
**פרופסור תלמה לויתן**

ביה"ס למדעי המתמטיקה

אוניברסיטת תל-אביב

# מתאם וניבוי

# קשרים בין משתנים

כתיבה: **פרופסור תלמה לויתן**, ביה״ס למדעי המתמטיקה, אוניברסיטת תל-אביב

מנהל פרויקט **כדאי לדעת**: פרופסור אילון סולן

עריכת לשון: ד״ר דנה ברעם

ריכוז אדמיניסטרטיבי: גלית הרצברג

עיצוב גרפי ועימוד: ניצן-שמיר מעצבים

איור ועטיפה: shutterstock

# תוכן עניינים

# הקדמה

## ➢ למה סטטיסטיקה?

הסטטיסטיקה כיום היא תחום מעניין ודינמי. המפגש בין כוח המחשוב המואץ לבין שיטות ניתוח מתקדמות מעצים מאוד את תרומתה של הסטטיסטיקה להבנת העולם שסביבנו. כמעט כל תחום של חיינו קשור לרעיונות סטטיסטיים – רפואה, כלכלה, פסיכולוגיה, תעשייה, היי-טק ועוד ועוד. גם בחיי היום-יום אנו נתקלים יותר ויותר במושגים הלקוחים משפת הסטטיסטיקה. ללא רקע מתאים אין לנו דרך להבין את המשמעות, להפיק תועלת מהמידע ולהגיע להחלטות נכונות. הסטטיסטיקה מעניקה לנו כלים מועילים להבנה טובה יותר של העולם סביבנו. יש, על כן, תועלת רבה ברכישת השפה הסטטיסטית ובהבנת הרעיונות היסודיים של הסטטיסטיקה מוקדם ככל האפשר.

לימודי הסטטיסטיקה שלנו עד כה התמקדו בניתוח נתונים הנוגעים למשתנה יחיד: גובה, משקל, גיל, ציון במתמטיקה וכדומה. רכשנו כלים שונים: למדנו לקרוא ולהכין טבלאות שכיחות של נתונים, לצייר דיאגרמות שכיחות, לחשב ערכים מרכזיים כגון ממוצע כמדד למיקום ערכי המשתנה, שונות כמדד לפיזור, ועוד.

נושא חשוב שבו עוסקת הסטטיסטיקה הוא **זיהוי קשרים** בין המרכיבים השונים של מערכות מורכבות, **מדידת חוזק הקשרים**, וניסיון **לנבא** את ערכו של מרכיב מסוים בעזרת מרכיבים אחרים הניתנים למדידה. בכך נעסוק בספר זה.

יחידת הלימוד 'מתאם וניבוי' עוסקת בזוג משתנים כמותיים ובאופן מדידת **הכיוון והעוצמה של הקשר** ביניהם. למשל, הקשר בין גובה ומשקל; בין מנת המשכל (IQ) ובין הציון במתמטיקה; בין תוצאות סקרי הבחירות ובין תוצאות הבחירות עצמן.

בספר נענה על סוגיות כגון אלה:

– האם הצלחה בלימודים בבית הספר התיכון יכולה לנבא הצלחה בלימודים באוניברסיטה?

– האם יש קשר בין בעיות במהלך הלידה של תינוק ובין הפרעות קשב בהמשך חייו?

– האם יש קשר בין מידת המעורבות של תלמידי בית ספר ברשתות חברתיות ובין ציוניהם?

– האם יש קשר בין עישון ובין מחלות ריאה? בין רמת הכולסטרול למחלות לב?

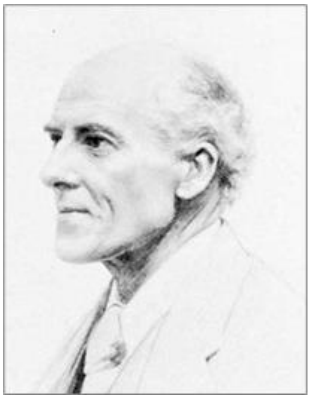
אם יימצאו קשרים כאלו, נרצה בהמשך גם **לנבא** את ערכו של אחד המשתנים מתוך ידיעת ערכו של המשתנה האחר. בדוגמאות שנציג בספר זה ננתח נתוני אמת וכך נכיר את השימושים המעשיים המגוונים של המושגים החדשים.

## ➢ מתאם וניבוי – היסטוריה על קצה המזלג

הראשון שעסק בקשרים בין משתנים היה סר פרנסיס גלטון (Francis Galton, 1822–1911), שהתעניין בתכונות העוברות בתורשה בקרב בני אדם ובמידת הדמיון בין ילדים להוריהם. בהמשך ההקדמה נציג את הנתונים המקוריים שאסף – גבהים של 1078 אבות ושל בניהם הבוגרים. את הנתונים הללו ניתח ממשיכו

קרל פירסון (Carl Pearson, 1857–1936) – מאבות הסטטיסטיקה – שגילה תופעות מעניינות ומפתיעות. ניתוחיו של פירסון היו הבסיס לתחום חשוב בסטטיסטיקה – שנקרא **רגרסיה**.

