סטטיסטיקה לשאלון 471
לפי התכנית החדשה

עומר גלעדי

מהי סטטיסטיקה?

הגדרה אינטואיטיבית

סטטיסטיקה היא תחום במתמטיקה העוסק באיסוף, ארגון, הצגה וניתוח של נתונים, כדי להסיק מסקנות על תופעות בעולם האמיתי.

מטרות הסטטיסטיקה

- **תיאור נתונים** – סיכום מסודר של מידע רב באמצעות טבלאות, גרפים ומדדים.
- **השוואה** – השוואת קבוצות שונות (למשל: שכבות, בתי ספר, מגמות).
- **קשר וחיזוי** – חקר קשר בין משתנים וניסיון לחזות התנהגות עתידית.

אוכלוסייה, מדגם ותצפית

מושגים בסיסיים

- **אוכלוסייה** – כל קבוצת הפריטים/האנשים שאנו מעוניינים לחקור.
  דוגמה: כל תלמידי כיתה י"א ברשת בתי הספר.
- **מדגם** – קבוצה קטנה יותר שנלקחת מתוך האוכלוסייה ונבדקת בפועל.
  דוגמה: 40 תלמידים שנבחרו באקראי מתוך כלל תלמידי הרשת.
- **תצפית** – פריט אחד בודד במדגם או באוכלוסייה.
  דוגמה: תלמיד יחיד, יחד עם הנתונים שנאספו עליו.

מאפיין ומשתנה סטטיסטי

- **מאפיין** – תכונה שמעניינת אותנו באוכלוסייה (ציון, גיל, מגדר, מגמה).
- **משתנה סטטיסטי** – מאפיין שנמדד על כל תצפית ויכול לקבל ערכים שונים.

משתנה כמותי (נומרי)

הגדרה

משתנה כמותי הוא משתנה שערכיו הם מספרים בעלי משמעות מדידה, כך שניתן לבצע עליהם פעולות חשבוניות (כמו חישוב ממוצע, הפרשים, אחוזים וכדומה).

סוגי משתנים כמותיים

- **כמותי בדיד** – מקבל ערכים שלמים בלבד.
  דוגמאות: מספר אחים, מספר רכבים במשפחה, מספר שיעורי בית שהוגשו.
- **כמותי רציף** – יכול לקבל כל ערך בתחום מסוים, כולל שברים.
  דוגמאות: גובה, משקל, זמן ריצה ל־100 מטר, ציון בין 0 ל־100.

דוגמאות

- X – ציון בבחינת מתמטיקה (בין 0 ל־100) – משתנה כמותי רציף.
- Y – מספר שעות לימוד בשבוע – משתנה כמותי בדיד.

משתנה איכותי (קטגורי)

הגדרה

משתנה איכותי הוא משתנה שערכיו הם קטגוריות או שמות, ולא מספרים בעלי משמעות חשבונית.

דוגמאות למשתנים איכותיים

- מגדר: זכר / נקבה.
- מגמה: פיזיקה / כימיה / מדעי המחשב / ביולוגיה.
- רמת לימוד במתמטיקה: 3 יח"ל / 4 יח"ל / 5 יח"ל.
- אזור מגורים: צפון / מרכז / דרום / ירושלים.

הערה חשובה

לעיתים נותנים לקטגוריות **קוד מספרי** (למשל 1=זכר, 2=נקבה), אך המשתנה עדיין נחשב **איכותי**, כי אין משמעות לחשב עליהם ממוצע או סטיית תקן.

דוגמה מסכמת: סיווג משתנים

נבחן את הטבלה הבאה, המתארת נתונים על תלמידים:

תלמיד	ציון במתמטיקה	מגמה	מספר אחים
א	92	פיזיקה	1
ב	75	כימיה	3
ג	88	מדעי המחשב	0

שאלה

סווג עבור כל עמודה בטבלה האם המשתנה הוא כמותי (בדיד/רציף) או איכותי.

פתרון מלא

- **ציון במתמטיקה** – משתנה כמותי (בד"כ נתפס כרציף בטווח 0–100).
- **מגמה** – משתנה איכותי (קטגוריות: פיזיקה, כימיה, מדעי המחשב).
- **מספר אחים** – משתנה כמותי **בדיד** (ערכים שלמים: $0, 1, 2, \ldots$).

איור: כמותי מול איכותי

סיווג משתנים סטטיסטיים

איכותי משתנה

דוגמאות: מגורים אזור מגדר, מגמה,

כמותי משתנה

דוגמאות: אחים מספר גובה, גיל, ציון,

הבחנה נכונה בין משתנה כמותי לאיכותי היא **בסיס** לכל ניתוח סטטיסטי תקין.

למה להציג נתונים בייצוגים שונים?

רעיון מרכזי

אותם נתונים יכולים להיות מוצגים בדרכים שונות: טבלה, דיאגרמת מקלות (עמודות), גרף קו, דיאגרמת עוגה, היסטוגרמה ועוד. כל ייצוג מדגיש היבט אחר של המידע.

מטרות עיקריות של ייצוג הנתונים

- **סיכום** – ריכוז מידע רב בצורה קצרה וברורה.
- **השוואה** – השוואת קבוצות או קטגוריות (כיתות, מגמות, שנים).
- **זיהוי מגמות** – עליה/ירידה, ריכוזים קיצוניים, חריגים.
- **תקשורת** – הצגת תוצאות באופן נגיש למקבלי החלטות ולתלמידים.

דוגמה: טבלת שכיחויות

נניח שנמדדו ציוני מתמטיקה של 20 תלמידים (בין 0 ל־100). כדי לסכם את הנתונים, מחלקים את הציונים לקטגוריות (טווחים):

טווח ציונים	שכיחות (מספר תלמידים)
59--50	2
69--60	5
79--70	7
89--80	4
100--90	2

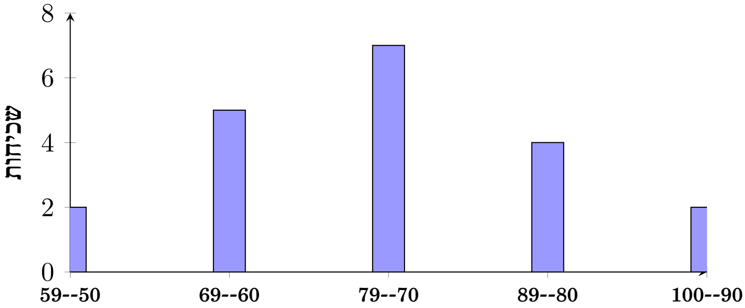
הסבר

- כל שורה מתארת קטגוריית ציונים ואת מספר התלמידים שנמצאים בטווח זה.
- סכום השכיחויות הוא $2+5+7+4+2=20$ — מספר התלמידים הכולל.
- טבלה זו משמשת בסיס לשרטוט דיאגרמות שונות.

דיאגרמת מקלות (עמודות)

רעיון

דיאגרמת מקלות (או עמודות) מתאימה במיוחד להצגת שכיחויות של קטגוריות נפרדות: טווחי ציונים, מגמות, ערים וכדומה.



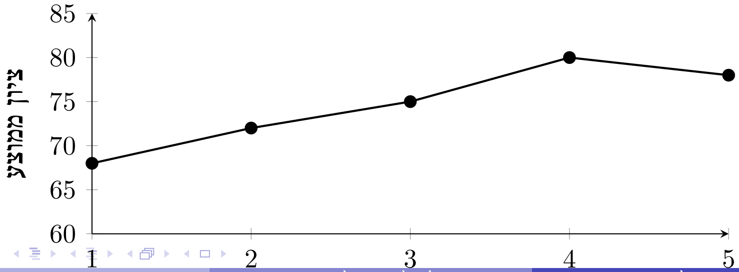
גובה כל עמודה מסמל את מספר התלמידים בטווח ציונים מסוים.

