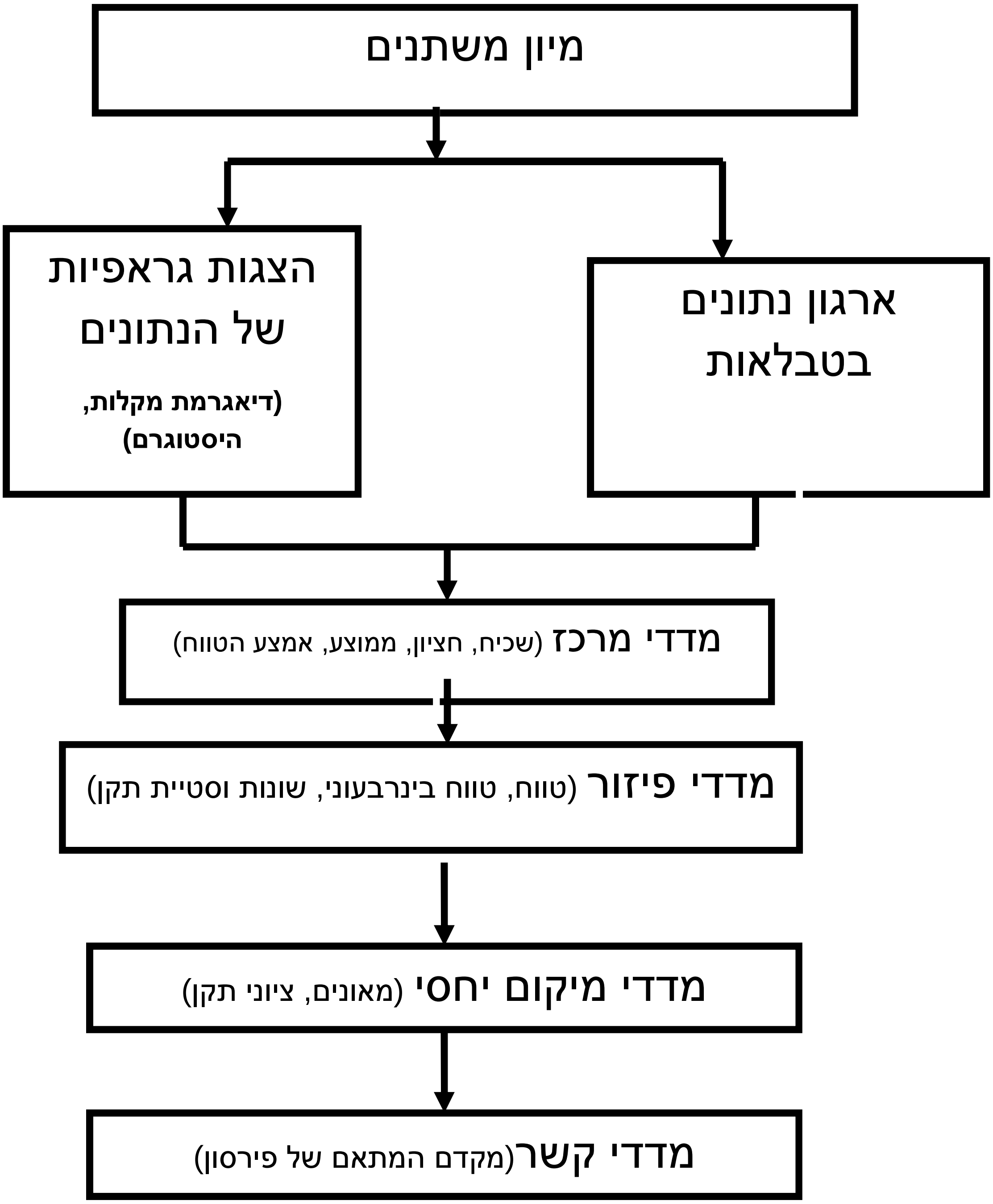
סיכום יחידה 5 – סטטיסטיקה תיאורית

מבנה היחידה



הצגת נתונים בטבלאות ובתרשימים גרפיים

מושגים

משתנה בדיד - משתנה כמותי היכול לקבל מספר סופי של ערכים בין כל שני ערכים אפשריים. למשל: מספר הנפשות במשפחה, מספר מכשירי T.V בדירה וכדומה.