פרנסיס גלטון (Francis Galton, 1822–1911)



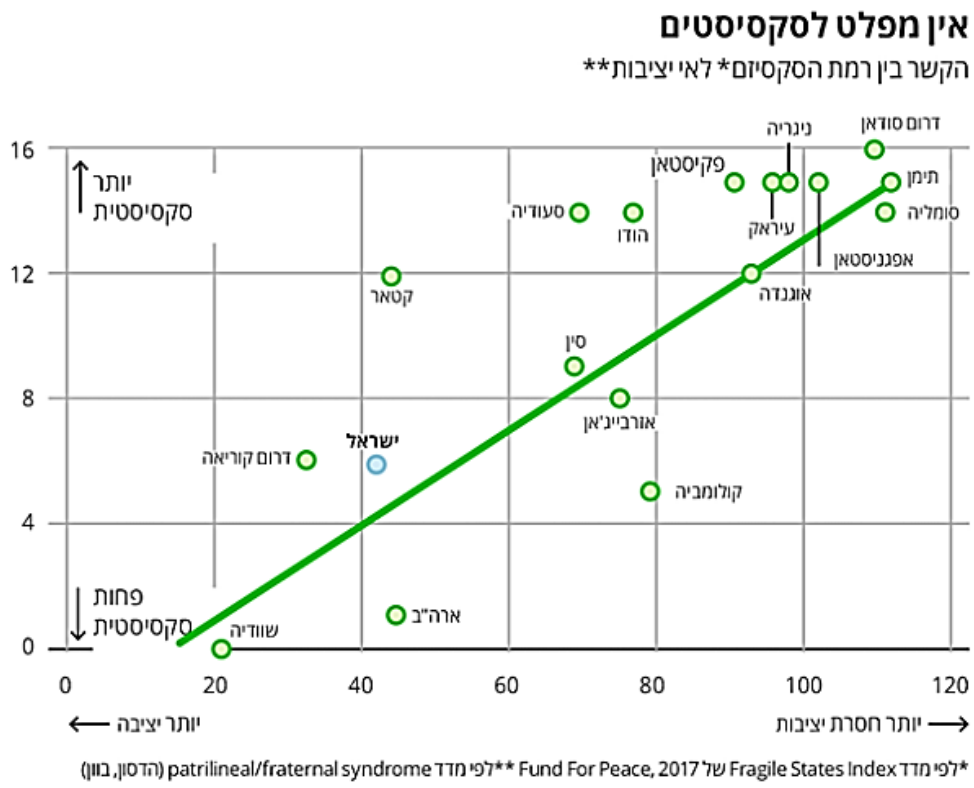
קרל פירסון (Carl Pearson, 1857–1936)

### ➢ מתאם וניבוי – מהיסטוריה לאקטואליה

כתבות פופולריות בתקשורת כוללות פעמים רבות גרפים המציגים קו ישר מעל פיזור של נקודות. המודל הקווי המוצג בגרפים כאלו הוא חלק מהכלים של טכניקה סטטיסטית רבת עוצמה, הנקראת **רגרסיה ליניארית**. רגרסיה ליניארית עוסקת ב**קשרים בין משתנים**, ומטרתה לזהות קשרים ומגמות ולחזות ערכים עתידיים של משתנים.

הגרף הבא לקוח מכתבה בעיתון הארץ (14.9.21), שתורגמה מהאקונומיסט. הגרף המקורי מדרג 176 מדינות בעולם בסולם 'סקסיזם', המשלב גורמים כמו אפליית נשים בחוק, פשיעה אלימה נגד נשים, גיל נישואים מוקדם לנשים, מנהגי המוהר ועוד.

הגרף שלפנינו חוקר קשרים בין רמת הסקסיזם בחברה ובין יציבות המדינה.




ביחידה זו נכיר את המונחים החשובים והעיקריים בתחום: מקדם מתאם בין שני משתנים, קשר קווי (ליניארי) עולה או יורד, קשר סטטיסטי, קשר סיבתי. כמו כן נעסוק בבעיות ניבוי באמצעות עקום הממוצעים ובאמצעות קו הרגרסיה, ואף נגלוש מעט לניבוי באמצעות עקומת ניבוי כלשהי. בסוף הספר נציג גם מדד לטיב הניבוי.

פרק חשוב בספר עוסק **בכשלי חשיבה**, שאת חלקם תיאר כבר פירסון.

# מדריך למורי מתמטיקה

הספר שלפנינו נכתב בעבור מורי מתמטיקה (ותלמידיהם), במטרה להקנות הבנה עמוקה בנושאים החשובים שנעסוק בהם. בספר הקפדנו על כתיבה ידידותית ובסיסית מבלי לוותר על הדיוק המתמטי.