גרף קו – מעקב אחרי שינוי לאורך זמן

דוגמה

בית ספר עוקב אחרי **ממוצע הציונים** במתמטיקה בכיתה מסוימת לאורך חמישה מבחנים:

מספר מבחן	ממוצע ציון
1	68
2	72
3	75
4	80
5	78



דיאגרמת עוגה

דוגמה

התפלגות מגמות בקרב 40 תלמידים:

- פיזיקה – 10 תלמידים
- כימיה – 12 תלמידים
- מדעי המחשב – 14 תלמידים
- ביולוגיה – 4 תלמידים

היסטוגרמה – נתונים כמותיים רציפים

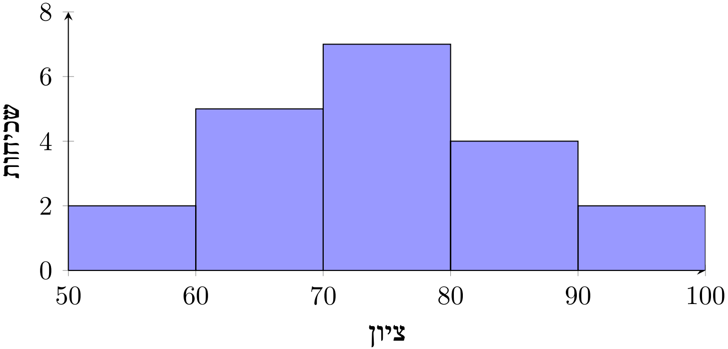
רעיון

היסטוגרמה מתאימה להצגת **התפלגות** של משתנה כמותי רציף (כגון גובה, משקל, ציון). ציר ה־x מחולק לקטעים (טווחים), וגובה כל עמודה מסמל את השכיחות בטווח.

נשתמש שוב בטבלת טווחי הציונים:

טווח ציונים	שכיחות
59--50	2
69--60	5
79--70	7
89--80	4
100--90	2

היסטוגרמה – שרטוט



שימו לב: בהיסטוגרמה הטווחים צמודים זה לזה, כדי להדגיש שמדובר במשתנה **רציף**.

סיכום – הצגת נתונים בייצוגים שונים

- **טבלה** – בסיס מסודר לכל ניתוח; מציגה שכיחויות בצורה טקסטואלית.
- **דיאגרמת מקלות** – טובה להשוואת קטגוריות ולטווחי ציונים.
- **גרף קו** – מדגיש **מגמה** לאורך זמן (עליות/ירידות).
- **דיאגרמת עוגה** – מדגישה חלק יחסי מכלל האוכלוסייה.
- **היסטוגרמה** – מתארת התפלגות של משתנה כמותי רציף לפי טווחים.

הערה פדגוגית

חשוב לתרגל עם התלמידים מעבר מאותה טבלת נתונים לייצוגים שונים, ולשאול בכל פעם: **מה קל לראות בייצוג הזה? ומה פחות בולט?**

מעבר מייצוג לייצוג – למה זה חשוב?

רעיון מרכזי

אותו מידע סטטיסטי יכול להופיע בצורות שונות: טבלה, גרף מקלות, עוגה, היסטוגרמה ועוד.
יכולת המעבר בין הייצוגים היא מהותית להבנה:

- מה הנתונים "אומרים" לנו?
- אילו תובנות בולטות בכל ייצוג?
- מה היתרון של כל דרך הצגה?

חשיבות לבגרות ולחיים

מעבר בין ייצוגים מאפשר:

- השוואה בין קבוצות
- זיהוי מגמות
- קבלת החלטות מושכלת

דוגמה 1: מטבלה לדיאגרמת עמודות

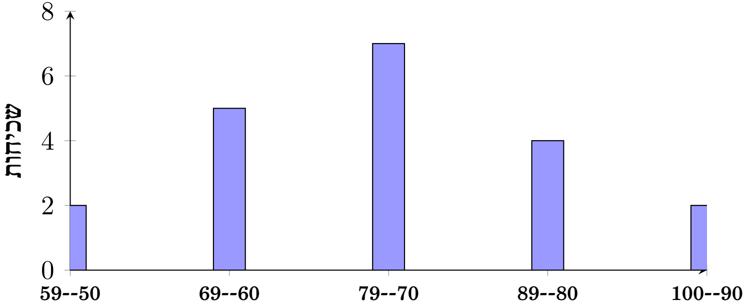
נתונים על התפלגות ציונים של 20 תלמידים:

טווח ציונים	שכיחות
59--50	**2**
69--60	**5**
79--70	**7**
89--80	**4**
100--90	**2**

מה ניתן לראות בטבלה?

- הטווח 79--70 הוא השכיח ביותר.
- הסה"כ הוא 20.

דוגמה 1: הייצוג הגרפי



בדיאגרמת מקלות – קל לראות מיד מי גבוה יותר ומי נמוך יותר, ללא חישובים.

דוגמה 2: מטבלה לדיאגרמת עוגה

הנתונים הקודמים:

טווח ציונים	שכיחות
59--50	**2**
69--60	**5**
79--70	**7**
89--80	**4**
100--90	**2**

כדי לצייר עוגה

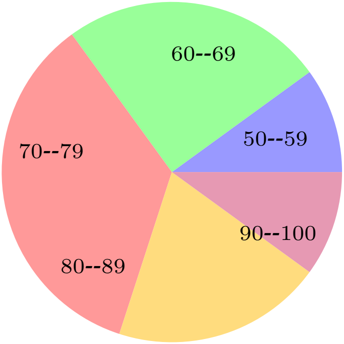
יש לחשב את החלק היחסי של כל טווח:

$$\frac{2}{20}=0.1,\ \frac{5}{20}=0.25,\ \frac{7}{20}=0.35,\ \frac{4}{20}=0.2,\ \frac{2}{20}=0.1$$

או באחוזים:

$$10\%, 25\%, 35\%, 20\%, 10\%$$

דיאגרמת עוגה – שרטוט

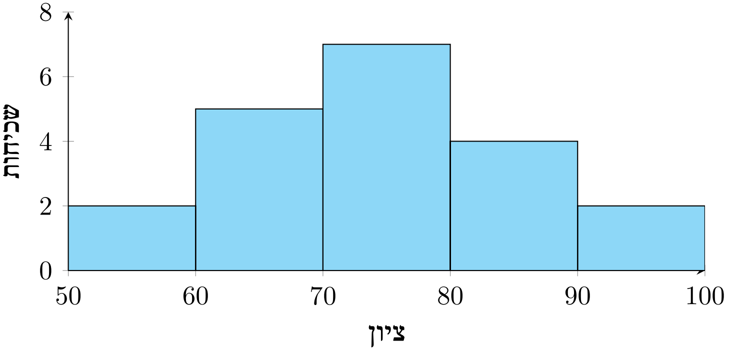


דיאגרמת עוגה מדגישה את החלק היחסי של כל טווח ציונים.

דוגמה 3: מטבלה להיסטוגרמה

רעיון

היסטוגרמה מתאימה כאשר הנתונים הם **כמותיים רציפים**. הטבלה שלנו כבר מאורגנת לפי טווחים – ולכן יכולה לשמש להיסטוגרמה.



שימו לב: בהיסטוגרמה **אין רווחים בין הטווחים**, בניגוד לדיאגרמת מקלות.

דוגמה 4: השפעת הייצוג על הפרשנות

ניקח את אותם נתונים ונציג בייצוגים שונים:

- בדיאגרמת מקלות – ברור שהטווח 70--79 הוא הגבוה ביותר.
- בהיסטוגרמה – בולטת צורת "גוש" במרכז.
- בדיאגרמת עוגה – ניתן להעריך במהירות את האחוז היחסי.

מסקנה

כל ייצוג מדגיש היבט אחר של הנתונים, ולכן נכון להשתמש במספר ייצוגים כדי להבין את התמונה המלאה.