משתנה רציף - משתנה כמותי היכול לקבל אינסוף ערכים בין כל שני ערכים אפשריים. למשל: גובה, משקל, טמפרטורה וכדומה.

שכיחות - מספר הפעמים שהתקבל ערך של המשתנה. השכיחות מסומנת ב - $f(x)$

התפלגות שכיחויות - טבלה שבה מצוינים ערכי המשתנה ולצד כל ערך שכיחות הופעתו.

שכיחות יחסית של מחלקה - שכיחות המחלקה מחולקת בגודל המדגם, $f(x)/n$.

שכיחות מצטברת - סה"כ השכיחות עד לגבול העליון של המחלקה. מסומנת ב- $F(x)$.

צפיפות- צפיפות היא השכיחות ליחידה של המשתנה הנחקר. הצפיפות שווה לשכיחות המחלקה המחולקת ברוחבה. הצפיפות מסומנת ב- d_i. בהיסטוגרמה הצפיפות היא גובה המלבן.

דיאגרמת מקלות - הצגה גראפית המתאימה למשתנה איכותי או למשתנה כמותי בדיד. על הציר האופקי X יירשמו ערכי המשתנה הנחקר, ועל הציר האנכי Y נציין את השכיחות המוחלטת של הערכים או שכיחות יחסית באחוזים . מעל כל ערך שעל ציר ה – X יוצב 'מקל' שגובהו פרופורציוני לשכיחות המקרים שהערך מייצג.

היסטוגרמה- הצגה גרפית של התפלגות שכיחויות עבור משתנה המקובץ במחלקות. הצגה זו בנויה ממלבנים כאשר המחלקה מיוצגת ע"י אורך של קטע – בסיס המלבן, והשכיחות מיוצגת ע"י שטח המלבן.

תיאור גרפי של צורת ההתפלגות:

	דיאגרמת מקלות	**היסטוגרם**
הנתונים	רשימת תצפיות	רשימת מחלקות
על ציר X על ציר Y	הערכים המדוייקים של התצפיות <u>שכיחויות</u> הערכים	התחומים של ערכי התצפיות <u>הצפיפות</u> של המחלקות
השכיחות	מתוארת ע"י גובה המקל	מתוארת ע"י שטח המלבן

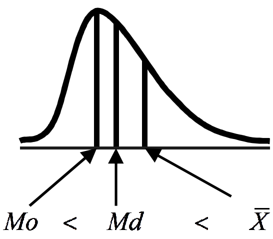
מדדי למיקום מרכזי (עמודים 22-34)

<table>
<tr><th rowspan="2">שם המדד</th><th rowspan="2">סימון</th><th colspan="3">חישוב המדד</th></tr>
<tr><th colspan="2">ברשימת תצפיות/טבלת שכיחויות משתנה בדיד</th><th>בטבלת שכיחויות - משתנה מקובץ במחלקות</th></tr>
<tr><td>שכיח</td><td>Mo</td><td colspan="2">ערך המשתנה בעל השכיחות הגבוהה ביותר</td><td>במחלקות שוות רוחב: השכיח הוא אמצע המחלקה בעלת השכיחות הגדולה ביותר.
במחלקות שונות רוחב: השכיח הוא אמצע המחלקה בעלת הצפיפות הגדולה ביותר.</td></tr>
<tr><td>חציון</td><td>Md</td><td colspan="2">$Md(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{n אי זוגי} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & \text{n זוגי} \end{cases}$</td><td>$Md = L_0 + \dfrac{\frac{n}{2} - F(x_{m-1})}{f(x_m)} \cdot (L_1 - L_0)$
ראה פירוט למטה</td></tr>
<tr><td>ממוצע חשבוני</td><td>$\bar{X}$</td><td><u>רשימת תצפיות</u>
$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$</td><td><u>טבלת שכיחויות</u>
$\bar{x} = \dfrac{\sum_x x f(x)}{n}$</td><td>$\bar{x} = \dfrac{\sum_x x_{mid} f(x)}{n}$
x_{mid} - נקודות האמצע של המחלקות</td></tr>
<tr><td>אמצע הטווח</td><td>MR</td><td colspan="2">$MR = \dfrac{x_{\min} + x_{\max}}{2}$</td><td>$MR = \dfrac{L_{\min} + L_{\max}}{2}$</td></tr>
</table>