– עושר הדוגמאות והמשימות הרבות אומנם מְשַוֶה לספר אופי מסורבל מעט, אך בו בזמן הוא מאפשר למורה לבחור דוגמאות לניתוח ולתרגול על פי צרכיו וטעמו האישי.

– הטקסט כולו מלווה בשאלות חשיבה שאותן מומלץ לברר במשותף עם התלמידים. כבר בפתח הספר מובאת דוגמה היסטורית שהובילה לפיתוח התורה כולה. עיון משותף בכיתה בדוגמה מאלפת זו ודיון בשאלות החשיבה המוצגות בה יכול להיות פתיח יעיל ליחידת ההוראה כולה.

– עיקר הכלים החישוביים כלולים בפרק 3, שמוצג בו מקדם המתאם, ובפרק 5, שבו מוצג קו הרגרסיה ונדון אופן הניבוי. שאר הפרקים מניחים את היסוד להבנת משמעותם וחשיבותם המעשית של כלים אלו.

– בדוגמאות ובמשימות מוצגים נתונים ממחקרי אמת, כדי להדגיש את השימושים המעשיים המגוונים של הכלים הנלמדים.

– בסוף כל פרק מובא אוסף עשיר של משימות שכל אחת מהן היא תרגיל בית יעיל להטמעת הרעיונות והכלים של הפרק. בסוף הספר צירפנו קובץ של פתרונות מפורטים למרבית המשימות, שהתלמידים יקבלו במועד המתאים למורה. שימוש נבון בפתרונות אלו מאפשר למורה לספק לתלמידים תוצאות של חלק מהחישובים ובכך להקל על המשימות.

בסוף כל פרק מובאים גם תרגילים קצרים ופשוטים יותר, שלחלקם מצורפת תשובה סופית. תרגילים נוספים – ללא פתרונות – אפשר למצוא בחוברת התרגילים הנלווית לספר זה.

– חלק גדול מהדוגמאות והמשימות בספר הן דוגמאות 'מתגלגלות', אשר ילוו אותנו שוב ושוב. אנו חוזרים לאותם נתונים ומתקדמים בניתוח שלהם על פי הרעיונות והכלים החדשים שנרכשים. באופן זה, תשומת הלב של הלומדים אינה מוסטת מהמאמץ הכרוך בעיבוד מושגים חדשים אל הצורך לנתח נתונים חדשים. אנו מקווים שהדוגמאות המעשיות המגוונות יעודדו את המורים (ואת התלמידים בהדרכת המורים) לחפש נתונים בנושאים נוספים שמעניינים אותם, ולנתח נתונים אלה באופן דומה.

– בחרנו דוגמאות שהנתונים בהן אינם רבים מדי, ולכן יכולנו להציג את כלל נתוני הבעיה, לתת להם תיאור גרפי ולערוך **ידנית** את החישובים שנדרשו. עם זאת, בבעיות מעשיות החישובים הם מייגעים והעיסוק בהם מיותר. היכרות עם מספר מצומצם מאוד של פקודות בתוכנה פשוטה כגון **אקסל** מייתר את העבודה הטכנית. בנספח ב בסוף הספר ריכזנו את כל הפקודות הנדרשות וגם הדגמנו את השימוש בהן מפעם לפעם.

לטעמנו הדרך המומלצת ללימוד הנושא כולו מבוססת על שימוש בתוכנה חישובית דוגמת אקסל.

– בסעיף 5.4 ובנספח ד מוצג מדד מקובל ונוח ליישום ל**טיב הניבוי.** נושא זה אינו כלול בתוכנית הלימודים המקובלת, אך עם זאת התרגילים שבו פשוטים מאוד להתמודדות וקל לכלול אותם במבחנים.

– תוכני **הנספחים** אינם מכוּוָנים להוראה הישירה. מטרתם להעמיק את הידע ואת ההבנה של מורי התוכנית וכן לסייע להם לנהל שיח ער עם תלמידים סקרנים וחקרנים.

היחידה 'מתאם וניבוי' היא המשך ישיר ליחידת הלימוד בסטטיסטיקה לכיתה י. עם זאת, כדי לא להסתמך על ידע מוקדם הוספנו בפרק 1 חזרה על כל המושגים ביסודות הסטטיסטיקה שנזדקק להם בהמשך. את החומר הכלול ביחידת הלימוד המקדימה אפשר למצוא בספר הדיגיטלי "סטטיסטיקה תיאורית בגובה העיניים" בהוצאת כותר, שכתבו אלונה רביב ותלמה לויתן.

## פירוט נושאי הפרקים

**פרק 1** פותח בסקירה קצרה של משתנה כמותי יחיד. נציג בו את הממוצע כמדד למיקום ערכי המשתנה ואת סטיית התקן כמדד לפיזורם. בהמשך מוצג גם רעיון ה**תְּקָנון** של משתנה, שמטרתו לקבל ערכים שאינם תלויים ביחידת המדידה של המשתנה (ציוני תקן). בפרקים הבאים יתברר שרעיון זה שימושי מאוד.