שאלה ופתרון – מעבר ייצוג

שאלה

לפניך טבלת שכיחויות:

טווח	שכיחות
40--49	3
50--59	7
60--69	5
70--79	1

- א. צייר דיאגרמת מקלות.
- ב. מהו הטווח השכיח ביותר?
- ג. האם זה מייצג ציונים "גבוהים"?

פתרון מלא

א. דיאגרמת מקלות – מילולית

טווח 50--59 יהיה העמודה הגבוהה ביותר, כי שכיחותו 7.

ב. הטווח השכיח ביותר

50--59

ג. משמעות

אף שהטווח השכיח הוא 50--59, זה **נחשב ציון נמוך יחסית**. כלומר – הטבלה מציגה קושי לימודי כללי של הקבוצה.

סיכום – מעבר מייצוג לייצוג

- אפשר להציג אותם נתונים בצורות שונות – טבלה, מקלות, עוגה, היסטוגרמה.
- לכל ייצוג יתרונות וחסרונות.
- **חשוב לשאול: מה יותר קל לראות כאן? מה פחות בולט?**
- **בפרשנות סטטיסטית – אין ייצוג "נכון", אלא מתאים לתפקיד.**

חשיבות לתלמידי בגרות

מעבר מיומן בין ייצוגים הוא מיומנות נדרשת בשאלון 471: הבנה, תיאור, השוואה והסקת מסקנות.

מהם מדדי מרכז?

רעיון מרכזי

מדדי מרכז הם מספרים המתארים את "מרכז הנתונים", או את הערך המייצג ביותר של קבוצת תצפיות.

מדדי מרכז נפוצים

- **ממוצע (Mean)**
- **ממוצע משוקלל (Weighted Mean)**
- **שכיח (Mode)**
- **חציון (Median)**

שימושיות

מדדי מרכז משמשים:

- לתיאור קבוצה של נתונים
- להשוואה בין קבוצות
- להסקת מסקנות על התנהגות אופיינית

הממוצע – הגדרה וחישוב

הגדרה מתמטית

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

כאשר x_i הם ערכי התצפיות ו-n הוא מספר התצפיות.

דוגמה

ציוני מבחן של חמישה תלמידים:

$$72,\ 85,\ 90,\ 68,\ 80$$

חישוב:

$$\bar{x} = \frac{72+85+90+68+80}{5} = \frac{395}{5} = 79$$

הממוצע הוא 79.

השכיח – ההגדרה הפשוטה ביותר

הגדרה

שכיח הוא הערך שמופיע הכי הרבה פעמים בנתונים.

דוגמה

$$3,\ 5,\ 7,\ 5,\ 8,\ 5,\ 9$$

הערך 5 מופיע 3 פעמים – יותר מכל ערך אחר. לכן השכיח הוא:

$$\text{Mode} = 5$$

הערה חשובה

ייתכן שלא יהיה שכיח אחד בלבד – או שלא יהיה כלל:

- $1, 2, 3, 4$ – ללא שכיח.
- $2, 2, 3, 3$ – שני שכיחים.

החציון – ערך האמצע

הגדרה

חציון הוא הערך "האמצעי" כאשר הנתונים מסודרים מהקטן לגדול. הוא מחלק את הנתונים ל-50% תחתונים ו-50% עליונים.

דוגמה 1 – מספר אי-זוגי של ערכים

$$4,\ 7,\ 1,\ 9,\ 3$$

סידור:

$$1,\ 3,\ 4,\ 7,\ 9$$

הערך האמצעי הוא 4 — זהו החציון.

דוגמה 2 – מספר זוגי של ערכים

$$6,\ 2,\ 9,\ 3$$

סידור:

$$2,\ 3,\ 6,\ 9$$

החציון:

ממוצע משוקלל

הגדרה

כאשר לערכים שונים יש **חשיבות שונה**, הממוצע מחושב לפי משקל:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

דוגמה

ציון תלמיד בשני חלקי מבחן:

- חלק א' – ציון 80, משקל 30%
- חלק ב' – ציון 92, משקל 70%

חישוב:

$$\bar{x} = \frac{0.3 \cdot 80 + 0.7 \cdot 92}{0.3 + 0.7} = \frac{24 + 64.4}{1} = 88.4$$

הציון המשוקלל: 88.4

השוואה בין מדדי מרכז

למי מתאים כל מדד?

- **ממוצע** – מתאים כאשר אין חריגות קיצוניות.
- **חציון** – עמיד לחריגים; מתאים ל"פיזור לא סימטרי".
- **שכיח** – טוב לנתונים איכותיים או בדידים.
- **ממוצע משוקלל** – מתאים כאשר חלקים שונים חשובים בצורה שונה.

דוגמה קצרה

ציונים:

$$50,\ 55,\ 60,\ 95$$

- **ממוצע:** $\frac{260}{4} = 65$
- **חציון:** $\frac{55+60}{2} = 57.5$
- **שכיח:** אין

המסקנה: הממוצע גבוה יחסית – בגלל הציון החריג **95**.

שאלה – מדדי מרכז

לפניך ציונים של 8 תלמידים:

$$72,\ 85,\ 90,\ 68,\ 80,\ 60,\ 90,\ 95$$

- א. חשב את הממוצע.
- ב. מצא את השכיח.
- ג. מצא את החציון.

פתרון מלא

א. ממוצע

סכום:

$$72+85+90+68+80+60+90+95=640$$

ממוצע:

$$\frac{640}{8}=80$$

ב. שכיח

הערך 90 מופיע פעמיים – הכי הרבה. לכן השכיח הוא:

$$90$$

ג. חציון

סידור:

$$60,\ 68,\ 72,\ 80,\ 85,\ 90,\ 90,\ 95$$

שני ערכי האמצע:

$$80,\ 85$$

חציון:

סיכום – מדדי מרכז

- מדדי מרכז מסכמים נתונים למספר בודד מייצג.
- ממוצע – רגיש לחריגים.
- חציון – עמיד לחריגים.
- שכיח – מתאים לקטגוריות.
- ממוצע משוקלל – מתאים כאשר יש הבדלי חשיבות.

למה זה חשוב לבגרות?

בשאלון 471 נדרשים:

- חישובים מדויקים
- הסבר מילולי ברור
- פרשנות מושכלת של תוצאות

מהם מדדי פיזור?

רעיון מרכזי

מדדי פיזור מתארים עד כמה הנתונים מפוזרים סביב מדד המרכז (בדרך כלל ממוצע או חציון).

מדדי פיזור נפוצים

- **טווח** (Range)
- **שונות** (s^2)
- **סטיית תקן** (s)

שימושים

מדדי פיזור מאפשרים:

- לזהות אחידות או שונות בין תצפיות
- להשוות בין קבוצות
- לאפיין סיכון/אי-ודאות

טווח

הגדרה

הטווח הוא ההפרש בין הערך הגבוה ביותר לערך הנמוך ביותר:

$$\text{Range} = \max - \min$$

דוגמה

$$72,\ 85,\ 90,\ 68,\ 80$$

$$\max = 90, \quad \min = 68$$

$$\text{Range} = 90 - 68 = 22$$

חיסרון

הטווח מתבסס רק על שני ערכים, ולכן עלול להטעות.

שונות – רעיון

רעיון אינטואיטיבי

שונות מודדת כמה רחוקים הנתונים **בממוצע** מהממוצע.

נוסחה

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

הערה

כל ההפרשים בריבוע, ולכן:

- ערכים רחוקים מהממוצע משפיעים יותר
- אין ביטול בין חיוביים לשליליים

סטיית תקן

הגדרה

סטיית תקן היא השורש הריבועי של השונות:

$$s = \sqrt{s^2}$$

משמעות

סטיית תקן מודדת **באותן יחידות של הנתונים**, ולכן קלה יותר להבנה.

כלל אצבע

ככל שסטיית התקן גדולה יותר – הנתונים **מפוזרים** יותר.

