18*
80	*4*	*22*
90	*3*	*25*

n=25

$F\geq 12.5$ ולכן Me= 70

חישוב חציון עבור משתנה כמותי רציף

שלבים בחישוב חציון עבור משתנה כמותי רציף:

1. לבנות עמודה של שכיחות מצטברת (מספר המקרים עד לאותו ערך)
2. לחשב את הערך n/2
3. לבדוק מתי בפעם הראשונה השכיחות המצטברת גדולה או שווה לחצי $F \geq n * 1/2$
4. נציב את הנתונים בנוסחה הבאה:

$$ME = L + \left(\frac{C}{f}\right) * \left(n * \frac{1}{2} - F_{(i-1)}\right)$$

L= גבול תחתון מחלקת חציון

C= רוחב מחלקה

f = שכיחות מחלקת חציון

n= גודל מדגם

$F_{(i-1)}$= השכיחות המצטברת של המחלקה שלפני מחלקת החציון

דוגמא:

x	f	F
50-60	2	2
60-65	3	5
65-70	5	10
70-80	8	18
80-90	4	22
90-100	3	25

n=25

$F \geq 12.5$

מחלקה (70-80)

$$ME = L + \left(\frac{C}{f}\right) * \left(n * \frac{1}{2} - F_{(i-1)}\right) = 70 + \left(\frac{10}{8}\right) * \left(25 * \frac{1}{2} - 10\right) = 73.125$$

חישוב חציון עבור משתנה איכותי

לא ניתן לחשב חציון עבור משתנה איכותי (אין משמעות לחישוב החציון במשתנה זה).

ג. **ממוצע**

הממוצע מתאר ברמה כללית את התפלגות המשתנה הנחקר והוא מחושב כסכום הערכים חלקי מספר הערכים.

נוסחה:

$$\bar{X}=\frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

ניתן לרשום זאת גם כך:

$$\bar{X}=\frac{\sum_{i=1}^{n} x_i f_i}{n}$$

Xi- ערכי המשתנה

fi- שכיחות המשתנה

חישוב ממוצע עבור משתנה כמותי בדיד

X	f
50	2
60	3
65	5
70	8
80	4
90	3

n=25

$$\bar{X}=\frac{50\cdot 2+60\cdot 3+65\cdot 5+70\cdot 8+80\cdot 4+90\cdot 3}{25}=70.2$$

חישוב ממוצע עבור משתנה כמותי רציף

X	f
50-60	*2*
60-65	*3*
65-70	*5*
70-80	*8*
80-90	*4*
90-100	*3*

n=25

נחשב אמצע קטע לכל מחלקה על ידי חישוב ממוצע לכל מחלקה ולאחר מכן נחשב את הממוצע (לדוגמא: במחלקה הראשונה נחשב 50+60 לחלק ל2 ונקבל X אמצע 55).

X אמצע	f
55	2
62.5	3
67.5	5
75	8
85	4
95	3

$$\bar{X} = \frac{55 \cdot 2 + 62.5 \cdot 3 + 67.5 \cdot 5 + 75 \cdot 8 + 85 \cdot 4 + 95 \cdot 3}{25} = 74.4$$

חישוב ממוצע עבור משתנה איכותי (שמי)

לא ניתן לחשב ממוצע עבור משתנה איכותי (אין משמעות לחישוב הממוצע במשתנה זה).

תכונות הממוצע

1.ניתן לחישוב רק למשתנה כמותי.

2.מתאר ברמה כללית את המשתנה אך הוא לא בהכרח ואף לרוב אינו ערך הקיים בסדרה.

3.הממוצע מושפע מערכים קיצוניים

<u>**שאלות לתרגול (2)**</u>

<u>**שאלה 1**</u>

לפניך סדרה של 7 מספרים;

11 , 5 , 20 , 13 , 4 , 5 , 11

חשב את החציון והשכיח של סדרה זו

<u>**פתרון**</u>

שכיח:

בסדרת מספרים זו נוכל למצוא שני שכיחים

בעלי אותה תדירות והם 5, 11 = MO

בעלי תדירות (2)

x	f	F
4	1	1
5	2	3
11	2	5
13	1	6
20	1	7

n = 7

חציון:

על מנת למצוא את החציון נחשב:

א. $7 * {}^{1}\!/_{2} = 3.5$

ב. $F \geq 3.5$

ג. ME = 11

<u>**שאלה 2**</u>

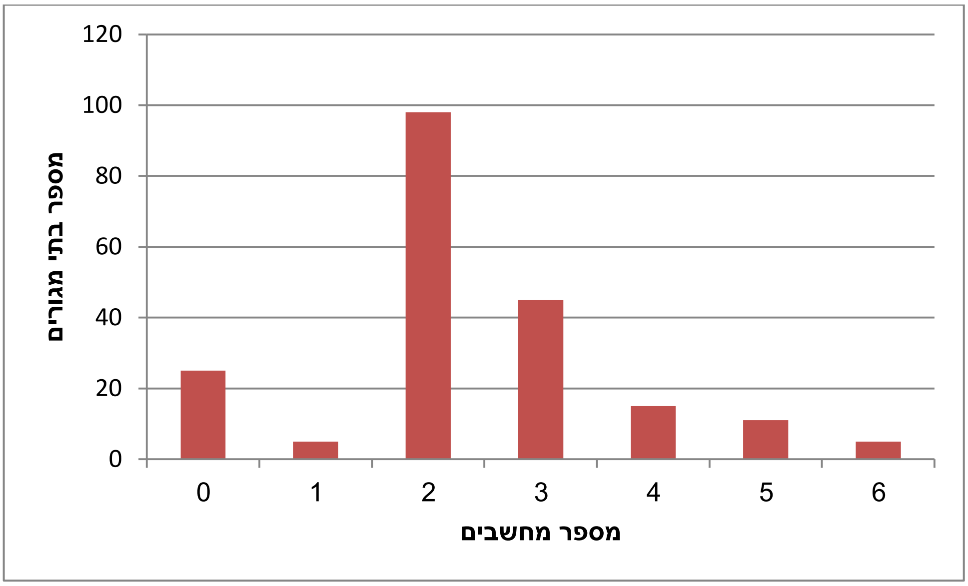
להלן נתונים של מספר המחשבים הנייחים בבית מתוך מדגם מקרי של בתי מגורים ממרכז הארץ.

מספר מחשבים	מספר בתי מגורים
0	25
1	5
2	98
3	45
4	15
5	11
6	5

א. הציגו את הנתונים בדיאגרמה מתאימה
ב. מהו השכיח, החציון והממוצע של מספר המחשבים?

<u>פתרון</u>

א. נשרטט דיאגראמת מקלות



ב. שכיח:

השכיח הוא 2 בתדירות (98) = MO .

x	f	F
0	25	25
1	5	30
2	98	128
3	45	173
4	15	188
5	11	199
6	5	204

n = 204

חציון:

על מנת למצוא את החציון נחשב:

א. $204 * {}^{1}/_{2} = 102$

ב. $F \geq 102$

ג. $ME = 2$

ממוצע:

$$\bar{X} = \frac{0*25 + 1*5 + 2*98 + 3*45 + 4*15 + 5*11 + 6*5}{204} = \frac{481}{204} = 2.36$$

$$\bar{X} = 2.36$$

שאלה 3

להלן התפלגות ההכנסה החודשית למשפחה בשכונה מסוימת באזור רמת גן.

הכנסה חודשית	שכיחות
4,500-9,500	11
9,500-14,500	52
14,500-19,500	24
19,500-29,500	17
29,500-34,500	3

א. חשב את הממוצע, החציון והשכיח של ההוצאות

ב. שרטט היסטוגרמה וסמן עליה את החציון אותו חישבת בסעיף הקודם

פתרון

א.

$\bar{x}$	X	*f*	F	*f'*
7,000	4,500-9,500	11	11	11 : 5,000 = 0.0022
12,000	9,500-14,500	52	63	52 : 5,000 = 0.0104
17,000	14,500-19,500	24	87	24 : 5,000 = 0.0048
24,500	19,500-29,500	17	104	17 : 10,000 = 0.0017
32,000	29,500-34,500	3	107	3 : 5,000 = 0.0006

n = 107

ממוצע:

$$\bar{X}=\frac{7{,}000*11+12{,}000*52+17{,}000*24+24{,}500*17+32{,}000*3}{107}=\frac{1{,}621{,}500}{107}$$

$$\bar{X}=\frac{1{,}621{,}500}{107}=15{,}154.2$$

חציון:

$$n*{}^{1}\!/_{2}=107*{}^{1}\!/_{2}=53.5$$

$$F\geq 53.5$$

החציון נמצא במחלקה 14,500-9,500 :

$$ME=L+\left({}^{C}\!/_{f}\right)*\left(n*{}^{1}\!/_{2}-F_{(i-1)}\right)$$

$$ME=9{,}500+\frac{5{,}000}{52}*(107*{}^{1}\!/_{2}-11)$$

$$ME=13{,}586.53$$

שכיח:

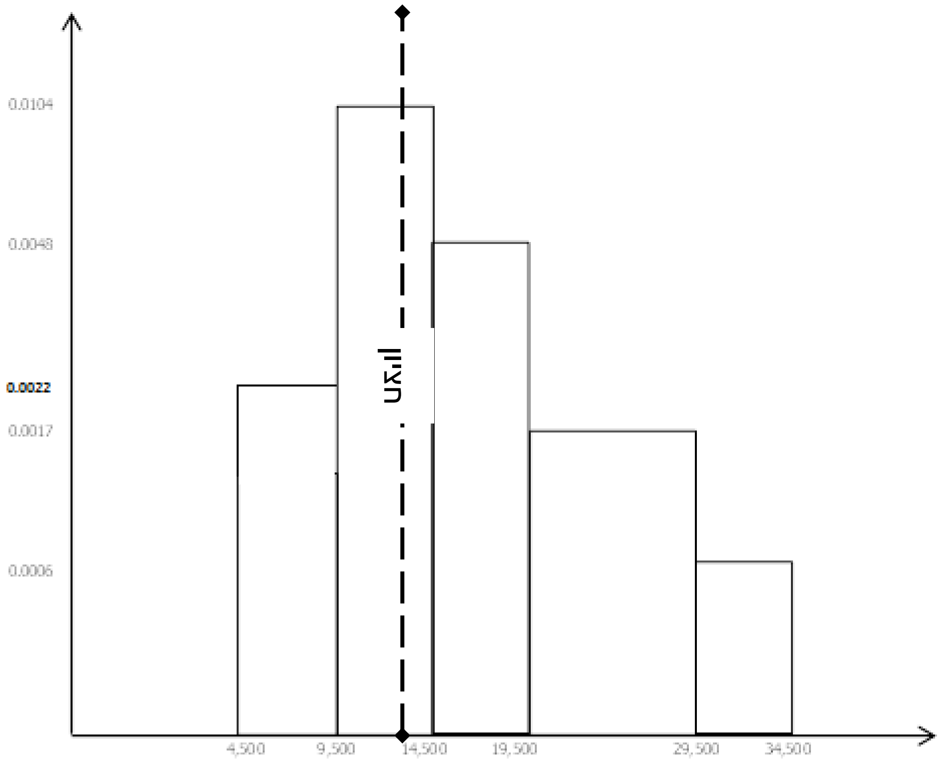
$$MO=L+\frac{f'_i-f'_{(i-1)}}{\left[f'_i-f'_{(i-1)}\right]+\left[f'_i-f'_{(i+1)}\right]}*C$$

השכיח נמצא במחלקה 14,500-9,500 :

$$MO=9{,}500+\left(\frac{0.0104-0.0022}{0.0104-0.0022+0.0104-0.0048}\right)*5{,}000$$

$$MO=12{,}471.01$$

ב. שרטט היסטוגרמה וסמן עליה את החציון אותו חישבת בסעיף הקודם



מדדי מיקום

אחוזון / מאון - הערך ש- P אחוז מהמקרים נמוכים ממנו או שווים לו.

לדוגמא : אחוזון 25 הוא הערך ש 25% מהמקרים יהיו מתחתיו או שווים לו ו 75% יהיו מעליו.

אחוזון/מאון נסמן ב- $X_{\frac{p}{100}}$ (למשל אחוזון 25 יסומן $X_{\frac{25}{100}}$).

חישוב אחוזון עבור משתנה כמותי בדיד

<u>שלבים בחישוב אחוזון עבור משתנה כמותי בדיד</u>

1. חישוב שכיחות מצטברת F
2. חישוב $n\bullet\frac{p}{100}$
3. נמצא מתי $F\geq n\bullet\frac{p}{100}$ בפעם הראשונה

<u>דוגמא:</u>

להלן התפלגות ציונים בכיתה מסוימת. חשבו את האחוזון ה- 80.

x	f	F
50	2	2
60	3	5
65	5	10
70	8	18
80	4	22
90	3	25

n=25

פתרון:

25*80/100= 20

F≥20

$X_{\frac{80}{100}}=80$ – 80% מהציונים מתחת לציון 80.

חישוב חציון עבור משתנה כמותי רציף

שלבים בחישוב אחוזון עבור משתנה כמותי רציף:

1. חישוב שכיחות מצטברת F
2. חישוב $n \bullet \frac{p}{100}$
3. נמצא מתי $F \geq n \bullet \frac{p}{100}$
4. הצבה בנוסחה

$$X_{\frac{P}{100}} = L + \left(\frac{C}{f}\right) * \left(n * \frac{P}{100} - F_{(i-1)}\right)$$

דוגמא:

חשבו על פי הטבלה מעלה המציגה את התפלגות הציונים בכיתה מסוימת את האחוזון ה – 80.

x	f	F
50-60	2	2
60-65	3	5
65-70	5	10
70-80	8	18
80-90	4	22
90-100	3	25

n=25

פתרון:

1. יש לחשב את הערך $n \bullet \frac{p}{100}$

$$n * \frac{p}{100} = 25 * \frac{80}{100} = 20$$

מכאן מיקום האחוזון ה 80 הינו 20

2.נמצא מתי $F \geq n \bullet \frac{p}{100}$

$$F \geq 20$$

לכן המחלקה הרלוונטית היא (80-90)

3. יש להציב בנוסחה הבאה:

$$X_{\frac{80}{100}} = L + \left(\frac{C}{f}\right) * \left(n * \frac{P}{100} - F_{(i-1)}\right) = 80 + \left(\frac{10}{4}\right) * \left(25 * \frac{80}{100} - 18\right) = 85$$

כלומר 80% מהסטודנטים קיבלו מתחת לציון 85, ו-25% קיבלו מעל ציון זה.

<u>דוגמאות נוספות (אחוזונים)</u>

<u>דוגמא 1</u>

x	f	F
50	2	2
60	3	5
65	5	10
70	8	18
80	4	22
90	3	25

א. מהו הציון ש 77% מהתלמידים נמצאים מתחתיו?
ב. מהו הציון ש 65% מהתלמידים נמצאים מעליו?
ג. מהו אחוז הסטודנטים שקיבלו ציון הנמוך מ 60 (כולל 60) ?
ד. מהו אחוז התלמידים אשר קיבלו ציון גובה מ 80 (לא כולל)?

<u>פתרון:</u>

א. יש למצוא את האחוזון ה 77.

נמצא את המיקום לפי:

$$F \geq n * \frac{p}{100} = 25 * \frac{77}{100} = 19.25$$

$$F \geq 19.25$$

$$X_{\frac{77}{100}} = 80$$

ב. יש למצוא את האחוזון ה 35 (100% - 65%=35%). נמצא את המיקום לפי :

$$F \geq n * \frac{p}{100} = 25 * \frac{35}{100} = 8.75$$

$$X_{\frac{35}{100}} = 65$$

ג. ניתן ישירות לפי הטבלה מעלה ניתן לראות שמספר הסטודנטים שקיבלו ציון נמוך מ 60 הוא 5. ובאחוזים 5/25=20%. ולכן מספר הסטודנטים שקיבלו ציון נמוך מ 60 מייצגים האחוזון ה 20%.

ד. לפי הטבלה ניתן לראות שמספר הסטודנטים שקיבלו ציון גבוה מ 80 הוא 3, ובאחוזים 3/25=12%.

<u>דוגמא 2</u>

x	f	F
50-60	2	2
60-65	3	5
65-70	5	10
70-80	8	18
80-90	4	22
90-100	3	25

n=25

א. מהו הציון ש 77% מהתלמידים נמצאים מתחתיו?

ב. מהו הציון ש 35% מהתלמידים נמצאים מעליו?

ג. מהו אחוז התלמידים שקיבלו מתחת לציון 67 (כולל)?

ד. מה אחוז התלמידים שקיבלו ציון בין 58 ל 67 ?

<u>פתרון:</u>

א. נמצא את האחוזון ה -77:

$$n*\frac{p}{100}=25*\frac{77}{100}=19.25$$

$\mathrm{F}\geq 19.25$ מכאן המחלקה היא 80-90

נציב בנוסחה של האחוזון:

$$\mathrm{X}_{\frac{77}{100}}=\mathrm{L}+\left(\frac{\mathrm{C}}{\mathrm{f}}\right)*\left(\mathrm{n}*\frac{P}{100}-\mathrm{F}_{(\mathrm{i}\text{-}1)}\right)=80+\left(\frac{10}{4}\right)*\left(25*\frac{77}{100}-18\right)=83.125$$

ב. נמצא את האחוזון ה 65:

1. $n*\frac{p}{100}=25*\frac{65}{100}=16.25$

2. $F\geq 16.25$ מכאן המחלקה היא 70-80:

3. יש להציב בנוסחה של האחוזון:

$$X_{\frac{65}{100}}=L+\left(\frac{C}{f}\right)*\left(n*\frac{P}{100}-F_{(i-1)}\right)=70+\left(\frac{10}{8}\right)*\left(25*\frac{65}{100}-10\right)=77.81$$

ג. נמצא את האחוזון המתאים לציון 67 על ידי חילוץ P מהנוסחה.
הציון 67 נופל במחלקה (65-70) ולכן:

$$X_{\frac{P}{100}}=67$$

$$67=65+\left(\frac{5}{5}\right)*\left(25*\frac{p}{100}-5\right)$$

$$p=28\%$$

ד. נמצא קודם את אחוז התלמידים שקיבלו ציון נמוך מ 58.
הציון 58 נופל מחלקה (50-60) ולכן:

$$X_{\frac{P}{100}}=58$$

$$58=50+\left(\frac{10}{2}\right)*\left(25*\frac{p}{100}-0\right)$$

$$p=6.4\%$$

על מנת למצוא את אחוז תלמידים שקיבלו ציון בין 58 ל 67 :

אחוז התלמידים עד 58 הינו 6.4%.

אחוז התלמידים עד 67 הינו 28%

על מנת לחשב את אחוז התלמידים בין שני הציונים (58-67) יש לחסר בין האחוזים:

$$p=28\%-6.4\%=21.6\%$$

שאלות לתרגול (3)

שאלה 1

לפניך התפלגות השכר לשעה של עובדים בחברה מסוימת:

שכר לשעה	שכיחות
25 – 32	30
32 – 45	2
45 – 53	55
53 – 60	21
60 – 70	9

א. חשב את הממוצע, החציון והשכיח.

ב. חשב את האחוזון ה- 16

ג. מצא את אחוז העובדים שהשכר לשעה שלהם בין 51.4 לבין 65.3

<u>פתרון:</u>

א.

f'	F	f	X	אמצע קטע
30 : 7 = 4.286	30	30	25-32	28.5
2 : 13 = 0.154	32	2	32-45	38.5
55 : 8 = 6.875	87	55	45-53	49
21 : 7 = 3	108	21	53-60	56.5
9 : 10 = 0.9	117	9	60-70	65

n = 117

ממוצע:

$$\bar{X} = \frac{28.5*30+38.5*2+49*55+56.5*21+65*9}{117} = \frac{5{,}398.5}{117}$$

$$\bar{X} = 46.141$$

חציון:

$$n*{}^{1}/_{2} = 117*{}^{1}/_{2} = 58.5$$

$$F \geq 58.5$$

החציון נמצא במחלקה 45-53

$$ME = L + \left(C/f\right) * \left(n * 1/2 - F_{(i-1)}\right)$$

$$ME = 45 + \frac{8}{55} * \left(117 * 1/2 - 32\right)$$

$$ME = 48.854$$

שכיח:

$$MO = L + \frac{f'_i - f'_{(i-1)}}{\left[f'_i - f'_{(i-1)}\right] + \left[f'_i - f'_{(i+1)}\right]} * C$$

השכיח נמצא במחלקה 45-53

$$MO = 45 + \frac{6.875 - 0.154}{(6.875 - 0.154) + (6.875 - 3)} * 8$$

$$MO = 50.074$$

ב. אחוזון 16 :

$$n * \frac{p}{100} = 117 * \frac{16}{100} = 18.72$$

$$F \geq 18.72$$

האחוזון ה-16 נמצא במחלקה 25-32

$$X_{\frac{P}{100}} = L + \left(\frac{C}{f}\right) * \left(n * \frac{P}{100} - F_{(i-1)}\right)$$

$$X_{\frac{16}{100}} = 25 + \frac{7}{30} * (117 * \frac{16}{100} - 0)$$

$$X_{\frac{16}{100}} = 29.368$$

ג. מצא את אחוז העובדים שהשכר לשעה שלהם בין 51.4 לבין 65.3

הערך 51.4 נמצא במחלקה (45-53):

$$51.4 = 45 + \frac{8}{55} * (117 * \frac{p}{100} - 32)$$

$$51.4 - 45 = \frac{8}{55} * (1.17p - 32)$$

$$6.4 = 0.17p - 4.654$$

$$11.054 = 0.17p$$

$$p = 65.026\%$$

הערך 65.3 נמצא במחלקה (60-70):

$$65.3 = 60 + \frac{10}{9} * (117 * \frac{p}{100} - 108)$$

$$65.3 - 60 = \frac{10}{9} * (1.17p - 108)$$

$$5.3 = 1.3p - 120$$

$$125.3 = 1.3p$$

$$p = 96.384\%$$

$$96.384\% - 65.026\% = 31.358\%$$

<u>שאלה 2</u>

במפעל ליצור לוחות פלסטיק נמדדו אורכי הלוחות (בס"מ) של 600 לוחות, הנתונים קובצו לטבלת השכיחויות הבאה;

אורך	מספר יחידות
0-35	
35-62	
62-90	
90-130	
130-210	

בעיבוד נתונים ראשוני, נמצא כי העשירון הראשון הינו 35 ס"מ, הרבעון הראשון הינו 62, הרבעון השלישי הינו 90 והעשירון העליון הינו 130.
השלם את טבלת השכיחויות שלעיל.

פתרון:

X	f	F
0 – 35	60	60
35 – 62	90	150
62 – 90	300	450
90 – 130	90	540
130 – 210	60	600

נגדיר:

עשירון תחתון (ראשון) – $X_{10\%}$

עשירון עליון – $X_{90\%}$

רבעון ראשון – $X_{25\%}$

רבעון שלישי – $X_{75\%}$

$X_{10\%} = 30$

F=$10\% * 600 = 60$

$X_{25\%} = 62$

F=$25\% * 600 = 150$

$X_{75\%} = 90$

F=$75\% * 600 = 450$

$X_{90\%} = 130$

F=$90\% * 600 = 540$

לאחר שמצאנו את השכיחות המצטברת F לכל מחלקה נשלים בטבלה את טור השכיחות f.

<u>מדדי פיזור</u>

מדדי המרכז נותנים לנו תמונה כוללת שלא תמיד משקפת את המצב בפועל. ייתכנו התפלגויות שונות עם אותם מדדי מרכז אך הם שונים במידת הפיזור שלהם.

לדוגמא:

כיתה (1)

X	f
70	1
80	1
90	1

n=3

כיתה (2)

X	f
80	3

n=3

לשתי הכיתות אותו ממוצע השווה ל 80, אבל יש שוני בין הכיתות. כלומר, הממוצע לא נותן תמונה מלאה ולכן יש לחשב גם מדדי פיזור כדי לקבל תמונה שלמה יותר לגבי ההתפלגות.

א. <u>תחום R</u>

ההפרש בין המספר המקסימאלי למספר המינימאלי.

נוסחה:

$$R= X_{max}- X_{min}$$

<u>חישוב תחום עבור משתנה כמותי בדיד</u>

דוגמא:

X	f
30	10
40	20
60	5
70	8
80	2

$$R= 80-30=50$$

<u>חישוב תחום עובר משתנה כמותי רציף</u>

דוגמא:

X	f
50-60	2
60-65	3
65 -70	5
70-80	8
80-90	4
90-100	3

$$R= 100-50=50$$

ככל שהתחום גדול יותר המשמעות היא שהפיזור גדול יותר. התחום הינו מדד קל לחישוב אך החיסרון שמתייחס לערכים קיצוניים בלבד ולא לוקח בחשבון את השכיחויות.

ב. <u>שונות וסטיית תקן</u>

מתארים את פיזור הערכים סביב הממוצע. אנו בעצם מבקשים לדעת מה הם המרחק הממוצע של הערכים מהממוצע.

דוגמא:

X	f
60	1
70	1
80	1

n=3

60← מרחק 10-

70← מרחק 0

80← מרחק 10

אם נחבר את המרחקים ונחלק בשלוש נקבל אפס. ולכן מעלים בריבוע כדי לתקן זאת. ולפי נוסחת השונות נקבל:

$$\mathrm{Sx}^2 = \frac{\sum_{i=1}^{n} \mathrm{f_i}(\mathrm{X_i} - \bar{\mathrm{X}})^2}{n}$$

$$\mathrm{Sx}^2 = \frac{(60-70)^2 + (70-70)^2 + (80-70)^2}{3} =$$

$$\mathrm{Sx}^2 = 66.66$$

השונות מחושבת על ידי חישוב ממוצע ריבועי הסטיות מהממוצע.

כאשר רוצים לחשב סטיית תקן נבצע שורש לשונות – ביטול פעולת הריבוע שעשינו קודם.

סטיית תקן $S_x = \sqrt{S_x^2}$

$$\mathrm{Sx} = \sqrt{66.66} = 8.16$$

סטיית תקן מוגדרת כרמת פיזור הערכים סביב הממוצע , או ממוצע המרחקים מהממוצע.

ככל שסטיית התקן או השונות גדולות יותר אזי הפיזור גדול יותר.

חישוב שונות וסטיית תקן עבור משתנה כמותי בדיד

דוגמא:

X	f
50	2
60	3
65	5
70	8
80	4
90	3

n= 25

$$\bar{X} = \frac{50 \cdot 2 + 60 \cdot 3 + 65 \cdot 5 + 70 \cdot 8 + 80 \cdot 4 + 90 \cdot 3}{25} = 70.2$$

$$S_x^2 = \frac{2(50-70.2)^2 + 3(60-70.2)^2 + 5(65-70.2)^2 + 8(70-70.2)^2 + 4(80-70.2)^2 + 3(90-70.2)^2}{25} = 112.96$$

$$S_x = \sqrt{112.96} = 10.62$$

חישוב שונות וסטיית תקן עבור משתנה כמותי רציף

דוגמא:

X	f
50-60	2
60-65	3
65 -70	5
70-80	8
80-90	4
90-100	3

n=25

ראשית יש לחשב אמצע מחלקה:

X אמצע קטע	f
55	2
62.5	3
67.5	5
75	8
85	4
95	3

$$\overline{X}=\frac{55\cdot 2+62.5\cdot 3+67.5\cdot 5+75\cdot 8+85\cdot 4+95\cdot 3}{25}=74.4$$

$$S_x^2=\frac{2(55-74.4)^2+3(62.5-74.4)^2+5(67.5-74.4)^2+8(75-74.4)^2+4(85-74.4)^2+3(95-74.4)^2}{25}=125.64$$

$$S_x=\sqrt{125.64}=11.21$$

<u>נוסחה נוספת לחישוב השונות</u>

$$\mathrm{S_x}^2=\frac{\sum_{i=1}^{n}\mathrm{f_i}*\mathrm{X_i}^2}{n}-\overline{\mathrm{X}}^2$$

כעת נחשב את השונות לפי הנוסחה השנייה ונקבל:

$$S_x{}^2=\frac{2\cdot 55^2+3\cdot 62.5^2+5\cdot 67.5^2+8\cdot 75^2+4\cdot 85^2+3\cdot 95^2}{25}-74.4^2=125.64$$

$$\mathrm{S_x}^2=\sqrt{125.64}=11.21$$

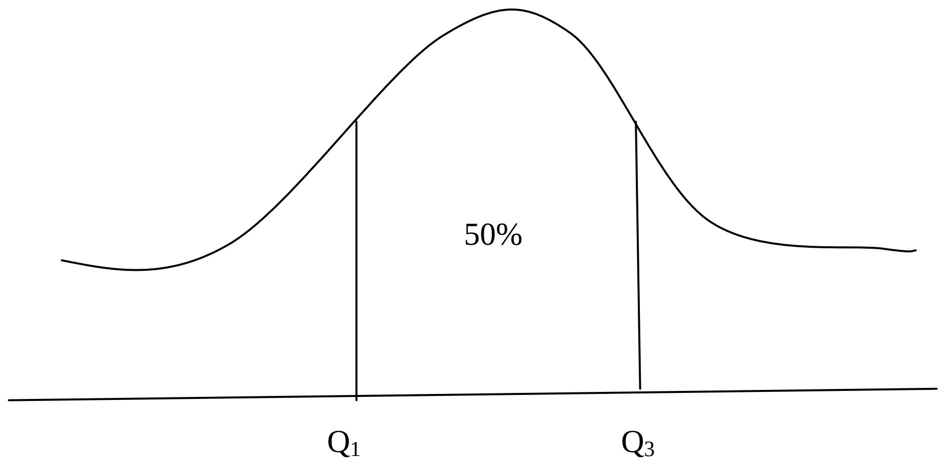
ג. <u>תחום בין רבעוני</u>

רבעון ראשון – נסמן ב - Q_1. הערך ש 25% מהמקרים קטנים ממנו או שווים לו ו 75% גדולים ממנו. הרבעון הראשון הוא גם האחוזון ה 25.

רבעון שני - נסמן ב Q_2. הערך ש 50% מהמקרים קטנים ממנו או שווים לו ו 50% גדולים ממנו. הרבעון השני הוא גם החציון.

רבעון שלישי – נסמן – Q_3. הערך ש 75% מהמקרים קטנים ממנו או שווים לו ו 25% גדולים ממנו. הרבעון השלישי הוא גם האחוזון ה 75.

תחום בין רבעוני – Q_3-Q_1 התחום הבין רבעוני מתאר את מחצית המקרים שבמרכז ההתפלגות. הוא ההפרש בין הרבעון השלישי לרבעון הראשון.



התחום הבין-רבעוני נותן לנו מרכז ההתפלגות שם נמצאים 50% מהמקרים.

דוגמא:

להלן התפלגות הציונים בכיתה מסוימת:

X	f	F
50	2	2
60	3	5
65	5	10
70	8	18
80	4	22
90	3	25
סה"כ	25	

חשבו את אורך התחום הבין –רבעוני

פתרון:

יש לחשב את האחוזון ה 75 ואת האחוזון ה 25 ולהפחית בין התוצאות.

$$n \cdot \frac{P}{100} = 25 \cdot \frac{75}{100} = 18.75$$
$$F \geq 18.75$$
$$X_{\frac{75}{100}} = 80$$

$$n \cdot \frac{P}{100} = 25 \cdot \frac{25}{100} = 6.25$$
$$F \geq 6.25$$
$$X_{\frac{25}{100}} = 65$$

$$X_{\frac{75}{100}} - X_{\frac{25}{100}} = 80 - 65 = 15$$

אורך התחום הבין רבעוני הוא 15.

שאלות לתרגול (4)

שאלה 1

להלן נתוני שנות וותק בעבודה של 11 עובדים בפירמה קטנה בצפון הארץ:

15, 3, 21, 18, 10, 13, 25, 5, 7, 11, 20

א. מצא את הציון הממוצע
ב. חשב את סטיית התקן
ג. מהו תחום (טווח) הנתונים?
ד. נעשתה טעות בעת חישוב שנות הוותק של שלושת העובדים אשר היו בעלי הוותק הנמוך ביותר (7, 5, 3). לאחר בדיקה חוזרת נמצא כי שלושתם בעלי וותק של 5 שנים. כיצד ישפיע תיקון זה על המדדים שחושבו בסעיפים הקודמים?

פתרון:

X	*f*	F
3	1	1
5	1	2
7	1	3
10	1	4
11	1	5
13	1	6
15	1	7
18	1	8
20	1	9
21	1	10
25	1	11

n = 11

א.