כל שאר פרקי הספר עוסקים בזוג משתנים כמותיים ו**בקשרים** ביניהם.

**פרק 2** עוסק בהצגות גרפיות. הכלי העיקרי של ההצגה הגרפית הוא **דיאגרמת פיזור**.

ב**פרק 3** מודגם כיצד שימוש בציוני תקן ובדיאגרמת פיזור מתוקננת מאפשר להגדיר הגדרה קלה ובהירה את המושג **מקדם מתאם** בין משתנים. מוצעות בו נוסחאות חישוב מקובלות, וכן נסקרות תכונות חשובות של מקדם המתאם.

ב**פרק 4** מנוסחת בעיית **ניבוי** ערכו העתידי של משתנה $Y$ על בסיס ערכו של משתנה אחר – משתנה מנבא, שערכיו ידועים לנו. בפרק מוצגת אבן הבוחן לטיב הניבוי, שהיא **עקרון הריבועים הפחותים**.

**פרק 5** עוסק כולו ב**קו הרגרסיה** – הניבוי הקווי הטוב ביותר למשתנה $Y$ על בסיס ערכו של משתנה מנבא. נציג נוסחאות חישוב מקובלות לקו הרגרסיה, נציע תיאור גרפי ונדגים את השימוש בקו המתקבל לצורך ניבוי ערכו העתידי של המשתנה $Y$.

ב**פרק 6** נעסוק בכשלים **לוגיים** בחיי היום-יום שעניינם קשרים בין משתנים, וכן נעסוק בקריאה ביקורתית של דיווחים בתקשורת ושל מאמרים פופולריים.

**הנספחים** מכוּוָנים, כאמור, למורים עצמם. הם יחליטו אם לשלב תכנים מתוכם בהוראה ובאיזה אופן.

בספר השתמשנו בסמלים אלה:

❤ שימו לב – מדגיש את עיקרי הדברים;

מסמן מוקשים חשיבתיים נפוצים;

מציין נקודות עקרוניות עדינות שיש לתת עליהן את הדעת;

✓ מציין תזכורת לדברים שכבר הוצגו בעבר;

מציין שימוש בתוכנת אקסל.

תודה לאלונה רביב על איורים רבים בספר. תודה גם על האישור הנדיב להשתמש בחומרים מתוך הספר שכתבו אלונה רביב ותלמה לויתן, "סטטיסטיקה תיאורית בגובה העיניים". תודה לאורית גת על עזרה בפתרון המשימות.

תודה גם לבית הספר למתמטיקה באוניברסיטת תל אביב על העזרה בעת הכנת הספר.

הערות, הארות, תיקונים והצעות לשיפור הספר יתקבלו בתודה.

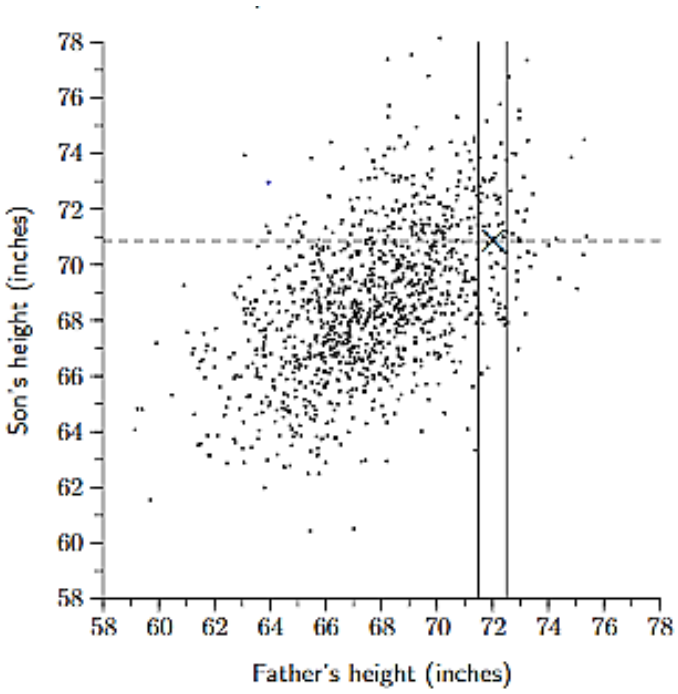
***תלמה לויתן***, talmale@tauex.tau.ac.il

## על תורשה ונסיגה – שאלות חקר[1]

**שאלה**: האם יש קשר בין גובה אבות לגובה הבנים הבוגרים שלהם?

ניסיון החיים שלנו מורה שאכן יש קשר כזה – קשר עולה: **הנטייה** היא שלהורים גבוהים ייוולדו ילדים גבוהים.