דוגמה – חישוב שונות וסטיית תקן

לוח ציונים:

$$4,\ 8,\ 6$$

חישוב ממוצע:

$$\bar{x} = \frac{4+8+6}{3} = 6$$

הפרשים מהממוצע:

$$-2,\ +2,\ 0$$

ריבועי הפרשים:

$$4,\ 4,\ 0$$

שונות:

$$s^2 = \frac{4+4+0}{3} = \frac{8}{3} \approx 2.67$$

סטיית תקן:

$$s = \sqrt{\frac{8}{3}} \approx 1.63$$

פרשנות – למה זה חשוב?

דוגמה

בשתי כיתות ממוצע הציונים זהה: 80.
אבל בכיתה א':

$$78,\ 82,\ 79,\ 81,\ 80$$

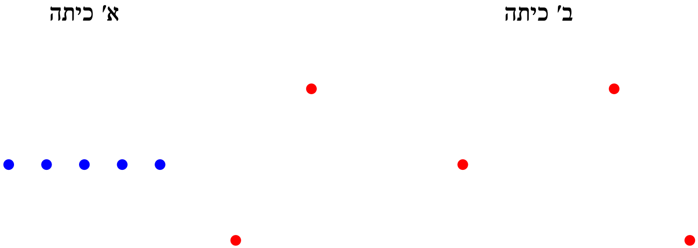
ובכיתה ב':

$$60,\ 95,\ 50,\ 100,\ 95$$

מסקנה

- הממוצע אינו מספר את כל הסיפור.
- בכיתה ב' יש פיזור גדול – רמות שונות מאוד.
- בכיתה א' – ביצועים אחידים.

איור – פיזור שונה סביב אותו ממוצע



בשתי הכיתות הממוצע זהה, אבל הפיזור שונה מאוד.

שאלה – מדדי פיזור

לפניך ציונים:

$$70,\ 80,\ 80,\ 90,\ 100$$

- א. חשב את הטווח.
- ב. מצא את השונות.
- ג. מצא את סטיית התקן.

פתרון מלא

א. טווח

$$100 - 70 = 30$$

ב. שונות

ממוצע:

$$\bar{x} = \frac{70 + 80 + 80 + 90 + 100}{5} = 84$$

הפרשים:

$$-14, \ -4, \ -4, \ 6, \ 16$$

ריבועים:

$$196, \ 16, \ 16, \ 36, \ 256$$

סכום:

$$196 + 16 + 16 + 36 + 256 = 520$$

שונות:

$$s^2 = \frac{520}{5} = 104$$

ג. סטיית תקן

סיכום – מדדי פיזור

- פיזור מתאר כמה הנתונים רחוקים ממדד המרכז.
- טווח – פשוט אך רגיש לקיצוניים.
- שונות – מודדת את הפיזור בריבוע.
- סטיית תקן – מדד מרכזי ומובן, באותן יחידות כמו הנתונים.

הערה פדגוגית

לימוד מדדי מרכז בלי מדדי פיזור נותן תמונה חלקית בלבד. שני הסוגים נחוצים כדי להבין נתונים.

מהי התפלגות נורמלית?

הגדרה אינטואיטיבית

התפלגות נורמלית היא התפלגות רציפה, סימטרית בצורת פעמון, המתארת את האופן שבו נתונים רבים בעולם האמיתי מתרכזים סביב ערך מרכזי.

מאפיינים מרכזיים

- סימטרית סביב הממוצע μ
- רוב התצפיות קרובות לממוצע
- ערכים קיצוניים – נדירים יותר

פונקציה נורמלית – נוסחה מלאה

פונקציית צפיפות ההסתברות

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

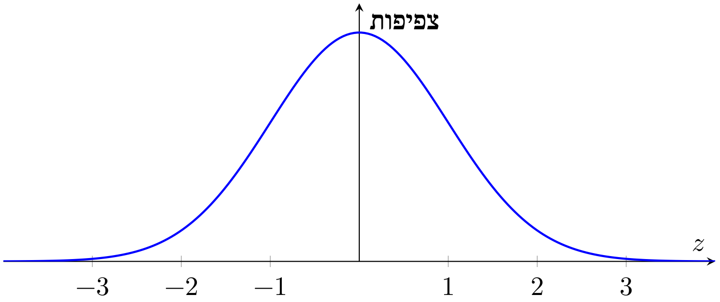
כאשר:

- μ – ממוצע
- σ – סטיית תקן

הערה לתלמידים

הנוסחה אינה נדרשת לחישוב בבגרות – אך חשוב להבין את משמעות μ ו־σ.

איור – התפלגות נורמלית סטנדרטית



התפלגות נורמלית סטנדרטית: $\mu = 0,\ \sigma = 1$

תכונת הפעמון

פירוש חזותי

- האזור הגבוה במרכז – רוב התצפיות
- ככל שמתרחקים מהממוצע – ההסתברות יורדת

חשיבות

התכונה הזו מאפשרת:

- זיהוי נתונים טיפוסיים
- איתור ערכים קיצוניים (Outliers)
- חישוב הסתברויות

כלל 68-95-99.7

מה הכלל אומר?

לכל התפלגות נורמלית מתקיימים בקירוב:

68% מהנתונים נמצאים בתוך $\mu \pm 1\sigma$
95% מהנתונים נמצאים בתוך $\mu \pm 2\sigma$
99.7% מהנתונים נמצאים בתוך $\mu \pm 3\sigma$

משמעות

רוב העצום של הנתונים נמצא קרוב מאוד לממוצע.

דוגמה – גובה תלמידים

גובה תלמידים מתפלג בקירוב נורמלית עם:

$$\mu = 170\text{ס"מ} \;, \qquad \sigma = 8\text{ס"מ}$$

שאלה

מה ניתן לומר על רוב התלמידים לפי כלל 68–95–99.7?

פתרון מילולי

- טווח $\mu \pm 1\sigma$:

$$170 - 8 = 162, \qquad 170 + 8 = 178$$

- לכן: **כ-68% מהתלמידים** בגובה שבין 162 ל־178 ס"מ.

התפלגות נורמלית וסטיית תקן – פרשנות

האם סטיית תקן גדולה זה טוב?

תלוי מה מודדים:

- ציונים – סטיית תקן גדולה משמעה שונות גדולה בין תלמידים.
- ייצור – סטיית תקן גדולה משמעה חוסר יציבות ואיכות נמוכה.
- ספורט – סטיית תקן גדולה משמעה ביצועים לא עקביים.

מסקנה

ללא הבנת הפיזור – הממוצע לבדו מטעה.

שאלה ברמת בגרות

גובה תלמידים מתפלג נורמלית:

$$\mu = 170, \quad \sigma = 6$$

- א. מה חלק התלמידים שגובהם בין 164 ל-176?
- ב. האם מדובר ברוב התלמידים? הסבר.

פתרון מלא

א. חישוב

נבדוק:

$$164,\ \ 176 = \mu \pm \sigma$$

כי:

$$170 - 6 = 164, \quad 170 + 6 = 176$$

לכן:

$$\approx 68\%$$

ב. הסבר מילולי

בערך שני שלישים מהתלמידים נמצאים בטווח הזה - זהו הרוב.

סיכום – מבוא להתפלגות נורמלית

- התפלגות סימטרית בצורת פעמון
- רוב הנתונים קרובים לממוצע
- סטיית תקן קובעת את רוחב הפיזור
- כלל 68–95–99.7 – כלי חשוב להערכה מהירה

למה זה חשוב לבגרות?

יש צורך:

- לפרש משמעות של ממוצע וסטיית תקן
- להעריך אחוזים בטווחים שונים
- לענות תשובות מילוליות

למה צריך טבלת נורמלית?

בעיה

פונקציית הצפיפות הנורמלית:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

אין לה אינטגרל פשוט – כלומר, קשה לחשב הסתברויות בצורה ישירה.