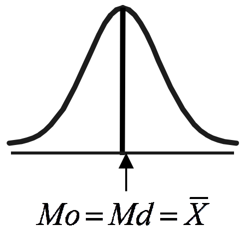
<table>
<tr><th>*</th><th><u>מרכיבי הנוסחה לחישוב חציון כשהמשתנה מקובץ במחלקות</u></th></tr>
<tr><td>n</td><td>מספר התצפיות (הנתונים)</td></tr>
<tr><td>x_m</td><td>המחלקה בה נמצא החציון (זו המחלקה שבה לראשונה $F(x_m) \geq n/2$)</td></tr>
<tr><td>L_0</td><td>גבול אמיתי תחתון של המחלקה x_m</td></tr>
<tr><td>$f(x_m)$</td><td>שכיחות המחלקה x_m</td></tr>
<tr><td>$F(x_{m-1})$</td><td>השכיחות המצטברת עד למחלקה x_{m-1} (המחלקה הקודמת ל- x_m)</td></tr>
<tr><td colspan="2">בנוסחה הנ"ל אין הבחנה בין n זוגי לאי זוגי. כדי להשתמש בנוסחה יש לעבוד בגבולות אמיתיים.</td></tr>
</table>

מיקום מדדי מיקום מרכזי (שכיח, חציון וממוצע) בצורות התפלגות שונות

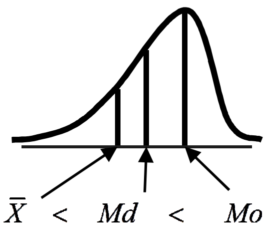
אסימטרית חיובית



סימטרית פעמונית



אסימטרית שלילית



תכונות מדדי מיקום מרכזי

המדד	תכונות
שכיח	• השכיח קל לחישוב. • תיתכן התפלגות עם יותר משכיח אחד (1,1,2,2,3), ותיתכן אף התפלגות שאין לה ערך שכיח (1,2,3,7,9) • השכיח אינו מושפע מערכים קיצוניים. • מוסר מעט אינפורמציה על הנתונים.
אמצע הטווח	• אמצע הטווח רגיש מאוד לערכים הקיצוניים בהתפלגות ורק להם.
חציון	• החציון אינו מושפע מערכים קיצוניים. • החציון מושפע מסדר הערכים ולא מהערכים עצמם. לכן החציון לא ישתנה כל עוד ה"מאזן" 50:50 מתחתיו ומעליו לא יופר. • כאשר התפלגות הנתונים **אסימטרית** עם נטיה חזקה לכיוון מסוים נעדיף את השימוש בחציון כמדד מרכזי על פני השימוש בממוצע.
ממוצע חשבוני	• הממוצע אינו חייב להיות אחד מערכי הסדרה. למשל: מס' הילדים הממוצע במשפחה ישראלית הוא 3.5 , אך לא קיימת משפחה שיש לה 3.5 ילדים. • הממוצע אינו יכול להימצא מחוץ לטווח הנתונים. למשל ממוצע המספרים 20,42,38,50,29 חייב להיות בין 20 ל – 50. • הממוצע מושפע מכל הערכים ובפרט מערכים קיצוניים. • אם מוסיפים לסדרת ערכים ערך קטן מהממוצע, הממוצע יקטן, אם מוסיפים ערך גדול מהממוצע, הממוצע יגדל, אם מוסיפים ערך השווה לממוצע, הממוצע לא ישתנה. • סכום ההפרשים של כלל הערכים בסדרת נתונים מהממוצע שלהם שווה תמיד ל**אפס** כלומר $\sum_{i=1}^{n}\left(x_i-\overline{x}\right)=0$. במילים אחרות: סכום ההפרשים מן הממוצע בסימן חיובי שווה לסכום ההפרשים מן הממוצע בסימן שלילי (בערך מוחלט).