ממוצע:

$$\overline{X} = \frac{3+5+7+10+11+13+15+18+20+21+25}{11} = \frac{148}{11}$$

$$\overline{X} = 13.4545$$

ב. שונות וסטיית תקן:

$$S_X^2 = \frac{(3)^2+(5)^2+(7)^2+(10)^2+(11)^2+(13)^2+(15)^2+(18)^2+(20)^2+(21)^2+(25)^2}{11} - (13.4545)^2$$

$$S_x^2 = 45.158$$

$$S_X = 6.72$$

ג. טווח הנתונים הוא בין R=25-3=22

ד. הממוצע לא ישתנה היות והוספנו שנתיים לוותק הנמוך ביותר 3 ומנגד, הורדנו שנתיים מהוותק הגבוה יותר 7, לכן בסה"כ המונה והמכנה נשארו ללא שינוי.
סטיית התקן תקטן כי הפיזור קטן.

שאלה 2

לפנייך נתוני התפלגות רמת המשכל של תושבי יישוב מסוים בדרום הארץ:

רמת משכל	שכיחות
70 – 90	33
90 – 100	21
100 – 110	5
110 – 130	40
130 – 140	15

א. חשב את סטיית התקן ואת התחום הבינרבעוני

ב. מצא את המאון ה- 67 ואת העשירון הרביעי

ג. מצא את אחוז התושבים עד לרמת משכל 132

ד. התברר כי חלה טעות ברישום רמות המשכל לגבי המחלקה האחרונה שצריכה להיות (155 – 130) ולא כפי שנרשמה (140 – 130). נמק בקצרה, ללא חישוב, איך התיקון של גבולות המחלקה האחרונה ישפיע על הממוצע, ועל סטיית התקן

<u>פתרון:</u>

אמצע קטע	X	*f*	F	P	P	*f'*
80	70 – 90	33	33	$^{33}/_{114}$	$^{33}/_{114}$	33 : 20 = 1.65
95	90 – 100	21	54	$^{21}/_{114}$	$^{54}/_{114}$	21 : 10 = 2.10
105	100 – 110	5	59	$^{5}/_{114}$	$^{59}/_{114}$	5 : 10 = 0.50
120	110 – 130	40	99	$^{40}/_{114}$	$^{99}/_{114}$	40 : 20 = 2.00
135	130 – 140	15	114	$^{15}/_{114}$	$^{114}/_{114}$	15 : 10 = 1.50
		n = 114		1		

א. כדי למצוא את סטיית התקן, ראשית עלינו למצוא את הממוצע;

$$\overline{X} = \frac{80*33+95*21+105*5+120*40+135*15}{114} = \frac{11{,}985}{114}$$

$$\overline{X} = 105.131$$

סטיית תקן:

$$S_X^2 = \frac{33*80^2+21*95^2+5*105^2+40*120^2+15*135^2}{114} - 105.131^2 = \frac{1{,}305{,}225}{114} - 11{,}052.53$$

$$S_X^2 = 396.81$$

$$S_X = \sqrt{396.81}$$

$$S_X = 19.92$$

תחום בינרבעוני:

$$n*{}^{25}\!/_{100} = 114*0.25 = 28.5$$

$F \geq 28.5 \rightarrow$ (70 – 90)

$$Q1 = 70 + \left(\frac{20}{33}\right)*(28.5-0) = 87.2727$$

$$n*{}^{75}\!/_{100} = 114*0.75 = 85.5$$

$F \geq 85.5 \rightarrow$ (110 – 130)

$$Q3 = 110 + \left(\frac{20}{40}\right)*(85.5-59) = 123.25$$

התחום הבינרבעוני הוא; 35.97 = 87.2727 – 123.25

ב. המאון ה- 67:

$$X_P = L + \left(C/f\right) * \left(n * P/100 - F_{(i-1)}\right)$$

$$n * 67/100 = 114 * 67/100 = 76.38$$

$F \geq 76.38$ → (110 – 130)

$$X_{67/100} = 110 + \left(\frac{20}{40}\right) * (76.38 - 59)$$

$$X_{67/100} = 118.69$$

העשירון ה- 4:

$$X_P = L + \left(C/f\right) * \left(n * P/100 - F_{(i-1)}\right)$$

$$n * 4/10 = 114 * 4/10 = 45.6$$

$F \geq 45.6$ → (90 – 100)

$$X_{4/10} = 90 + \left(\frac{10}{21}\right) * (45.6 - 33)$$

$$X_{4/10} = 96$$

ג. הערך 132 נמצא במחלקה 140 – 130, ולכן נציב;

$$132 = 130 + \left(10/15\right) * \left(114 * P/100 - 99\right)$$

$$132 - 130 = 10/15(1.14P - 99)$$

$$2 = 0.76P - 66$$

$$68 = 0.76P$$

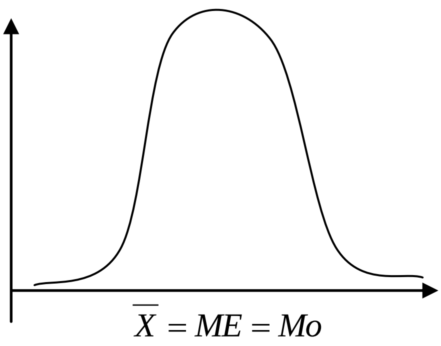
$$P = 89.473\%$$

ד. הממוצע וסטיית התקן, שניהם ישתנו, שכן אלו הם מדדים המושפעים מערכים קיצוניים, מכיוון שהגדלנו את ערכי המחלקה יגדל הממוצע, ומכיוון שהגדלנו את הפיזור והמחלקה התרחבה, תגדל סטיית התקן.

סוגי התפלגויות

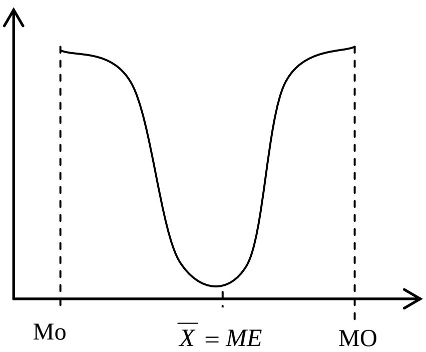
1. התפלגות נורמלית

- התפלגות סימטרית חד שיאית (שיא אחד)
- ממוצע = שכיח = חציון. שלושתם באותה נקודה באמצע ההתפלגות.
- ההתפלגות הינה סימטרית . כלומר תמונת ראי, כל מה שקורה מימין קורה גם משמאל. לדוגמא: צפיפות המחלקה הראשונה שווה לצפיפות המחלקה האחרונה
- ריכוז המקרים הוא באמצע והצפיפות הולכת ופוחתת בקצוות בצורה סימטרית.



2. התפלגות סימטרית אך לא נורמלית (דו שיאית)

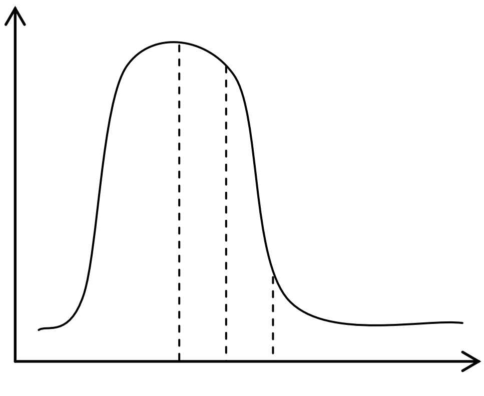
- במקרה זה יש לנו שני שכיחים ולכן ההתפלגות נקראת דו שיאית.
- הממוצע והחציון שווים- שניהם באמצע ההתפלגות
- ההתפלגות סימטרית



3. אסימטרית ימנית (חיובית)

קיים ריכוז של מקרים בערכים הנמוכים וערכים קיצוניים בערכים הגבוהים שמושכים את הזנב לכיוון ימין (ומגדילים את הממוצע)

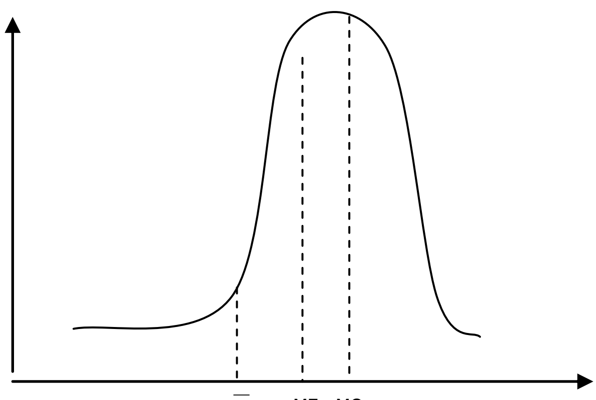
$$\overline{X} > ME > Mo$$



4. אסימטרית שמאלית (שלילית)

קיים ריכוז של מקרים בערכים הגבוהים וערכים קיצוניים בערכים הנמוכים שמושכים את הזנב לכיוון שמאל (ומקטינים את הממוצע)

$$\overline{X} < ME < Mo$$



<u>גבול אמיתי וגבול מדומה</u>

גבול מדומה הינו מצב שיש הפרש בין הגבול העליון של מחלקה אחת לבין הגבול התחתון של המחלקה הבאה. במצב זה יש ערכים להם אין מחלקה רלוונטית ולכן יש לעבור לגבול אמיתי.

<u>דוגמא:</u>

להלן התפלגות של משקל בטונות של משלוחי חצץ.

שכיחות	טונות של חצץ
3	5 – 9
10	10 – 12
14	13 – 16
25	17 – 20
17	21 –24
9	25 –28
22	29 –32

א. איזה אחוזון מהווה משלוח של חצץ שמשקלו 14 טונות?

לפני שנענה על השאלה ראשית נשים לב שיש "קפיצה" בין הגבולות של המחלקות. דהיינו הגבול העליון של מחלקה אחת לא שווה לגבול התחתון של המחלקה הבאה, ולכן אין מחלקה מתאימה לערכים מסוימים כגון 9.5

על מנת לעבור מגבול מדומה לגבול אמיתי יש לעבוד על פי אחת משתי השיטות הבאות:

אפשרות 1: נחשב ממוצע לגבול עליון מחלקה ראשונה וגבול תחתון מחלקה שנייה ונקבל 9.5 =2/(10+9). מספר זה, 9.5 , יהיה הגבול העליון החדש של מחלקה ראשונה והגבול התחתון חדש של מחלקה שנייה. נמשיך באותו האופן גם למחלקות הבאות.

אפשרות 2 : נחשב את ההפרש בין הגבולות . בדוגמא שלנו 9-10=1

חצי מן ההפרש (0.5) נוריד מהגבול התחתון וחצי מההפרש נוסיף לגבול העליון של כל מחלקה.

לאחר עדכון הגבולות נקבל את הטבלה הבאה:

טונות של חצץ	שכיחות
4.5 – 9.5	3
9.5 – 12.5	10
12.5 – 16.5	14
16.5 – 20.5	25
20.5 –24.5	17
24.5 –28.5	9
28.5 –32.5	22

שימו לב שגם הוספנו 0.5 למספר האחרון (32.5) והפחתנו 0.5 מהמספר הראשון (4.5)

וזאת משום שהשינוי צריך להיות עקבי בכל המחלקות.

כעת בטבלה זו לכל ערך יש מחלקה מתאימה כעת ניתן להמשיך ולענות על השאלות.

<u>פתרון:</u>

F	f	X
3	3	4.5 – 9.5
13	10	9.5 – 12.5
27	14	12.5 – 16.5
52	25	16.5 – 20.5
69	17	20.5 –24.5
78	9	24.5 –28.5
100	22	28.5 –32.5

א.

14 מתאים למחלקה (12.5 – 16.5), ולכן נציב;

$$14 = 12.5 + \left(\frac{4}{14}\right) \cdot \left(100 \cdot \frac{P}{100} - 13\right)$$

$$P = 18.25\%$$

<u>שינוי לינארי</u>

כיצד ישתנו הממוצע וסטיית התקן אם מקטינים או מגדילים את כל ערכי ההתפלגות בקבוע או אחוז מסוים?

<u>דוגמא:</u>

להלן התפלגות הציונים בכיתה. כאשר הממוצע שווה ל 80 וסטיית תקן שווה ל 8.16

X	f
70	1
80	1
90	1

נניח והמרצה החליט להעלות כל ציון ב 5% ולהוסיף 2 נקודות. כיצד ישתנה הממוצע וכיצד תתשנה סטיית התקן?

X	f
70*1.05+ 2= 75.5	1
80*1.05+2= 86	1
90*1.05+ 2=96.05	1

ממוצע חדש 86 , סטיית תקן חדשה 8.57

הממוצע גדל ב 86 =2+ 80*1.05

סטיית התקן גדלה ב 8.57 = 1.05*8.16

כלומר הממוצע הושפע מתוספת של אחוזים ושל קבוע, אך סטיית התקן הושפעה רק מהשינוי באחוזים.

מכאן:

<u>תוספת של קבוע:</u>

- תוספת/הורדה של קבוע לכל אחד מערכי הסדרה תגדיל את הממוצע בקבוע
- תוספת של קבוע לא תשנה את סטיית התקן או השונות- לא תשנה את ההפרשים מן הממוצע (החדש).

הכפלה בקבוע:

- הכפלה בקבוע (או תוספת/הורדה של אחוזים) תכפיל גם את הממוצע פי אותו קבוע
- הכפלה בקבוע תכפיל גם את סטיית התקן בקבוע
- השונות תוכפל בקבוע בריבוע

הנוסחה לחישוב ממוצע חדש:

$$\bar{y} = a\bar{x} + b$$

$\bar{x}$ - ממוצע לפני שינוי

$\bar{Y}$ - ממוצע אחרי שינוי

a- תוספת / הורדה באחוזים (הכפלה בקבוע)

b- תוספת / הורדה של קבוע

הנוסחה לחישוב סטיית התקן החדשה:

$$Sy = a * Sx$$

Sx- סטיית תקן ישנה

Sy- סטיית תקן חדשה

a- תוספת / הורדה באחוזים (הכפלה בקבוע)

דוגמא:

בחברה לייצור מזון המשכורת הממוצעת היא 6000 ₪ וסטיית התקן שווה ל 1000. להלן שתי אלטרנטיבות אפשריות:

א. המנהל החליט להוסיף לכל עובד 10% למשכורת ויחלק שי חודשי בשווי 100 ₪

ב. המנהל החליט לקצץ את המשכורות ב 10% ולנכות 200 ₪ מכל משכורת.

מה הממוצע החדש וסטיית התקן החדשה בכל אלטרנטיבה?

פתרון:

א. הממוצע החדש $\bar{Y} = a\bar{x} + b = 1.1 \cdot 6000 + 100 = 6700$

סטיית התקן החדשה $Sy = a * Sx = 1.1 \cdot 1000 = 1100$

ב. הממוצע החדש $\bar{Y} = a\bar{x} + b = 0.9 \cdot 6000 - 200 = 5200$

סטיית התקן החדשה $Sy = a * Sx = 0.9 \cdot 1000 = 900$

<u>ציון תקן</u>

מתאר מיקום יחסי של תצפית X בהתפלגות מסוימת. נסמן ציון תקן ב –Z.

ציון תקן מייצג את המרחק של תצפית X מהממוצע של ההתפלגות ביחידות של סטיית תקן.

ציון תקן הוא מספר ללא יחידות ולכן ניתן להשוות אותו להתפלגויות שונות.

נוסחה לחישוב ציון תקן:

$$Z = \frac{X_i - \bar{X}}{S_x}$$

כאשר,

Z – ציון תקן, X_i - הערך של תצפית i, $\bar{X}$ - ממוצע , S_X - סטיית תקן

תכונות ציון התקן:

1. ככל שציון התקן רחוק יותר מנקודת האפס, התצפית חריגה יותר (ללא קשר לסימן).
2. ממוצע ציוני התקן תמיד שווה לאפס.
3. סטיית התקן של ציוני התקן תמיד שווה ל-1.

<u>דוגמא:</u>

להלן הנתונים לגבי הציון של רמי והתפלגות הציונים בסטטיסטיקה ובכלכלה. באיזה מקצוע רמי יותר טוב? במימון או בסטטיסטיקה?

<u>סטטיסטיקה</u>

ציון רמי 86

ממוצע כיתה 83

סטיית תקן 4

<u>כלכלה</u>

ציון רמי 79

ממוצע כיתה 70

סטיית תקן 3

פתרון:

ציון תקן של רמי בסטטיסטיקה:

$$Z = \frac{X_i - \bar{X}}{S_x} = \frac{86-83}{4} = 0.75$$

ציון תקן של רמי בכלכלה:

$$Z = \frac{X_i - \bar{X}}{S_x} = \frac{79-70}{3} = 3$$

אמנם רמי קיבל בסטטיסטיקה ציון גבוה יותר לעומת ציונו בכלכלה (86 לעומת 79), אבל יחסית לכיתה נבדוק לפי ציון תקן.

ציון תקן רמי סטטיסטיקה- 0.75 (נמצא 0.75 סטיות תקן מעל הממוצע)

ציון תקן רמי כלכלה- 3 (נמצא 3 סטיות תקן מעל הממוצע)

מסקנה: ביחס לכיתה רמי טוב יותר בכלכלה.

התוצאה של ציון התקן יכולה להיות חיובית /שלילית /אפס:

ציון תקן חיובי- התצפית מעל הממוצע

ציון תקן שלילי- התצפית מתחת לממוצע

ציון תקן אפס- התצפית מתלכדת עם הממוצע

שאלות לתרגול (5)

שאלה 1

להלן נתוני מדגם לגבי גובה של קבוצת סטודנטים ממרכז הארץ אשר נבחרו באקראי:

גובה בס"מ	שכיחות
160 – 170	15
170 – 180	205
180 – 190	66
190 – 200	3
200 – 210	1

א. המדגם ייחשב כ"תקין" אם סטודנטים אשר גובהם 173 ס"מ או פחות הוא מעל 10%. האם המדגם תקין?

ב. חשב את הממוצע, החציון והשכיח ואפיין בעזרתם את צורת ההתפלגות

פתרון:

f'	F	f	x	אמצע קטע
15:10=1.50	15	15	160 – 170	165
205:10=20.5	220	205	170 – 180	175
66:10=6.60	286	66	180 – 190	185
3:10=0.30	289	3	190 – 200	195
1:10=0.10	290	1	200 – 210	205

N=290

א. 173 נופל במחלקה (180 – 170), ולכן נציב;

$$173=170+\left(10/205\right)*\left(290*P/100-15\right)$$

$$P=26.38\%$$

המדגם תקין שכן סטודנטים אשר גובהם 173 ס"מ או פחות אכן מעל 10% ועומד על 26.38%.

ב. ממוצע:

$$\overline{X}=\frac{165*15+175*205+185*66+195*3+205*1}{290}=177.069$$

חציון:

$$n*\frac{1}{2}=290*\frac{1}{2}=145$$

$F\geq 145$ החציון נמצא במחלקה 180 – 170

$$ME=170+\left(\frac{10}{205}\right)*\left(290*\frac{1}{2}-15\right)$$

$$ME=176.341$$

שכיח:

השכיח נמצא במחלקה 180 – 170

$$MO=170+\frac{20.5-1.5}{(20.5-1.5)+(20.5-6.6)}*10$$

$$MO=175.77$$

אפיון צורת ההתפלגות:

$\overline{X}\geq ME\geq MO$

מדובר בהתפלגות אסימטרית ימנית (חיובית)

<u>שאלה 2</u>

להלן התפלגות רכב נוסעים פרטי על פי גיל הרכב (בשנים) בקרית אונו:

גיל הרכב (בשנים)	**שכיחות**
1 – 3	53
4 – 7	13
8 – 11	9
12 – 15	25
16 – 19	4

א. חשב את הגיל הממוצע והשונות

ב. חשב את המאון ה- 67

<u>פתרון:</u>

ראשית נבנה את טבלת ההתפלגויות כאשר נעבור מגבול מדומה לגבול אמיתי:

$\bar{X}$	X	*f*	F	P	P	*f '*
2	0.5 – 3.5	53	53	$^{53}/_{104}$	$^{53}/_{104}$	53 : 3 = 17.666
5.5	3.5 – 7.5	13	66	$^{13}/_{104}$	$^{66}/_{104}$	13 : 4 = 3.25
9.5	7.5 – 11.5	9	75	$^{9}/_{104}$	$^{75}/_{104}$	9 : 4 = 2.25
13.5	11.5 – 15.5	25	100	$^{25}/_{104}$	$^{100}/_{104}$	25 : 4 = 6.25
17.5	15.5 – 19.5	4	104	$^{4}/_{104}$	$^{104}/_{104}$	4 : 4 = 1

n = 104

א. ממוצע:

$$\overline{X} = \frac{2*53+5.5*13+9.5*9+13.5*25+17.5*4}{104} = \frac{670.5}{104}$$

$$\overline{X} = 6.45$$

שונות:

$$S_{x^2} = \frac{53*2^2+13*5.5^2+9*9.5^2+25*13.5^2+4*17.5^2}{104} - 6.45^2$$

$$S_{X^2} = \frac{7{,}198.75}{104} - 41.6025$$

$$S_{X^2} = 27.616$$

ב. חשב את המאון ה- 67

$$n*{}^{67}\!/_{100} = 104*{}^{67}\!/_{100} = 69.68$$

$F \geq 69.68 \rightarrow$ (7.5-11.5)

$$X_{67/100} = 7.5 + \left({}^{4}\!/_{9}\right)*(69.68-66) = 9.135$$

<u>שאלה 3</u>

להלן נתונים על המשכורת החודשית המשולמת לעובדים במפעל לייצור שטיחים;

שכיחות	**הכנסה (אלפי ₪)**
3	0 – 3
12	3 – 6
12	6 – 12
12	12 – 15
3	15 – 18

א. צייר היסטוגרמה לנתונים. מהי צורת ההתפלגות?

ב. חשב ממוצע, חציון, שכיח, סטיית תקן ותחום בינרבעוני להכנסות

ג. מהי ההכנסה החודשית שציון התקן שלה הוא 1?

ד. המפעל החליט לחלק מתנות לחג בתלושים בהתאם לגובה ההכנסה. כל שכיר יקבל 15% ממשכורתו בתלושי קנייה כלליים, ועוד תלוש נוסף לקניית ביגוד בסך 300 ₪. מה יהיה ערך התלוש הממוצע ומה תהיה סטיית התקן של ערכו?

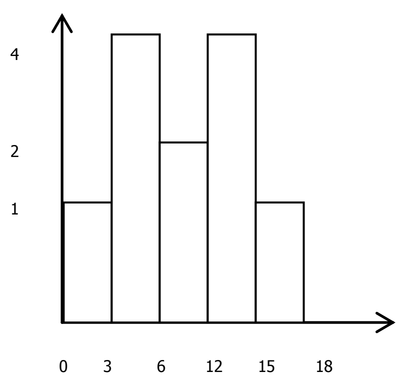
ה. למפעל הצטרפו חמישה עובדים חדשים שמשכורתם בין 18,500 ל- 19,500 ₪. כיצד ישתנו סטיית תקן וממוצע שחישבת בסעיף ב'? יש לתאר את כיוון השינוי להסבירו.

<u>פתרון:</u>

f '	**P**	**P**	**F**	***f***	**X**	$\bar{x}$
3 : 3 = 1	$3/42$	$3/42$	3	3	0 – 3	1.5
12 : 3 = 4	$15/42$	$12/42$	15	12	3 – 6	4.5
12 : 6 = 2	$27/42$	$12/42$	27	12	6 – 12	9
12 : 3 = 4	$39/42$	$12/42$	39	12	12 – 15	13.5
3 : 3 = 1	$42/42$	$3/42$	42	3	15 – 18	16.5

n = 42

א.



צורת ההתפלגות סמטרית דו שיאית

ב. ממוצע:

$$\overline{X} = \frac{1.5*3+4.5*12+9*12+13.5*12+16.5*3}{42} = 9$$

חציון:

$$n*\tfrac{1}{2} = 42*0.5 = 21$$

$F \geq 21$ → (6 – 12)

$$ME = 6+\left(\frac{6}{12}\right)*(21-15) = 9$$

שכיח:

השכיח נמצא במחלקה (6 – 3), נציב;

$$MO = 3+\frac{4-1}{(4-1)+(4-2)}*3$$

$$Mo = 4.8$$

ובמחלקה (15 – 12), נציב;

$$MO = 12+\frac{4-2}{(4-2)+(4-1)}*3$$

$$MO = 13.2$$

סטיית תקן:

$$S_x^2 = \frac{3*1.5^2+12*4.5^2+12*9^2+12*13.5^2+3*16.5^2}{42} - 9^2 = \frac{4,225.5}{42} - 81$$

$$S_x^2 = 19.607$$

$$S_X = 4.427$$

תחום בינרבעוני:

$$n*\tfrac{25}{100} = 42*0.25 = 10.5$$

$F \geq 10.5$ → (3 – 6)

$$Q_1 = 3 + \left(\frac{3}{12}\right) * (10.5 - 3) = 4.875$$

$$n * 75/100 = 42 * 0.75 = 31.5$$

$F \geq 31.5 \rightarrow$ (12 – 15)

$$Q_3 = 12 + \left(\frac{3}{12}\right) * (31.5 - 27) = 13.125$$

אורך התחום הבינרבעוני הוא: 8.25 = 4.875 – 13.125

ג.

$$Z = \frac{X_i - \bar{X}}{S_X}$$

$$1 = \frac{X_i - 9}{4.427}$$

$$X_i = 13.427$$

ד. הממוצע החדש – מושפע גם מהשינוי באחוזים וגם מהתוספת הקבועה, ולכן ערך <u>התלוש</u> הממוצע יהיה שווה ל:

$$\bar{Y} = 0.15 * 9 + 0.3 = 1.65$$

סטיית התקן החדשה – מושפעת רק מהשינוי באחוזים, ולכן סטיית התקן של ערך התלוש:

$$S_y = 0.15 * 4.427 = 0.66405$$

ה. הממוצע הוא מדד המושפע מערכים קיצוניים, ומכיוון שהוספנו ערכים קיצוניים, הממוצע יגדל. סטיית תקן- תגדל היות והפיזור גדל. הטווח יגדל מאחר והערך המקסימלי גדל.

שאלה 4

להלן התפלגות ההכנסות של עובדים במפעל פלסטיק במרכז הארץ;

הכנסה	שכיחות
3,500 – 4,000	?
4,000 – 4,500	13
4,500 – 5,000	18
5,000 – 6,000	25
6,000 – 7,000	8
7,000 – 9,000	15
9,000 – 12,000	9

א. ידוע כי ההכנסה הממוצעת הייתה 5,897.5 ₪. מצא את מספר העובדים שהיו בקטגוריה הראשונה

ב. חשב תחום, חציון, שכיח, עשירון עליון וסטיית תקן להכנסות .

ג. על בסיס חישובך בסעיפים הקודמים, אפיין את צורת ההתפלגות.

ד. עובד מסוים קיבל ציון תקן של (1-). מהי משכורתו?

ה. בחודש ינואר התווספו עוד 10 עובדים חדשים שמשכורתם נעה בין 3,500 – 2,000 ₪. כיצד ישתנו הממוצע וסטיית התקן? אין צורך לחשבם מחדש, יש לתאר את כיוון השינוי הכללי ולהסבירו.

ו. המפעל החליט לבצע קיצוץ שכר רוחבי (לצרכי התייעלות) בהתאם לגובה הכנסתם של עובדיו. ירידת שכר של 15% ממשכורתו של כל עובד, בנוסף, ירידה קבועה של 500 ש"ח. מה תהיה ההכנסה הממוצעת החדשה ומהי סטיית התקן?

פתרון:

$\bar{X}$	X	*f*	F	*f '*
3,750	3,500 – 4,000	?	12	12 : 500 = 0.024
4,250	4,000 – 4,500	13	25	13 : 500 = 0.026
4,750	4,500 – 5,000	18	43	18 : 500 = 0.036
5,500	5,000 – 6,000	25	68	25 : 1,000 = 0.025
6,500	6,000 – 7,000	8	76	8 : 1,000 = 0.008
8,000	7,000 – 9,000	15	91	15 : 2,000 = 0.0075
10,500	9,000 – 12,000	9	100	9 : 3,000 = 0.003

n = 100

א. ידוע כי ההכנסה הממוצעת הייתה 5,897.5 ₪. מצא את מספר העובדים שהיו בקטגוריה הראשונה

$$\overline{X}=\frac{3{,}750*x+4{,}250*13+4{,}750*18+5{,}500*25+6{,}500*8+8{,}000*15+10{,}500*9}{88+x}$$

$$5{,}897.5=\frac{3{,}750x+544{,}750}{88+x}$$

$$x=12$$

ב. תחום: R=12,000-3,500=8,500

חציון:

$$n*\tfrac{1}{2}=100*0.5=50$$

$F\geq 50$ → (5,000 – 6,000)

$$ME=5{,}000+\left(\frac{1{,}000}{25}\right)*(50-43)=5{,}280$$

שכיח:

השכיח נמצא במחלקה (4,500 – 5,000), נציב;

$$MO=4{,}500+\frac{0.036-0.026}{(0.036-0.026)+(0.036-0.025)}*500$$

$$MO=4{,}738.095$$

עשירון עליון:

$$n*\tfrac{90}{100}=100*0.9=90$$

$F\geq 90$ → (7,000 – 9,000)

$$X_{90/100}=7{,}000+\left(\frac{2{,}000}{15}\right)*(90-76)=8{,}866.666$$

סטיית תקן:

$$S^2{}_x=\frac{12*3{,}750^2+13*4{,}250^2+18*4{,}750^2+25*5{,}500^2+8*6{,}500^2+15*8{,}000^2+9*10{,}500^2}{100}-5{,}897.5^2=$$

$$S_{X^2}=3{,}781{,}368.75$$

$$S_X=1{,}944.574$$

ג. הממוצע גדול מהחציון שגדול מהשכיח ולכן מדובר בהתפלגות אסימטרית ימנית (חיובית).

ד. עובד מסוים קיבל ציון תקן של (-1). מהי משכורתו?

$$-1 = \frac{X_i - 5{,}897.5}{1{,}944.574}$$

$$-1{,}944.574 = X_i - 5{,}897.5$$

$$X_i = 3{,}952.925$$

ה. הממוצע הוא מדד המושפע מערכים קיצוניים, ומכיוון שהוספנו ערכים קיצוניים, הממוצע יקטן . *סטיית תקן-* תגדל היות והפיזור גדל. לפני שהתווספו עשרה עובדים נע תחום הערכים בין 3,500-12,000 , כעת תחום הערכים הוא בין 2,000-12,000.

ו. נמצא את הממוצע החדש – מושפע גם מהשינוי באחוזים וגם מהשינוי הקבוע:

$$\overline{Y} = 0.85 * 5{,}897.5 - 500 = 4{,}512.875$$

נמצא את סטיית התקן החדשה – מושפעת רק מאחוזים:

$$S_Y = 0.85 * 1{,}944.574 = 1{,}652.887$$

הסתברות

תורת ההסתברות עוסקת בניסוי מקרי. ניסוי מקרי האו ניסוי שאין לדעת בוודאות מראש את תוצאות הניסוי, אלא קיים מרחב של תוצאות אפשריות.

לדוגמא: בהטלת קובייה מרחב התוצאות האפשריות הן: {1,2,3,4,5,6}

על מנת לבנות מודל הסתברותי נעבור קודם על מספר מושגים בסיסיים בתורת הקבוצות.

מושגים בסיסיים בתורת הקבוצות

- **אומגה** – Ω או W – מרחב המדגם. אוסף של כל התוצאות האפשריות של הניסוי המקרי.

  בהטלת קובייה מרחב המדגם : $\Omega=W = \{1, 2, 3, 4, 5, 6\}$

- **מאורע פשוט** – e_i – תוצאה בודדת במרחב המדגם.

  לדוגמא : קבלת את התוצאה 3 בהטלת קובייה $e_i = \{3\}$

- **מאורע** – מאורע A הינו אוסף כל התוצאות המקיימות את A

  לדוגמא: נניח שבהטלת קובייה מאורע A הינו קבלת תוצאה זוגית ולכן

  $A = \{2, 4, 6\}$

- **מאורע משלים**- מאורע משלים ל A נסמן ב- $\bar{A}$. אוסף כל התוצאות השייכות למרחב המדגם אך לא ל- A. (כל מספר המופיע ב- W אך לא ב- A)

  לדוגמא: אם מאורע A הינו קבלת תוצאה זוגית אז המאורע המשלים, $\bar{A}$ הוא כל התוצאות האי זוגיות $\bar{A} = \{1, 3, 5\}$

- **מאורע ריק** Ø – מאורע בו התוצאה הינה בלתי אפשרית.