כבר לקראת סוף המאה ה-19 רותקו המדענים לגנטיקה – חקר תכונות שעוברות בתורשה. בשנת 1903 פרסם הסטטיסטיקאי החשוב פירסון נתונים של 1,078 משפחות (מספר עצום ורב לתקופתו); לכל משפחה ציין זוג מספרים – גובה הבן הבכור וגובה האב, שנמדדו באינצ'ים[2]. הוא הציג כל אחד מזוגות הנתונים שהתקבל כנקודה במישור. התקבלה דיאגרמה זאת:




גובהי הבנים הבכורים וגובה האבות על פי מחקרו של פירסון

מה אנו למדים מדיאגרמה זו? האם היא תומכת בהשערה הראשונית שלנו?

לפנינו ענן של נקודות שמסתמנת בו **נטייה כללית לעלייה** של גובה הבן עם העלייה בגובה האב. עם זאת, ברור לנו מהדיאגרמה – וגם מניסיון החיים שלנו – שהקשר בין שני המשתנים אינו מוחלט: נתבונן למשל באבות שגובהם כ-72 אינצ' (180 ס"מ) ונראה שיש פיזור של גובהי הבנים המתאימים; ממוצע הגבהים מסומן באיור ב-X.

---

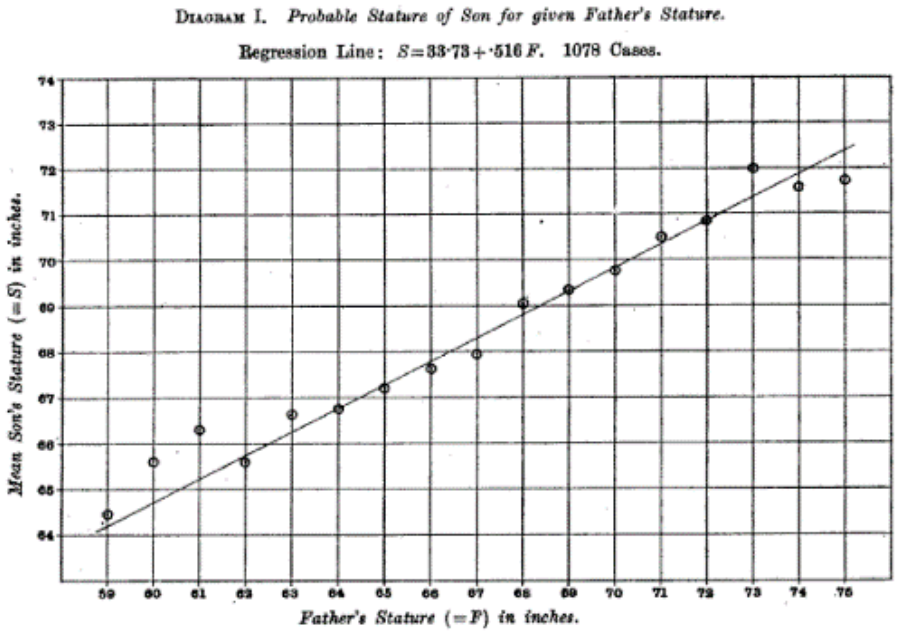
[1] מומלץ לפתוח את הלימוד של יחידה זו בדיון קבוצתי בשאלות אלו. כדאי ללוות את הדיון בדיאגרמות ההיסטוריות המוצגות כאן.

[2] אינצ' היא יחידת מידה מקובלת בבריטניה, והיא שווה ל- 2.54 ס"מ.

**חומר למחשבה**: ציירו **אליפסה** הדוקה שתקיף את מרבית נקודות הדיאגרמה שלמעלה. האם מדובר באליפסה צרה, כלומר קשר חזק, או באליפסה רחבה – כלומר קשר רופף יותר?

**שאלה**: האם על פי הדיאגרמה ממוצע גובהי הבנים עולה עם העלייה בגובהי האבות?

באיור הבא מוצגים הממוצעים של גובהי הבנים (מסומנים בעיגולים) על פי גובה האבות, ואנו רואים בו נטייה ברורה של **עלייה קווית**. באיור מופיע גם **קו הניבוי**, שעליו נדון בהרחבה בחוברת זו.




ממוצע גובהי הבנים ביחס לגובה האבות על פי ממצאיו של פירסון

**שאלה**: האם נצפה שהבנים של אבות גבוהים מאוד יהיו גם הם גבוהים מאוד? מה באשר לאבות נמוכים מאוד? כאן טמונה הפתעה. כפי שניווכח בהמשך, לאבות גבוהים מאוד צפויים אומנם בנים גבוהים מאוד – אבל יחסית גבוהים פחות מהאבות; ובדומה לכך, לאבות נמוכים מאוד צפויים בנים נמוכים מאוד – אבל פחות נמוכים מהאבות. במילים אחרות, בגובה הבנים התגלתה תופעה של **נסיגה מערכי קיצון לכיוון ממוצע** הגובה. באיור רואים בבירור שממוצע גובה הבנים כשגובה האבות הוא 74 או 75 אינץ' נמוך יותר מהממוצע שלהם כשגובה האבות הוא 73 אינץ'.