מדדי פיזור (עמודים 35-42)

מדדי הפיזור מציינים את מידת פיזור הנתונים בהתפלגות בשני אופנים: טווח הפיזור או פיזור ביחס לממוצע.

- **טווח** (Range) - ההפרש בין הערך הגבוה ביותר בהתפלגות לבין הערך הנמוך ביותר התפלגות.
- **טווח בינרבעוני** (Interquartile Range) - טווח הכולל את 50% הערכים הנמצאים במרכז ההתפלגות בין הרבעון התחתון (Q_1) לרבעון העליון (Q_3).
- **שונות** (Variance) - ממוצע של ריבועי הסטיות מן הממוצע.
- **סטיית תקן** (Standard Deviation) - השורש הריבועי (החיובי) של השונות.

<table dir="rtl">
<tr><th rowspan="2">שם המדד</th><th rowspan="2">סימון</th><th colspan="3">חישוב המדד</th></tr>
<tr><th colspan="2">רשימת תצפיות/טבלת שכיחויות
משתנה בדיד</th><th>טבלת שכיחויות - משתנה מקובץ
במחלקות</th></tr>
<tr><td>טווח</td><td>R</td><td colspan="2">$R = x_{\max} - x_{\min}$</td><td>$R = L_{\max} - L_{\min}$</td></tr>
<tr><td>טווח בין רבעוני</td><td>IQR</td><td colspan="2"></td><td>$Q_1 = L_0 + \dfrac{\frac{n}{4} - F(x_{m-1})}{f(x_m)} \cdot (L_1 - L_0)$
$Q_3 = L_0 + \dfrac{\frac{3n}{4} - F(x_{m-1})}{f(x_m)} \cdot (L_1 - L_0)$
$\Downarrow$
$IQR = Q_3 - Q_1$</td></tr>
<tr><td>שונות</td><td>S^2</td><td><u>רשימת תצפיות</u>
$S_x^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$
או
$S_x^2 = \dfrac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$</td><td><u>טבלת שכיחויות</u>
$S_x^2 = \dfrac{\sum_x (x - \bar{x})^2 f(x)}{n}$
או
$S_x^2 = \dfrac{\sum_x x^2 f(x)}{n} - \bar{x}^2$</td><td><u>טבלת שכיחויות</u>
$S_x^2 = \dfrac{\sum_x (x_{mid} - \bar{x})^2 f(x)}{n}$
או
$S_x^2 = \dfrac{\sum_x x_{mid}^2 f(x)}{n} - \bar{x}^2$
x_{mid} - נק' האמצע של המחלקות</td></tr>
<tr><td>סטיית תקן</td><td>S</td><td colspan="2">$S_x = +\sqrt{S_x^2}$</td><td>$S_x = +\sqrt{S_x^2}$</td></tr>
</table>

תכונות של מדדי הפיזור

המדד	תכונות
טווח	• הטווח מושפע **רק** מהערכים הקיצוניים בהתפלגות (min/max) , ולא מתחשב בפיזור בין יתר הערכים.
טווח בינרבעוני	• הטווח הבינרבעוני מתחשב רק ב- 50% מהנתונים הנמצאים במרכז ההתפלגות (ומתעלם ממחצית מנתוני ההתפלגות הנמצאים בקצוות). • כאשר ההתפלגות אסימטרית עם נטיה חזקה לכיוון מסוים נעדיף את השימוש בטווח הבינרבעוני על פני סטיית התקן.
שונות /סטיית תקן	• השונות נמדדת ביחידות של המשתנה **בריבוע**. למשל במדידת גובה בס"מ יחידות המדידה של השונות הן ס"מ^2. מאידך סטיית התקן נמדדת ביחידות של המשתנה. • סטיית התקן מבטאת **בקירוב** את המרחק הממוצע של הנתונים מממוצע ההתפלגות. • סטיית התקן מתחשבת בסטיות **כל** הנתונים מן הממוצע ונותנת משקל יחסי מתאים לכל ערך בהתאם למרחקו מן הממוצע ובהתאם לשכיחותו. כל סטיה תורמת כריבוע גודלה, כך שלסטיות של ערכים קיצוניים יש השפעה גדולה יותר על המדד.