פתרון

במקום לחשב בעצמנו – משתמשים בטבלת התפלגות נורמלית סטנדרטית, המבוססת על משתנה Z עם:

$$\mu = 0, \quad \sigma = 1$$

סטנדרטיזציה – המרת ערך ל-Z

הגדרה

כדי להשתמש בטבלה, ממירים כל ערך x לערך תקני:

$$z = \frac{x - \mu}{\sigma}$$

משמעות

הערך z אומר:

- כמה סטיות תקן x רחוק מהממוצע
- האם מעל הממוצע $(z > 0)$ או מתחתיו $(z < 0)$

מה נותנת הטבלה?

הטבלה נותנת

את ההסתברות:

$$P(Z \leq z)$$

כלומר – שטח משמאל לערך z.

השלמה

אם צריך הסתברות מימין:

$$P(Z > z) = 1 - P(Z \leq z)$$

דוגמה בסיסית – מפרשים מהטבלה

נתון

בטבלה:

$$P(Z \leq 1.00) = 0.8413$$

משמעות

- **ההסתברות להיות מתחת ל־1 סטיית תקן מעל הממוצע: 84.13**
- **ההסתברות להיות מעל ערך זה:**

$$P(Z > 1.00) = 1 - 0.8413 = 0.1587$$

דוגמה 1 – המרת X ל-Z

גובה תלמידים:

$$\mu = 170, \quad \sigma = 6$$

שאלה

מה ערך ה־z של תלמיד שגובהו 178?

פתרון

$$z = \frac{178 - 170}{6} = \frac{8}{6} \approx 1.33$$

פירוש

הציון 178 **נמצא 1.33 סטיות תקן מעל הממוצע.**

דוגמה 2 – הסתברות בעזרת טבלה

גובה תלמידים התפלג נורמלית:

$$\mu = 170, \quad \sigma = 6$$

שאלה

מה ההסתברות שתלמיד יהיה נמוך מ־178 ס"מ?

פתרון

המרה ל-z:

$$z = \frac{178 - 170}{6} \approx 1.33$$

מהטבלה:

$$P(Z \leq 1.33) \approx 0.9082$$

תשובה: כ־90.8%

דוגמה 3 – הסתברות בין שני ערכים

נניח ש:

$$\mu = 70, \quad \sigma = 10$$

שאלה

מה ההסתברות שציון X יהיה בין 60 ל-85?

פתרון

המרות:

$$z_1 = \frac{60-70}{10} = -1$$

$$z_2 = \frac{85-70}{10} = 1.5$$

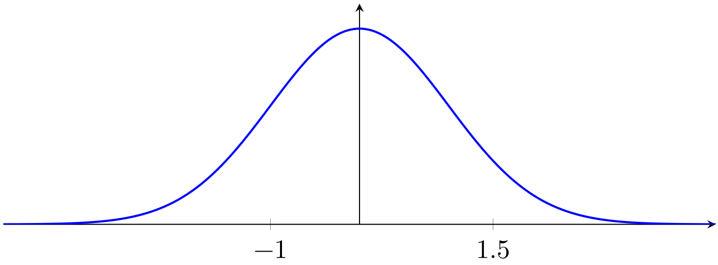
מהטבלה:

$$P(Z \leq -1) = 0.1587$$

$$P(Z \leq 1.5) = 0.9332$$

לכן:

איור – הסתברות בין שני ערכים



האזור המוצלל = ההסתברות בין שני ערכים.

ערכים מעל ערך מסוים

שאלה

מה ההסתברות שציון X יהיה **מעל 85**?

תשובה כללית

$$P(X > a) = 1 - P(Z \leq z_a)$$

הערה

חישוב פשוט – ללא אינטגרלים.

שאלה ברמת בגרות

ציון במבחן מתפלג נורמלית:

$$\mu = 78, \quad \sigma = 12$$

- א. מה ההסתברות שתלמיד יקבל מעל 90?
- ב. האם מדובר בציון נדיר? הסבר.

פתרון מלא

א. חישוב

המרה:

$$z = \frac{90-78}{12} = \frac{12}{12} = 1$$

מהטבלה:

$$P(Z \leq 1) = 0.8413$$

לכן:

$$P(X > 90) = 1 - 0.8413 = 0.1587$$

תשובה: כ־15.9%

ב. פירוש

ההסתברות נמוכה – פחות מ־20%. לכן ציון מעל 90 הוא יחסית נדיר.

סיכום – שימוש בטבלת נורמלית

- ממירים ערך x לערך תקני z
- משתמשים בטבלה כדי לקבל הסתברות
- ההסתברות בטבלה היא תמיד מימין **לאפס ועד** z
- הסתברות מעל ערך = השלמה ל-1
- הסתברות בין ערכים = הפרש בין שתי ההסתברויות

למה זה חשוב לבגרות?

יכולת חישוב והסקת מסקנות מילוליות הן חלק מרכזי בשאלון 471.

מהי רגרסיה?

רעיון מרכזי

רגרסיה היא שיטה סטטיסטית לחקר קשר בין משתנים, כאשר אחד מהם נחשב **משתנה מסביר** (עצמאי) והשני **משתנה מוסבר** (תלוי).

מטרות

- לבדוק האם קיים קשר בין משתנים
- לאפיין את עוצמת הכיוון שלו
- לבצע **חיזוי** של ערכים עתידיים

משתנה מסביר ומשתנה מוסבר

הגדרות

- **משתנה מסביר (עצמאי)** – גורם משפיע/מנבא, מסומן לרוב ב־X
- **משתנה מוסבר (תלוי)** – משתנה שתלוי בערכי X, מסומן לרוב ב־Y

דוגמה

מחקר על הקשר בין:

- X – מספר שעות לימוד בשבוע
- Y – ציון במבחן מתמטיקה

הערה פדגוגית

חשוב להדגיש לתלמידים: **עובדה שקיים קשר סטטיסטי אינה מוכיחה סיביתיות.**

דוגמה מספרית – נתונים

נבדקו 10 תלמידים, ונרשמו:

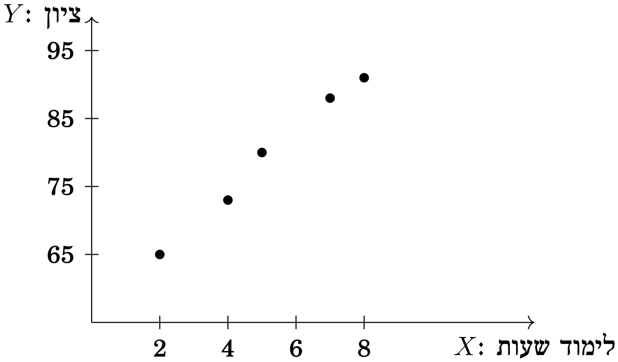
שעות לימוד : X,   ציון במבחן : Y

ציון Y	שעות לימוד X
65	2
72	4
80	5
88	7
90	8

שאלה

האם נראה שקיים קשר חיובי בין שעות לימוד לציון?

דיאגרמת פיזור – המחשה גרפית



ככל ש-X גדל – Y נוטה לעלות: **קשר חיובי**.

פרשנות – מה אפשר לראות?

מסקנות מהגרף

- ככל ששעות הלימוד רבות יותר – הציון גבוה יותר
- הנקודות מצביעות על מגמה חיובית
- הקשר אינו מושלם, יש "רעש" וסטייה

חשיבות

מציאת קו שמתאר את הקשר מאפשר:

- תיאור הנתונים
- חיזוי ציונים עתידיים

קו מגמה

רעיון

במקום לסמוך רק על נקודות, נרצה למצוא קו שמייצג את המגמה הכללית, והוא נקרא:

קו רגרסיה

שימושים

- חיזוי
- הערכה של השפעת X על Y
- זיהוי ערכים חריגים

רגרסיה – לא רק קו

רעיון חשוב

רגרסיה רלוונטית גם:

- כשיש יותר ממשתנה מסביר אחד
- כשקשר אינו ליניארי
- כשצריך מודל לחיזוי בזמן אמת

אבל בבגרות

במסגרת שאלון 471, אנו עוסקים בעיקר:

- בקשר בין שני משתנים
- בקו רגרסיה ליניארי
- בחיזוי פשוט

שאלה ברמת בגרות

בטבלת נתונים ניתנו X שעות לימוד ו-Y ציון במבחן.