  לדוגמה: בהטלת קובייה מאורע C הינו קבלת התוצאה 8 ולכן C הינו מאורע ריק.

  $C = \{\emptyset\}$

- **איחוד מאורעות U** – A U B משמעו אוסף כל התוצאות שכלולות ב A או ב B או בשניהם (המילה "או" מייצגת פעולת איחוד).

- **חיתוך מאורעות ∩** - חיתוך של המאורעות. $A \cap B$ משמעו כל התוצאות שכלולות ב A וגם ב B, כלומר תוצאות משותפות (המילה "וגם" מייצגת פעולת חיתוך).

דוגמא:

W = {1, 2, 3, 4, 5, 6, 7, 8}

A = {2, 4, 5, 6, 7}

B = {1, 2}

C = { 6, 8}

D = {1, 2, 5, 7, 8}

חשבו:

א. $A \cup B$

ב. $A \cap B$

ג. $\overline{D}$

ד. $\overline{A \cap B}$

פתרון:

$A \cup B = \{1, 2, 4, 5, 6, 7 \}$

$A \cap B = \{2 \}$

$\overline{D} = \{3, 4, 6\}$

$\overline{A \cap B} = \{1, 3, 4, 5, 6, 7, 8\}$

הקשר בין מאורע ומאורע משלים לו

- חיתוך של מאורע ומאורע משלים לו הם מאורעות זרים $\bar{A} \cap A = \varnothing$
- האיחוד של מאורע ומאורע משלים לו הוא מרחב המדגם $\bar{A} \cup A = \Omega = W$

שאלות לתרגול (6)

שאלה 1

נתונות הקבוצות הבאות:

W = {a,b,c,d,e,f,g,h,i,j}

M = {b,d,f,h,j}

R = {a,c,e,g,i}

S = {a,b,g}

P = {c,d,e,f}

מצא את הקבוצות הבאות:

א. $M \cup R$ ב. $\left(\overline{M \cap P}\right)$ ג. $M \cap R$ ד. $\left(\overline{R \cup S}\right)$

ה. $\overline{S}$ ו. $\left(\overline{S} \cap P\right) \cup R$

פתרון:

א. $M \cup R = (a,b,c,d,e,f,g,h,i)$

ב. $\left(\overline{M \cap P}\right) = (a,b,c,e,g,h,i,j)$

ג. $M \cap R = (\phi)$

ד. $\left(\overline{R \cup S}\right) = (d,f,h,j)$

ה. $\overline{S} = (c,d,e,f,h,i,j)$

ו. $\left(\overline{S} \cap P\right) \cup R = (a,c,d,e,f,g,i)$

<u>שאלה 2</u>

לפניך הגדרותיהן של מספר קבוצות:

W – מרחב המדגם הינו כל תושבי ניו יורק

M – גברים

L – צמחוני / צמחונית

P – בעלי / בעלות חיית מחמד

N – נשואים / נשואות

<u>בעזרת סימונים אלו, ציין את הקבוצות הבאות:</u>

א. הגברים בעלי חיית המחמד שאינם צמחונים

ב. צמחוניות נשואות

ג. הגברים שאינם נשואים ובעלי חיית מחמד או צמחוניות

ד. גברים שאינם בעלי חיית מחמד

ה. נשים בעלות חיית מחמד

פתרון:

א. הגברים בעלי חיית המחמד שאינם צמחונים

$M \cap P \cap \overline{L}$

ב. צמחוניות נשואות

$\overline{M} \cap L \cap N$

ג. הגברים שאינם נשואים ובעלי חיית מחמד או צמחוניות

$(M \cap \overline{N} \cap P) \cup (\overline{M} \cap L)$

ד. גברים שאינם בעלי חיית מחמד

$M \cap \overline{P}$

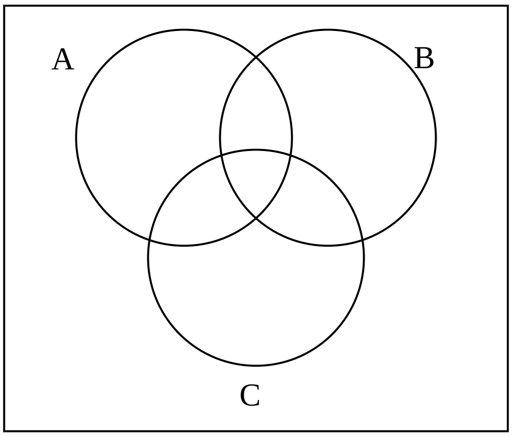
ה. נשים בעלות חיית מחמד

$\overline{M} \cap P$

<u>פעילות בין קבוצות – דיאגרמת וואן</u>

בדיאגרמת וואן מאורעות מיוצגים בעיגולים ומרחב המדגם מתואר על ידי מלבן או ריבוע.

לדוגמא: מאורעות A, B, C במרחב המדגם.

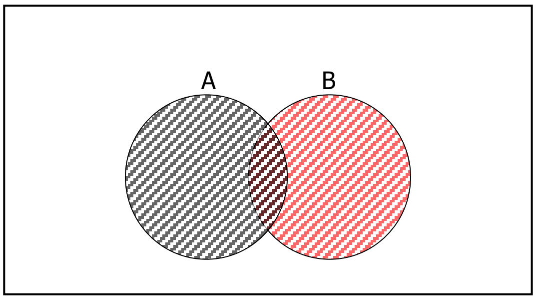


פעולת איחוד- בפעולת איחוד הפתרון בדיאגרמת וואן הינו כל השטח שסומן (בצבע אחד לפחות) וזאת כיוון שפעולת איחוד בין A ל B למשל היא כל המאורעות שכלולים ב A או ב B או בשניהם.

<u>דוגמא:</u>

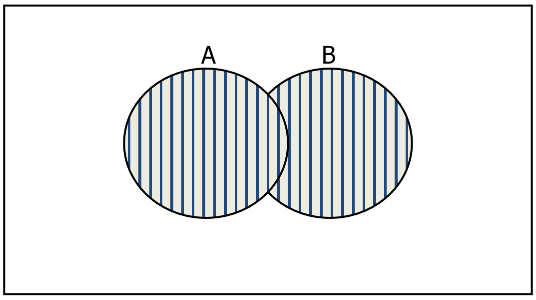
הצג A U B בדיאגרמת וואן.

בשלב הראשון: נסמן כל אחד מהקטעים בצבע נפרד. את A נסמן בצבע השחור ואת B בצבע האדום.



כעת יש להציג את האיחוד (כל השטח המסומן לפחות אחד משני הצבעים), השטח המבוקש הסופי הינו :

סופי

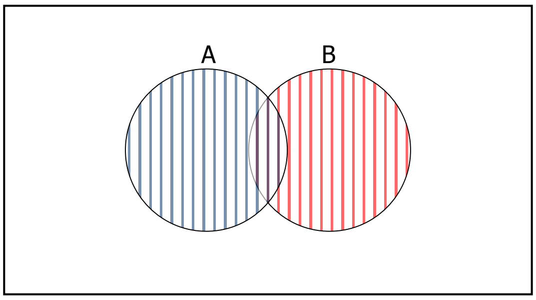


פעולת חיתוך- הפתרון בדיאגרמת וואן הינו השטח שסומן בכל הצבעים יחד, וזאת כיוון שפעולת חיתוך מייצגת את החלק המשותף.

<u>דוגמא:</u>

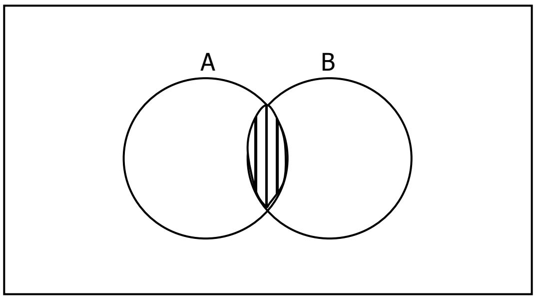
הציגו $B \cap A$ בדיאגרמת וואן.

בשלב הראשון : נסמן כל אחד מהקטעים בצבע נפרד. את A נסמן בצבע השחור ואת B בצבע האדום.



כעת יש להציג את החיתוך- כלומר את האזור בו כל הצבעים נמצאים יחד. ולכן התוצאה הסופית היא כלדקמן:

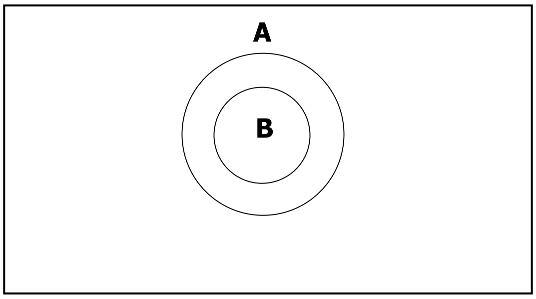
סופי



תצורות שונות של דיאגרמות וואן

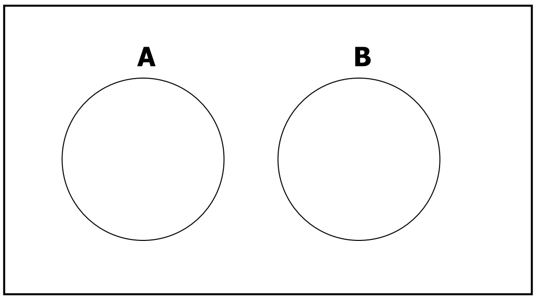
1. קבוצות חלקיות

אם B היא תת קבוצה של A, כלומר A מכילה את B, הקשר ביניהן יראה כך בדיאגרמת וואן:



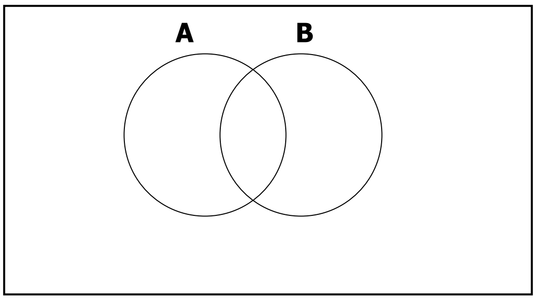
2. קבוצות זרות

אם ל A ול B אין איברים משותפים , הקשר בין הקבוצות יראה כך:



3. קבוצה לא זרות

קבוצות שיש להן איברים משותפים אך אינן קשורות בקשר של הכלה , הקשר בניהן יראה כך:



חוקים בתורת הקבוצות

- **חוק החילוף**

$A \cap B = B \cap A$

$A \cup B = B \cup A$

- **חוק הקיבוץ**

$(A \cap B) \cap C = A \cap (B \cap C)$

$(A \cup B) \cup C = A \cup (B \cup C)$

- **חוק הפילוג**

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

- **כללי דה-מורגן**

מטרת חוקי דה מורגן להפוך חיתוך באיחוד ולהפך על ידי שימוש במשלים. חוקי דה מורגן קובעים שלכל שתי מאורעות A ו- B מתקיים:

$\bar{A} \cup \bar{B} = \overline{A \cap B}$

$\bar{A} \cap \bar{B} = \overline{A \cup B}$

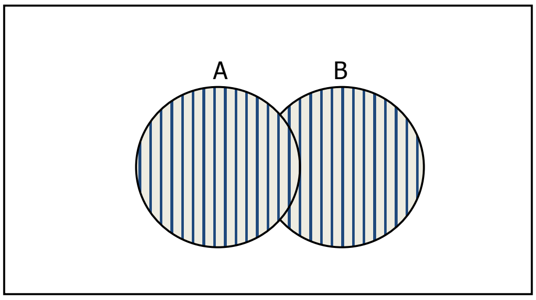
דוגמא:

שרטטו בדיאגרמת וואן את $\bar{A} \cap \bar{B}$

פתרון:

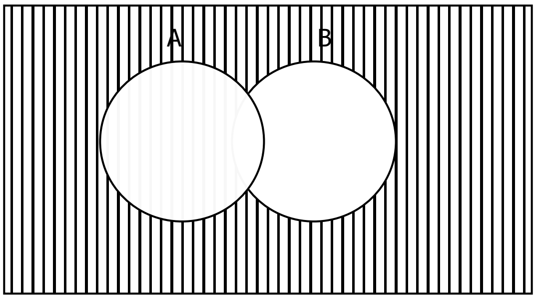
לפי כלל דה מורגן ניתן לומר ש $\bar{A} \cap \bar{B} = \overline{A \cup B}$

ראשית נשרטט את $A \cup B$



בשלב הסופי נציג את המשלים לאיחוד ונקבל:

סופי



.

<u>כתיב קבוצות ודיאגרמת וואן לשתי קבוצות</u>

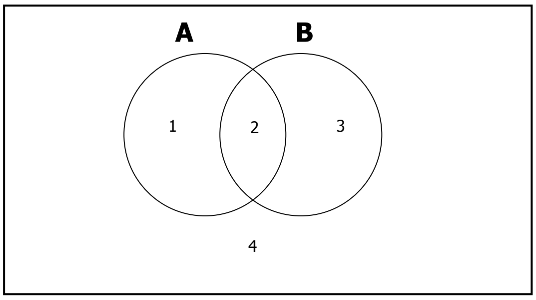
<u>נניח:</u>

(A) קבוצת האנשים שקוראים את עיתון ידיעות אחרונות

(B) קבוצת האנשים שקוראים את עיתון מעריב

(1) רק A

קבוצת האנשים שקוראים רק את ידיעות אחרונות



$A \cap \bar{B}$

(2) גם A וגם B:

$A \cap B$

קבוצת האנשים שקוראים את ידיעות אחרונות ומעריב

(3) רק B:

$\bar{A} \cap B$

קבוצת האנשים שקוראים רק את מעריב

(4) לא A ולא B:

$\bar{A} \cap \bar{B}$ או $\overline{A \cup B}$

קבוצת האנשים שלא קוראים מעריב ולא את ידיעות (לא קוראים אף עיתון).

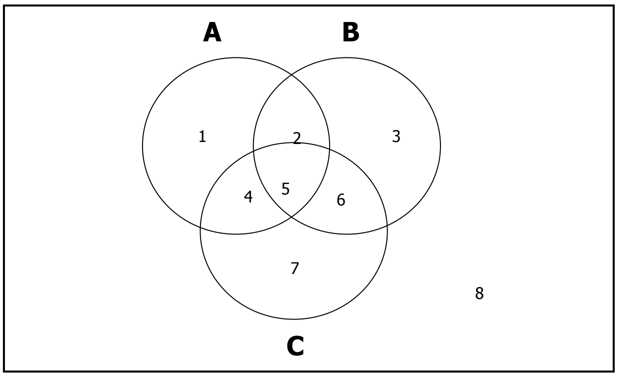
כתיב קבוצות ודיאגרמת וואן, שלוש קבוצות

נניח:

(A) קבוצת האנשים שקוראים את עיתון ידיעות אחרונות

(B) קבוצת האנשים שקוראים את עיתון מעריב

(C) קבוצת האנשים שקוראים את עיתון גלובס



(1) רק A:

רק ידיעות אחרונות

$A \cap \bar{B} \cap \bar{C}$

(2) רק A ו- B:

רק ידיעות ומעריב (ללא גלובס)

$A \cap B \cap \bar{C}$

(3) רק B:

רק מעריב

$\bar{A} \cap B \cap \bar{C}$

(4) רק A ו- C:

רק ידיעות וגלובס (ללא מעריב)

$A \cap C \cap \bar{B}$

(5) A וגם B וגם C:

ידיעות וגם מעריב וגם גלובס

$A \cap B \cap C$

(6) רק B ו- C:

רק מעריב וגלובס (ללא ידיעות)

$\bar{A} \cap B \cap C$

(7) רק C :

רק גלובס

$\bar{A} \cap \bar{B} \cap C$

(8) לא A ולא B ולא C:

לא קוראים אף עיתון

$\bar{A} \cap \bar{B} \cap \bar{C}$ או $\overline{A \cup B \cup C}$

<u>דוגמא 1</u>

900 איש נשאלו לגבי הבילוי המועדף עליהם בשעות הפנאי, ידוע ש:

600 איש אוהבים קולנוע

400 איש אוהבים תיאטרון

300 איש אוהבים מוזיאונים

200 איש אוהבים גם קולנוע וגם תיאטרון אבל לא אוהבים מוזיאונים

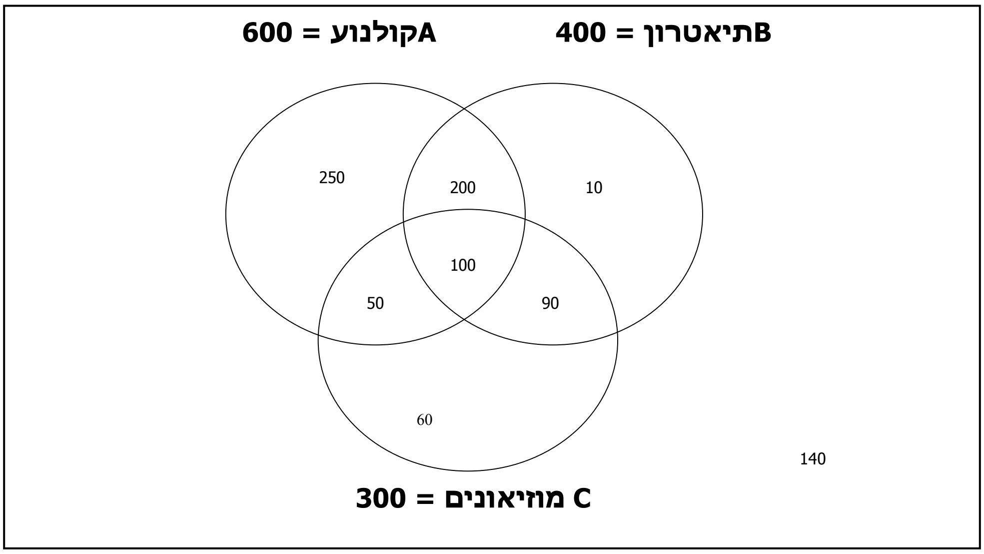
50 איש אוהבים קולנוע ומוזיאונים אבל לא אוהבים תיאטרון

90 איש אוהבים תיאטרון ומוזיאונים אבל לא אוהבים קולנוע

100 איש אוהבים את שלושת הבילויים

<u>פתרון:</u>

א.



א. כמה אנשים אוהבים רק מוזיאונים?

60

ב. כמה אנשים אוהבים בדיוק סוג בילוי אחד?

250+60+10=320

ג. כמה אנשים לא אף אחד מהבילויים?

140

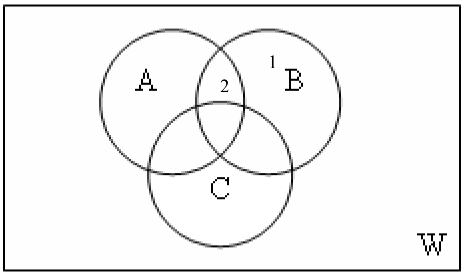
ד. כמה אוהבים לפחות בילוי אחד?

900-140=760

ה. כמה אוהבים בדיוק שני סוגי בילויים ?

200+50+90=340

<u>דוגמא 2</u>



תלמיד בתיכון מסוים נשלח לוועדת משמעת במידה ו:

A – התלמיד התחצף למורה

B – התלמיד העתיק עבודת בית

C – התלמיד העתיק במבחן

ענו על השאלות הבאות

א. איזה מאורע מייצג שטח 1?

ב. איזה מאורע מייצג שטח 1 ו-2?

פתרון:

א. שטח 1 מייצג את המאורע B בלבד. ולכן מייצג את המאורע שהתלמיד העתיק עבודת בית (ולא התחצף למורה ולא העתיק במבחן).

ב. על מנת לענות על סוג מעין זה של שאלה – נגדיר כל אחד מהשטחים המבוקשים ונמצא את המשותף ביניהם.

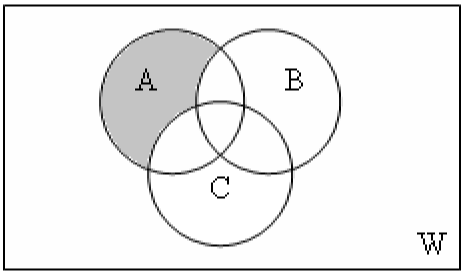
(1) ~~לא התחצף למורה~~, העתיק עבודת בית, לא העתיק במבחן

(2) ~~התחצף למורה~~, העתיק עבודת בית, לא העתיק במבחן

כלומר השטח מייצג את המאורע שהתלמיד העתיק עבודת בית ולא העתיק במבחן.

שאלות לתרגול (7)

שאלה 1



בכיתה מסוימת נגדיר את הקבוצות הבאות:

A – קבוצת בעלי רכב

B – קבוצת העולים החדשים

C – קבוצת תושבי העיר תל אביב

השטח המסומן בדיאגראמת וון מייצג את המאורע:

א. קבוצת תושבי העיר תל אביב שהינם עולים חדשים

ב. קבוצת תושבי העיקר תל אביב שאינם בעלי רכב והם עולים חדשים

ג. קבוצת בעלי הרכב, שאינם עולים חדשים ואינם תושבי העיר תל אביב

ד. קבוצת העולים החדשים הגרים בתל אביב

פתרון:

השטח המסומן בדיאגראמת וון מייצג את המאורע:

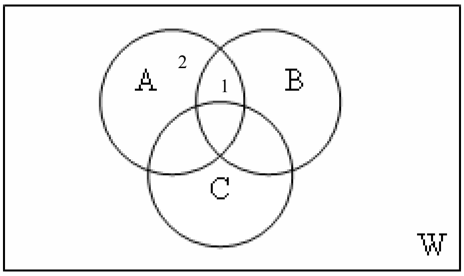
א. קבוצת תושבי העיר תל אביב שהינם עולים חדשים

ב. קבוצת תושבי העיקר תל אביב שאינם בעלי רכב והם עולים חדשים

ג. קבוצת בעלי הרכב, שאינם עולים חדשים ואינם תושבי העיר תל אביב

ד. קבוצת העולים החדשים הגרים בתל אביב

שאלה 2



עבור שגרת יומו של מתמחה נגדיר את המאורעות הבאים:

A – הכין קפה

B – יש הרבה עבודה

C – הגיע בבוקר בזמן

השטח המסומן בספרות 1 ו- 2 בדיאגראמת וון מייצג את המאורע:

א. יש הרבה עבודה

ב. הגיע בבוקר בזמן ויש הרבה עבודה

ג. הכין קפה או הגיע בבוקר בזמן

ד. הכין קפה ולא הגיע בבוקר בזמן

<u>פתרון:</u>

השטח המסומן בספרות 1 ו- 2 בדיאגראמת וון מייצג את המאורע:

a. יש הרבה עבודה

b. הגיע בבוקר בזמן ויש הרבה עבודה

c. הכין קפה או הגיע בבוקר בזמן

d. הכין קפה ולא הגיע בבוקר בזמן

על מנת לענות על סוג מעין זה של שאלה – נגדיר כל אחד מהשטחים המבוקשים ונמצא את המשותף בניהם;

(1) הכין קפה, יש ~~הרבה עבודה~~, לא הגיע בבוקר בזמן

(2) הכין קפה, אין ~~הרבה עבודה~~, לא הגיע בבוקר בזמן

<u>שאלה 3</u>

נגדיר את הקבוצות הבאות לגבי תלמידים מצטיינים בבית הספר בהרצליה:

G – מרכיב משקפיים, M – לומד ביולוגיה.

בטא באמצעות G, M והסמלים שלמדת את הקבוצות הבאות;

א. תלמידים שאינם מרכיבים משקפיים ואינם לומדים ביולוגיה

ב. תלמידים המרכיבים משקפיים ואינם לומדים ביולוגיה

ג. תלמידים הלומדים ביולוגיה ומרכיבים משקפיים או תלמידים שאינם מרכיבים משקפיים ואינם לומדים ביולוגיה

<u>פתרון:</u>

א. תלמידים שאינם מרכיבים משקפיים ואינם לומדים ביולוגיה

$\overline{M}\cap\overline{G}$ או $\overline{M\cup G}$

ב. תלמידים המרכיבים משקפיים ואינם לומדים ביולוגיה

$\overline{M}\cap G$

ג. תלמידים הלומדים ביולוגיה ומרכיבים משקפיים **או** תלמידים שאינם מרכיבים משקפיים ואינם לומדים ביולוגיה

$(M\cap G)\cup(\overline{M}\cap\overline{G})$

שאלה 4

מתוך 100 סטודנטים בפקולטה לכלכלה, נסמן ב- S את אלו שעברו בהצלחה את הקורס סטטיסטקה ו- M את אלו שעברו את הקורס מיקרו-כלכלה. נתון:

מספר הסטודנטים בקבוצה $\bar{S}\cap M$ הינו 37

מספר הסטודנטים בקבוצה $S\cap \bar{M}$ הינו 21

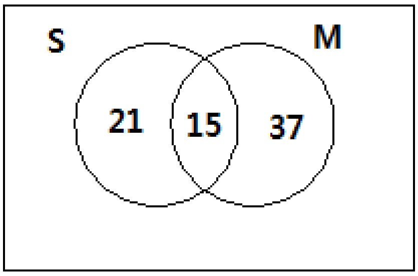
מספר הסטודנטים בקבוצה $S\cap M$ הינו 15

א. מצא את מספר הסטודנטים שעברו את הקורס מיקרו-כלכלה

ב. מצא את מספר הסטודנטים שעברו בהצלחה את הקורס סטטיסטקה

ג. כמה איברים בקבוצה $\bar{S}\cap\bar{M}$

<u>פתרון:</u>



נדרש:

א. מצא את מספר הסטודנטים שעברו את הקורס מיקרו-כלכלה

כלומר – M

15 + 37 = 52

ב. מצא את מספר הסטודנטים שעברו בהצלחה את הקורס סטטיסטקה

כלומר – S

21 + 15 = 36

ג. כמה איברים בקבוצה $\bar{S}\cap\bar{M}$

100 – 21 – 15 – 37 = 27

<u>שאלה 5</u>

350 חלקאים נשאלו איזה יבול מגדלים, נתון ש:

100 חקלאים מגדלים תמרים

93 חקלאים מגדלים מנגו

240 חקלאים מגדלים בננות

22 חקלאים מגדלים תמרים ומנגו אבל לא בננות

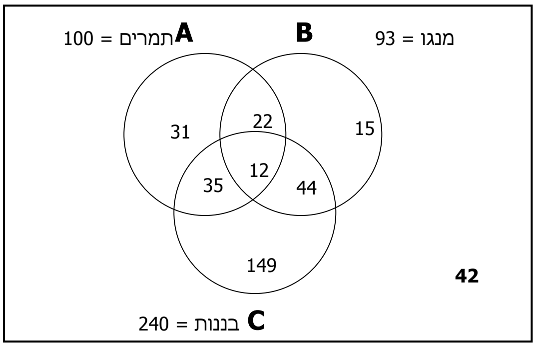
35 חקלאים מגדלים תמרים ובננות אבל לא מנגו

44 חקלאים מגדלים מנגו ובננות אבל לא תמרים

12 חקלאים מגדלים את שלושת הפירות

א. כמה חקלאים מגדלים בדיוק סוג פרי אחד?
ב. כמה חקלאים מגדלים בדיוק שני סוגי פירות?
ג. כמה חקלאים מגדלים <u>לפחות</u> שני סוגי פירות?
ד. כמה חקלאים מגדלים רק בננות?
ה. כמה חקלאים לא מגדלים פירות כלל?

<u>פתרון:</u>



א. כמה חקלאים מגדלים בדיוק סוג פרי אחד?

31+15+149=195

ב. כמה חקלאים מגדלים בדיוק שני סוגי פירות?

22 + 35 + 44 = 101

ג. כמה חקלאים מגדלים <u>לפחות</u> שני סוגי פירות?

22 + 35 + 44 + 12 = 113

ד. כמה חקלאים מגדלים רק בננות?

149

ה. כמה חקלאים לא מגדלים פירות כלל?

42

יסודות ההסתברות

הסתברות- הינה הסיכוי שמאורע מסוים יתרחש לאחר ביצוע הניסוי. נגדיר הסתברות ל A :

$$P(A) = \frac{\text{מספר המאורעות הפשוטים הכלולים ב } A}{\text{מספר המאורעות הפשוטים במרחב המדגם}}$$

לדוגמא: בהטלת קובייה:

- הסיכוי לקבל 2 שווה ל- $\frac{1}{6}$
- הסיכוי לקבל מס זוגי (2, 4, 6) שווה ל - $\frac{1}{2} = \frac{3}{6}$

תכונות ההסתברות

- תחום ההגדרה של ההסתברות הוא בין 1 ל 0.

$0 \le P(A) \le 1$

- ההסתברות של מרחב המדגם שווה ל – 1.

$\mathrm{p(w)} = \mathrm{p}(\Omega) = 1$

- ההסתברות לקבוצה ריקה שווה לאפס

$P(\varnothing) = 0$

- ההסתברות של חיתוך מאורעות זרים שווה לאפס

$P(A \cap B) = 0$

ההסתברות של איחוד מאורעות זרים

$$P(A\cup B)=P(A)+P(B)$$

דוגמא:

בהטלת קובייה מאורע A הוא קבלת מספר ריבועי.

מאורע B הוא קבלת מספר גדול מ- 4.

חשבו: $P(A\cup B)$

פתרון:

A={1,4}

B={5,6}

$$P(A)=\frac{2}{6}=\frac{1}{3}$$

$$P(B)=\frac{2}{6}=\frac{1}{3}$$

$$P(A\cup B)=P(A)+P(B)=\frac{1}{3}+\frac{1}{3}=\frac{2}{3}$$

ההסתברות של איחוד מאורעות לא זרים

ההסתברות שלפחות אחד מהמאורעות יקרה:

$$\boxed{P(A\cup B)=P(A)+P(B)-P(A\cap B)}$$

במידה ונעביר אגפים נקבל:

ההסתברות ששני המאורעות יתרחשו בו זמנית

$$\boxed{P(A\cap B)=P(A)+P(B)-P(A\cup B)}$$

דוגמא:

30% מהאוכלוסייה קוראת עיתון גלובס (מאורע A)

60% מהאוכלוסייה קוראת עיתון מעריב (מאורע B)

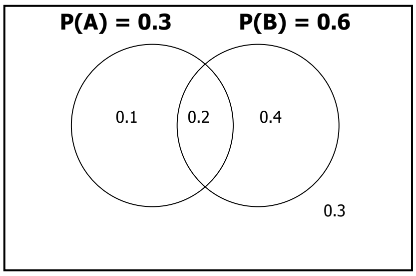
20% מאוכלוסייה קוראת גלובס ומעריב

א. אדם נבחר באופן מקרי, מה ההסתברות שהוא קורא לפחות אחד משני העיתונים?

ב. אדם נבחר באופן מקרי, מה הסיכוי שהוא קורא גם גלובס וגם מעריב?

ג. אדם נבחר באופן מקרי, מה הסיכוי שהוא לא קורא מעריב ולא קורא גלובס?

פתרון:



א. $P(A\cup B)=P(A)+P(B)-P(A\cap B)=0.3+0.6-0.2=0.7$

ב. $P(A\cap B)=0.2$

ג. $1-0.7=0.3$

שאלות לתרגול (8)

שאלה 1

F ו- G מאורעות זרים.

נתון כי: $P(F)=0.3$, $P(G)=0.4$, חשב את $P(F\cup G)$

א. 0.7

ב. 0.4

ג. 0.05

ד. 0

פתרון

תשובה א' נכונה 0.7 , כיוון שאיחוד של מאורעות זרים הינו החיבור שלהם – 0.3+0.4.

שאלה 2

A, B ו- C מאורעות זרים. נתון כי: $P(A)=0.15$, $P(B)=0.22$, $P(C)=0.38$, חשב את $P(A\cap B\cap C)$

א. 0

ב. 0.75

ג. 0.008

ד. 1

פתרון

תשובה א' נכונה כיוון שחיתוך של מאורעות זרים שווה לאפס.

שאלה 3

מדען מבצע שני ניסויים. ההסתברות להצלחה בניסוי הראשון היא 0.7, ההסתברות בניסוי השני היא 0.6. הסיכוי להצלחה של לפחות אחד הניסויים היא 0.75.