השפעת טרנספורמציה ליניארית על מדדי המיקום המרכזי ומדדי הפיזור

נתונה סדרת ערכים : $x_1, x_2, x_3, ...x_n$. אם נבצע על כל ערך בסדרה את הפעולה הבאה $bx_i + a$ (הכפלה בקבוע b והוספת קבוע a) נקבל את הסדרה: $bx_1 + a,\ \ bx_2 + a,\ \ bx_3 + a,\ ....,\ bx_n + a$.

הפעולה שביצענו היא: $x' = bx + a$ נקראת **טרנספורמציה ליניארית** (המרה קווית).

- כאשר $b = 1$ השינוי הליניארי כולל הוספת[1] קבוע a בלבד לכל ערך בסדרה $(x' = x + a)$.
- כאשר $a = 0$ השינוי הליניארי כולל הכפלת כל ערך בסדרה פי b בלבד[2] $(x' = bx)$.

בביצוע שינוי ליניארי מהצורה $x' = bx + a$ נוכל לקבל את המדדים של x' מתוך המדדים של x מבלי לבנות את ההתפלגות של x' באופן הבא:

מדדי מיקום מרכזי	מדדי פיזור
$Mo_{X'} = b \cdot Mo_X + a$	$R_{X'} = b \cdot R_X$
$Md_{X'} = b \cdot Md_X + a$	$S_{X'} = \|b\| \cdot S_X$
$\overline{x'} = b \cdot \overline{x} + a$	$S^2_{X'} = b^2 \cdot S^2_X$
$MR_{X'} = b \cdot MR_X + a$	

- **הוספת** קבוע a ל**כל** הערכים בהתפלגות תגרום להגדלת כל מדדי המיקום המרכזי ב $- a$.
  הפחתת קבוע a מ**כל** הערכים בהתפלגות תגרום להקטנת כל מדדי המיקום המרכזי ב $- a$.
  מדדי הפיזור **אינם מושפעים** מהוספה או הפחתת קבוע.
- **הכפלת** כל הערכים בהתפלגות בקבוע חיובי b תגרום **להכפלת** כל מדדי המיקום המרכזי פי b .
  גם מדדי הפיזור, למעט השונות, יוכפלו פי b . (השונות תוכפל פי b^2)

[1] או הפחתת קבוע (a שלילי)
[2] למשל, אם x מייצג משקל בק"ג ו- x' משקל בגרמים אזי $x' = 1000x$.

חישוב ממוצע וסטיית התקן כאשר מאחדים שתי קבוצות או יותר

נתונות שתי קבוצות וידועים הגדלים הבאים:

	I	II
מספר נתונים	n_1	n_2
ממוצע	$\bar{x}_1$	$\bar{x}_2$
שונות	s_1^2	s_2^2

מאחדים את שתי הקבוצות לקבוצה אחת. מטרתנו היא לחשב את הממוצע המשוקלל והשונות המצורפת של הקבוצה המאוחדת כשידועים לנו הנתונים לעיל.

ממוצע משוקלל של שתי הקבוצות: $$\boxed{\bar{x}=\frac{n_1\cdot\bar{x}_1+n_2\cdot\bar{x}_2}{n_1+n_2}}$$

שונות מצורפת של שתי הקבוצות: $$\boxed{S_c^2=\frac{n_1\cdot\left(\bar{x}_1^2+s_1^2\right)+n_2\cdot\left(\bar{x}_2^2+s_2^2\right)}{n_1+n_2}-\bar{x}^2}$$ $$S_c=+\sqrt{S_c^2}$$

שים לב!