- א. תאר במילים את הקשר בין X ל-Y
- ב. האם הקשר חיובי, שלילי, או שאין קשר?
- ג. האם הקשר מושלם? הסבר

פתרון מלא

א. תיאור מילולי

ככל שמספר שעות הלימוד עולה – הציון במבחן נוטה לעלות.

ב. סוג הקשר

הקשר **חיובי**.

ג. האם הקשר מושלם?

לא. הנקודות אינן על קו ישר אחד. קיימות סטיות, כך שיש **אי-סדירות** ו"רעש" בנתונים.

סיכום – מבוא לרגרסיה

- רגרסיה חוקרת קשר בין משתנים
- מטרות: תיאור, הבנה, חיזוי
- גרף פיזור הוא כלי חזותי ראשוני
- קשר חיובי/שלילי/ללא קשר
- קו רגרסיה – תיאור הקשר

למה זה חשוב לבגרות?

דרישה מרכזית בשאלון 471: **לפרש את הקשר ולהסבירו במילים.**

מהי דיאגרמת פיזור?

הגדרה

דיאגרמת פיזור היא ייצוג גרפי של זוגות נתונים (x, y) באמצעות נקודות על מערכת צירים.

מטרות

- לזהות כיוון של קשר
- להעריך עוצמה של קשר
- לגלות ערכים חריגים
- להחליט אם מתאים לבצע רגרסיה ליניארית

כיווני קשר בדיאגרמת פיזור

קשר חיובי

כאשר X גדל – Y גדל. הנקודות נוטות לעלות משמאל לימין.

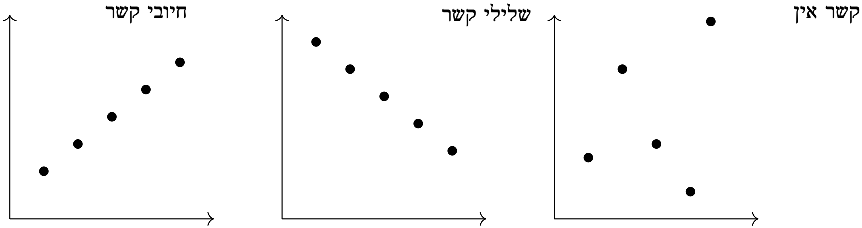
קשר שלילי

כאשר X גדל – Y קטן. הנקודות נוטות לרדת משמאל לימין.

אין קשר

הנקודות מפוזרות ללא תבנית עקבית.

איור: סוגי קשר



עוצמת קשר

רעיון

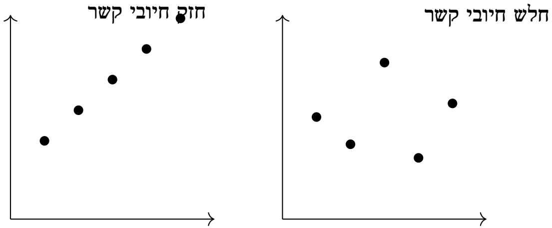
גם אם הכיוון ברור, העוצמה יכולה להשתנות:

- קשר חזק – נקודות קרובות לקו ישר
- קשר חלש – נקודות מרוחקות מהקו

השלכות

- קשר חזק □ אפשר לנבא היטב
- קשר חלש □ תחזיות פחות מדויקות

איור: קשר חזק מול חלש



קשר ליניארי ולא ליניארי

קשר ליניארי

הקשר נראה כמו קו ישר – עולה או יורד באופן עקבי.

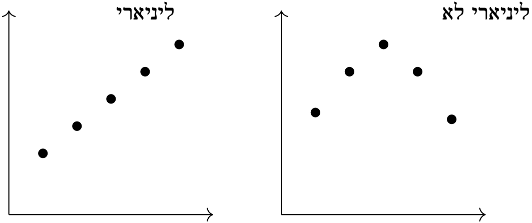
קשר לא ליניארי

הקשר עקמומי/מעוקל – דוגמה: צורת U, או עלייה ואחריה ירידה.

חשוב!

רגרסיה ליניארית מתאימה רק כאשר הקשר ליניארי.

איור: ליניארי מול לא ליניארי



ערכים חריגים בדיאגרמת פיזור

מה זה ערך חריג?

נקודה שנמצאת רחוק מאוד משאר הנקודות.

השלכות

- עלול לעוות את הקשר
- משפיע חזק על הממוצע
- משפיע על שיפוע קו הרגרסיה

מסקנה

לפני בניית מודל – לזהות ולהתייחס לחריגים.

שאלה ברמת בגרות

בדיאגרמת פיזור בין X ל־Y:

- הנקודות עולות משמאל לימין
- אך פזורות יחסית סביב הקו

שאלה

- א. האם הקשר חיובי או שלילי?
- ב. האם הקשר חזק או חלש?
- ג. מה ניתן לומר על יכולת החיזוי של X?

פתרון מלא

א. סוג הקשר

חיובי.

ב. עוצמת הקשר

חלש–בינוני, כי הנקודות מפוזרות.

ג. יכולת חיזוי

חלשה – קושי לבצע תחזית מדויקת.

סיכום – קשר בדיאגרמות פיזור

- דיאגרמת פיזור – כלי חזותי לזיהוי קשר
- כיוון: חיובי / שלילי / אין קשר
- עוצמה: חזק / חלש
- צורה: ליניארית / לא ליניארית
- זיהוי ערכים חריגים – חשוב מאוד

הערה לבגרות

בשאלון 471 נדרש לפרש דיאגרמות פיזור במילים, ולא רק לצייר אותן.

מהו מקדם המתאם?

הגדרה

מקדם המתאם הליניארי r הוא מספר בין -1 ל-1 המודד את עוצמת וכיוון הקשר הליניארי בין שני משתנים.

משמעות הערך

- $r > 0$ - קשר חיובי
- $r < 0$ - קשר שלילי
- $|r|$ קרוב ל-1 - קשר חזק
- $|r|$ קרוב ל-0 - קשר חלש/אין קשר

טווח הערכים האפשריים

הקצוות

- $r = 1$ - קשר חיובי מושלם
- $r = -1$ - קשר שלילי מושלם
- $r = 0$ - אין קשר ליניארי

הערה חשובה

$r = 0$ אינו אומר שאין קשר כלל – ייתכן קשר לא-ליניארי.

נוסחת חישוב (באופן רעיוני)

בבגרות לא נדרש לבצע חישוב מלא באמצעות הנוסחה, אבל חשוב להבין את הרעיון:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$

- מכנה – "נורמליזציה"
- מונה – מתאם בין סטיות של X לסטיות של Y

רעיון מרכזי

אם X עולה ו־Y עולה ביחד – $r > 0$ אם X עולה ו־Y יורד – $r < 0$

דוגמה – נתונים

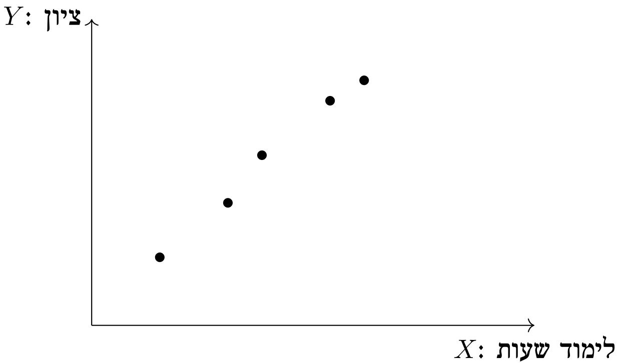
נבחן את הנתונים הבאים:

Y	X
65	2
72	4
80	5
88	7
90	8

שאלה

האם הקשר חזק או חלש?

דוגמה – הערכה חזותית של r



הנקודות קרובות לקו עולה – קשר חיובי **חזק**.