א. מה ההסתברות שהמדען יצליח בניסוי?

ב. מה ההסתברות שהמדען לא יצליח בשני הניסויים?

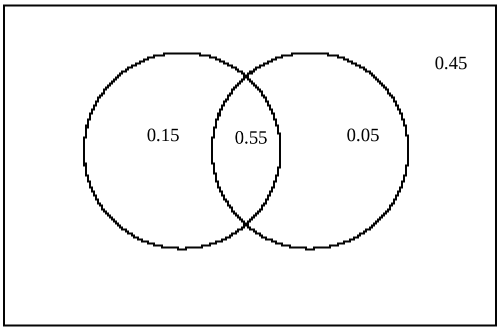
<u>פתרון</u>

$$P(A\cup B)=0.7+0.6-P(A\cap B)=0.75$$
$$P(A\cap B)=0.55$$

-

א. $P(B\bigcap A)=1.3-0.75=0.55$

ב. $1-0.55=0.45$

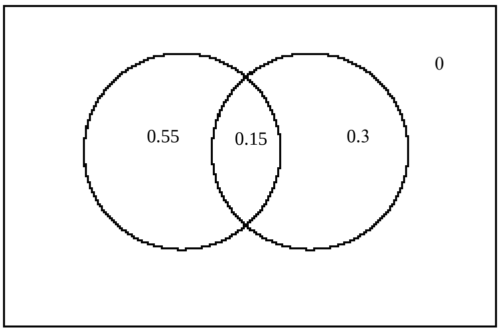


<u>שאלה 4</u>

כל הקיבוצים בצפון הארץ מגדלים **לפחות** פרי אחד. ישנם שני סוגי פירות – בננות ורימונים. 70% מגדלים בננות ו- 15% מגדלים את שני הסוגים. איזה אחוז מהקיבוצים בצפון מגדלים רק פרי אחד?

<u>פתרון</u>

*מאחר וכל הקיבוצים מגדלים לפחות פרי אחד (כלומר **אין** קיבוץ שאינו מגדל פירות) השטח מחוץ לעיגולים שווה לאפס.



$$55\%+15\%=70\%$$

<u>שאלה 5</u>

בסקר מסוים התברר כי 35% אוהבים עוגת תפוחים, 60% אוהבים עוגת גבינה ו- 25% אוהבים עוגת שוקולד. ידוע כי 15% אוהבים את שלושת העוגות. לכל זוג עוגות יש 20% שאוהבים אותם. מה ההסתברות שאדם שנבחר באקראי מהאוכלוסייה:

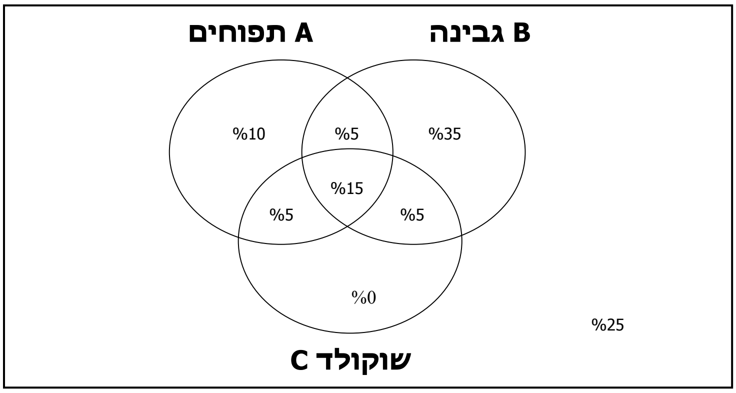
א. אוהב רק עוגה אחת?

ב. אוהב רק עוגת שוקולד?

ג. אוהב בדיוק שתי עוגות?

ד. לא אוהב אף אחת משלושת העוגות?

<u>פתרון</u>



מה ההסתברות ש:

א. אוהב רק עוגה אחת? 45% =10%+35%+0%

ב. אוהב רק עוגת שוקולד? 0

ג. אוהב בדיוק שניים? % 15 =% 5 +% 5 + 5%

ד. לא אוהב אף אחת מהעוגות? 25%

הסתברות מותנית ושימוש בעצי החלטות

הסתברות מותנית וההסתברות לחיתוך מאורעות

$P(B/A)$ ההסתברות שיקרה מאורע B אם ידוע ש A כבר קרה.

$$\boxed{P(B/A)=\frac{P(A\cap B)}{P(A)}}$$ - הסתברות מותנית

ניתן להציג נוסחה זו של הסתברות מותנה גם כך:

$$\boxed{P(A\cap B)=P(B/A)\cdot P(A)}$$ - ההסתברות לחיתוך מאורעות

דוגמא:

בהטלת שתי קוביות נגדיר את המאורעות הבאים:

A- התקבלה התוצאה 5 בקובייה הראשונה

B- סכום התוצאות בשתי הקוביות 11 לפחות

א. חשבו את ההסתברות לקבל 5 בקובייה הראשונה

ב. חשבו את ההסתברות לקבל סכום 11 לפחות בשתי הקוביות.

ג. חשבו את ההסתברות לקבל בהטלת שתי הקוביות סכום 11 לפחות אם ידוע כי בקובייה הראשונה התקבלה התוצאה 5.

פתרון:

א. להלן האפשרויות לקבל 5 בקובייה הראשונה (5,1) (5,2) (5,3) (5,4) (5,5), (5,6) כלומר 6 אפשריות מתוך מרחב מדגם של 36 וזאת כיוון שמרחב המדגם של הטלת קובייה אחת הוא 6 אפשרויות ומרחב המדגם של הטלת 2 קוביות הוא 6X6 =36 אפשרויות.

$$P(A)=\frac{6}{36}$$

ב. להלן האפשרויות לקבל 11 לפחות בשתי הקוביות.

(5,6) (6,5) (6,6)

כלומר $P(B)=\frac{3}{36}$

ג. בסעיף זה מרחב המדגם הצטמצם ל 6 אפשרויות כדלקמן: (5,1) (5,2) (5,3) (5,4) (5,5), (5,6), וזאת כיון שידוע שבקובייה הראשונה התקבל המספר 5.

מתוך אפשרויות אלה יש רק אופציה אחת שנותנת לנו סכום 11 לפחות (5,6) ולכן התשובה היא הסתברות של 1 מתוך 6 אפשרויות . ובנוסחה:

$$P(B/A)=\frac{P(A\cap B)}{P(A)}=\frac{1}{6}$$

<u>מאורעות תלויים ובלתי תלויים</u>

A ו- B מאורעות בלתי תלויים כאשר הסתברות המותנה שווה להסתברות ללא התנאי.

כלומר כאשר: $P(B/A)=P(B)$

דוגמא:

בכד 10 כדורים , 2 צהובים ו – 8 אדומים

נגדיר:

מאורע A– הוצאת כדור ראשון צהוב

מאורע B - הוצאת כדור שני אדום

א. מוציאים באקראי 2 כדורים ללא החזרה, מה ההסתברות להוציא כדור ראשון צהוב וכדור שני אדום?

ב. מוציאים באקראי 2 כדורים עם החזרה , מה ההסתברות להוציא כדור ראשון צהוב וכדור שני אדום?

<u>פתרון:</u>

א. יש לחשב את החיתוך בין שני המאורעות ולכן:

$$P(A)=\frac{2}{10}$$

$$P(B/A)=\frac{8}{9}$$

$$P(A\cap B)=P(A)\cdot P(B/A)=\frac{2}{10}\cdot\frac{8}{9}=\frac{16}{90}$$

מצב זה נכון כאשר אין החזרה , כלומר כאשר המאורעות תלויים.

ב. כעת יש לענות על השאלה כאשר יש החזרה, כלומר כאשר אין תלות בין המאורעות.

$$P(A)=\frac{2}{10}$$

$$P(B/A)=P(B)=\frac{8}{10}$$

$$P(A\cap B)=P(A)\cdot P(B)=\frac{2}{10}\cdot\frac{8}{10}=\frac{16}{100}$$

מצב זה נכון כאשר אין תלות בין המאורעות.

ונגדיר:

A ו – B מאורעות בלתי תלויים אם ורק אם:

$$P(A\cap B)=P(A)\cdot P(B)$$

אם:

$P(A\cap B)\neq P(A)\cdot P(B)\Leftarrow$ **המאורעות תלויים**

פתרון בעיות באמצעות שימוש בעצי החלטות

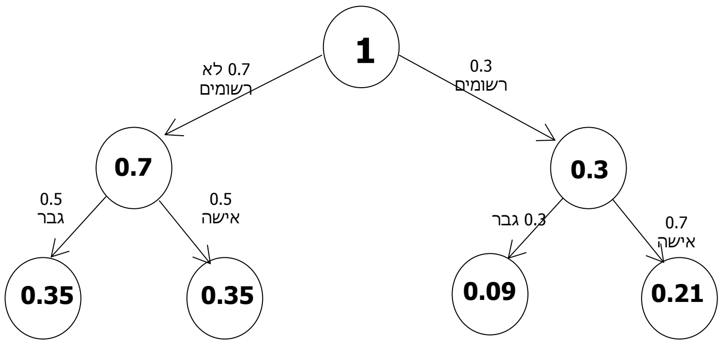
דוגמא:

30% מהאוכלוסייה הבוגרת רשומה למכון כושר. בין אלו הרשומים למכון כושר 70% נשים. מתוך אלו שאינם רשומים למכון כושר 50% נשים. אדם נבחר באקראי.

א. אדם נבחר באקראי, מה ההסתברות שמדובר בגבר?

ב. אם ידוע שנבחר גבר, מה ההסתברות שהוא אינו רשום למכון כושר?

פתרון:



א. אדם נבחר באקראי, מה ההסתברות שמדובר בגבר?

$$0.35+0.09=0.44$$

ב. אם ידוע שנבחר גבר, מהי ההסתברות שהוא אינו רשום לחדר כושר?

$$\frac{0.35}{0.35+0.09}=0.79$$

נשים לב למספר חוקים והנחיות בבניית עץ החלטות:

1. שאלות שאנו פותרים באמצעות עץ הן שאלות שיש בהן מספר שלבים. מבנה העץ צריך לשקף את השלבים המוצגים בשאלה.
2. מספר הענפים היוצאים מכל עיגול הוא כמספר האפשרויות. סכום ההסתברות על הענפים היוצאים מכל עיגול שווה ל -1.
3. סכום העיגולים באפשרות מסוימת צריך להיות שווה לעיגול ממנו יצאו העיגולים. (לדוגמא, בתרגיל מעלה 0.7 = 0.35+0.35)
4. בכל שלב בעץ סכום סך העיגולים חייב להשלים לאחד. לדוגמא , בתרגיל מעלה 1=0.35+0.35+0.21+0.09.

שאלות לתרגול (9)

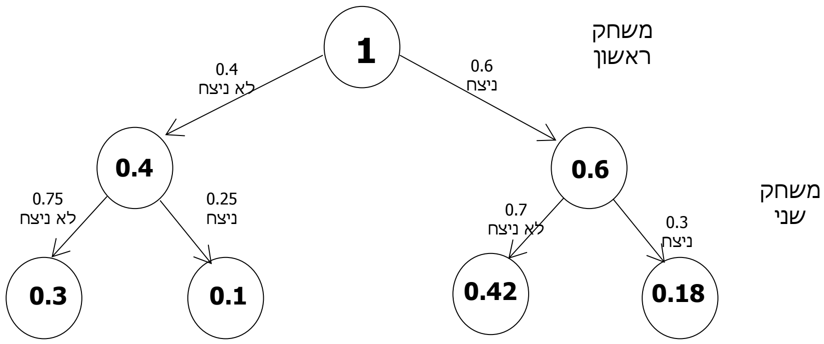
שאלה 1

אדם משחק "זוג או פרד" פעמיים. ההסתברות לנצח בפעם הראשונה היא 0.6, אם ניצח בפעם הראשונה, ההסתברות שלו לנצח שוב היא 0.3. אם הפסיד בפעם הראשונה ההסתברות שלו לנצח בפעם השנייה היא 0.25.

א. מה ההסתברות שהשחקן ניצח במשחק אחד בלבד?

ב. אם ידוע שניצח פעם אחת בלבד מה ההסתברות שניצח במשחק השני?

פתרון:



א. מה ההסתברות שהשחקן ניצח במשחק אחד בלבד?

$0.1 + 0.42 = 0.52$

ב. אם ידוע שניצח פעם אחת בלבד מה ההסתברות שניצח במשחק השני?

$$\frac{0.1}{0.1 + 0.42} = 0.192$$

שאלה 2

ההסתברות שלירן ייקח מטריה בבוקר היא 0.8 וההסתברות שהוא יגיע לעבודה בזמן היא 0.6. ההסתברות שלירן גם ייקח מטריה וגם יגיע בזמן לעבודה היא 0.5.

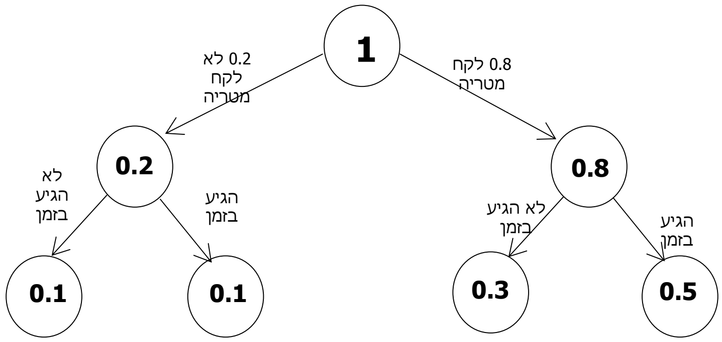
א. מה ההסתברות שלירן לקח מטריה ולא הגיע בזמן ?

ב. מה ההסתברות שלירן יגיע בזמן אם ידוע שלקח מטריה?

ג. מה ההסתברות שלירן לא לקח מטריה אם ידוע שהגיע בזמן?

ד. ידוע שלירן לא לקח מטריה , מה ההסתברות שגם לא יגיע בזמן לעבודה?

<u>פתרון</u>



לקח מטריה והגיע בזמן+ לא לקח מטריה והגיע בזמן = 0.6

$0.6 = x + 0.5$

$x = 0.1$

א. מה ההסתברות שלירן לקח מטריה ולא הגיע בזמן ?

0.3

ב. מה ההסתברות שלירן יגיע בזמן אם ידוע שלקח מטריה?

$$\frac{0.5}{0.3 + 0.5}$$

ג. מה ההסתברות שלירן לא לקח מטריה אם ידוע שהגיע בזמן?

$$\frac{0.1}{0.1 + 0.5}$$

ד. ידוע שלירן לא לקח מטריה , מה ההסתברות שגם לא יגיע בזמן לעבודה?

$$\frac{0.1}{0.1 + 0.1}$$

<u>רגרסיה לינארית</u>

רגרסיה לינארית הינה שיטה סטטיסטית המשמשת למטרות חיזוי ובדיקת קשר בין משתנים.

דוגמאות: הקשר בין גובה ומשקל, עישון ומחלת לב.

המקרה הפשוט ביותר הוא זה שקיימים 2 משתנים:

X- משתנה מסביר (בלתי תלוי)

Y- משתנה מוסבר (תלוי)

כלומר X מסביר את Y.

מטרת קו הרגרסיה הוא לנבא את Yהמשתנה התלוי באמצעות המשתנה הבלתי תלוי.

<u>דוגמא:</u>

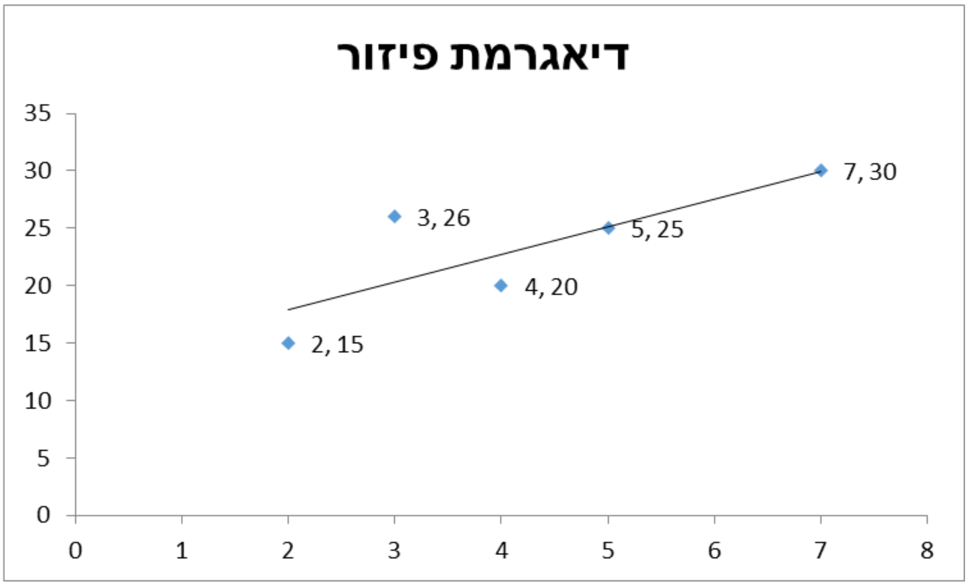
להלן נתונים עבור 5 תצפיות של מספר האנשים בתור בקופת החולים וזמן ההמתנה בדקות.

תצפית	מספר האנשים בתור – X	זמן ההמתנה – Y
1	2	15
2	5	25
3	7	30
4	4	20
5	3	26

מכאן, המשתנה המסביר הוא מספר האנשים בתור והמשתנה המוסבר זה זמן ההמתנה.

רגרסיה לינארית היא בעצם מדגם של זוגות של 2 משתנים X ו Y . שלכל תצפית X מתאימה תצפית Y אחת ולהפך.

מדגם של זוגות ניתן לתאר באמצעות דיאגרמת פיזור:



נבנה קו יש שהוא הקו שהסטיות ממנו הן מינימליות, זהו קו הרגרסיה.

<u>נבחין בין המצבים הבאים:</u>

1. קשר לינארי חיובי – המשתנים נוטים לנוע לאותו כיוון. בממוצע כאשר משתנה אחד עולה גם המשתנה השני יעלה. ולהפך, כאשר משתנה אחד יורד משתנה שני גם יורד.
2. קשר לינארי שלילי- המשתנים נוטים לנוע בכיוונים מנוגדים. כלומר, בממוצע כאשר משתנה אחד עולה משתנה שני יורד ולהפך.
3. קיים קשר אך הוא אינו לינארי- יש קשר כלשהו בין המשתנים אך הקשר אינו ניתן להגדרה כקשר לינארי (קו ישר)
4. אין קשר בין המשתנים- שינוי במשתנה אחד אינו משפיע על המשתנה השני

<u>מקדם המתאם של פירסון</u>

מדד סטטיסטי אשר בודק את הקשר בין שני משתנים. את הכיוון של הקשר (קשר חיובי או שלילי) ואת רמת הדיוק או העוצמה של הקשר (חזק או חלש).

$$r_{xy} = \frac{Sxy}{Sx * Sy}$$

תחום ההגדרה של מקדם המתאם הוא בין 1+ ל 1-.

r= 1 קשר לינארי חיובי מושלם

r= -1 קשר לינארי שלילי מושלם

r= 0 לא קיים קשר לינארי

<u>כללי אצבע לפירוש עוצמת הקשר:</u>

קשר חיובי	תיאור	קשר שלילי
0	אין קשר לינארי	0
0.01- 0.19	קשר חלש מאוד	[-0.19]- [-0.01]
0.20- 0.39	קשר חלש	[-0.39]- [-0.20]
0.40- 0.69	קשר בינוני	[-0.69]-[0.40]
0.70- 0.89	קשר חזק	[-0.89]- [-0.70]
0.90- 0.99	קשר חזק מאוד	[- 0.99]- [-0.90]
1	קשר מושלם	-1

<u>שונות משותפת</u>

על מנת לחשב את מקדם המתאם יש לחשב ראשית שונות משותפת. שונות משותפת היא מדד סטטיסטי המראה את כיוון הקשר בין המשתנים (חיובי או שלילי).

הנוסחה לחישוב שונות משותפת:

$$S_{XY}=\frac{\sum(X_i-\overline{X})\cdot(Y_i-\overline{Y})}{n}$$

<u>המשך פתרון דוגמא:</u>

ראשית נחשב את הממוצע ל X ול Y. ונקבל:

$$\overline{X}=\frac{2+5+7+4+3}{5}=4.2$$

$$\overline{Y}=\frac{15+25+30+20+26}{5}=23.2$$

כעת נחשב שונות וסטיית תקן ל X ול Y:

$$S^2{}_X=\sum\frac{X^2{}_i}{n}-\overline{X}^2$$

שימו לב!
בסטטיסטיקה תיאורית יש בנוסחה fi, כאן כיוון שכל תצפית היא אחת השכיחות תמיד 1.

ונקבל :

$$S^2_X = \sum \frac{X^2_i}{n} - \bar{X}^2 = \frac{2^2+5^2+7^2+4^2+3^2}{5} - 4.2^2 = 2.96$$
$$S_X = \sqrt{2.96} = 1.72$$

$$S^2_y = \sum \frac{Y^2_i}{n} - \bar{Y}^2 = \frac{15^2+25^2+30^2+20^2+26^2}{5} - 23.2^2 = 26.96$$
$$S_y = \sqrt{26.96} = 5.19$$

כעת נחשב את השונות משותפות:

$$S_{XY} = \frac{\sum (X_i - \bar{X})\cdot(Y_i - \bar{Y})}{n} =$$
$$\frac{(2-4.2)(15-23.2)+(5-4.2)(25-23.2)+(7-4.2)(30-23.2)+(4-4.2)(20-23.2)+(3-4.2)(26-23.2)}{5} = 7.16$$

כעת ניתן לחשב את מקדם המתאים על פי הנוסחה:

$$r_{xy} = \frac{\text{Sxy}}{\text{Sx*Sy}} = \frac{7.16}{1.72*5.19} = 0.80$$

מקדם המתאם 0.80 – קיים קשר לינארי חיובי חזק

כעת ניתן לבנות את משוואת קו ישר כדלקמן:

קו לחיזוי Y- $\hat{Y}$ (תחזית Y לפי X)

נוסחת קו הרגרסיה

$$\hat{y} = b_{y|x} \cdot x + a_{y|x}$$ [1]

$\hat{y}$ - תחזית Y

$b_{y|x}$ - השיפוע של הקו (של Y לפי X) – מהו השינוי הצפוי ב – Y עם שינוי ביחידת X.

$a_{y|x}$ - חותך הקו עם ציר ה- Y. מהו Y הצפוי כאשר X שווה לאפס.

חישוב שיפוע

שתי נוסחאות אפשריות:

$$b_{y|x} = \frac{S_{xy}}{S^2_{\ x}}$$

או

$$b_{y|x} = \frac{R \bullet S_y}{S_x}$$

חישוב חותך

$$a_{y|x} = \bar{y} - b_{y|x} \cdot \bar{x}$$

לפי דוגמא:

$$b_{y|x} = \frac{S_{xy}}{S^2_{\ x}} = \frac{7.16}{2.96} = 2.42$$

[1] ניתן להתייחס למשוואת הקו הישר גם כ y=ax+b כפי שנעשה בפרק 1, נושא 9 בנושא שינוי לינארי. בספר זה נתייחס למשוואת הקו הישר של רגרסיה לינארית כמוצג במשוואה מעלה כאשר b מייצג את השיפוע ו- a את החותך.

$$a_{y|x}=\bar{y}-b_{y|x}\cdot\bar{x}=23.2-2.42\bullet 4.2=13$$

ולכן הקו הוא:

$$\hat{Y}=2.42X+13$$

בהסתמך על קו הרגרסיה שבניתם ענו על השאלות הבאות:

א. מה צפוי להיות זמן ההמתנה בתור כאשר אין אנשים בתור בכלל?

תשובה:

$\hat{Y}=2.42X+13=2.42\cdot 0+13=13$ זמן ההמתנה הצפוי כאשר X=0 הינו 13 דקות

ב. מה צפוי להיות זמן ההמתנה כאשר יש 10 אנשים בתור?

תשובה:

$\hat{Y}=2.42X+13=2.42\cdot 10+13=37.2$ זמן ההמתנה כאשר x=10 הינו 37.2 דקות.

שאלות לתרגול (10)

שאלה 1

בבניין משרדים הממוקם בתל אביב בדקו את הקשר בין מספר האנשים הנמצאים בבניין לבין מספר החניות **הפנויות** בכל יום. התקבלו הנתונים הבאים עבור 5 ימים:

מספר חניות פנויות Y	מספר האנשים X
40	60
35	65
25	35
15	80
0	110

א. חשב את מקדם המתאם בין מספר האנשים לחניות הפנויות. הסבר/י את התוצאה שהתקבלה.

ב. כמה חניות פנויות צפויות להיות עבור 90 אנשים?

פתרון

א.

$$\bar{X} = \frac{60+65+35+80+110}{5} = 70$$

$$S_X = \sqrt{\frac{60^2+65^2+35^2+80^2+110^2}{5} - 70^2} = 24.69$$

$$\bar{Y} = \frac{40+35+25+15+0}{5} = 23$$

$$S_Y = \sqrt{\frac{40^2+35^2+25^2+15^2+0^2}{5} - 23^2} = 14.35$$

$$S_{xy} = \frac{(60-70)*(40-23)+(65-70)*(35-23)+(35-70)*(25-23)+(80-70)*(15-23)+(110-70)*(0-23)}{5} = -260$$

$$r_{xy} = \frac{Sxy}{Sx*Sy} = \frac{-260}{24.69*14.35} = -0.733$$

- קיים קשר לינארי שלילי וחזק.

ב. נמצא את משוואת קו הרגרסיה:

$$b = \frac{Sxy}{S^2x} = \frac{-260}{609.59} = -0.4265$$

$$a = \bar{y} - b * \bar{x} = 23 - (-0.4265) * 70 = 52.85$$

$$Y = -0.4265 * X + 52.85$$

ומכאן אם נציב 90 במקום X (מספר האנשים בבניין), אז המסקנה שמספר החניות הפנויות יהיה 14.47

<u>שאלה מספר 2</u>

חוקר מעוניין לבדוק האם קיים תוצאת הציון הפסיכומטרי משפיעה על ממוצע הציונים בסיום התואר הראשון.

החוקר בחר 10 בוגרים באקראי וקיבל את התוצאות הבאות:

X- ציון פסיכומטרי	Y- ממוצע התואר הראשון
550	60
600	85
650	90
620	92
660	95
640	90
710	100
680	96
690	98
720	100

א. חשב את מקדם המתאם בין X ל Y

ב. בנה את משוואת הרגרסיה של Y כפונקציה של X

פתרון

א. מקדם המתאם:

$$\bar{X} = 652$$

$$S_X = 49.55$$

$$\bar{Y} = 90.6$$

$$S_Y = 11.18$$

$$S_{xy} = 502.8$$

$$r_{xy} = \frac{Sxy}{Sx*Sy} = \frac{502.8}{49.55*11.18} = 0.907$$

ב. משוואת קו הרגרסיה:

$$b = \frac{Sxy}{S^2x} = \frac{502.8}{2,455.2} = 0.2047$$

$$a = \bar{y} - b*\bar{x} = 90.6 - 0.2047*652 = -42.87$$

$$Y = 0.2047*X - 42.87$$

ניסוי בינומי

הקדמה

לפני שנסביר על ההתפלגות הבינומית. להלן הקדמה לגבי מספר חישובים ומונחים רלוונטיים:

עצרת (!)

!n מייצג את מספר אפשרויות לסידור n פריטים בשורה.

החישוב מתבצע על ידי מכפלת כל המספרים עד 1 :

$n!=n*(n-1)*(n-2)……2*1$

דוגמאות:

$100! = 100 * 99 * 98 * 97 * …. * 1$

$5! = 5 * 4 * 3 * 2 * 1$

$1! = 1$

כדי לחשב $\binom{n}{K}$ דהיינו את מספר האפשרויות של K מתוך n (כאשר הסדר לא חשוב) נשתמש בנוסחה הבאה:

$$\binom{n}{K}=\frac{n!}{(n-k)!k!}$$

דוגמא:

מתוך 7 אנשים נבחר קבוצה של 3, ללא חשיבות לסדר הבחירה. בכמה קבוצות שונות ניתן לבחור?

פתרון:

$$\binom{7}{3}=\frac{7!}{4!*3!}=\frac{7*6*5*4*3*2*1}{4*3*2*1*3*2*1}=35$$

לחילופין ניתן לחישוב בעזרת הכפתור NCR במחשבון ונקבל:

(7NCR3)=35

ניסוי בינומי הוא ניסוי הבנוי מ n ניסויים בלתי תלויים, כאשר בכל ניסוי קיימות שתי אפשרויות: הצלחה או כישלון. ההסתברות להצלחה היא p וההסתברות לכישלון היא q.

נסמן ב- k את מספר ההצלחות ו n-k את מספר הכישלונות.

$$P(x=K)=\binom{n}{K}p^{K}\cdot(q)^{n-k}$$

כאשר X מתפלג בינומית נרשום כך: X~Bin(n,p)

K = מספר ההצלחות

n= מספר הניסויים

p=ההסתברות להצלחה

q= ההסתברות לכישלון (כאשר q+p=1)

<u>דוגמא:</u>

ההסתברות של שחקן שחמט מסוים לנצח במשחק היא 0.6

א. מה ההסתברות שינצח 2 משחקים מתוך 4?

X~B(4,0.6)

n = 4

P = 0.6

q=1-0.6=0.4

K = 2

$$P(x=2)=\binom{4}{2}0.6^{2}\cdot(0.4)^{2}=$$

$$\left(4NCR2\right) * 0.6^{2} * 0.4^{2}=$$

$$6 * 0.6^{2} * 0.4^{2}=0.3456$$

ב. מה ההסתברות שינצח פעמיים אם ישחק 5 משחקים ?

X~Bin(5,0.6)

$$P(x=2)=\binom{5}{2}*0.6^{2}*0.4^{3}=5NCR2*0.6^{2}*0.4^{3}=0.2304$$

ג. בהמשך לסעיף ב, מה ההסתברות שינצח לפחות 4 משחקים?

P(x=4) + P(x=5)

$$P(x=4) = \binom{5}{4} * 0.6^4 * 0.4^1 = 5NCR4 * 0.6^4 * 0.4^1 = 0.2592$$

$$P(x=5) = \binom{5}{5} * 0.6^5 * 0.4^0 = 5NCR5 * 0.6^5 * 0.4^0 = 0.07776$$

0.33696=0.2592+0.07776

ד. מה ההסתברות שינצח לכל היותר 4 משחקים?

אפשרות ראשונה היא לחשב :

$$P(x \le 4) = P(x=1) + P(x=2) + P(x=3) + P(x=4)$$

אפשרות שניה היא לחשב את המשלים:

$$P(x \le 4) = 1 - P(x=5) = 1 - \binom{5}{5} * 0.6^5 * 0.4^0 = 1-0.07776=0.92224$$

ה. מה ההסתברות שלא ינצח בכלל?

$K = 0$

$$P(x=0) = \binom{5}{0} * 0.6^0 * 0.4^5 = 5NCR0 * 0.6^0 * 0.4^5 = 0.01024$$

<u>התוחלת והשונות של ההתפלגות</u>

התוחלת (ממוצע) של משתנה מקרי בינומי= הסתברות כפול גודל המדגם (נסמן ב – E).[2]

$$E(x) = n \cdot P$$

השונות:

$$S^2 = n \cdot P \cdot (1-P)$$

[2] כיוון שמדובר בהתפלגות תאורטית (הסתברות של תוצאות צפויות) ולא אמפירית נגדיר את הממוצע כתוחלת ונסמן אותו E.

דוגמא:

בהמשך לדוגמא הקודמת כאשר X~Bin(5,0.6), חשבו תוחלת ושונות.

פתרון:

$$E(x) = n \cdot P = 5 \cdot 0.6 = 3$$

המשמעות היא שתוחלת או ממוצע הניצחונות הוא 3 . כלומר אם השחקן ישחק את המשחק אין סוף פעמים בממוצע ינצח 3 מתוך 5 משחקים.

כעת נחשב את השונות:

$$S^2 = n \cdot P \cdot (1-P) = 5 \cdot 0.6 \cdot (1-0.6) = 1.2$$

ולכן סטיית התקן שווה ל-

$$S = \sqrt{1.2} = 1.095$$

סטיית התקן מתארת את מידת הפיזור של הערכים סביב התוחלת.