פירסון בחן דוגמאות נוספות ובכולן התגלתה תופעה דומה. לדוגמה:

– רץ ששבר שיא אולימפי, בריצה הבאה ההישג שלו יהיה כנראה קרוב יותר לממוצע ההישגים שלו.

– טיפול רפואי שהתגלה שיש לו תוצאות נפלאות בניסוי, כנראה יהיה פחות נפלא בניסוי הבא.

– מדריכי טיס טענו שכאשר הם משבחים את חניכיהם על ביצוע טוב במיוחד, הביצוע הבא נופל ממנו, וכאשר הם נוזפים בחניכים על ביצוע גרוע במיוחד, הביצוע שאחריו טוב יותר. מסקנת המדריכים הייתה ששבח מוביל לפגיעה בביצוע בעתיד ואילו גערה מובילה לשיפור. האומנם?

את התופעה המפתיעה שתיארנו בדוגמאות הללו כינה גלטון בשם "רגרסיה (נסיגה) אל הממוצע", וזוהי גם הסיבה **שקו הניבוי** כונה "קו רגרסיה". בהמשך נגלה שהתופעה היא למעשה טבעית והיא מתגלית שוב ושוב בתחומים רבים. חשוב אפוא להכיר את התופעה ולהבין אותה כדי להימנע **מכשלי חשיבה** נפוצים.

**חומר למחשבה**: חִשבו על זוגות משתנים נוספים שעל פי דעתכם יש ביניהם קשר עולה או יורד. חִשבו על תופעת הנסיגה לממוצע בקשר למשתנים שבחרתם. עוד נחזור ונדון בכך בהרחבה בפרק 6.

ביחידת לימוד זו נעסוק בכל השאלות שהעלינו כאן, נכיר גם מדד לחוזק הקשר בין משתנים. כמו כן נעסוק באפשרות לנבא את ערכו העתידי של משתנה כלשהו על בסיס משתנה מנבא שערכו ידוע לנו.

# פרק 1
# ידע מקדים

## 1.1 מדדי מיקום ופיזור של משתנה כמותי

בידינו $n$ מדידות של **משתנה כמותי** כלשהו $X$. תוצאות $n$ המדידות מסומנות $x_1,...,x_n$.

נאמר לעיתים ש- $x_1,...,x_n$ הן **תצפיות** על המשתנה $X$. הסטטיסטיקה מציעה מדדים שונים לאפיון תכונות של המשתנה:

- המדד המקובל ביותר ל**מיקום** ערכי המשתנה שבידינו הוא ***הממוצע*** שלהם $\bar{x}$:

$$\bar{x} = \frac{1}{n} \cdot \underbrace{\left(x_1 + \cdots + x_n\right)}_{\downarrow}$$

**סכום כל המספרים ברשימה החל ב- $x_1$ וכלה ב- $x_n$**

- המדד השימושי ביותר ל**פיזור** של ערכי המשתנה הוא **השונות**, או למעשה **סטיית התקן**:

***השונות*** (מסומנת $\sigma^2$, קרי סיגמה בריבוע) של רשימת מספרים $x_1,...,x_n$ היא **הממוצע של ריבועי הסטיות של ערכי הרשימה מהממוצע שלהם**:

$$\sigma^2 = \frac{1}{n}\underbrace{\left[(x_1-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right]}_{\downarrow} = \underbrace{\frac{1}{n}\left[x_1^2 + \cdots + x_n^2\right]}_{\downarrow} - \underbrace{\bar{x}^2}_{\downarrow} \qquad (0)$$

**סכום ריבועי הסטיות מהממוצע $\bar{x}$** **ממוצע הריבועים** **ריבוע הממוצע**

מכיוון שהשונות נמדדת ביחידות שהן ריבוע היחידה של המשתנה המקורי, נהוג לחשב את השורש של השונות, וכך מקבלים את ***סטיית התקן***: $\sigma = \sqrt{\sigma^2}$.

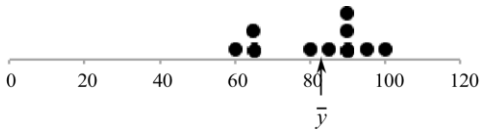
הממוצע והשונות מתקבלים מייד בעזרת מחשבון כיס.

**דוגמה 1: ציוני פרויקט וציוני תעודה.** מורה הציע לתלמידיו להכין גם פרויקט, נוסף על המבחן הסופי. הוא הבטיח שציון הפרויקט יהיה ציון מגן, כלומר בתעודה יופיע הגבוה מבין שני הציונים. בלוח 1 מפורטים הציונים של עשרת התלמידים.