הערכת r - כלל אצבע

ערך r	פרשנות
1–0.8	קשר חזק מאוד
0.8–0.5	קשר בינוני-חזק
0.5–0.3	קשר חלש
0.3–0	כמעט אין קשר

הערה

ההערכות אינן מדויקות - אבל מקובלות בהוראה וביישומים בסיסיים.

דוגמה – פרשנות מילולית

נתון

נמצא שמקדם המתאם בין שעות לימוד לציון הוא:

$$r = 0.83$$

פרשנות

- הקשר חיובי חזק
- ככל ששעות הלימוד עולות – הציון נוטה לעלות
- ניתן לבצע חיזוי סביר

דוגמה – קשר שלילי

נתון

נמצא שמקדם המתאם בין שעות צפייה בטלוויזיה לציון במבחן:

$$r = -0.65$$

פרשנות

- הקשר שלילי בינוני-חזק
- ככל שמבלים יותר זמן מול הטלוויזיה – הציון יורד
- הקשר אינו מושלם – יש שונות בין התלמידים

שאלה ברמת בגרות

נמצא שמקדם המתאם בין משתנים X ו־Y הוא:

$$r = 0.12$$

- א. האם הקשר חזק או חלש?
- ב. מה הכיוון?
- ג. האם ניתן לבצע חיזוי טוב? הסבר.

פתרון מלא

א. עוצמה

קשר חלש מאוד.

ב. כיוון

חיובי.

ג. חיזוי

לא. הקשר חלש ולכן X לא מנבא היטב את Y.

סיכום – מקדם מתאם

- r מודד כיוון ועוצמת קשר
- בין -1 ל־1
- קרוב ל־1 או ל־-1 – קשר חזק
- קרוב ל־0 – קשר חלש
- לא להניח שהקשר ליניארי – לבדוק גרף!

לבגרות

נדרש:

- לתת **פרשנות מילולית** לערך r
- להסביר את משמעותו

זיהוי קשר מטבלה – הרעיון

כאשר הנתונים ניתנים בטבלה של זוגות (x, y), אפשר לזהות קשר ליניארי לפי:

- האם Y נוטה לעלות כאשר X עולה ⇐ קשר חיובי
- האם Y נוטה לרדת כאשר X עולה ⇐ קשר שלילי
- האם אין דפוס עקבי ⇐ אין קשר או קשר חלש

הערה פדגוגית

בבגרות על התלמידים לכתוב **פרשנות מילולית**, גם ללא גרף.

דוגמה 1 – טבלת נתונים

X	Y
1	10
2	14
3	18
4	22
5	25

שאלה

מה ניתן לומר על הקשר בין X ל-Y?

פתרון – דוגמה 1

- ככל ש-X עולה, גם Y עולה.
- העלייה יחסית עקבית.

מסקנה

קיים קשר **חיובי חזק** בין המשתנים.

הסבר

עלייה עקבית ב-Y מרמזת על קשר ליניארי חיובי.

דוגמה 2 – טבלת נתונים

Y	X
20	1
18	2
16	3
15	4
12	5

שאלה

מה ניתן לומר על הקשר בין X ל-Y?

פתרון – דוגמה 2

- ככל ש-X עולה, Y יורד.
- הירידה עקבית למדי.

מסקנה

קשר שלילי חזק.

משמעות

ניתן לחזות את Y לפי X ברמת דיוק טובה.

דוגמה 3 – טבלת נתונים

X	Y
1	30
2	45
3	40
4	47
5	42

שאלה

האם קיים קשר ברור בין X ל-Y?

פתרון – דוגמה 3

- התוצאות אינן עולות או יורדות באופן עקבי
- יש עליות וירידות אקראיות

מסקנה

הקשר חלש או לא קיים.

משמעות

קשה לבצע חיזוי או רגרסיה משמעותית.

דוגמה ברמת בגרות

לפניך נתונים:

X	Y
10	90
12	94
15	99
20	98

- א. תאר במילים את הקשר בין X ל-Y
- ב. האם הקשר חזק או חלש?
- ג. האם X מנבא היטב את Y?

פתרון מלא

א. תיאור הקשר

בדרך כלל, ככל ש-X עולה, Y נוטה לעלות, אך לא בצורה עקבית.

ב. עוצמת הקשר

קשר חיובי חלש-בינוני.

ג. יכולת חיזוי

חיזוי חלקי בלבד. יש סטיות רבות ולכן קיים קושי לנבא ערכים באופן מדויק.

כללי זיהוי מהירים לתלמידים

כיצד לזהות קשר בקלות?

- אם Y תמיד עולה כש-X עולה □ קשר חיובי
- אם Y תמיד יורד כש-X עולה □ קשר שלילי
- אם Y "קופץ" למעלה ולמטה □ קשר חלש
- ככל שהשינוי עקבי יותר □ קשר חזק יותר

טיפ לבגרות

שימו לב למילים "נוטה", "בדרך כלל", "באופן עקבי" – אלו מונחים שמעריכים!

סיכום – קשר בטבלאות

- ניתן לזהות קשר גם ללא ציור גרף
- ביטויים מילוליים חשובים מאוד
- זיהוי עוצמה – לפי עקביות
- זיהוי כיוון – לפי מגמה כללית

למה זה חשוב בבגרות?

לעיתים מוצגת רק טבלה – והתלמיד נדרש להסיק מסקנות מילוליות בלבד.

מהו קו רגרסיה?

הגדרה

קו רגרסיה ליניארית הוא קו המתאר את הקשר בין X ל-Y:

$$\hat{Y} = a + bX$$

כאשר:

- a – חיתוך הציר (ערך Y כש-$X = 0$)
- b – שיפוע (שינוי ב-Y עבור כל יחידה של X)

מטרה

הקו מאפשר חיזוי ערכים של Y על סמך ערכים של X.

משמעות השיפוע והחיתוך

שיפוע b

- $b > 0$ – קשר חיובי
- $b < 0$ – קשר שלילי
- $|b|$ גדול – שינוי חד

חיתוך a

הערך של Y כאשר $X = 0$ (לא תמיד יש לו משמעות פיזית).

נוסחאות חישוב

בבגרות לרוב ניתנים:

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{r \cdot s_y}{s_x}$$

כאשר:

- r – מקדם המתאם
- s_y – סטיית תקן של Y
- s_x – סטיית תקן של X

הערה חשובה לבגרות

ברוב המקרים נותנים את a ו־b או מאפשרים לחשב אותם בקלות, ולא מצפים לבצע חישוב מלא עם הנתונים.

דוגמה – נתונים

נבדקו 8 תלמידים ונמצא:

$$a = 30, \quad b = 5$$

שאלות

1. כתוב את משוואת קו הרגרסיה
2. חשב את הציון הצפוי לתלמיד שלמד 6 שעות

פתרון – דוגמה

1. משוואת הקו

$$\hat{Y} = 30 + 5X$$

2. חיזוי

עבור $X = 6$:

$$\hat{Y} = 30 + 5 \cdot 6 = 30 + 30 = 60$$

תשובה: הציון הצפוי הוא 60.

דוגמה – פרשנות מילולית

נניח שמצאנו:

$$\hat{Y} = 50 + 4X$$

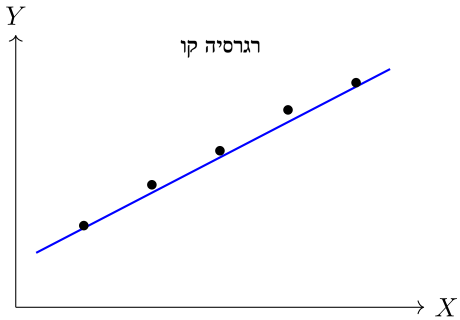
פרשנות

- כל שעה נוספת של לימוד מוסיפה **בממוצע 4 נקודות** לציון.
- תלמיד שאינו לומד כלל צפוי לקבל ציון: 50.