שאלות לתרגול (11)

שאלה 1

מתלמד בלוחמת קרטה מנסה מזלו בניסיון לשבור לוחות עץ 4 פעמים. ידוע כי הסיכוי של מתלמד להצליח לשבור לוח הינו 35%. הנח שהניסיונות השונים אינם תלויים ביניהם.

נגדיר: X – משתנה מקרי המונה את מספר הפעמים שהצליח לשבור לוח.

מבין 4 הניסיונות, מה ההסתברות ש ב– 2 או יותר יצליח לשבור לוח?

פתרון

$X \sim B(4,\ 0.35)$

$$\mathrm{P}(X \geq 2) = 1 - P(X=0) - P(X=1) =$$
$$= 1 - \binom{4}{0} * 0.35^0 * (1-0.35)^4 - \binom{4}{1} * 0.35^1 * (1-0.35)^3 = 0.437$$

שאלה 2

מכונה חדשה שפותחה מעלה את הסיכוי לייצור מוצרים תקינים מ-0.6 ל-0.8, המכונה החדשה תייצר 6 מוצרים.

א. מהי התפלגות , תוחלת ושונות של מספר המוצרים התקינים שתייצר המכונה החדשה?

ב. מהי ההסתברות שלפחות 5 מוצרים שתייצר המכונה החדשה יהיו תקינים?

ג. מהי התפלגות , תוחלת ושונות מספר המוצרים אשר יוצרו במכונה החדשה ואינם תקינים?

ד. ייצרו במכונה <u>הישנה</u> 6 מוצרים גם כן, מהי תוחלת ושונות מספר המוצרים התקינים?

פתרון

א. $X \sim B(6,\ 0.8)$

$$E(X) = 6 * 0.8 = 4.8$$
$$\nu(X) = 6 * 0.8 * (1 - 0.8) = 0.96$$

ב. $P(X \geq 5) = P(X = 5) + P(X = 6) = 0.65536$

ג. $X \sim B(6,\ 0.2)$

$$E(X) = 6 * 0.2 = 1.2$$
$$\nu(X) = 6 * 0.2 * (1 - 0.2) = 0.96$$

ד.

$$E(X) = 6 * 0.6 = 3.6$$
$$\nu(X) = 6 * 0.6 * (1 - 0.6) = 1.44$$

<u>התפלגות פואסונית</u>

מתאימה למקרים בהם יש התרחשות של אירועים בקצב קבוע בפרק זמן מסוים. התפלגות זו מתארת את ההסתברות כי אירוע מסוים יקרה K פעמים בזמן נתון.

נגדיר X מתפלג פואסנית (λ) X~P

כאשר λ- קצב האירועים ליחידת זמן

נגדיר את ההסתברות שאירוע מסוים יקרה K פעמים כדלקמן:

$$\mathrm{p(x=k)} = \ell^{-\lambda} * \frac{\lambda^k}{\mathrm{k!}}$$

<u>התוחלת והשונות של ההתפלגות</u>

בהתפלגות זו התוחלת שווה לשונות ושווה ל λ.

$$E(x) = S^2 = \lambda$$

<u>דוגמא:</u> מספר המכוניות הממוצע שעוברות בכביש לא סלול מסוים בדרום הארץ הוא 3 מכוניות לדקה. מספר המכוניות העוברות בכביש זה מתפלג פואסונית. $\lambda = 3$ - קצב האירועים הינו 3 מכוניות לדקה

א. מהי ההסתברות כי בדקה מסוימת לא תעבור אף מכונית?

K=0

$$\mathrm{p(x=0)} = \ell^{-\lambda} * \frac{\lambda^k}{\mathrm{k!}} = \ell^{-3} * \frac{3^0}{0!} = 0.049$$

ב. מהי ההסתברות כי במשך 4 דקות תעבורנה בכביש 8 מכוניות?

הערה: יש להביא את ה- K ואת ה- λ למכנה משותף , ולכן עבור 4 דקות

k=8 $\lambda = 3*4 = 12$

$$\mathrm{p(x=8)} = \ell^{-\lambda} * \frac{\lambda^k}{\mathrm{k!}} = \ell^{-12} * \frac{12^8}{8!} = 0.065$$

ג. מהי התוחלת, השונות וסטיית התקן של מספר המכוניות שעוברות לדקה?

$$E(x) = S^2 = \lambda = 3$$

סטיית התקן שווה לשורש השונות $S = \sqrt{3} = 1.73$

<u>שאלות לתרגול (12)</u>

<u>שאלה 1</u>

המספר הילדים הממוצע לדקה המתגלשים על מגלשה מסוימת במגרש המשחקים בשעות היום הוא 2 לדקה ומתפלג פואסונית.

א. מהי התוחלת ושונות מספר הילדים המתגלשים על המגלשה בדקה מסוימת?

ב. מה הסיכוי שבדקה מסוימת לא יתגלש אף ילד במגלשה?

ג. מה הסיכוי שבדקה מסוימת לפחות ילד אחד יתגלש במגלשה?

ד. מה הסיכוי שבמשך 10 דקות יתגלשו על המגלשה 15 ילדים?

<u>פתרון</u>

א. $E(x) = S^2 = \lambda = 2$

ב. $\mathrm{p(x = 0)} = \ell^{-2} * \dfrac{2^0}{0!} = 0.135$

ג. $\mathrm{p(x \geq 1)} = 1 - \mathrm{p(x = 0)} = 1 - (\ell^{-2} * \dfrac{2^0}{0!}) = 1 - 0.135 = 0.86$

ד. $\lambda = 2 * 10 = 20$

$$\mathrm{p(x = 15)} = (\ell^{-20} * \frac{20^{15}}{15!}) = 0.051$$

<u>שאלה 2</u>

מספר התקלות ליום במערכת הרמזורים שלאורך דרך נמיר הוא משתנה מקרי פואסוני.
ממוצע התקלות במערכת עומד על 4 תקלות בחודש.

א. מה ההסתברות שבמהלך חודשיים מסוימים תהיה לפחות תקלה אחת?

ב. מה ההסתברות שבמהלך שבועיים מסוימים (הנח/י בחודש 4 שבועות) תהיינה 2 תקלות?

פתרון

λ = 4 לחודש

א. מה ההסתברות שבמהלך חודשיים מסוימים תהיה לפחות תקלה אחת?

$$\lambda = 4 * 2 = 8$$

$$p(x \geq 1) = 1 - p(x = 0) = 1 - e^{-8} * \frac{8^0}{0!} = 0.9996$$

ב. מה ההסתברות שבמהלך שבועיים מסוימים (הנח/י בחודש 4 שבועות) תהיינה 2 תקלות?

$$\lambda = \frac{4}{2} = 2$$

$$p(x = 2) = e^{-2} * \frac{2^2}{2!} = 0.2706$$

התפלגות נורמאלית

ההתפלגות הנורמלית הינה התפלגות חשובה ביותר בשיטה הסטטיסטית. ההתפלגות ידועה גם בשם התפלגות פעמון או התפלגות גאוס.

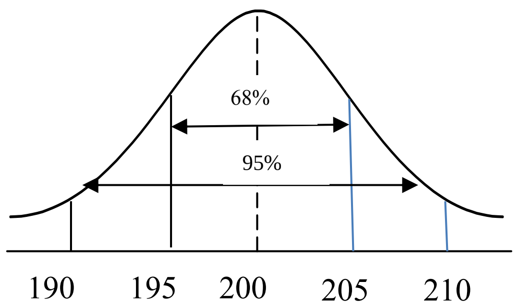
להלן תכונות ההתפלגות

- התפלגות של משתנה מקרי כמותי רציף
- התפלגות סימטרית ,חד שיאית
- ממוצע= שכיח = חציון
- רוב המקרים נמצאים באמצע ושכיחות המקרים הולכת ויורדת

להלן אחוזי ההתפלגות סביב התוחלת לפי סטיות התקן עבור ההתפלגות הנורמלית:

- במרחק סטיית תקן אחת משני הצדדים של הממוצע נמצאים 68% מהמקרים
- במרחק של שתי סטיות תקן משני הצדדים של התוחלת נמצאים 95% מהמקרים
- במרחק של שלוש סטיות תקן משני הצדדים של התוחלת נמצאים 99.7% מהמקרים.

לדוגמא: משתנה X מתפלג נורמלית עם ממוצע 200 ושונות 25.



השונות היא 25 ולכן סטיית התקן היא 5.

כלומר במרחק סטיית תקן אחת מהממוצע , דהיינו בין 195 ל 200 נמצאים 68% מהמקרים.

במרחק שתי סטיות תקן מהתוחלת, דהיינו בין 190 ל 210 נמצאים 95% מהמקרים.

במרחק שלוש סטיות תקן מהתוחלת, דהיינו הין 215 ל 185 נמצאים 99.7% מהמקרים.

ערכי Z ושימוש בלוח ההתפלגות הנורמלית הסטנדרטית

ההתפלגות הנורמלית הסטנדרטית ,היא ההתפלגות של משתנה מקרי Z (ציון תקן) .
התפלגות זו היא התפלגות נורמלית שהתוחלת שלה היא 0 והשונות שווה לאחד (גם סטיית התקן שווה לאחד). עבור ההתפלגות Z קיימת טבלה המספקת לכל ערך Z את אחוז התצפיות או ההסתברות המצטברת. כלומר את השטח מתחת לעקומה <u>משמאל</u> לערך Z.

ציון תקן

מתאר מיקום יחסי של תצפית X בהתפלגות. ציון תקן מייצג את המרחק של תצפית X מהממוצע של ההתפלגות ביחידות של סטיית תקן. ציון תקן הוא מספר ללא יחידות ולכן ניתן להשוות אותו להתפלגויות שונות.

נוסחה לחישוב ציון תקן[3]:

$$Z = \frac{X_i - \bar{X}}{S_x}$$

כאשר

X_i - הערך של תצפית i

$\bar{X}$ - ממוצע

S_x - סטיית התקן

[3] בסטטיסטיקה נהוג להבדיל בין מדד סטטיסטי למדד פרמטר. ממד סטטיסטי הוא מדד במדגם. כגון ממוצע , שונות וסטיית תקן של המדגם. מדד פרמטר הוא מדד באוכלוסייה. כגון ממוצע, שונות וסטיית תקן באוכלוסייה. נהוג לעיתים להבדיל בין מדדים סטטיסטיים לפרמטרים ולסמן $\overline{X}$ כממוצע המדגם ו µ כממוצע באוכלוסייה, S_x כסטיית התקן במדגם לעומת σ_x התקן באוכלוסייה. למען הפשטות, בספר זה נסמן תמיד את הממוצע כ- $\bar{x}$ ואת סטיית התקן כ S_x.

עבור כל התפלגות נורמלית ניתן לראות את ההסתברות המצטברת עד לנקודה מסוימת לפי ציון תקן Z. בהתפלגויות שונות אם ה- Z שווה, כלומר המרחק בסטיות תקן מהממוצע שווה, אז ההסתברות המצטברת עד הנקודה שווה.

<u>דוגמא:</u>

התפלגות 1- משתנה X מתפלג נורמלית עם ממוצע 5 וסטיית תקן 10.

חשבו את ציון תקן עבור תצפית 25.

$$\mathrm{Z}=\frac{\mathrm{X_i}-\bar{X}}{S_{\mathrm{x}}}=\frac{25-5}{10}=2$$

התצפית רחוקה 2 סטיות תקן מהממוצע

התפלגות 2- Y הינו משתנה מקרי המתפלג נורמלית עם סטיית תקן 5 ותוחלת 100.

חשבו את ציון תקן עבור תצפית 110.

$$\mathrm{Z}=\frac{\mathrm{X_i}-\bar{X}}{S_{\mathrm{x}}}=\frac{110-100}{5}=2$$

גם בהתפלגות 2 התצפית רחוקה 2 סטיות תקן מהממוצע.

ולכן אחוז התצפיות עד לתוצאה 25 בהתפלגות ראשונה שווה לאחוז התצפיות עד לתוצאה 110 בהתפלגות השנייה.

לוח ההתפלגות הנורמלית הסטנדרטית[4] – Z

לוח ההתפלגות הנורמלית הסטנדרטית (ראו לוח מטה בעמוד הבא) משמש כבסיס לחישובים עבור כל ההתפלגויות הנורמליות, כאשר עבור כל ערך Z , ניתן לחשב את השטח המתאים (כלומר אחוז התצפיות או ההסתברות המצטברת) עד ערך ה- Z.

בטבלת ההתפלגות בטור השמאלי מופיעים ערכי ה- Z עם ספרה אחת אחרי הנקודה ובשורה העליונה הספרה השנייה אחרי הנקודה. בגוף הלוח מופיע השטח או אחוז התצפיות המצטבר מראשית ההתפלגות ועד הערך Z.

דוגמא:

א. מהו אחוז התצפיות המייצג את השטח עד הערך Z=1.21 ?

ב. מהו ה- Z המתאים לאחוז התצפיות של 0.516 ?

פתרון:

א. כדי למצוא את ההסתברות המתאימה לערך 1.21 נמצא את ה- Z המתאים לפי צידי הטבלה. בטור הראשון מצד שמאל נסתכל על הערך 1.2, על מנת להגיע ל 1.21 נסתכל על 0.01. ניתן לראות שהאחוז המתאים בגוף הטבלה הינו 0.8869 ולכן אחוז התצפיות עד הערך Z=1.21 הינו 0.8869

ב. כעת יש למצוא את ה- Z המתאים לפי האחוז. לכן כעת נסתכל בגוף הטבלה ונחפש את הערך 0.516. ניתן לראות ש ערך ה- Z המתאים הוא 0.04.

[4] לעיתים תמצאו הבדלים קטנים בין הלוחות השונים של ההתפלגות הנורמלית הסטנדרטית. ההבדלים נובעים ממספר הספרות לאחר הנקודה שנלקחו בחשבון.

ההתפלגות הנורמלית הסטנדרטית

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0046	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0135	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0227	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0238	0.0233
-1.8	0.0359	0.0350	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0550	0.0540	0.0530	0.0520	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0670	0.0650	0.0640	0.0630	0.0620	0.0610	0.0590	0.0580	0.0570	0.0560
-1.4	0.0810	0.0790	0.0780	0.0760	0.0750	0.0740	0.0720	0.0710	0.0690	0.0680
-1.3	0.0970	0.0950	0.0930	0.0920	0.0900	0.0890	0.0870	0.0850	0.0840	0.0820
-1.2	0.1150	0.1130	0.1110	0.1090	0.1070	0.1060	0.1040	0.1020	0.1000	0.0980
-1.1	0.1360	0.1340	0.1310	0.1290	0.1270	0.1250	0.1230	0.1210	0.1190	0.1170
-1.0	0.1590	0.1560	0.1540	0.1520	0.1490	0.1470	0.1450	0.1420	0.1400	0.1380
-0.9	0.1840	0.1810	0.1790	0.1760	0.1740	0.1710	0.1680	0.1660	0.1630	0.1610
-0.8	0.2120	0.2090	0.2060	0.2030	0.2000	0.1980	0.1950	0.1920	0.1890	0.1870
-0.7	0.2420	0.2390	0.2360	0.2330	0.2300	0.2270	0.2240	0.2210	0.2180	0.2150
-0.6	0.2740	0.2710	0.2680	0.2640	0.2610	0.2580	0.2550	0.2510	0.2480	0.2450
-0.5	0.3080	0.3050	0.3010	0.2980	0.2950	0.2910	0.2880	0.2840	0.2810	0.2780
-0.4	0.3450	0.3410	0.3370	0.3340	0.3300	0.3260	0.3230	0.3190	0.3160	0.3120
-0.3	0.3820	0.3780	0.3750	0.3710	0.3670	0.3630	0.3590	0.3560	0.3520	0.3480
-0.2	0.4207	0.4168	0.4129	0.4009	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

חישוב הסתברויות בהתפלגות נורמלית

כאשר יש לחשב את אחוז התצפיות עבור ערך מסוים נעבוד על פי השלבים הבאים:

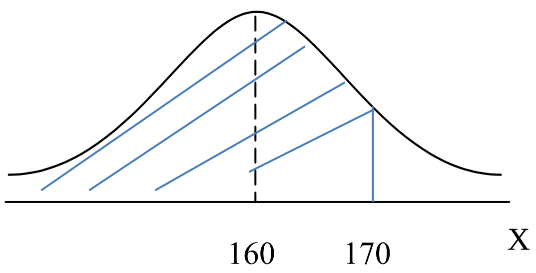
1. נחשב את ערך ה- Z על ידי הצבה בנוסחת סטיית התקן
2. נמצא את ערך ה – Z המתאים בטבלה (לפי הטור השמאלי ושהשורה העליונה) ואת האחוז המתאים בגוף הטבלה.

דוגמא:

גובה האוכלוסייה הבוגרת מתפלגת נורמלית עם ממוצע 160 ס"מ וסטיית תקן של 15.

א. חשבו את אחוז האנשים שגובהם מתחת ל 170
ב. חשבו את אחוז האנשים שגובהם נמוך יותר מ 140.
ג. חשבו את אחוז האנשים שגובהם מעל 140.
ד. חשבו את אחוז האנשים שגובהם בין 167 ל 155 ס"מ.
ה. חשבו את אחוז האנשים שגובהם רחוק מהממוצע בלפחות 40 ס"מ.

פתרון:



א. נחשב את ציון התקן המתאים

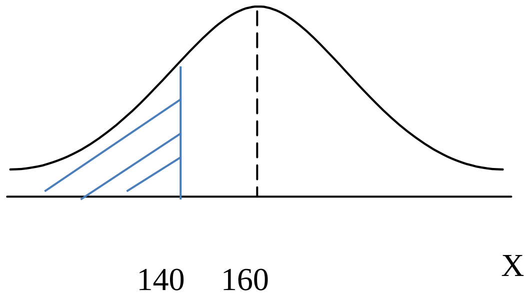
$$Z = \frac{X_i - \bar{X}}{S_x} = \frac{170-160}{15} = 0.67$$

כעת נמצא את הערך Z בטבלת ההתפלגות הנורמלית ואת האחוז המתאים

P(X<170)= 0.7486=74.86%

משמעות תוצאה זו היא שעד הערך 170 קיימים 74.86% מהמקרים באוכלוסייה.

ב. מהו אחוז האנשים שגובהם נמוך יותר מ 140 ס"מ ?



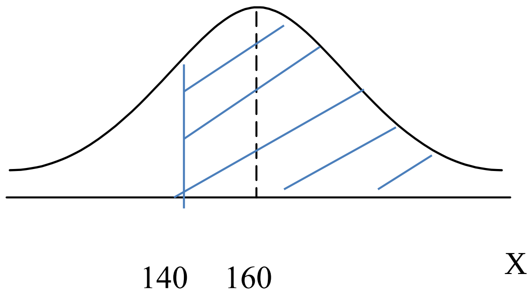
נחשב את ציון התקן המתאים:

$$Z = \frac{X_i - \bar{X}}{S_x} = \frac{140-160}{15} = -1.33$$

נמצא את האחוז המתאים בטבלה ונקבל :

על פי הטבלה האחוז המתאים הוא 0.92 , ולכן P(X<140) =0.092=9.2%

א. אחוז האנשים שגובהם מעל 140



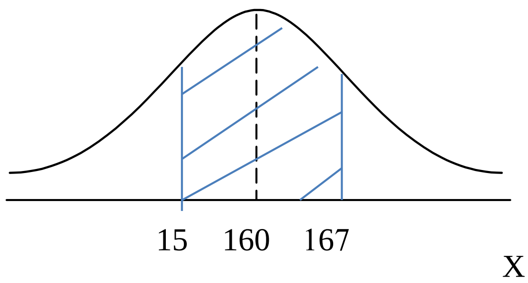
שימו לב!! בטבלה תמיד נקבל את האחוזים עד הערך (כלומר משמאל לערך), ולכן אם אנחנו מעוניינים למצוא את האחוזים מימין לערך, התוצאה תהיה אחד פחות האחוזים משמאל לערך. וזאת כיוון שסך השטח מתחת להתפלגות שווה ל 100% או אחד. ולכן:

P(X>140)=1-0.092=90.08%

כלומר ל 90.08% מהאוכלוסייה גבוהים יותר מ 140 ס"מ.

ב. אחוז האנשים שגובהם בין 167 ל 155 ס"מ.

כדי למצוא שטח בין שתי נקודות. נמצא את ערך ה- Z והאחוז המתאים בטבלה לכל אחת מהנקודות. לאחר מכן נחסיר בין שתי התוצאות.



נמצא את ציון התקן והאחוזים המתאימים עבור גובה 167

$$Z=\frac{X_i-\bar{X}}{S_x}=\frac{167-160}{15}=0.46$$

P(x<167)= 0.6772

כעת נחשב את ציון התקן והאחוזים עבור גובה 155

$$Z=\frac{X_i-\bar{X}}{S_x}=\frac{155-160}{15}=-0.33$$

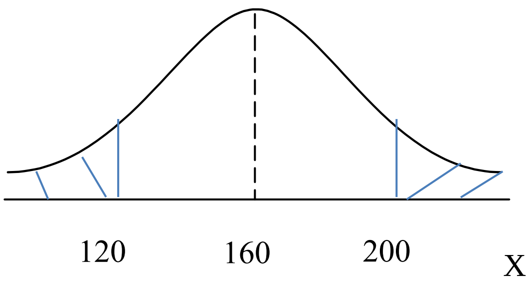
P(x<155)= 0.3710

כדי לקבל את השטח הרלוונטי נחסר בין התוצאות:

P= 0.6772-0.3710= 0.3062=30.62%

כלומר, 30.62% מהאוכלוסייה עם גובה בין 155 ל 167.

ג. מהו אחוז האנשים שגובהם רחוק ב 40 ס"מ או יותר מהממוצע?



נמצא את ערך ה – Z והשטח עבור גובה 120:

$$Z = \frac{X_i - \bar{X}}{S_x} = \frac{120-160}{15} = -2.67$$

P(x<100)= 0.0038

כעת נחשב את ערך ה Z והשטח עבור 200 :

$$Z = \frac{X_i - \bar{X}}{S_x} = \frac{200-160}{15} = 2.67$$

P(X>200)=1-0.9962=0.0038

סך ההסתברות שווה ל 0.0038+0.0038= 0.0076

*שימו לב! בגלל שהשטחים סימטריים הם גם שווים ולכן יכולנו לחשב רק שטח אחד ולהכפילו ב 2.

חישוב ערך מסוים לפי אחוז התצפיות בהתפלגות הנורמלית

כאשר אנו נדרשים לחשב את הערך של X, כאשר ידוע אחוז התצפיות שהתקבלו עד לערך זה, נעבוד לפי השלבים הבאים:

1. נמצא את האחוז הנתון הקרוב היותר בתוך טבלת ההתפלגות הנורמלית (גוף הטבלה)
2. נמצא את ערך ה- Z המתאים בטבלה.
3. נציב בנוסחת ציון התקן ונחלץ את- x.

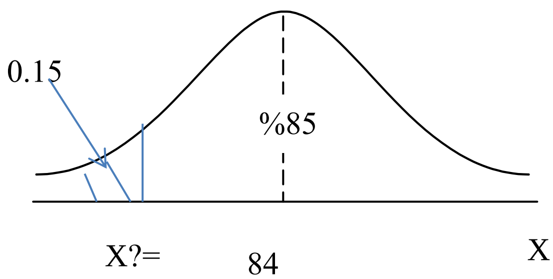
דוגמא:

התפלגות ציוני התלמידים בבחינות הבגרות שואפת לנורמלית עם ממוצע 84 וסטיית תקן של 6.

א. מהו הציון ש 85% מהתלמידים קיבלו ציון גבוה ממנו.
ב. תלמיד מצטיין הינו תלמיד שנמצא בין ה-10% שקיבלו את הציונים הכי גבוהים. מהו הציון המינימלי שיש לקבל כדי להיות מצטיין?

פתרון:

א.



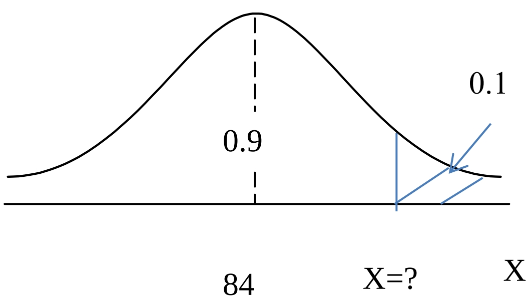
ה- Z המתאים ל 0.15 שווה בקירוב ל Z=-1.04.[5]

[5] על מנת לקבל קירוב מדויק יותר ניתן לחשב ממוצע של ערכי ה- Z הקרובים ביותר לאחוז הרלוונטי. במקרה שלנו ממוצע בין 1.04 ל 1.03 , כלומר 1.035 . בספר זה אנו נשתמש בערך הקרוב ביותר כפי שהתבצע בדוגמא.

נציב בנוסחה:

$$-1.04 = \frac{X_i - 84}{6}$$

$X_i = 75.68 \approx 76$, כלומר 85% מהסטודנטים קיבלו מעל ציון 76.



כיוון שהטבלה מראה את השטח משמאל לערך נחפש את ה- Z המתאים עבור 0.9 , ונקבל Z=1.28 בקירוב (ערך ה- Z בטבלה שהכי קרוב ל 0.9)

נציב כעת בנוסחת ה- Z ונחלץ את הציון (X):

$$1.28 = \frac{X_i - 84}{6}$$

מכאן $X_i = 91.68 \approx 92$, כלומר 10% מהסטודנטים קיבלו ציון מעל 92.

שאלות לתרגול (13)

שאלה 1

מפעל מזמין משלוח של חומרי גלם מידי יום. כמות חומר הגלם הינה משתנה מקרי X בעל התפלגות נורמלית עם תוחלת 25 טון וסטיית תקן 4 טונות. בוחרים באקראי משלוח של חומרי גלם. חשב את ההסתברויות הבאות:

א. הכמות היא לכל היותר 27 טון
ב. הכמות היא לכל היותר 29 טון
ג. הכמות היא לפחות 29 טון
ד. הכמות היא לכל היותר 21 טון

פתרון:

א.

$$Z=\frac{X_i-\overline{X}}{Sx}=\frac{27-25}{4}=0.5$$
$$p(x<27)=0.6915=69.15\%$$

ב.

$$Z=\frac{X_i-\overline{X}}{Sx}=\frac{29-25}{4}=1$$
$$p(x<29)=0.8413=84.13\%$$

ג.

$$Z=\frac{X_i-\overline{X}}{Sx}=\frac{29-25}{4}=1$$
$$p(x<29)=0.8413=84.13\%$$
$$p(x>29)=1-0.8413=0.1587=15.87\%$$

ד.

$$Z=\frac{X_i-\overline{X}}{Sx}=\frac{21-25}{4}=-1$$
$$p(x<21)=0.1590=15.90\%$$

שאלה 2

ההכנסה החודשית לאדם בעיר מסוימת מתפלגת נורמלית עם תוחלת 6,000 ₪ וסטיית תקן 200.

א. מה ההסתברות שההכנסה החודשית של אדם הנבחר באקראי תעלה על 6,140 ₪ ?

ב. מה ההסתברות שההכנסה החודשית של אדם הנבחר באקראי תהיה בין 5,900 ל 6,160 ש"ח ?

ג. מה צריכה להיות ההכנסה החודשית של אדם כך ש – 96% מהעובדים האחרים בעיר ירוויחו לחודש יותר ממנו?

<u>פתרון</u>

א.

$$Z=\frac{X_i-\bar{X}}{Sx}=\frac{6{,}140-6{,}000}{200}=0.7$$
$$p(x>6{,}140)=1-0.7580=24.2\%$$

ב.

$$Z=\frac{6{,}160-6{,}000}{200}=0.8\rightarrow P=0.7881$$
$$Z=\frac{5{,}900-6{,}000}{200}=-0.5\rightarrow P=0.3085$$
$$P=0.7881-0.3085=0.4796$$

ג.

$$-1.75=\frac{X_i-6{,}000}{200}$$
$$X_i=5{,}650$$

יישומים באקסל

מטרת חלק זה של הקורס הוא יישום הנושאים השונים שלמדנו כגון: הצגת דיאגרמות , חישוב מדדים וחישוב קו הרגרסיה באקסל.

בניית דיאגרמות באקסל

דיאגרמת מקלות

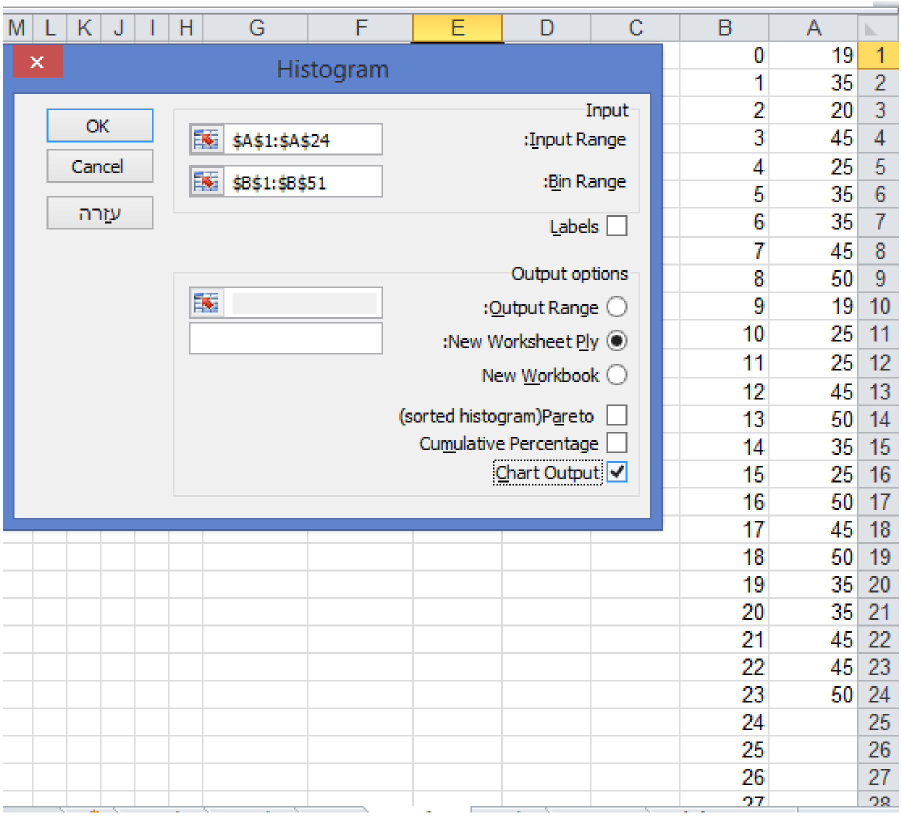
הדיאגרמה המתאימה לתיאור משתנה כמותי בדיד הינה דיאגרמת מקלות. כדי לראות כיצד ניתן לבנות דיאגרמת מקלות באקסל נסתכל על הדוגמא הבאה.

להלן התפלגות הגילאים במכון כושר מסוים:

19	45
35	50
18	35
45	25
25	50
35	45
35	50
45	35
50	35
18	45
25	45
25	50

<u>בניית דיאגרמת מקלות באקסל</u>

1. נחשב מספר מקסימלי על ידי שימוש בפונקציית Max.
   במקרה שלנו המספר מקסימלי שווה ל 50
2. נבנה טור נוסף החל מהערך אפס ועד לערך הגדול ביותר (ניתן לרשום שני מספרים לפחות ולאחר מכן לגרור עד המספר המקסימלי).
3. נלחץ נתונים ← Data analysis ←Histogram
4. בחלון שנפתח נקליד את הנתונים כמוצג מטה:
   Input Rang - נקליד את הנתונים הגולמיים
   Bin Range - נקליד את הטור המאורגן מהערך אפס עד לערך הגדול ביותר
5. נבחר באופציה Chart output



הערה: out put option מתייחס להיכן אנחנו רוצים לקבל את הפלט (בגיליון חדש או באזור מסוים בגיליון עליו אנחנו עובדים). בחירה זו נעשית לפי נוחות והעדפות.
לסיום נלחץ OK ולאחר סידור הכותרות והצירים נקבל את הדיאגרמה הבאה:

היסטוגרמה (דיאגרמת מלבנים)

היסטוגרמה הינה הדיאגרמה המתאימה לתיאור משתנה כמותי רציף, כלומר לטבלה שמסודרת במחלקות.

בניית היסטוגרמה באקסל

בהמשך לנתונים מהדוגמא הקודמת יש להציג היסטוגרמה, כאשר רוחב מחלקה שווה ל 10.