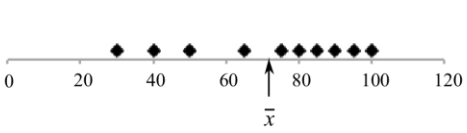
נסמן: $X$ – ציוני הפרויקט, $Y$ – ציוני התעודה.

מדדי המיקום: ממוצעי הציונים, שהתקבלו בעזרת מחשבון, הם: $\bar{x} = 71$ ו- $\bar{y} = 82$.

תיאור גרפי: באיור שמימין מוצגים ציוני הפרויקט ובאיור משמאל הציונים בתעודה. בכל איור סימנו בחץ את הממוצע, שהוא נקודת שיווי המשקל.




**ציוני תעודה**




**ציוני פרויקט**

**לוח 1**. נתוני ציון הפרויקט והציון בתעודה של עשרת התלמידים

| התלמיד | ציון פרויקט, $x$ | ציון תעודה, $y$ |
|---|---|---|
| 1 | 30 | 60 |
| 2 | 50 | 90 |
| 3 | 40 | 65 |
| 4 | 95 | 95 |
| 5 | 100 | 100 |
| 6 | 90 | 90 |
| 7 | 75 | 80 |
| 8 | 65 | 65 |
| 9 | 85 | 85 |
| 10 | 80 | 90 |
| סה"כ | 710 | 820 |
| ממוצע | $\bar{x} = 71$ | $\bar{y} = 82$ |

**מדדי הפיזור – חישוב ידני** : בטור האפור בלוח 2 בהמשך רשמנו את הסטיות של ערכי המשתנה $X$ מהממוצע המתאים. העלאה בריבוע וחישוב ממוצע ריבועי הסטיות הללו נותן את השונות של $X$. הוצאת שורש השונות נותנת את סטיית התקן $\sigma_X$.

**לוח 2**. חישוב ידני של השונות של שני המשתנים על פי שתי נוסחאות החישוב

| התלמיד | $x-71$ | $(x-71)^2$ | $y^2$ |
|---|---|---|---|
| 1 | -41 | 1681 | 360 |
| 2 | -21 | | |
| 3 | -31 | | |
| 4 | 24 | | |
| 5 | 29 | | |
| 6 | 19 | | |
| 7 | 4 | | |
| 8 | 6 | | |

| | | | |
|---|---|---|---|
| 9<br>10 | 14<br>9 | | |
| סה״כ | 0 | | |
| **ממוצע** | 0 | $\sigma_x^2 =$ | $\frac{1}{10}\left[y_1^2+\cdots+y_{10}^2\right]=$ |

## מטלות

**א.** השלימו בלוח 2 את שתי העמודות השמאליות, כולל הסכום והממוצע של הערכים בכל עמודה. היעזרו בעמודה האחרונה כדי לקבל את השונות של $Y$.

**ב.** חשבו את שתי סטיות התקן ישירות בעזרת מחשבון, והשוו לתוצאת החישוב הידני.

[תשובה: סטיות התקן הן: $\sigma_X = 22.78$ ו- $\sigma_Y = 13.27$. באיורים נשים לב שאכן ערכי הציונים בתעודה $Y$ מרוכזים יותר סביב הממוצע שלהם.]

## 1.2 ציוני תקן

בדוגמה 1 שהצגנו למעלה נמדדו שני המשתנים באותן יחידות, וכך מתאפשרת השוואה ביניהם. אך מה עלינו לעשות אם נרצה להשוות ערכי משתנים שנמדדו ביחידות שונות, כמו למשל ציון התעודה של תלמיד בהשוואה למספר השעות שהוא מבלה בצפייה במסכים? לצורך השוואות כאלה יש להשתחרר מהתלות ביחידת המדידה.

***ציון התקן*** של ערך $x$ של משתנה כמותי בעל ממוצע $\bar{X}$ וסטיית תקן $\sigma_X$ הוא $\boxed{\frac{x-\bar{x}}{\sigma_X}}$.

כדי לקבל את ציון התקן, חיסרנו מערך המשתנה $X$ את הממוצע וחילקנו בסטיית התקן. מכיוון שהמונה והמכנה הם בעלי אותה יחידה, המנה שלהם נטולת יחידה.

❤ ציון התקן מודד **בכמה סטיות תקן גבוה (או נמוך) ערך מסוים $x$ מממוצע ערכי המשתנה**.

✋ הערך של ציון התקן הוא **נטול יחידה**. מכאן, ציוני התקן מאפשרים להשוות את **המיקום היחסי של פרט מסוים** בתכונות שונות, גם אם נמדדו ביחידות מידה אחרות.

**דוגמה 2: משקל וגובה של אייל הפעוט.** אימו של אייל בן השנה מודאגת מעגלגלותו של הפעוט. במדידות שנערכו בטיפת חלב התברר שמשקלו של אייל הוא 11.9 קילוגרם, בעוד המשקל הממוצע בגיל זה הוא 10.2 קילוגרם. האם יש לאם סיבה לדאגה? חשבו ציון תקן של המשקל.