הערה

החיתוך אינו בהכרח מציאותי – אבל מייצג את המודל.

איור – קו רגרסיה



שאלה ברמת בגרות

הקו המתאר קשר בין X ל-Y הוא:

$$\hat{Y} = 40 + 3X$$

- א. מה משמעות השיפוע?
- ב. מה הציון הצפוי לתלמיד עם $X = 8$?
- ג. האם $a = 40$ מציאותי? הסבר.

פתרון מלא

א. משמעות השיפוע

כל שעה נוספת של לימוד מוסיפה בממוצע 3 נקודות לציון.

ב. חיזוי

$$\hat{Y} = 40 + 3 \cdot 8 = 64$$

תשובה: 64.

ג. מציאותיות

ייתכן שלא. המודל אומר שתלמיד שלא למד כלל יקבל 40 – אך זה לא בהכרח סביר במציאות. עם זאת, זו משמעות החיתוך במודל.

סיכום – קו רגרסיה

- קו רגרסיה: $\hat{Y} = a + bX$
- b – שינוי ב־Y לכל יחידת X
- a – ערך משוער של Y כאשר $X = 0$
- שימוש מרכזי – חיזוי

לבגרות

חשוב:

- לבצע חיזוי
- לפרש את a ואת b
- להסביר האם הפרשנות מציאותית

מהו ערך אמיתי ומהו ערך מנובא?

ערך אמיתי

הערך שנמדד בפועל בניסוי/סקר – מסומן לרוב Y.

ערך מנובא

הערך שחזינו באמצעות קו הרגרסיה – מסומן:

$$\hat{Y}$$

הוא אינו בהכרח זהה לערך האמיתי!

שגיאה (Residual)

הגדרה מתמטית

ההבדל בין הערך האמיתי למנובא נקרא:

$$e = Y - \hat{Y}$$

משמעות

- $e > 0$ - הערך גבוה מהתחזית
- $e < 0$ - הערך נמוך מהתחזית
- $e = 0$ - תחזית מדויקת

למה יש שגיאה?

- קו הרגרסיה הוא קירוב בלבד
- הנתונים מושפעים מרעש, משתנים נוספים, אקראיות
- **המטרה היא למזער את סכום השגיאות**

שיטה נפוצה: **ריבועי מינימום** – כדי למזער שגיאות באופן מיטבי.

דוגמה – קו רגרסיה

ניתן קו:

$$\hat{Y} = 50 + 4X$$

ונמדדו ערכים:

$$X = 5, \quad Y = 72$$

שאלה

חשבו את $\hat{Y}$ ואת השגיאה.

פתרון – דוגמה

1. **חיזוי**

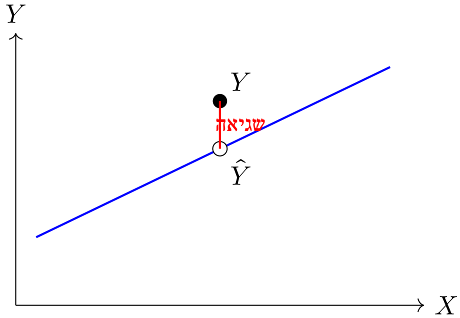
$$\hat{Y} = 50 + 4 \cdot 5 = 70$$

2. **שגיאה**

$$e = Y - \hat{Y} = 72 - 70 = 2$$

מסקנה: התחזית הייתה נמוכה ב־2 נקודות.

איור – שגיאה בגרף



שאלה ברמת בגרות

נתון קו רגרסיה:

$$\hat{Y} = 40 + 3X$$

ונמדדו:

$$X = 6, \quad Y = 70$$

- א. חשב את $\hat{Y}$
- ב. חשב את השגיאה
- ג. פירש את משמעות השגיאה

פתרון מלא

א. חיזוי

$$\hat{Y} = 40 + 3 \cdot 6 = 58$$

ב. שגיאה

$$e = Y - \hat{Y} = 70 - 58 = 12$$

ג. פירוש

התחזית הייתה נמוכה ב־12 נקודות מהערך בפועל. כלומר, המודל לא הצליח לנבא היטב את המקרה הזה.

סיכום – תצפיות אמיתיות ומנובאות

- Y – ערך אמיתי, $\hat{Y}$ – ערך מנובא
- שגיאה: $e = Y - \hat{Y}$
- ייתכן ש-e יהיה חיובי או שלילי
- מטרת הרגרסיה – למזער שגיאות

בבגרות

נדרש:

- לחשב $\hat{Y}$
- לחשב שגיאה
- לפרש אותה במילים

מהי טרנספורמציה?

הגדרה

טרנספורמציה היא פעולה מתמטית שמשנה את ערכי המשתנים, למשל:

- הכפלה במספר
- הוספת קבוע
- שינוי יחידות

שאלה מרכזית

כיצד טרנספורמציה משפיעה על:

- ערכי X או Y
- קו הרגרסיה: a, b

טרנספורמציה על Y

נניח שיש קו:

$$\hat{Y} = a + bX$$

ועושים טרנספורמציה:

$$Y' = kY$$

תוצאה

- החיתוך מוכפל ב־k
- השיפוע מוכפל ב־k

כלומר:

$$\hat{Y}' = ka + kbX$$

משמעות

הקשר נשמר, אך הסקאלה משתנה.

טרנספורמציה על X

נניח שיש קו:

$$\hat{Y} = a + bX$$

ועושים טרנספורמציה:

$$X' = kX$$

תוצאה

השיפוע משתנה כך:

$$b' = \frac{b}{k}$$

והחיתוך נשאר זהה:

$$a' = a$$

משמעות

שינוי יחידות ב-X משפיע על השיפוע בלבד.

טרנספורמציה – דוגמה

נניח:

$$\hat{Y} = 20 + 2X$$

והחלטנו למדוד X בשעות כפולות: $X' = 2X$

שאלה

מהו קו הרגרסיה החדש?

פתרון – דוגמה

לפי הכלל:

$$b' = \frac{b}{k} = \frac{2}{2} = 1$$

החיתוך:

$$a' = 20$$

תשובה

$$\hat{Y} = 20 + 1X'$$

הקשר נשאר זהה, אבל הערכים "נמתחו".

טרנספורמציה – פרשנות מילולית

רעיון מרכזי

טרנספורמציות לא משנות את כיוון ועוצמת הקשר, אלא רק את ייצוגו המספרי.

דוגמה

אם מדדנו משקל בגרמים במקום בקילוגרמים:

- השיפוע ישתנה
- הקשר הסטטיסטי לא ישתנה

מתי עושים טרנספורמציות?

- שינוי יחידות מידה
- התאמת נתונים להשוואה
- נרמול או סטנדרטיזציה
- שינוי סקאלה לצורך חיזוי

הערה

בבגרות – לא נדרש לבצע חישובים מסובכים, אבל חשוב להבין את ההיגיון.

שאלה ברמת בגרות

נתון קו רגרסיה:

$$\hat{Y} = 50 + 4X$$

נעשתה טרנספורמציה:

$$X' = 2X$$

- א. מצא את קו הרגרסיה החדש
- ב. הסבר במילים כיצד השתנה השיפוע

פתרון מלא

א. חישוב

$$b' = \frac{4}{2} = 2$$
$$a' = 50$$

לכן:

$$\hat{Y} = 50 + 2X'$$

ב. הסבר

כאשר מכפילים את X ב־2, השיפוע חצוי, כי כל יחידה ב־X' מייצגת שתי יחידות ב־X.

סיכום – טרנספורמציות ברגרסיה

- טרנספורמציות משנות סקאלה – לא את הקשר הסטטיסטי
- טרנספורמציה על Y – משפיעה על החיתוך והשיפוע
- טרנספורמציה על X – משפיעה על השיפוע בלבד
- חשוב להבין פרשנות מילולית ולא רק חישובית

לבגרות

נדרש להסביר **איך ומדוע** משתנים a ו־b משתנים. לא נדרש חישוב מורכב.