1. נחשב את הערך המקסימלי על ידי שימוש בפונקציית Max
2. נחליט על רוחב המחלקה הרצוי לנו (בדוגמא שלנו הוחלט על רוחב רצוי של 10).
3. נציג עמודת חלוקת מחלקות לפי גבול עליון של כל מחלקה החל מהערך אפס ועד הערך המקסימלי כדלקמן:

   מחלקה ראשונה = מינימום (ערך אפס) + רוחב מחלקה

   מחלקה שניה= סוף מחלקה ראשונה + רוחב מחלקה וכך הלאה

בדוגמא שלנו נקבל את המחלקות הבאות:

מחלקה	הסבר
10	0-10
20	11-20
30	21-30
40	31-40
50	41-50

4. באקסל נבחר נתונים ← data analysis ←Histogram

בחלון שנפתח נקליד את הנתונים כדלקמן:

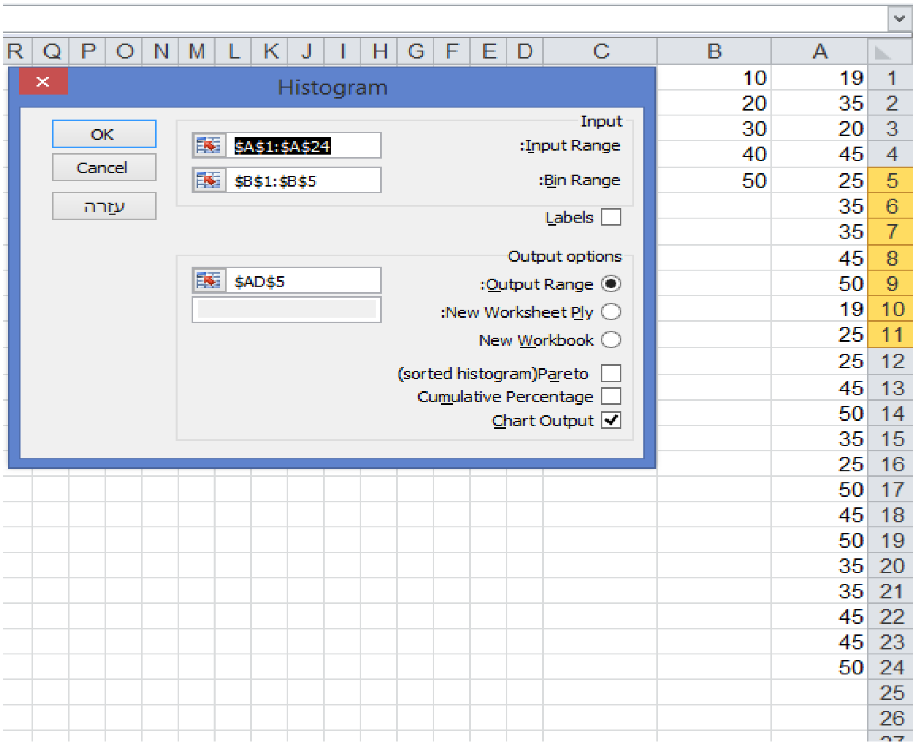
Input Range - נקליד את הנתונים הגולמיים

Bin Range - עמודת המחלקות שהכנו

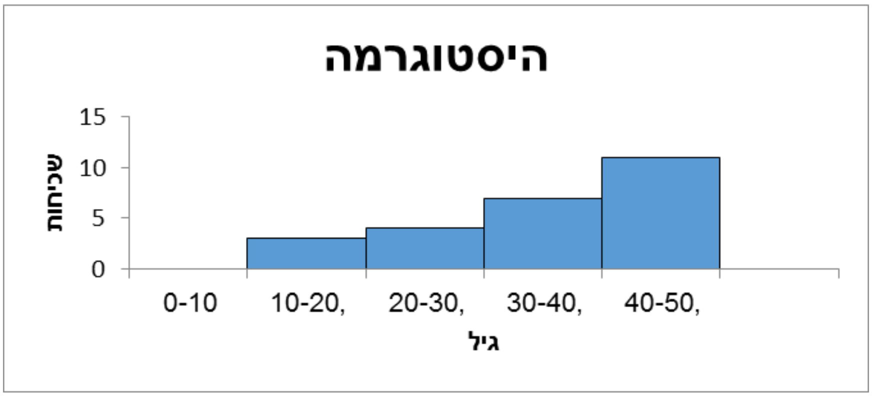
נסמן את האופציה:

☑ Chart output

הערה: במידה ונרצה להציג גם גרף שכיחות מצטברת , נסמן גם את האופציה Cumulative percentage



לסיום, יש לסדר ולעצב את הגרף הכותרות והצירים ונקבל:



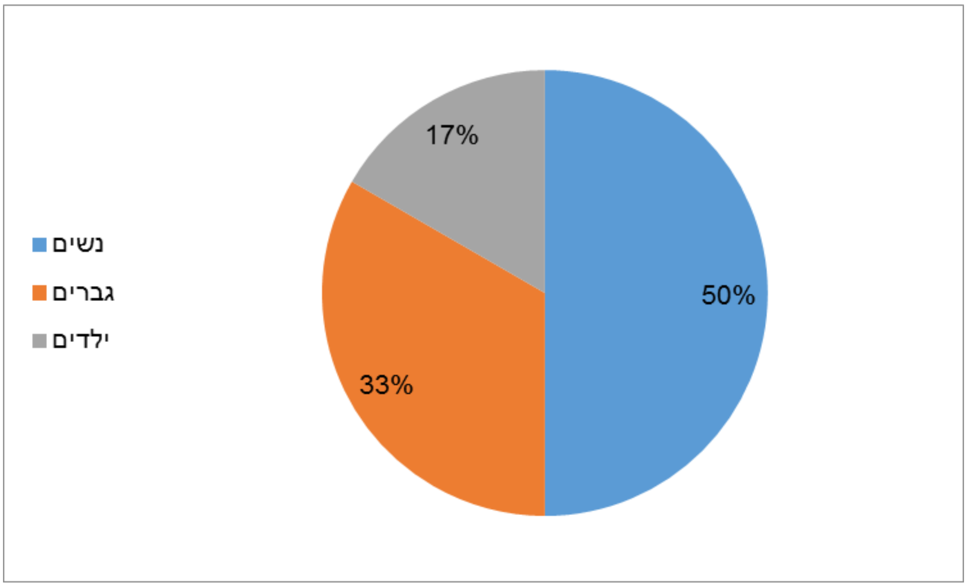
בניית דיאגרמת עיגול / עוגה באקסל

דוגמא:

להלן התפלגות האוכלוסייה בעיר מסוימת. יש לבנות דיאגרמת עוגה/ עיגול באקסל.

נשים	50%
גברים	33%
ילדים	17%

1. בתחילה יש לסמן את הטור עם האחוזים הרלוונטיים.
2. הוספה ← תרשים עוגה
3. לסיום יש לערוך ולסדר כותרות ונקבל את התרשים הבא



חישוב מדדים סטטיסטיים באמצעות אקסל

מדדי מרכז

ממוצע חשבונאי (פונקציית Average)

ממוצע חשבונאי (אריתמטי) – סכום התוצאות מחולק ב n

$$\bar{X}_A = \frac{X_1 + X_2 + ...X_n}{n}$$

כאשר:

$\bar{X}_A$ - ממוצע חשבונאי או אריתמטי

X1…Xn- תוצאה 1 עד תוצאה n

n- מספר התוצאות

הערה חשובה!!!

שימו לב שפונקציית Average באקסל מניחה משקל שווה לכל תוצאה. ולכן חישוב זה מתאים לממוצע כפי שחושב ברגרסיה לינארית (כאשר השכיחות לכל תוצאה היא 1) ולא כאשר השכיחויות או המשקל של כל תוצאה שונה.

דוגמא:

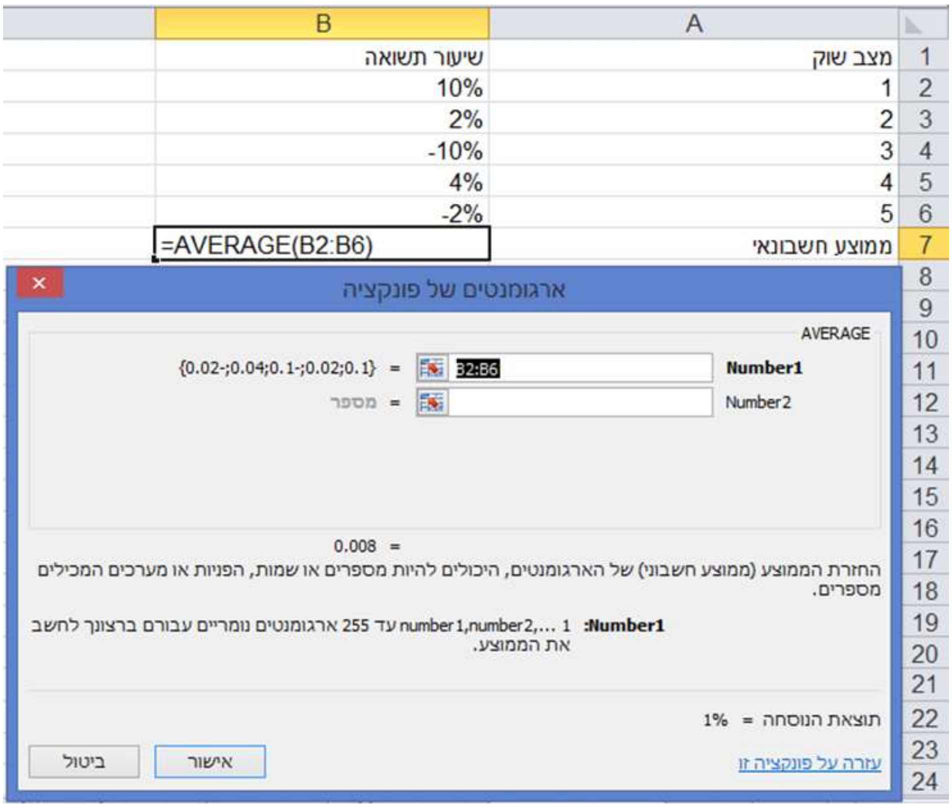
להלן שיעור התשואה החודשי הצפוי ב 5 מצבי שוק שונים (הסתברות לכל מצב שווה).

מצב שוק	1	2	3	4	5
שיעור תשואה	10%	2%	-10%	4%	-2%

חשבו ממוצע חשבונאי.

פתרון

יש לבחור בפונקציה Average ולהזין את הנתונים כפי שמוצגים בתרשים מטה.



בתוצאת הנוסחה ניתן לראות שנקבל שהממוצע שווה ל 1%.

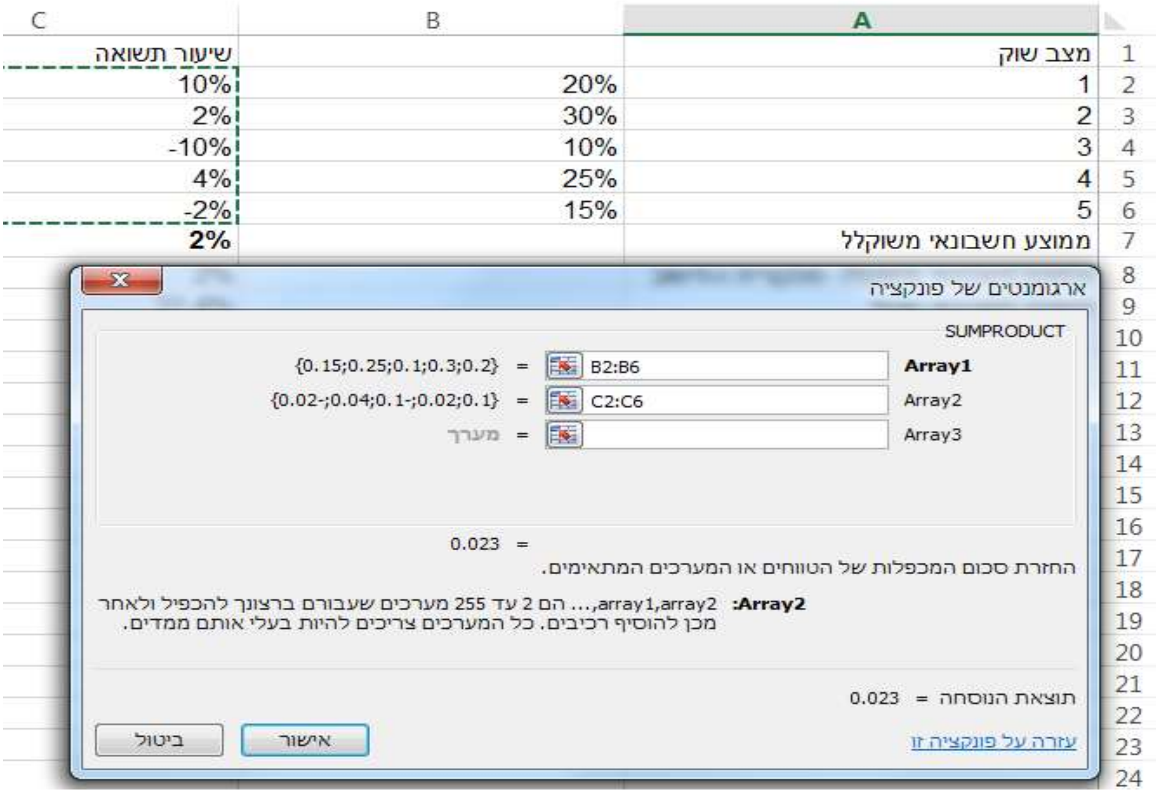
ממוצע משוקלל SUMPRODUCT

כאשר המשקל או ההסתברות לכל תוצאה שונה נשתמש בממוצע משוקלל.

דוגמא: להלן שיעור התשואה החודשי הצפוי בחמישה מצבי שוק שונים.

מצב שוק	1	2	3	4	5
הסתברות	20%	30%	10%	25%	15%
שיעור תשואה	10%	2%	-10%	4%	-2%

כאשר ההסתברויות או משקל כל תוצאה שונה, יש לבחור בפונקציית Sumproduct ולהזין את הנתונים כפי שמוצגים בתרשים מטה.



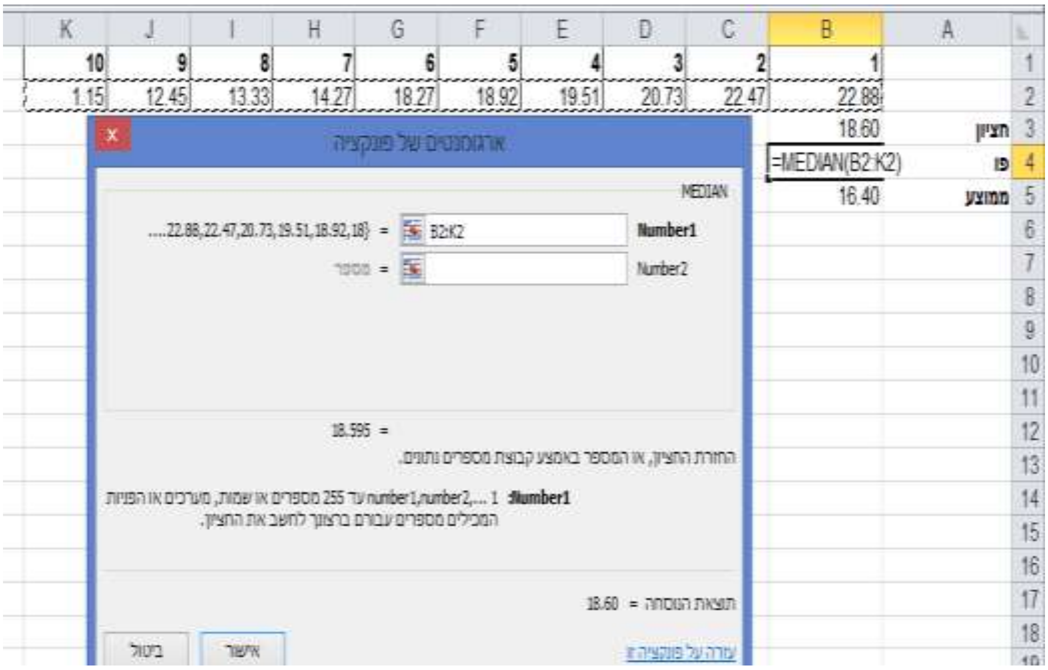
תוצאה: ממוצע חשבונאי משוקלל של 2% חודשי.

חציון Median

להלן שיעור תשואה שנתי ממוצע של 10 מניות.

1	2	3	4	5	6	7	8	9	10
22.88	22.47	20.73	19.51	18.92	18.27	14.27	13.33	12.45	1.15

על מנת לחשב את האיבר האמצעי נשתמש בפונקציית MEDIAN ונזין את הנתונים כמוצג בטבלה מטה:



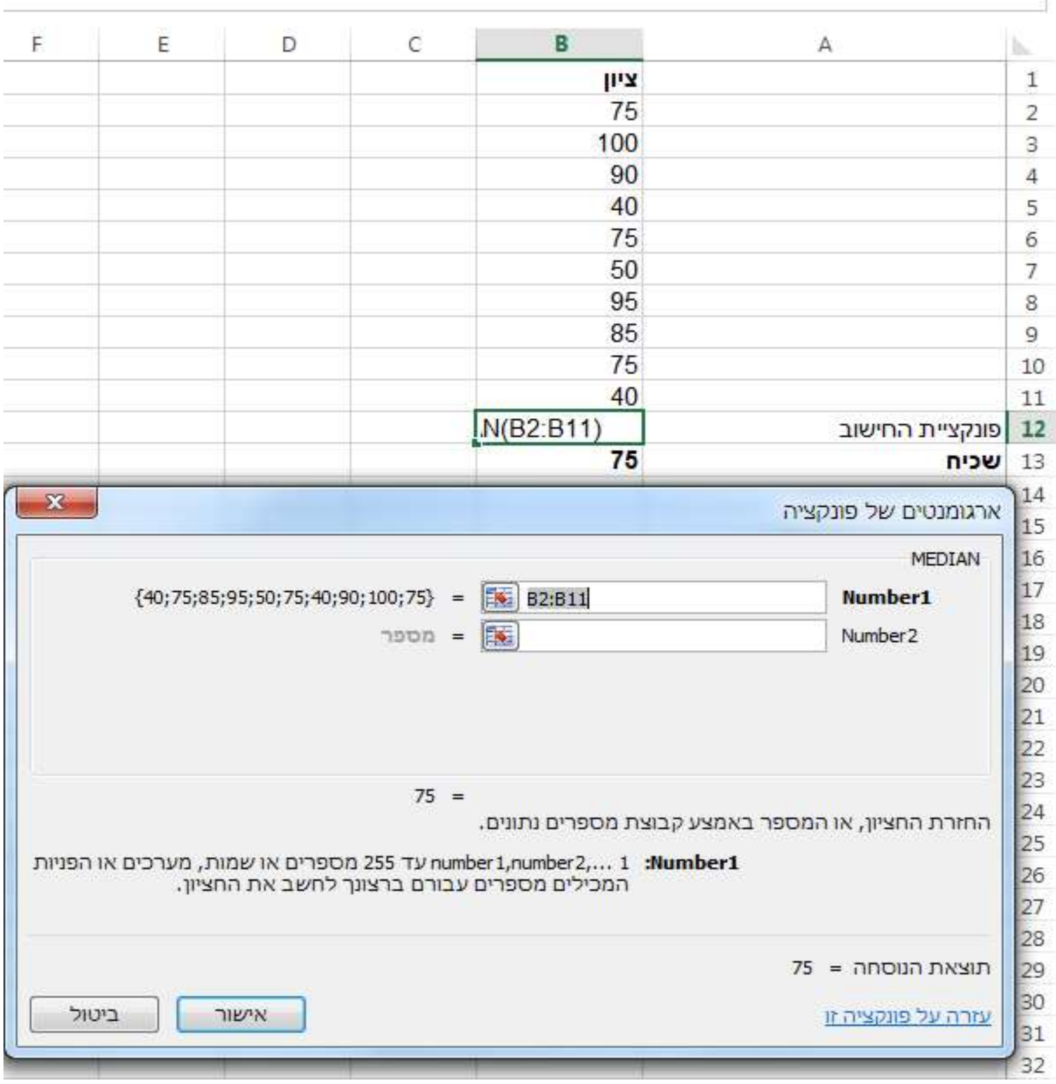
בתוצאת הנוסחה שנקבל שהחציון שווה ל 18.60.

שכיח MODE

שכיח הוא הערך המופיע מספר פעמים הרב ביותר בסדרת נתונים.

דוגמת חישוב באקסל:

להלן ציוני סטודנטים במבחן. על מנת למצוא את הציון שאותו קיבלו מספר הרב ביותר של הסטודנטים יש להזין את הנתונים ולהשתמש בפונקציית MODE:



בתוצאת הנוסחה התקבל שהציון השכיח שווה ל 75.

FREQUENCY - הצגת התפלגות השכיחות באקסל

נשתמש בפונקציה זו כאשר נרצה לחשב את השכיחות בטווחים שונים.

דוגמא: להלן מספר המכירות בחודש של 10 מוצרים שונים.

	מכירות
1	2,200
2	5,100
3	400
4	1,900
5	600
6	200
7	950
8	4,000
9	15,000
10	16,000

מהי שכיחות המוצרים שהמכירות שלהם עד ל 1,000, בין 1,000 -5,000, בין 5,000- 10,000 ומעל 10,000?

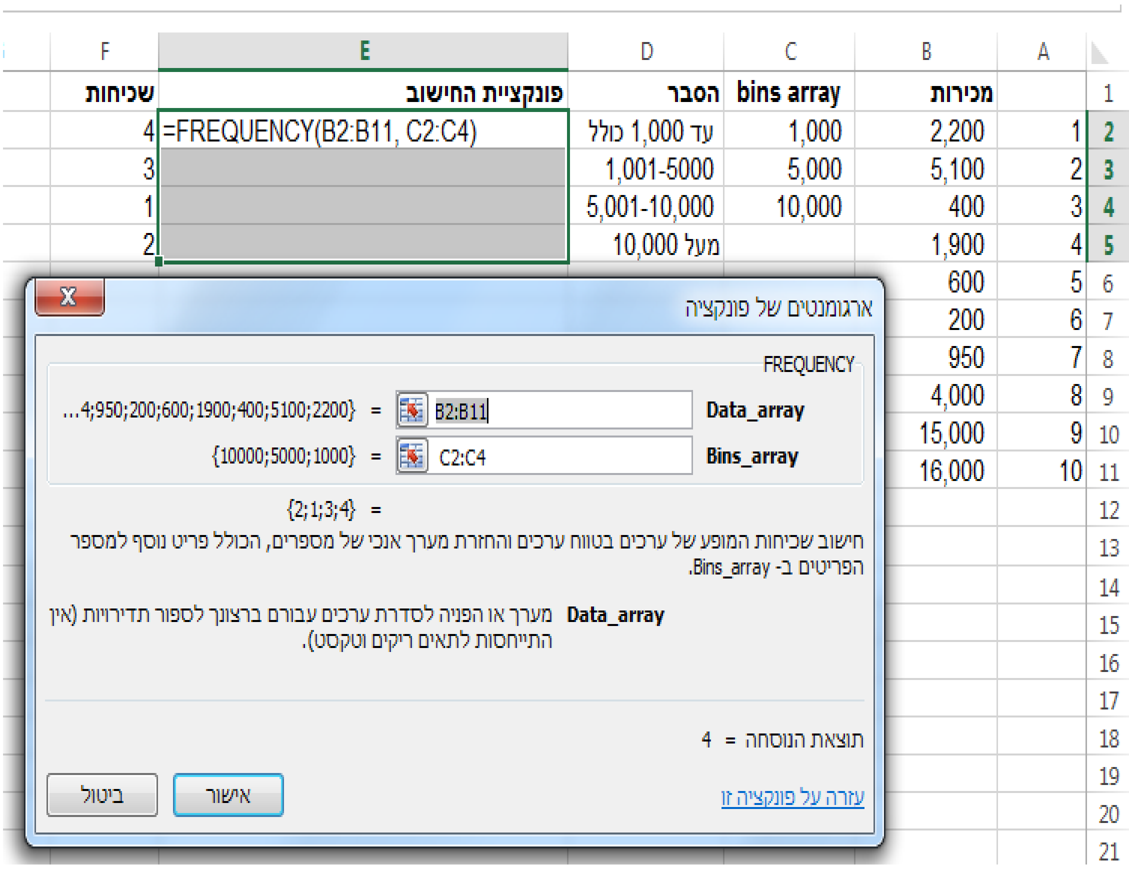
פתרון באקסל:

נשתמש בפונקציית Frequency, **יש לסמן את מספר התאים כמספר הטווחים שנדרשים לחישוב.** ונזין את הנתונים כמוצב בטבלה מטה, כאשר:

Data_array- טווח הנתונים לבדיקה.(נתונים גולמיים)

Bins_array- בעמודה זו נזין את הטווחים לפיהם נרצה לבדוק את שכיחות הכמות במלאי.

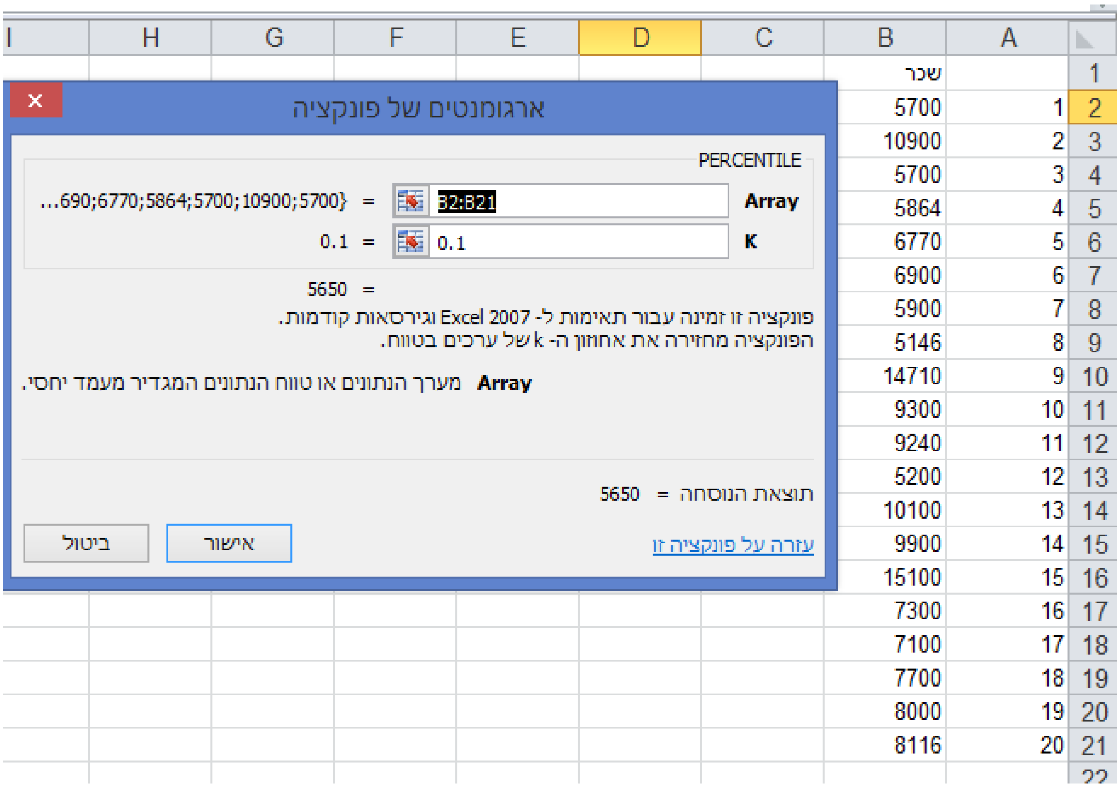
לסיום יש ללחוץ Ctrl+ Shift+ Enter



מהתוצאות ניתן לראות ש 4 מוצרים מוכרים מתחת ל 1,000 יחידות בחודש, 3 מוצרים בין 1,000 ל 5,000 , מוצר אחד בין 5,000 ל 10,000 ו 2 מוצרים מוכרים מעל 10,000 יחידות בחודש.

אחוזונים Percentile

על מנת לחשב באקסל אחוזון יש לבחור בפונקציה percentile ולהזין את הנתונים כמוצג מטה. במידה ונרצה לחשב לדוגמא אחוזון 10 ב K נרשום 10%, כלומר 0.1



במקרה זה קיבלנו שהאחוזון העשר שווה ל 5650. כלומר 90% מרוויחים מעל סכום זה ו 10% מתחת לסכום זה.

מדדי פיזור

סטיית תקן (STDEVP)[6]

$$S_X = \sqrt{\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+.....(X_n-\bar{X})^2}{n}}$$ - סטיית התקן עבור n תצפיות

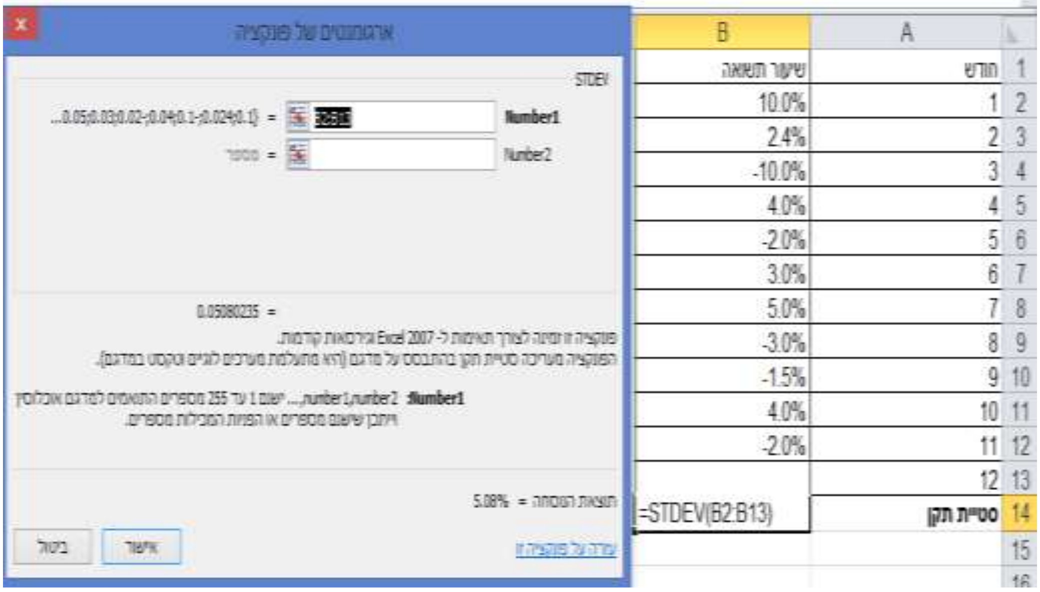
שימו לב שבחישוב סטיית התקן באקסל, כמו בפונקציית Average, יש הנחה של משקל שווה לכל תוצאה (לכל תצפית שכיחות 1) .

דוגמא:

להלן שיעורי התשואה של מניה מסוימת ב-12 החודשים האחרונים. חשבו את סטיית התקן.

פתרון

נבחר בפונקציית STDEVP ונזין את הנתונים כפי שמוצגים בתרשים מטה.



תוצאת הנוסחה: 5.08%

[6] כאשר מעוניינים להסיק מתוצאות מדגם לתוצאות כלל האוכלוסייה נהוג לחלק את ריבועי ההפרשים מהממוצע ב n-1 בנוסחת סטיית התקן (וזאת כדי לקבל מדד בלתי מוטה). במצבים אלו (כאשר מעוניינים לחלק ב n-1) יש לבחור בפונקציה STDEV. לעומת זאת כאשר מעוניינים לחלק את ריבועי ההפרשים ב- n בנוסחת סטיית התקן, יש להשתמש בפונקציה STDEVP. בכל מקרה, ההבדל בין התוצאות של שתי הפונקציות זניח יחסית כאשר מדובר במדגם גדול. בספר זה נדגים את חישובי סטיית התקן באקסל באמצעות הפונקציה STDEVP (כלומר חלוקה ב- n).