- כאשר שתי הקבוצות שוות גודל $\left(n_1=n_2\right)$, הממוצע המשוקלל הוא הממוצע ה**פשוט** של שתי הקבוצות - $\bar{x}=\frac{\bar{x}_1+\bar{x}_2}{2}$
- כאשר ממוצעי שתי הקבוצות שווים, השונות המצורפת היא ממוצע משוקלל של שונויות שתי הקבוצות - $S_c^2=\frac{n_1\cdot s_1^2+n_2\cdot s_2^2}{n_1+n_2}$
- ניתן להכליל את הנוסחאות הנ"ל גם ל- k קבוצות:

ממוצע משוקלל: $$\boxed{\bar{X}=\frac{\sum_{j=1}^{k}n_j\cdot\bar{x}_j}{n}}$$

שונות מצורפת: $$\boxed{S_c^2=\frac{\sum_{j=1}^{k}n_j\cdot\left(\bar{x}_j^2+s_j^2\right)}{n}-\bar{X}^2}$$

מקרא: k - מספר הקבוצות, j - אינדקס הקבוצות, n_j - מספר הנתונים בקבוצה ה- j, $\bar{x}_j$ - ממוצע הקבוצה ה- j. n - גודל הקבוצה המצורפת: $n=\sum_{j=1}^{k}n_j$

מדדי מיקום יחסי (עמודים 43-48)

ה**מאון** ה – k הוא ערך של המשתנה (C_K) ש $k\%$ מהנתונים בהתפלגות קטנים ממנו (או שווים לו).

נוסחאות לחישוב C_K ו- K_C כשנתונה טבלת שכיחויות של משתנה המקובץ במחלקות:

נוסחה לחישוב C_K (k ידוע) - $$\boxed{C_K=\frac{\frac{n\cdot k}{100}-F(x_{m-1})}{f(x_m)}\cdot(L_1-L_0)+L_0}$$

נוסחה לחישוב K_C (C_K ידוע) - $$\boxed{K_C=\left[\frac{C_K-L_0}{L_1-L_0}\cdot f(x_m)+F(x_{m-1})\right]\cdot\frac{100}{n}}$$

הערות:

- שיטת החישוב בעזרת הנוסחה זהה לזו שבנוסחת החציון בטבלת שכיחויות של משתנה מקובץ.
- כדי להשתמש בנוסחה יש לעבוד בגבולות אמיתיים.

ציוני תקן

ציון התקן של X מסומן ב – Z_x ומבטא בכמה סטיות תקן רחוק X מממוצע ההתפלגות.

נוסחה לחישוב - $$\boxed{Z_x=\frac{x-\bar{x}}{s_x}}$$

תכונות של ציוני תקן:

- ציוני התקן הם מספרים טהורים, בלתי תלויים ביחידות המדידה, ולכן ניתן להשוות בעזרתם מיקום יחסי של תצפיות מהתפלגויות של משתנים הנמדדים ביחידות שונות (למשל: משקל בק"ג עם גובה בס"מ או שכר בש"ח עם וותק בשנים).
- $Z_x>0$ כאשר התצפית נמצאת <u>מעל</u> הממוצע $x>\bar{X}$
  $Z_x<0$ כאשר התצפית נמצאת <u>מתחת</u> לממוצע $x<\bar{X}$
  $Z_x=0$ כאשר התצפית <u>שווה</u> לממוצע $x=\bar{X}$
- ציוני תקן הם שינוי ליניארי על ערכי הסדרה: $Z_x=\frac{1}{s_x}\cdot x-\frac{\bar{x}}{s_x}$ $\left(b=\frac{1}{s_x}\ ,\ a=-\frac{\bar{x}}{s_x}\right)$
- הממוצע של ציוני התקן שווה תמיד לאפס $(\bar{z}=0)$, והשונות/סטיית התקן שלהם שווה תמיד ל-1 $(S_Z^2=S_Z=1)$.