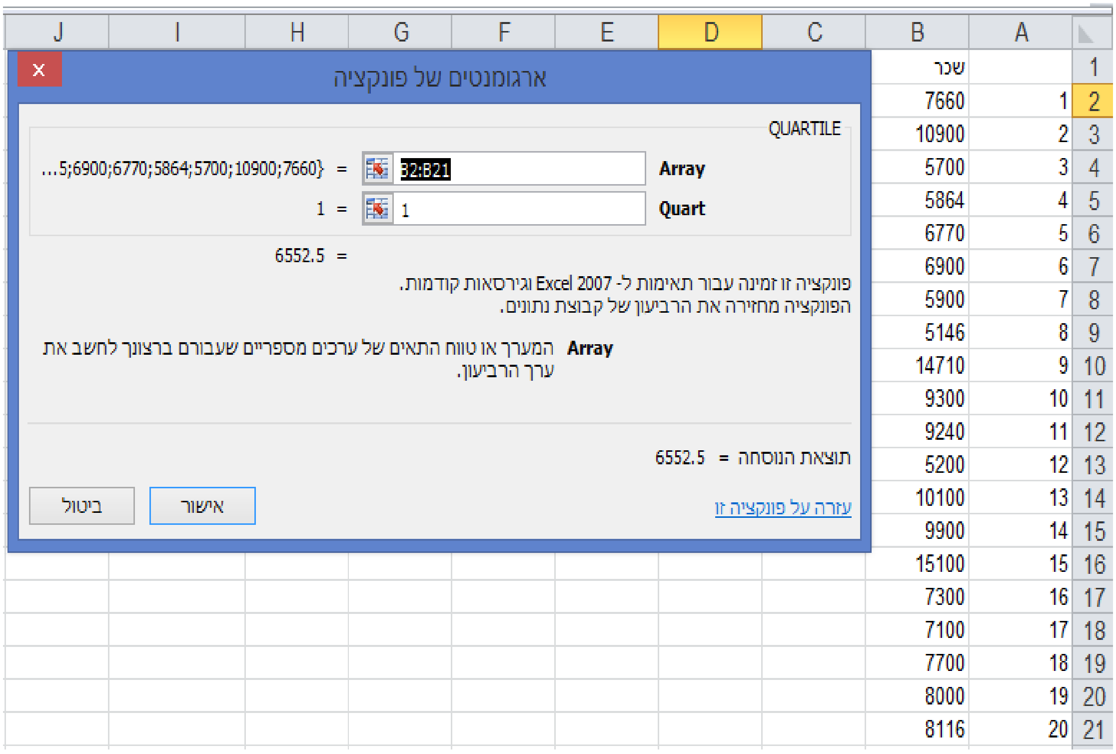
תחום בין-רבעוני וחישוב רבעונים (Quartile)

תחום בין רבעוני שווה לרבעון השלישי פחות הרבעון הראשון.

כלומר, תחום בין רבעוני = Q_3-Q_1. התחום הבין רבעוני מתאר את מחצית המקרים שבמרכז ההתפלגות. הוא ההפרש בין הרבעון השלישי לרבעון הראשון.

דוגמא:

נתונים המשכורות החודשיות של 20 עובדים. על מנת לחשב באקסל רבעונים יש לבחור בפונקציה quartile ולהזין את הנתונים כמוצג מטה. במידה ונרצה לחשב רבעון ראשון נזין 1 ב Quart. במידה ונרצה לדוגמא לחשב רבעון 3 , נזין Quart 3.

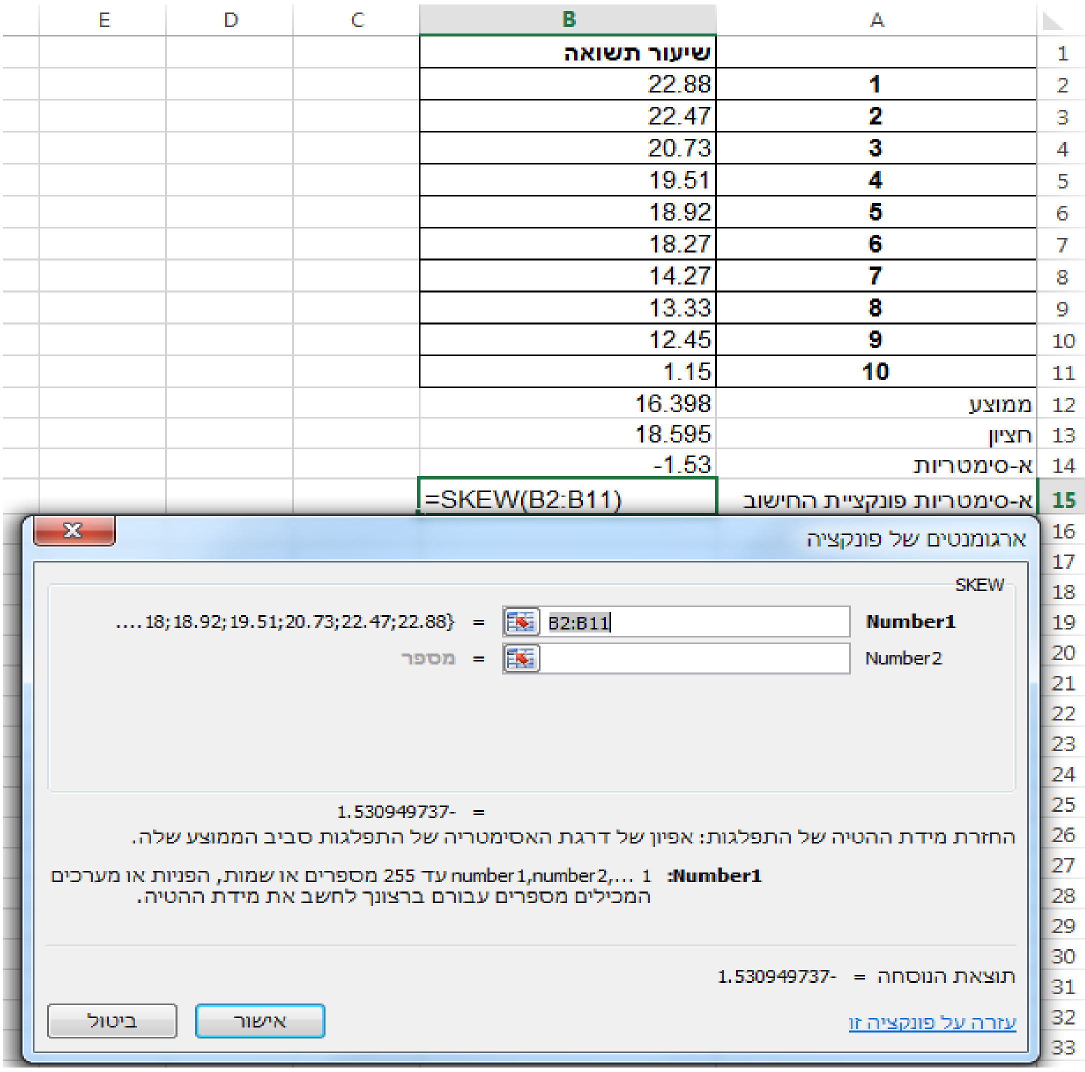


בדוגמא זו חושב הרבעון הראשון והתקבלה התוצאה של 6552.5. כלומר 25% מהעובדים מקבלים שכר נמוך יותר מ 6552.5 ו- 75% מקבלים מעל שכר זה.

צורת ההתפלגות

אסימטריות SKEW

להלן נתוני שיעורי תשואה שנתיים של עשר קרנות נאמנות. על מנת לחשב את אופן התפלגות התשואות (א-סימטריות שלילית או חיובית) , יש להזין את הנתונים באקסל ולבחור בפונקציית SKEW. ערך SKEW חיובי מעיד על התפלגות אסמיטרית חיובית וערך שלילי מעיד על התפלגות אסימטריות שלילית.



קיבלנו אסימטריות שלילית של 1.53-. נשים לב שהממוצע (16.398) נמוך מהחציון (18.595) וזאת משום שהממוצע ירד עקב המצאה של תוצאה קיצונית נמוכה (תשואה של 1.15% בקרן מספר 10).

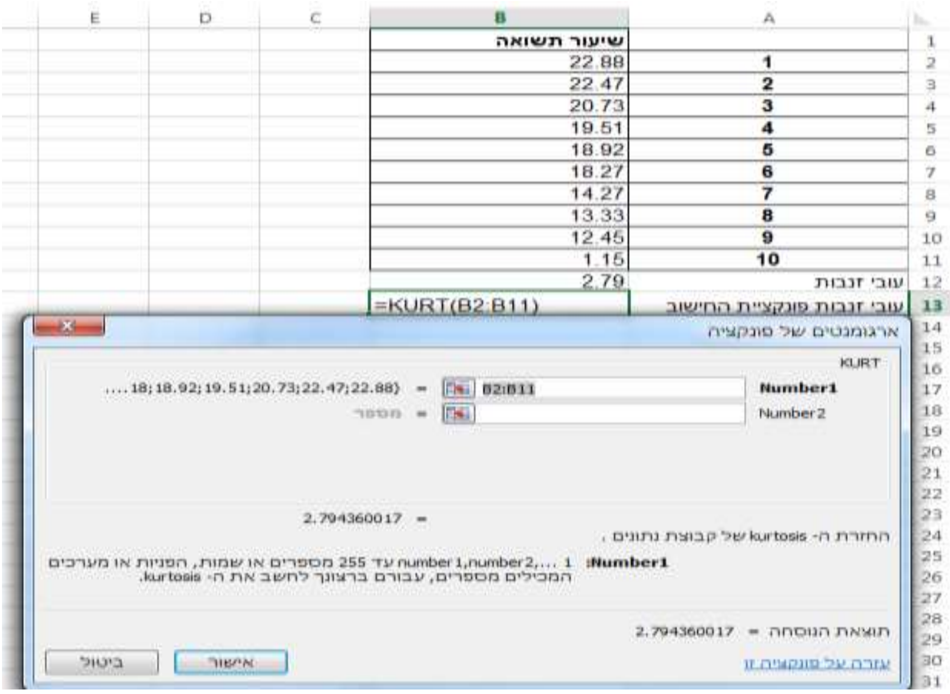
	שיעור תשואה
1	22.88
2	22.47
3	20.73
4	19.51
5	18.92
6	18.27
7	14.27
8	13.33
9	12.45
10	40.35
ממוצע	20.318
חציון	19.215
א-סימטריות	1.95
א-סימטריות פונקציית החישוב	1.95

לעומת זאת ניתן לראות כאשר יש תוצאה קיצונית חיובית, נקבל אסימטריה חיובית וכעת הממוצע גדל והוא גדול יותר מהחציון.

<u>עובי זנבות (קצוות) KURT</u>

עובי זנבות מציין האם קצוות ההתפלגות / זנבות עבים יותר או דקים יותר לעומת ההתפלגות הנורמלית. אם התוצאה שנקבל היא חיובית המשמעות היא שהזנבות דקים יותר (דהיינו ההתפלגות תלולה יותר) ואם התוצאה שלילית המשמעות היא שהזנבות עבים יחסית (דהיינו ההתפלגות שטוחה יחסית).

להלן נתוני שיעורי תשואה שנתיים של עשר קרנות נאמנות. על מנת לחשב את עובי זנבות , יש להזין את הנתונים באקסל ולבחור בפונקציית KURT:



עובי זנבות חיובי של 2.79 מעיד על התפלגות בעלת זנבות דקים יותר משל ההתפלגות הנורמלית.

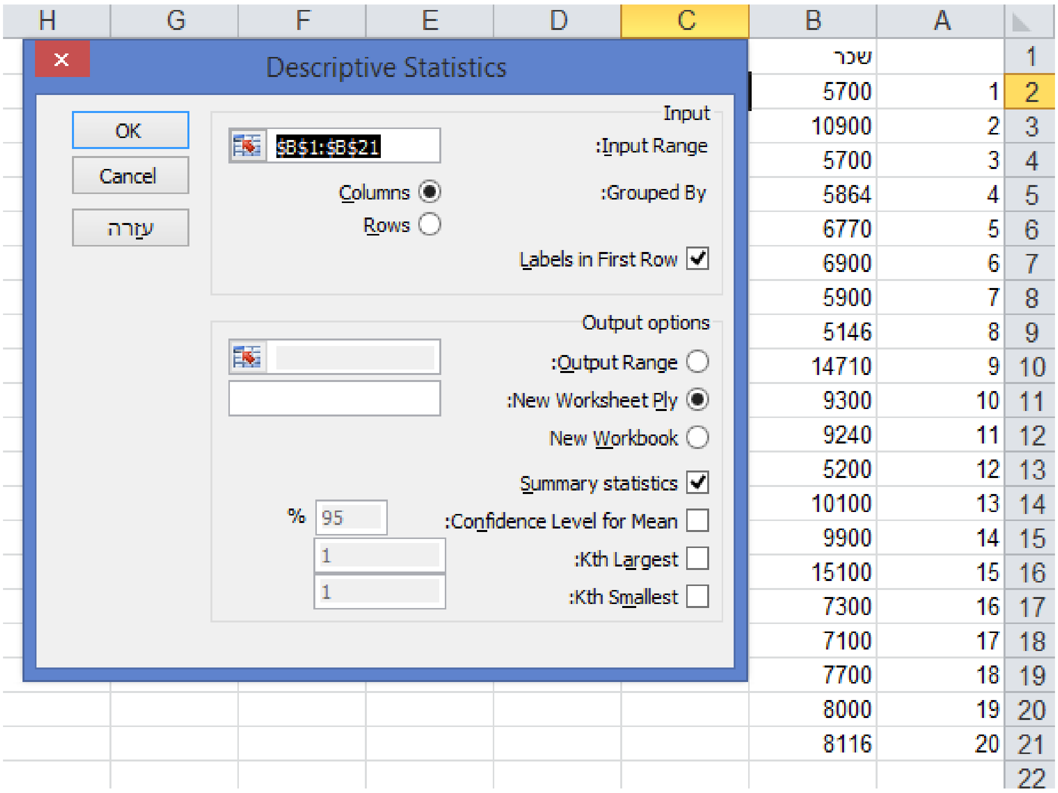
<u>סיכום מדדים באמצעות Descriptive Statistics</u>

להלן נתוני המשכורות של 20 עובדים.

על מנת לחשב את כל המדדים הסטטיסטיים באקסל יחד יש לעבוד על פי השלבים הבאים:

1.נתונים ← Data analysis ← Descriptive statistics

2. בחלון שנפתח נזין את הנתונים כדלקמן:



נלחץ OK ונקבל את התוצאות כדלקמן:

	שכר	
ממוצע	8232.3	Mean
	635.934	Standard Error
חציון	7500	Median
שכיח	5700	Mode
סטיית תקן	2843.984	Standard Deviation
שונות	8088242	Sample Variance
עובי זנבות	1.305013	Kurtosis
אסימטריות	1.282474	Skewness
טווח	9954	Range
מינימום	5146	Minimum
מקסימום	15100	Maximum
סה"כ	164646	Sum
מס תצפיות	20	Count

רגרסיה לינארית פשוטה

רגרסיה לינארית משמשת בעיקר למטרות חיזוי. כלומר, חיזוי תוצאות משתנה אחד (משתנה תלוי או מוסבר) על סמך ערכי משתנה אחר (משתנה בלתי תלוי או מסביר). המקרה הפשוט ביותר הוא כאשר יש לנו משתנה מסביר אחד ומשתנה מוסבר אחד.

להלן מודל קו הרגרסיה הלינארי הפשוט:

[7] $\hat{Y}_i = a + bX_i$

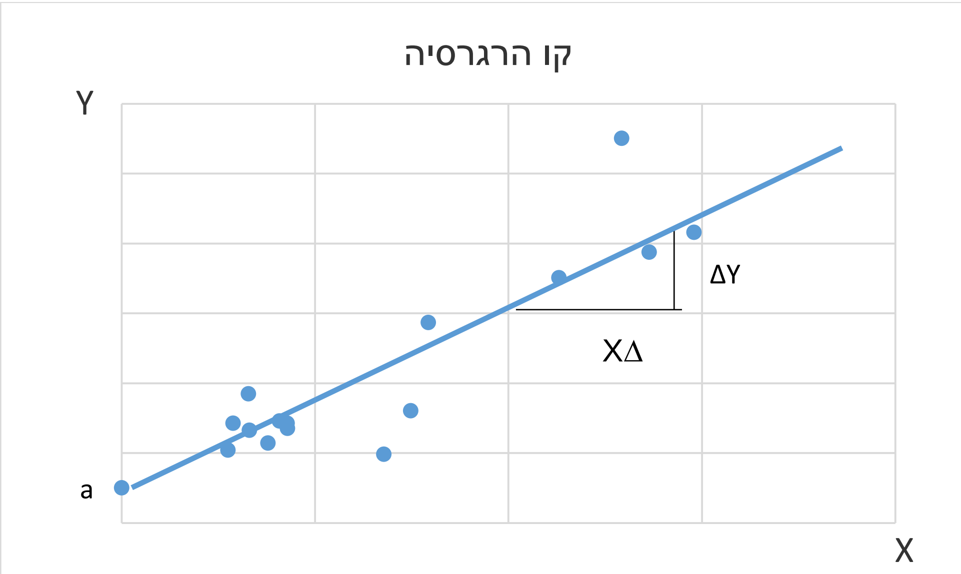
כאשר:

a - חותך. הערך הממוצע של Y כאשר X שווה לאפס.

b - מקדם הרגרסיה. מייצג את השינוי הצפוי ליחידת Y , עם שינוי ביחידת X.

$\hat{Y}_i$ - הערך המנובא של Y עבור תצפית i

X_i - הערך של X עבור תצפית i



b , מקדם הרגרסיה, מייצג את השיפוע של קו הרגרסיה, עד כמה שינוי ב –X משפיע על Y. הקורלציה את הדיוק שלו או עד כמה הסטיות קרובות או רחוקות מקו הרגרסיה. בגרף מעלה מוצג קו של רגרסיה לינארית פשוטה עם שיפוע חיובי המעיד על קשר חיובי בין המשתנים. כמובן שהקשר יכול להיות גם שלילי ואז קו הרגרסיה יהיה בעל שיפוע שלילי, דהיינו קו יורד כלפי מטה.

על מנת למצוא את קו הרגרסיה נשתמש ב "שיטת הרבועים הפחותים" אשר מביאה למינימום את סכום הסטיות (בריבוע) של קו הרגרסיה.

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

כאשר

$\hat{Y}_i$ - הערך המנובא של Y עבור תצפית i

Y_i- הערך בפועל של Y עבור תצפית i

<u>דוגמא למציאת קו הרגרסיה באמצעות אקסל:</u>

על מנת לבדוק האם יש קשר בין הוצאות השיווק לרווח השנתי ברשת למכירת בגדים, נדגמו 15 חנויות שונות. להלן הנתונים:

הוצאות שיווק	רווח לשנה
795	9332
800	8263
862	11000
842	9644
842	10260
881	12073
886	11922
950	12520
842	10360
844	10020
845	10280
852	10355
862	11012
876	11013
876	11491
881	12072

על פי קו הרגרסיה, מהי תחזית הרווח לחנות עם הוצאות שיווק בסך 800 ש"ח?

פתרון באקסל:

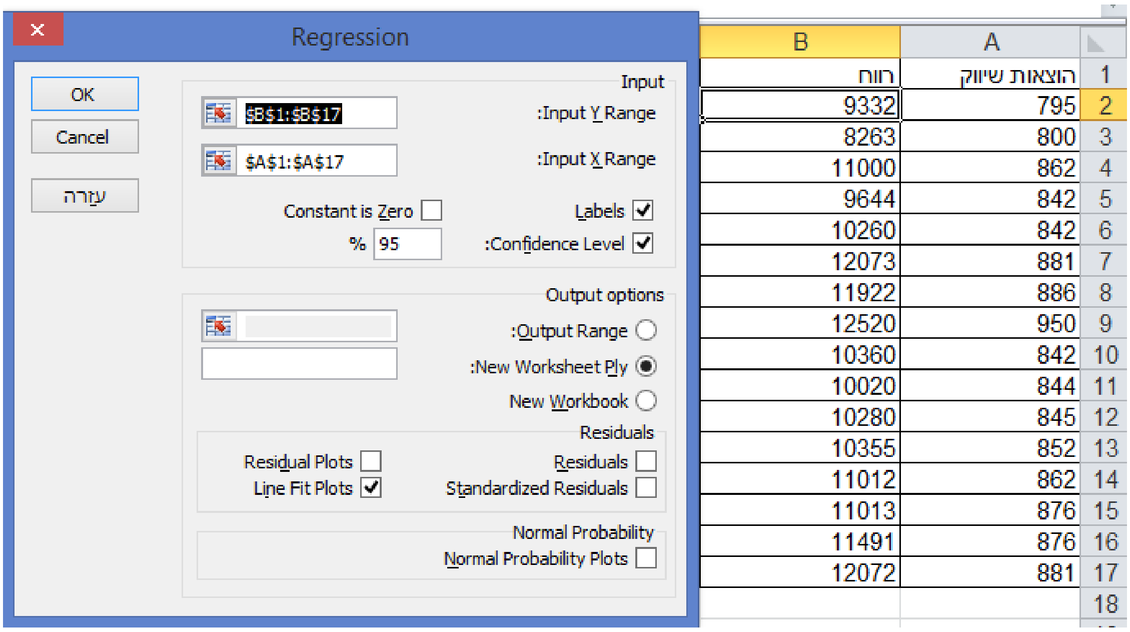
נתונים ← Data Analysis ←regression

(במידה ו Data Analysis אינו פעיל באקסל יש לבחור באפשרויות ← תוספות ← להוסיף את Analysis Tool Pack לתוספות הפעילות).

יש להזין את הנתונים כמוצג בחלון מטה:

Input Y range: משתנה תלוי

Input X range: משתנה בלתי תלוי



להלן תוצאות ניתוח הרגרסיה גרפית ובטבלה:

						SUMMARY OUTPUT
					Regression Statistics	
					0.910704	Multiple R
					0.829382	R Square
					0.817195	Adjusted R Square
					487.0635	Standard Error
					16	Observations
						ANOVA
	Significance F	*F*	*MS*	*SS*	*df*	
	9.56074E-07	68.05457	16144645	16144645	1	Regression
			237230.9	3321232	14	Residual
				19465877	15	Total
Upper 95%	*Lower 95%*	*P-value*	*t Stat*	*andard Err*	*Coefficients*	
-7556.095976	-20416.6427	0.000365	-4.66509	2998.095	-13986.4	Intercept
36.2695394	21.30164721	9.56E-07	8.249519	3.489366	28.78559	הוצאות שיווק

ניתוח פלט הרגרסיה

- **Multiple R-** מקדם המתאם
- **R Square-** מקדם ההסבר, מידת ההסבר של המשתנה המוסבר על ידי המשתנה המסביר.
- **Adjusted R square-** מקדם ההסבר המתוקנן, לוקח בחשבון את מספר המשתנים. בניגוד למקדם ההסבר שתמיד יגדל כאשר נוסיף עוד משתנים מסבירים,

מקדם ההסבר המתוקנן יקטן כאשר נוסיף עוד משתנים מסבירים שאינם מועילים למודל.

- **Standard Error**- אומדן לסטיית התקן של ההפרעה/ הסטיות.
- **Observations**- מספר התצפיות במדגם.
- **F**- סטטיסטי שבאמצעותו בוחנים האם ניתן לדחות או לקבל את השערות האפס.

$H0: b_i=0$

$H1: b\neq 0$

ערך F רחוק מאפס משקף מצב בו ניתן לדחות את השערת האפס לעומת ערך F קרוב לאפס המשקף מצב שיביא לאי דחיית השערת האפס.

- **Significance F**- ההסתברות לטעות באם נדחה את השערת האפס. לרוב נקבל סיכוי של עד 5% .
- **Regression ss**- השונות המוסברת
- **Residual SS**- השונות הלא מוסברת
- **Total SS**- סך השונות

להלן מספר כללים :

1. שונות מוסברת + שונות לא מוסברת = סך השונות
2. $R^2 = \frac{Regression\_SS}{Total\_SS}$, מקדם ההסבר
3. $F = \frac{Regression\_MS}{Residual\_MS}$

טבלת מקדמי המשתנים:

- **Coefficient**- ערכי האומדנים (חותך ומקדם)
- **Standard Error**- סטיות התקן המדגמיות של כל אחד מאומדנים
- **T stat**- סטטיסטי בעזרתו בוחנים אם המשתנה המסביר מתאים לדחיית השערת האפס. ערך T רחוק מהאפס מעיד על דחיית השערת האפס (משתנה הנבחר מסביר את Y).
- **P-Value**- בדומה ל F משקף את הסתברות לטעות.
- כאשר מדובר ברגרסיה למשתנה אחד , F Significance ו- P significance יהיו זהים.

- **Upper 95%, Lower 95%**- רווח סמך סימטרי בו מצטבר 95% משטח ההתפלגות t.

מתוצאות הניתוח נקבל את ערכי המקדמים (coefficients) כדלקמן:

-13986.4 = a

28.78 =b

ולכן משוואת התחזית היא:

$$\hat{Y}_i = -13986.4 + 28.78X_i$$

כאשר $\hat{Y}_i$ מייצג את הערך החזוי או הצפוי של Y עבור תצפית i, לעומת Yi אשר מייצג את הערך בפועל.

ניתן לראות שהרגרסיה מובהקת כיוון ש F significance קטן יותר מרמת מובהקות של 5% שהוגדרה . באותו אופן הקשר בין X ל Y הינו קשר מובהק (p value=0 עבור b).

משמעות R Square מציג את יכולות הניבוי של X להסביר את Y. ובמקרה שלנו 82.9% מהתנודות ב –X מסבירות את התנודות במשתנה התלוי, Y.

השיפוע שווה ל 28.78 ולכן הגדלת הוצאות השיווק ב 1 ₪ תגרום להגדלת הרווח הצפוי ב 28.78 ₪.

החותך שווה למינוס 13986.4 , ולכן ללא הוצאות שיווק (x=0), הרווח הוא שלילי ושווה למינוס 13986.4.

ולפיכך, תחזית הרווח לחנות בעלת הוצאות שיווק שנתיות בסך 800 ₪ שווה ל 9037.6₪ :

$$\hat{Y}_i = -13986.4 + 28.78 \bullet 800 = 9037.6$$

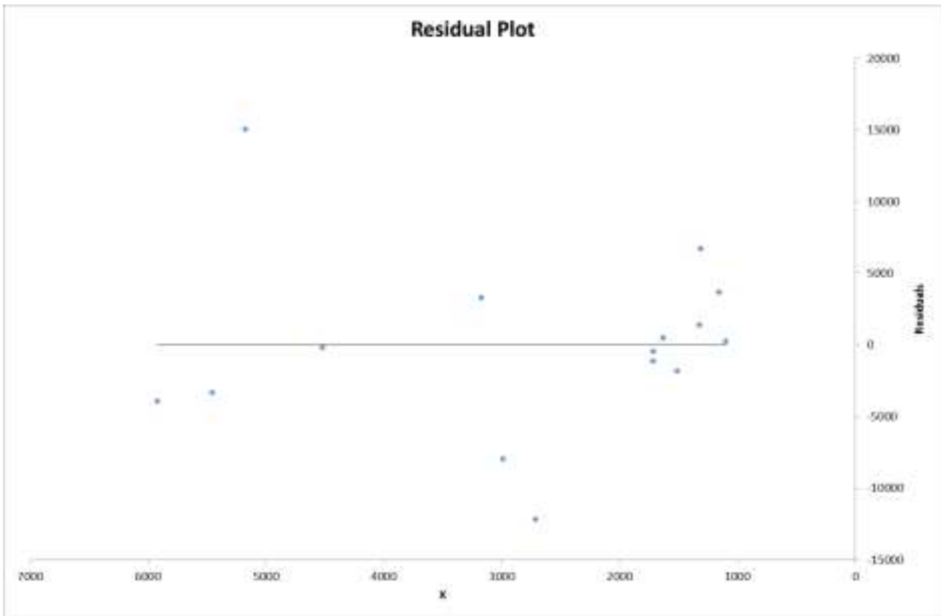
בעיות אפשריות ברגרסיה

1. **מתאם סדרתי**

אחת ההנחות ברגרסיה היא שקיימת אי תלות בין ההפרעות, כלומר בין הסטיות מקו הרגרסיה. בסדרות עיתיות, כלומר סדרות על פני זמן, בעיה זו עלולה להופיע כיוון שקיימים משתנים מסבירים הגדלים או קטנים על פני זמן. לדוגמא: המחיר או המדד של אתמול מסביר את המחיר או המדד של היום.

שיטות לאבחון מתאם סדרתי

- **פלט סטיות Residual Plot**



בדוגמא זו, לפי הגרף מעלה, לא ניתן לראות הוכחה לריכוז שאריות מצידו האחד של גרף האפס או מצידו השני. במידה וכן היה ניתן לראות פיזור שאינו מיקרי , זו הייתה הוכחה גרפית לקיום מתאם סדרתי.

- **מבחן Durbin Watson (DW)**

מבחן נפוץ לבדיקת מתאם סדרתי הינו מבחן דרבין ווטסון (DW). באקסל ניתן לבצע מבחן זה באמצעות תוספת PH stat.

	Durbin-Watson Calculations
7323387.8782	Sum of Squared Difference of Residuals
3321232.2535	Sum of Squared Residuals
2.2050	**Durbin-Watson Statistic**

תוצאת מבחן DW בדוגמא מעלה d=2.2050

תחום הגדרה של תוצאות DW: $0 \le d \le 4$.

על מנת להחליט האם הערך שהתקבל מעיד על בעיית מתאם יש לבדוק בטבלת ההתפלגות של d (ראו מטה). כאשר k מייצג את מספר המשתנים המסבירים (במקרה של רגרסיה פשוטה k=1) ו N את מספר התצפיות.

כללי החלטה לבדיקת מתאם סדרתי

- $d_U \le d \le 4-d_U$ - לא קיים מתאם סדרתי
- $d \ge 4-d_L$ - קיים מתאם סדרתי שלילי
- $d \le d_L$ - קיים מתאם סדרתי חיובי
- $4-d_U < d < 4-d_L$ - אין תשובה חד משמעית
- $d_L < d < d_U$ - אין תשובה חד משמעית

טבלת ערכים קריטיים של מבחן Durbin Watson

N	K = 1		K = 2		K = 3		K = 4		K = 5		K = 6		K = 7	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.81	1.75	0.69	1.97	0.56	2.21	0.45	2.47	0.34	2.73
16	1.11	1.37	0.98	1.54	0.86	1.73	0.73	1.93	0.62	2.15	0.50	2.39	0.40	2.62
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.66	2.10	0.55	2.32	0.45	2.54
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06	0.60	2.26	0.50	2.46
19	1.18	1.40	1.07	1.53	0.97	1.68	0.86	1.85	0.75	2.02	0.65	2.21	0.55	2.40
20	1.20	1.41	1.10	1.54	1.00	1.68	0.89	1.83	0.79	1.99	0.69	2.16	0.60	2.34
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96	0.73	2.12	0.64	2.29
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94	0.77	2.09	0.68	2.25
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92	0.80	2.06	0.72	2.21
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90	0.84	2.04	0.75	2.17
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89	0.87	2.01	0.78	2.14
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88	0.90	1.99	0.82	2.12
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.00	1.86	0.93	1.97	0.85	2.09
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85	0.95	1.96	0.87	2.07
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84	0.98	1.94	0.90	2.05
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83	1.00	1.93	0.93	2.03
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83	1.02	1.92	0.95	2.02
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82	1.04	1.91	0.97	2.00
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81	1.06	1.90	0.99	1.99
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.14	1.81	1.08	1.89	1.02	1.98
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80	1.10	1.88	1.03	1.97
36	1.41	1.52	1.35	1.59	1.30	1.65	1.24	1.73	1.18	1.80	1.11	1.88	1.05	1.96
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80	1.13	1.87	1.07	1.95
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.20	1.79	1.15	1.86	1.09	1.94
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79	1.16	1.86	1.10	1.93
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79	1.18	1.85	1.12	1.93
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78	1.24	1.84	1.19	1.90
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77	1.29	1.82	1.25	1.88
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.37	1.77	1.33	1.81	1.29	1.86
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77	1.37	1.81	1.34	1.85
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77	1.40	1.81	1.37	1.84
70	1.58	1.64	1.55	1.67	1.53	1.70	1.49	1.74	1.46	1.77	1.43	1.80	1.40	1.84
75	1.60	1.65	1.57	1.68	1.54	1.71	1.52	1.74	1.49	1.77	1.46	1.80	1.43	1.83
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77	1.48	1.80	1.45	1.83
85	1.62	1.67	1.60	1.70	1.58	1.72	1.55	1.75	1.53	1.77	1.50	1.80	1.47	1.83
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78	1.52	1.80	1.49	1.83
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78	1.54	1.80	1.51	1.83
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78	1.55	1.80	1.53	1.83

ניתן לראות לפי הטבלה מעלה שבמקרה של k=1 ו n=15 (בדוגמא שלנו משתנה מסביר אחד ו 16 תצפיות) ולכן du=1.11, dL=1.37.

לכן d=1.773 נמצא בתחום שבו אין מתאם סדרתי:

$$d_U \le d \le 4 - d_U$$
$$1.11 \le 2.2050 \le 4 - 1.11$$
$$1.08 \le 2.2050 \le 2.89$$

ולכן נסיק שאין מתאם סדרתי.

סיבתיות Causality

בעיה נוספת שעלולה להתקיים ברגרסיה היא בעיית סיבתיות. לעיתים לא ברור באופן חד משמעי האם X גורם ל – Y או Yגורם ל- X. "מי קודם למי? הביצה או התרנגולת? " . כאשר אנו בונים מודל אנו מניחים ששינוי במשתנה המסביר X גורם לשינוי במשתנה המוסבר Y, אך לעיתים כיוון הקשר אינו ברור. האם X גורם ל Y או שיתכן שY גורם ל- X. לדוגמא האם גובה המשכורת משפיע על המוטיבציה בעבודה או האם יתכן שגם המוטיבציה בעבודה משפיעה על גובה המשכורת? כלומר, הרבה פעמים ניתן להוכיח קשר בין המשתנים אך קשה להוכיח סיבתיות, דהיינו ששינוי ב – X הוא זה שגורם לשינוי ב-Y.[8]

[8] ישנן שיטות לבדיקת סיבתיות כגון שימוש במבחן Granger ורגרסיות lag עליהם לא נרחיב בספר זה.

רגרסיה מרובת משתנים

ברגרסיה לינארית פשוטה בדקנו כיצד משתנה אחד, משתנה בלתי תלוי או מסביר (X) משמש לניבוי משתנה שני, משתנה תלוי או מוסבר (Y). במקרים רבים, ניתן לפתח מודל טוב יותר על ידי התחשבות במספר משתנים בלתי תלויים (להגדיל את מקדם ההסבר).

משוואת רגרסיה רבת משתנים (עבור K משתנים) נראית כדלקמן:

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + ... + b_K X_{Ki}$$

כאשר :

$\hat{Y}_i$ - הערך המנובא של Y עבור תצפית i

a- החותך של Y . (התוצאה של Y כאשר כל יתר המשתנים שווים לאפס)

b_1- השינוי הצפוי ב- Y עקב שינוי ב X_1 כאשר יתר המשתנים נותרים קבועים (משתנים 1 עד K).

X_{1i} - הערך של משתנה X_1.

בשביל מה צריך יותר ממשתנה אחד?

לרוב משתנה אחד לא מסביר באופן מספק את השונות במשתנה המוסבר. כלומר, נוסיף עוד משתנים על מנת לשפר את יכולת הניבוי.

שיקולים בבחירת משתנים מנבאים

1. נרצה מתאם בין המשתנה הבלתי תלוי והמשתנה התלוי גבוה (בערך מוחלט) ככל שניתן. ככל שהמתאם גבוה יותר, המשתנה מועיל יותר בניבוי שלו.
2. רצוי מתאם נמוך ככל הניתן (בערך מוחלט) בין המשתנים הבלתי תלויים.

הסיבה היא שכאשר המתאם גבוה בין המשתנים הבלתי תלויים הוספתם פחות אפקטיבית וזאת כי בעצם יש ביניהם דמיון רב ולכן ואינם תורמים תוספת חדשה משמעותית למודל. בנוסף מתאם גבוה בין המשתנים הבלתי תלויים עלול לגרום לבעיית מולטיקולינאריות עליה נדבר בהמשך.

<u>דוגמא לרגרסיה מרובת משתנים</u>

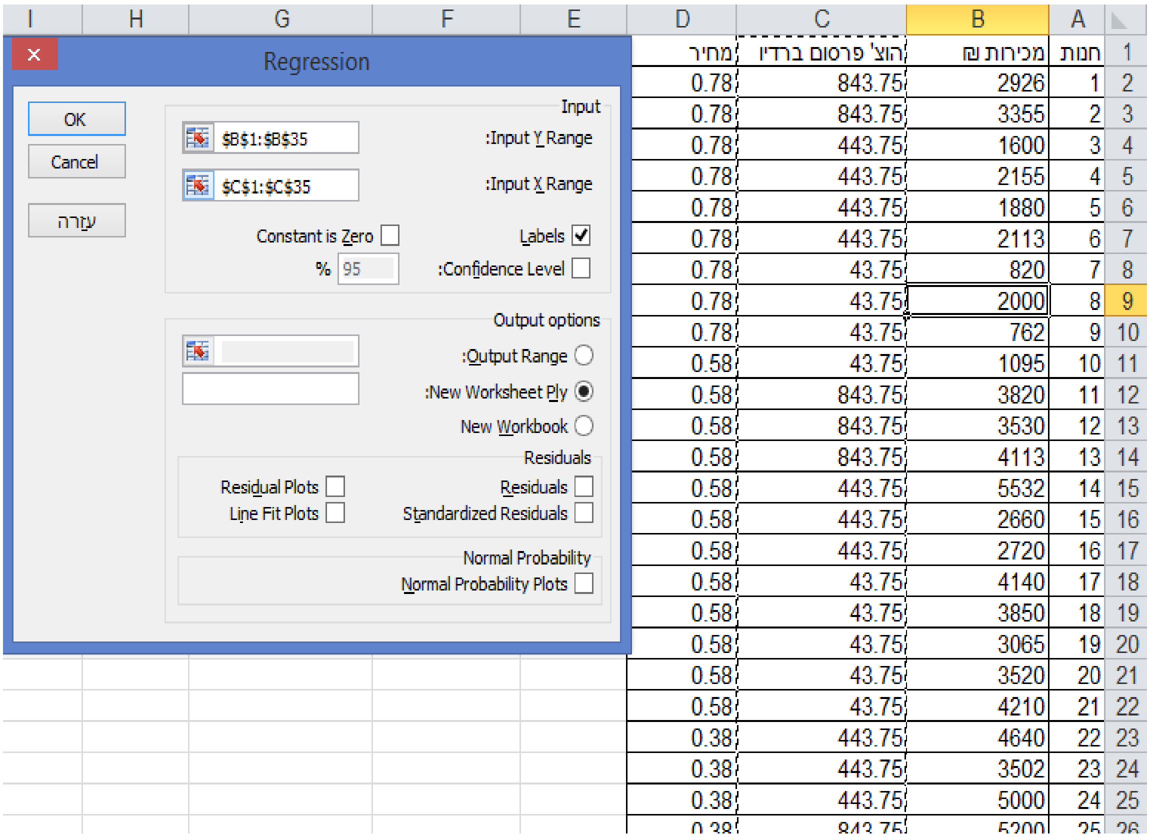
מנהל השיווק ברשת סופרמרקטים גדולה מעוניין להכניס חטיף חדש לחנויות הרשת ברחבי המדינה. המנהל מעוניין לבדוק את השפעת הוצאות עבור פרסום ברדיו והמחיר על המכירות. לשם כך החליט המנהל ראשית לערוך מחקר שוק ולדגום 34 חנויות בהן המוצר החדש נמכר לתקופת ניסיון. להלן התוצאות:

חנות	מכירות ₪	הוצ' פרסום ברדיו	מחיר
1	2926	843.75	0.78
2	3355	843.75	0.78
3	1600	443.75	0.78
4	2155	443.75	0.78
5	1880	443.75	0.78
6	2113	443.75	0.78
7	820	43.75	0.78
8	2000	43.75	0.78
9	762	43.75	0.78
10	1095	43.75	0.58
11	3820	843.75	0.58
12	3530	843.75	0.58
13	4113	843.75	0.58
14	5532	443.75	0.58
15	2660	443.75	0.58
16	2720	443.75	0.58
17	4140	43.75	0.58
18	3850	43.75	0.58
19	3065	43.75	0.58
20	3520	43.75	0.58
21	4210	43.75	0.58
22	4640	443.75	0.38
23	3502	443.75	0.38
24	5000	443.75	0.38
25	5200	843.75	0.38
26	4000	843.75	0.38
27	5020	843.75	0.38
28	1900	843.75	0.38
29	680	43.75	0.38
30	3642	43.75	0.38
31	3254	43.75	0.38
32	2298	43.75	0.38
33	780	43.75	0.38
34	1921	443.75	0.38

א. בצעו רגרסיה ליניארית פשוטה על מנת לבדוק את הקשר בין הוצאות עבור פרסום ברדיו למכירות

נבחר בנתונים ← Data Analysis ← Regression

ובחלון נציב את הנתונים כדלקמן (שימו לב שעבור משתנה X נכניס רק את הנתונים עבור דקות הפרסום):



	A	B	C	D
1	חנות	מכירות ₪	הוצ' פרסום ברדיו	מחיר
2	1	2926	843.75	0.78
3	2	3355	843.75	0.78
4	3	1600	443.75	0.78
5	4	2155	443.75	0.78
6	5	1880	443.75	0.78
7	6	2113	443.75	0.78
8	7	820	43.75	0.78
9	8	2000	43.75	0.78
10	9	762	43.75	0.78
11	10	1095	43.75	0.58
12	11	3820	843.75	0.58
13	12	3530	843.75	0.58
14	13	4113	843.75	0.58
15	14	5532	443.75	0.58
16	15	2660	443.75	0.58
17	16	2720	443.75	0.58
18	17	4140	43.75	0.58
19	18	3850	43.75	0.58
20	19	3065	43.75	0.58
21	20	3520	43.75	0.58
22	21	4210	43.75	0.58
23	22	4640	443.75	0.38
24	23	3502	443.75	0.38
25	24	5000	443.75	0.38

*הערה- במידה ולא נרשום כלום תחת confidence level, ברירת המחדל תהיה 95%.

ונקבל:

						SUMMARY OUTPUT
					Regression Statistics	
					0.395	Multiple R
					0.156	R Square
					0.129	Adjusted R Square
					1282.057	Standard Error
					34.000	Observations
						ANOVA
	Significance F	*F*	*MS*	*SS*	*df*	
	0.021	5.908	9711205.331	9711205.331	1.000	Regression
			1643669.103	52597411.287	32.000	Residual
				62308616.618	33.000	Total
Upper 95%	*Lower 95%*	*P-value*	*t Stat*	*Standard Error*	*Coefficients*	
3051.589	1659.852	0.000	6.896	341.626	2355.720	Intercept
3.035	0.267	0.021	2.431	0.679	1.651	הוצ' פרסום ברדיו

בתוצאות שקיבלנו ניתן לראות שהרגרסיה מובהקת. וזאת כיוון שה- F significance שווה ל 0.021 וקטן מרמת מובהקות של 0.05 שהוגדרה (במבחן רגרסיה פשוטה תוצאות P value של השיפוע, כלומר המקדם של הוצאות השיווק ו F significance יהיו זהות).

מקדם ההסבר (מקדם המתאם בריבוע) שווה ל 15.6% המעיד על כך ש 15.6% אחוז מהשינוי במכירות מוסבר על ידי שינוי בהוצאות הפרסום.

המשוואה שקיבלנו המציגה את הקשר בין המכירות להוצאות הפרסום ברדיו היא כדלקמן:

$$\hat{Y}_i = 2335.72 + 1.651X_i$$

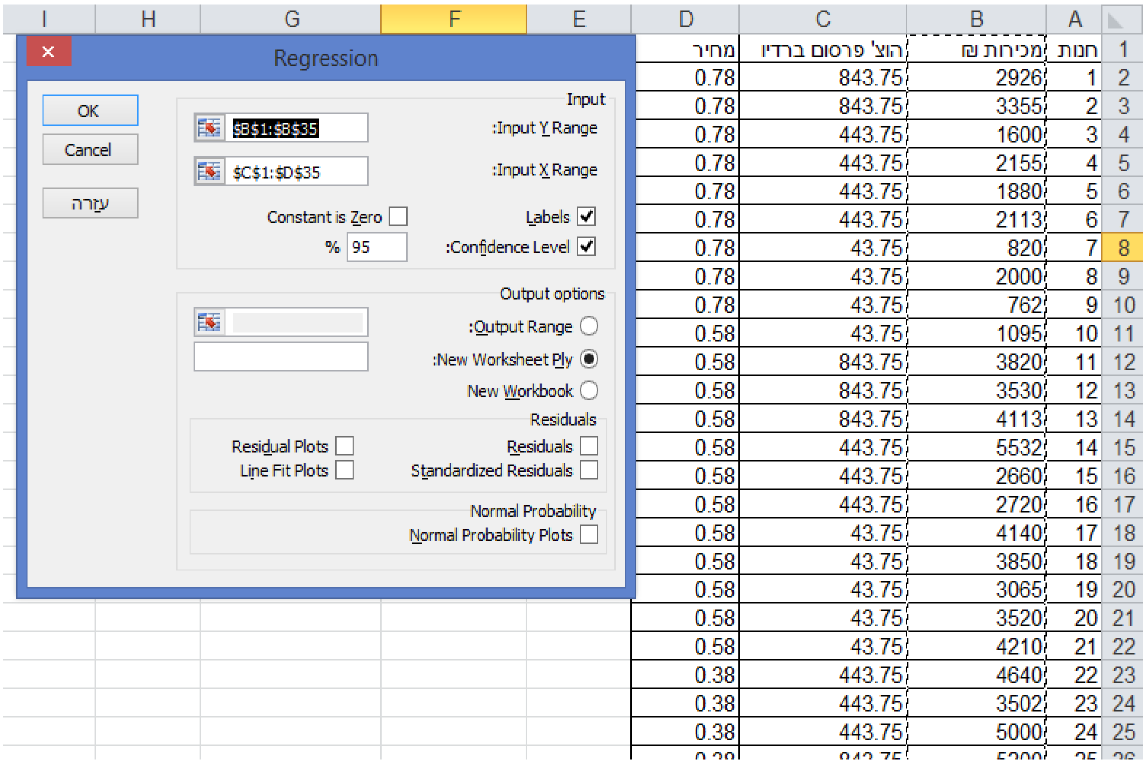
מקדם הרגרסיה של הוצאות השיווק שווה ל 1.651 , ולכן שינוי של 1 ₪ בהוצאות הפרסום יגרום לשינוי של 1.651 ₪ במכירות.

החותך שווה ל 2355.72 , כלומר ללא הוצאות פרסום ברדיו (X=0) המכירות שוות ל 2355.72

ב. בצעו רגרסיה לינארית רבת משתנים כדי להראות את הקשר בין המכירות לבין הוצאות השיווק והמחיר.

נבחר בנתונים ← Regression ← Data Analysis

ובחלון נציב את הנתונים כדלקמן:



	A	B	C	D
1	חנות	מכירות ₪	הוצ' פרסום ברדיו	מחיר
2	1	2926	843.75	0.78
3	2	3355	843.75	0.78
4	3	1600	443.75	0.78
5	4	2155	443.75	0.78
6	5	1880	443.75	0.78
7	6	2113	443.75	0.78
8	7	820	43.75	0.78
9	8	2000	43.75	0.78
10	9	762	43.75	0.78
11	10	1095	43.75	0.58
12	11	3820	843.75	0.58
13	12	3530	843.75	0.58
14	13	4113	843.75	0.58
15	14	5532	443.75	0.58
16	15	2660	443.75	0.58
17	16	2720	443.75	0.58
18	17	4140	43.75	0.58
19	18	3850	43.75	0.58
20	19	3065	43.75	0.58
21	20	3520	43.75	0.58
22	21	4210	43.75	0.58
23	22	4640	443.75	0.38
24	23	3502	443.75	0.38
25	24	5000	443.75	0.38

1. להלן תוצאות הרגרסיה:

ונקבל:

SUMMARY OUTPUT						
	Regression Statistics					
Multiple R	0.511					
R Square	0.261					
Adjusted R Square	0.213					
Standard Error	1218.746					
Observations	34.000					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2.000	16263030.168	8131515.084	5.475	0.009	
Residual	31.000	46045586.450	1485341.498			
Total	33.000	62308616.618				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3905.195	806.077	4.845	0.000	2261.191	5549.199
הוצ' פרסום ברדיו	1.615	0.646	2.500	0.018	0.297	2.932
מחיר	-2759.255	1313.784	-2.100	0.044	-5438.735	-79.776

בתוצאות שקיבלנו ניתן לראות שהרגרסיה מובהקת וזאת כיוון ש ה significance F קטן יותר מרמת מובהקות של 0.05 שהוגדרה. ניתן לראות שרמת המובהקות אף השתפרה. ברגרסיה הפשוטה קיבלנו significance F של 0.021 לעומת 0.009 במודל הרגרסיה המרובה.

מקדם ההסבר (מקדם המתאם בריבוע) שווה ל % 26.1 המעיד על כך ש 26.1% אחוז מהשינוי במכירות מוסבר על ידי שינוי בהוצאות השיווק ושינוי במחיר. נשים לב ששיפרנו את המודל, בהשוואה לרגרסיה הפשוטה כאשר אחוז השונות המוסברת גדל מ % 15.6 ל 26.1% וגם מקדם ההסבר המתוקנן גדל לעומת הרגרסיה הפשוטה ל 21.3% לעומת 12.9% בתוצאות הרגרסיה הפשוטה. הדבר המעיד על כך שהוספת משתנה נוסף שיפרה את תוצאות המודל. תוצאות ה- P value של המקדמים קטנות יותר מ 0.05 , ולכן ניתן להסיק שכל המקדמים מובהקים.

המשוואה שקיבלנו המציגה את הקשר בין המכירות לשיווק היא כדלקמן:

$$\hat{Y}_i = 3905.19 + 1.61X_{1i} - 2759.25X_{2i}$$

המקדם של הוצאות הפרסום ברדיו שווה ל 1.61 , ולכן הגדלת הוצאות הפרסום ב 1 ₪ יגרום לגידול של 1.61 ₪ במכירות.

המקדם של משתנה המחיר שווה ל -2759.25 , כלומר העלאת המחיר ב 1 ₪ יגרור קיטון של 2759.25 במכירות.

לשם ההדגמה , נחשב את המכירות הצפויות עבור הוצאות פרסום של 600 ₪ ומחיר של 0.5 ₪ לחטיף:

$$\hat{Y}_i = 3905.19 + 1.61 \cdot 600 - 2759.25 \cdot 0.5 = 3491.56$$

במקרה זה המכירות הצפויות שוות ל 3491.56 ₪ .

בעיית מולטי – קולינאריות

בעיה אפשרית ברגרסיה מרובה הינה מתאם גבוה בין המשתנים הבלתי תלויים. מתאם גבוה בין המשתנים הבלתי תלויים עלול גרום לבעיות במודל כגון חוסר מובהקות של מקדמי הרגרסיה והיפוך סימנים של מקדמי הרגרסיה (קשר חיובי הופך לשלילי ולהפך).

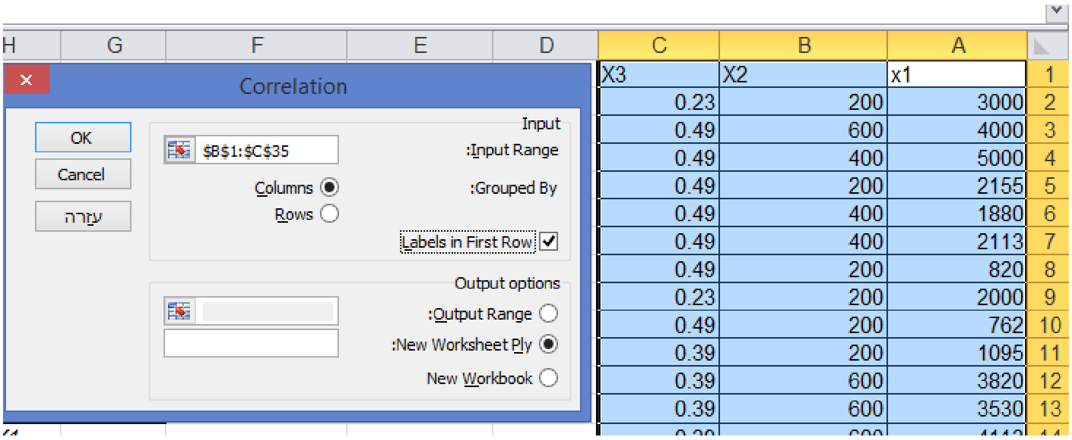
זיהוי בעיית מולטי קולינאריות באמצעות מטריצת קורלציות

דרך אפשרית לזיהוי בעיית מולטיקולינאריות היא הצגת ומציאת המתאםבין כל זוגות המשתנים המסבירים במודל באמצעות מטריצת קורלציות.

לדוגמא: עבור מודל רגרסיה מסוים נניח שאנו רוצים לבדוק האם יש מתאם גבוה בין 3 משתנים בלתי תלויים X1, X2, ו-X3.

על מנת להציג מטריצה זו אופן העבודה באקסל הוא כלדקמן:

נתונים ← data analysis ← Correlation



לאחר הקלדת הנתונים נקבל את פלט המטריצה הבאה:

	X1	X2	X3
X1	1		
X2	0.394787	1	
X3	-0.33476	-0.0268675	1

לפי פלט מטריצת מקדמי המתאם: המתאם בין X1 ל X2 הוא 0.39, בין X1 ל X3 המתאם הוא מינוס 0.33 ובין X2 ל X3 המתאם הוא מינוס 0.02.

ניתן לראות בפלט זה שבמקרה שלנו אין מתאם גבוה באופן חריג בין המשתנים התלויים ולכן אין חשש לבעיית מולטי קולינאריות.

במידה וקיימת בעיה של מולטי קולינאריות פתרונות אפשריים הינם:

1. השמטת חלק מהמשתנים – השמטת חלק מהמשתנים ביניהם הבעיה קיימת.
2. הגדלת המדגם- הגדלת גודל המדגם משפרת את דיוק המודל ולכן תקטין את השפעת בעיית המולטי קולינאריות.

<u>משתנה דמה</u>

עד כה התייחסנו למודלים עם משתנים כמותיים בלבד. אבל במקרים רבים יש צורך לקחת בחשבון גם משתנים איכותיים. לדוגמא נניח שבבדיקת הקשר בין שכר לשעות עבודה, נרצה לקחת בחשבון גם את השפעת המגדר (זכר לעומת נקבה).

אם למשתנה איכותי מסוים יש שתי קטגוריות נגדיר:

Xi=0 , אם התצפית מתאימה לקטגוריה 1

Xi=1 , אם התצפית מתאימה לקטגוריה 0.

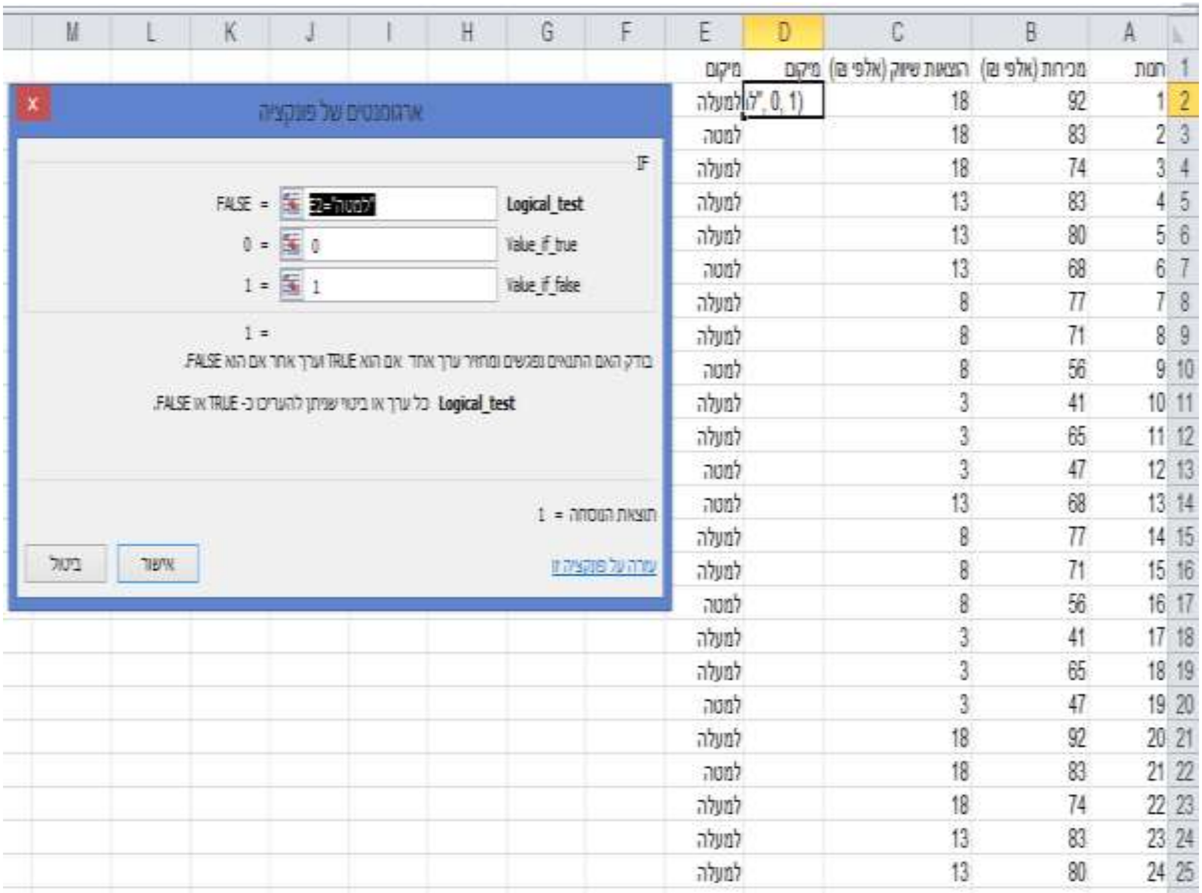
דוגמא :

מנהל שיווק של רשת סופרמרקטים מעוניין לבדוק את הקשר בין המכירות של מוצר מסוים לבין הוצאות השיווק על המוצר לבין מיקומו של המוצר בחנות (בקומה העליונה או התחתונה של החנות).

$X_2=0$, אם החנות הציבה את המוצר בקומה התחתונה

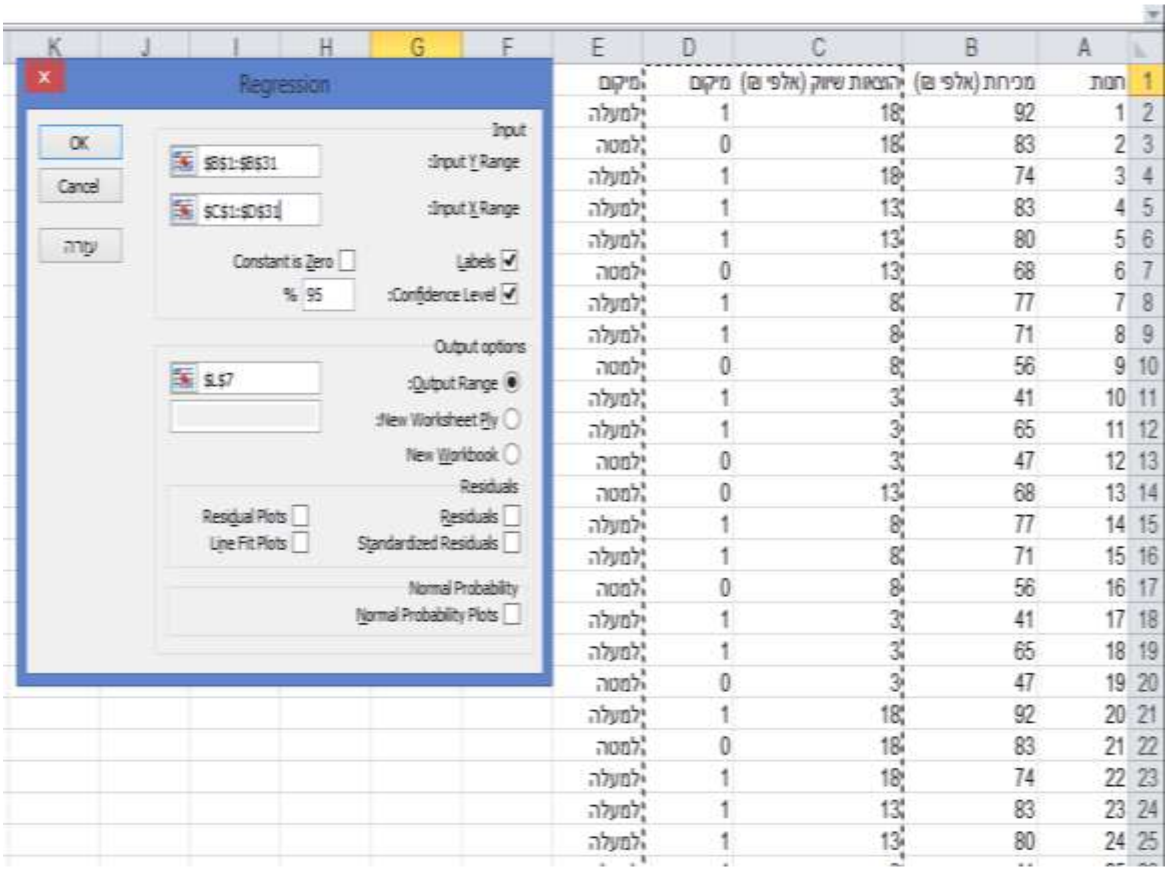
$X_2=1$, אם החנות הציבה את המוצר בקומה העליונה

על מנת לקדד את משתנה הדמי באקסל ניתן להשתמש בפונקציית If, כאשר נבנה טור חדש בו נרשום את הנוסחה כדלקמן:



	A	B	C	D	E
1	חנות	מכירות (אלפי ש"ח)	הוצאות שיווק (אלפי ש"ח)	מיקום	מיקום
2	1	92	18	", 0, 1)	למעלה
3	2	83	18		למטה
4	3	74	18		למעלה
5	4	83	13		למעלה
6	5	80	13		למעלה
7	6	68	13		למטה
8	7	77	8		למעלה
9	8	71	8		למעלה
10	9	56	8		למטה
11	10	41	3		למעלה
12	11	65	3		למעלה
13	12	47	3		למטה
14	13	68	13		למטה
15	14	77	8		למעלה
16	15	71	8		למעלה
17	16	56	8		למטה
18	17	41	3		למעלה
19	18	65	3		למעלה
20	19	47	3		למטה
21	20	92	18		למעלה
22	21	83	18		למטה
23	22	74	18		למעלה
24	23	83	13		למעלה
25	24	80	13		למעלה

לאחר מכן על ידי גרירה נשלים את יתר הטור וכעת ניתן להריץ את הרגרסיה כמוצג מטה:



SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.850
R Square	0.722
Adjusted R Square	0.702
Standard Error	8.183
Observations	30.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance*
Regression	2.000	4698.787	2349.393	35.087	0.000
Residual	27.000	1807.913	66.960		
Total	29.000	6506.700			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	40.659	3.744	10.861	0.000	32.978	48.340
הוצאות שיווק (אלפי ש)	2.141	0.271	7.914	0.000	1.586	2.696
מיקום	9.770	3.172	3.080	0.005	3.262	16.279

קיבלנו את קו הרגרסיה הבא:

$$\hat{Y}_i = 40.659 + 2.141X_{1i} + 9.770X_{2i}$$

עבור מוצר שהוצב בקומה התחתונה של החנות נקבל :

$X_2=0$ משום ש , $\hat{Y}_i = 40.659 + 2.141X_{1i}$

עבור מוצר שהוצב בקומה העליונה של החנות נקבל:

$X_2=1$ משום ש , $\hat{Y}_i = 50.429 + 2.141X_{1i}$

כיוון ש $X_2=1$, 9.770 מתווסף ל 40.659.

מתוצאות המודל ניתן להסיק שעבור מוצר (בין אם הוא ממוקם בקומה העליונה או התחתונה) גידול של 1 (אלפי ₪) בהוצאות השיווק יגדיל את המכירות ב 2.141 (אלפי ₪) . בנוסף ניתן להסיק שמיקום המוצר בקומה העליונה של החנות יגדיל את המכירות של המוצר בממוצע של 9.770 (אלפי ₪).