מדדי קשר (עמודים 49-58)

מקדם המתאם של פירסון – r

מדד לעוצמת הקשר הליניארי בין שני משתנים כמותיים

טווח הערכים של המדד: $\boxed{-1 \le r \le +1}$

חישוב r :

(I) $$\boxed{r = \frac{\operatorname{cov}(x,y)}{s_x \cdot s_y}}$$

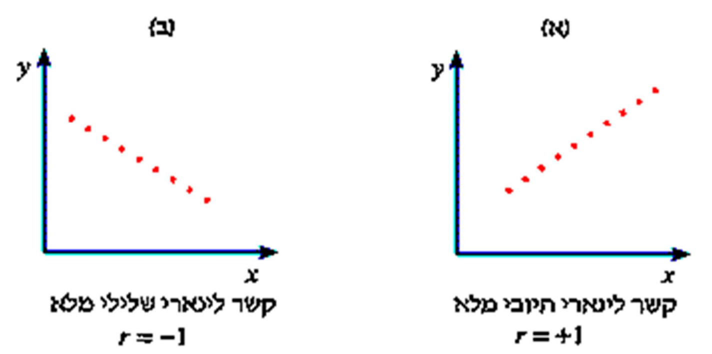
נוסחאות נוספות שקולות:

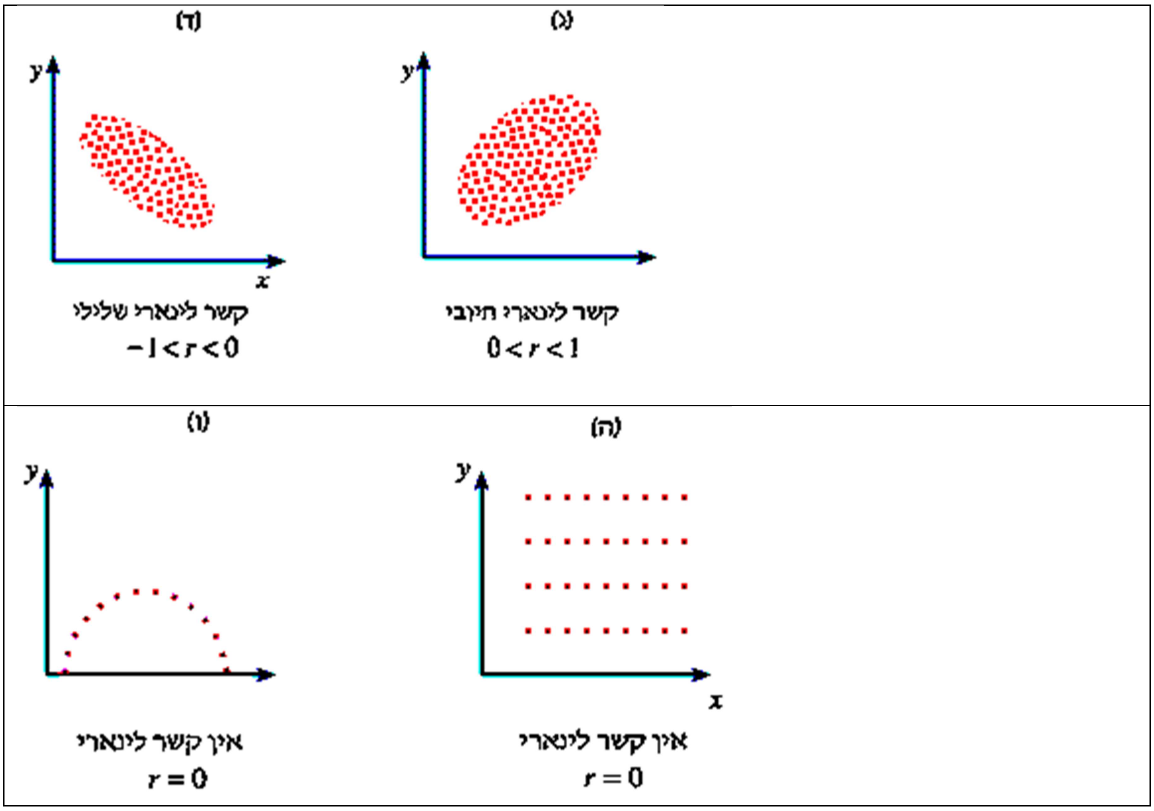
(II) $$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

(III) $$r = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

<u>דיאגרמות פיזור וערכו של R</u>

מטרת דיאגרמת הפיזור היא לקבל **רושם ראשוני** מהי **צורת הקשר** בין שני המשתנים (ליניארי/לא לינארי), מהי **מגמתו** (חיובית/שלילית) ומהי **עוצמתו** (חלשה/חזקה).
להלן מספר דוגמאות לדיאגרמות פיזור . לצד כל דיאגרמה מצוין ערכו המתאים של r .





במקרים א' ו-ב' כל התצפיות – הנקודות בדיאגרמת הפיזור – מונחות על קו ישר אחד. הקשר הוא קשר ליניארי מלא.

במקרה א' הקשר הוא במגמה חיובית ובמקרה ב' - במגמה שלילית.

במקרה ג' התצפיות מפוזרות סביב קו ישר במגמה חיובית. הקשר הוא ליניארי לא מלא.

במקרה ד' התצפיות מפוזרות סביב קו ישר במגמה שלילית. הקשר הוא ליניארי לא מלא. עוצמת הקשר הליניארי ב – ד' גבוהה מזו שב – ג' כיוון שהנקודות "קרובות" יותר לקו.

במקרה ה' אין קשר בין שני המשתנים

במקרה ו' אין קשר ליניארי בין שני המשתנים אך אין זה אומר שאין קשר בין שני המשתנים. מהתרשים אפשר לראות שיש קשר לא ליניארי (ממעלה שניה).

תכונות של r

1. טווח הערכים של המדד: $-1 \le r \le +1$

2. כאשר $r=+1$ הקשר בין x ל- y **חיובי מלא**. כל הנקודות בדיאגרמת הפיזור נמצאות על קו ישר אחד עולה. (ראה תרשים א). בין המשתנים מתקיים הקשר $b>0\ ;\ y=bx+a$.

   כאשר $r=-1$ הקשר בין x ל- y **שלילי מלא**. כל הנקודות בדיאגרמת הפיזור נמצאות על קו ישר אחד יורד. (ראה תרשים ב). בין המשתנים מתקיים הקשר $b<0\ ;\ y=bx+a$.

3. ככל ש – r גדל בערכו המוחלט ומתקרב ל- 1, הקשר הליניארי בין שני המשתנים **חזק יותר**.
4. אם $r=0$, פירוש הדבר ש**אין מתאם ליניארי**, אך ייתכן שיש קשר אחר לא ליניארי.

5. **השפעת טרנספורמציה לינארית על מקדם המתאם:**
   נתונים שני משתנים X ו- Y. ומקדם המתאם ביניהם: $r_{X,Y}$
   נגדיר : $X^{'}=b\cdot X+a$ ו - $Y'=c\cdot Y+d$ (טרנספורמציה לינארית)
   אזי מקדם המתאם ביניהם (לאחר הטרנספורמציה\יות) יהיה: $r_{X'Y'}=\frac{b\cdot c}{|b\cdot c|}\cdot r_{X,Y}$

6. מקדם המתאם הוא מספר טהור ואינו תלוי ביחידות המדידה.

שונות משותפת

השונות המשותפת של x ו- y - $\operatorname{cov}(x,y)$ **- מדד למידת ההשתנות המשותפת של שני משתנים כמותיים.**

נוסחאות חישוב:
$$\boxed{\operatorname{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n}=\frac{\sum_{i=1}^{n}x_iy_i}{n}-\bar{x}\cdot\bar{y}}$$

תכונות:

1. אם $\operatorname{cov}(x,y)>0$ אזי $r>0$
   אם $\operatorname{cov}(x,y)<0$ אזי $r<0$
   אם $\operatorname{cov}(x,y)=0$ אזי $r=0$
2. $\operatorname{cov}(x,x)=S_X^2$

3. השונות המשותפת תלויה ביחידות המדידה של המשתנים.