התבוננות בנתוני משרד הבריאות מעלה שסטיית התקן של המשקל בגיל זה היא 1.05 קילוגרם. ציון התקן של משקל הפעוט הוא אפוא $\frac{11.9-10.2}{1.05}=1.62$, דהיינו 1.62 סטיות תקן (די הרבה) מעל הממוצע.

לאימו של אייל יש נתון נוסף: אורכו 79 ס״מ. היא בדקה ומצאה שהאורך הממוצע של תינוק בגיל זה הוא 76 ס״מ. כלומר אייל גם ארוך מהממוצע. כיצד בודקים אם המשקל חריג גם בהתחשב באורך?

בדיקה העלתה שסטיית תקן של האורך בגיל זה היא 2.7 ס״מ. בעזרת כל הנתונים הללו היא חישבה את ציון התקן של האורך וקיבלה: $\frac{79-76}{2.7}=1.11$. דהיינו 1.11 סטיות תקן מעל הממוצע.

**שורה תחתונה**: הן משקל הפעוט הן אורכו חורגים כלפי מעלה. על כן, כשלוקחים בחשבון את האורך, המשקל אינו כה חריג. ובכל זאת, מידת החריגות של המשקל גבוהה יותר $1.62>1.11$, ולאייל יש משקל עודף מעט.

– העובדה שציון התקן הוא נטול יחידה אפשרה להשוות את מידת החריגות של המשקל לזו של האורך.

– מכיוון שגובה ומשקל הם בעלי התפלגות נורמלית, אפשר היה במקרה זה גם להשוות את **האחוזונים** של הערכים שבדקנו. בחוברת זו כלל לא נעסוק בהתפלגות של משתנים או באחוזונים.

## אז מה היה לנו? מדדי מיקום ופיזור, ציוני תקן

**כלים**

- ממוצע – מדד למיקום: $\bar{x}=\frac{1}{n}\cdot\left(x_1+\cdots+x_n\right)$.
- שונות – מדד לפיזור:

$$\sigma^2=\underbrace{\frac{1}{n}\left[x_1^{\,2}+\cdots+x_n^{\,2}\right]}_{\text{ממוצע הריבועים}}-\underbrace{\bar{x}^2}_{\text{ריבוע הממוצע}} \quad = \quad \frac{1}{n}\underbrace{\left[(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right]}_{\text{סכום ריבועי הסטיות מהממוצע } \bar{x}} \qquad (0)$$

**תובנות**

כדי לעבור מערך המשתנה $x$ **לציון התקן** שלו יש לערוך **פעולת תקנון** זו: $x\to\frac{x-\bar{x}}{\sigma_X}$.

– מתקבל מספר נטול יחידה.

– שינוי ליניארי בקנה המידה של המשתנה, למשל מקילומטר למייל או ממעלות צלזיוס למעלות פרנהייט, לא ישנה את ציון התקן.

– ציוני התקן מאפשרים להשוות את **מידת החריגות** של ערכי תכונות שנמדדו ביחידות שונות.

– ציון התקן יהיה **חיובי** לערכים שמעל ממוצע האוכלוסייה ו**שלילי** לערכים שמתחת לממוצע.

– ממוצע ציוני התקן של התכונה הוא 0, סטיית התקן היא 1.

– ככל שהערך חריג יותר כלפי מעלה או מטה, כך רחוק ציון התקן שלו מ-0.

– ציון התקן של פרט נותן מושג על מיקומו היחסי בהתפלגות המשתנה. נוסיף ונאמר,

– ברוב המקרים הערכים המתוקננים נעים בין (2-) ל-2. ערכים מתוקננים שמעל 3 או מתחת ל-(3-) הם נדירים.

**כלים חישוביים**

- מטרת הסקירה בפרק זה היא להיזכר בנוסחאות הבסיסיות של הסטטיסטיקה ובמשמעותן. עם זאת, בכל המשך הספר כדאי להימנע ככל האפשר משימוש בנוסחאות אלו לחישובים ידניים, אלא רצוי להשתמש במחשבונים כדי לקבל את הממוצע ואת השונות של רשימת נתונים.
- למעשה, הדרך המומלצת לפתרון המשימות שבפרקים הבאים היא להשתמש בתוכנות חישוביות פשוטות. בספר זה ההדרכה היא לשימוש בתוכנת אקסל, בשל פשטותה וזמינותה בכל מחשב (ראו נספח ב בעמ' 103).

## משימות חישוב וחשיבה: מיקום, פיזור וציוני תקן

(פתרונות בעמ' 108)

**משימה I**

מראיין נדרש לבחור אחד מבין שני מועמדים (נקרא להם מועמד 1 ומועמד 2) למשרה חדשה. ההכרעה אמורה להתבסס, בין השאר, על ציוני הגמר של התואר הראשון. התברר שציון הגמר של מועמד 1 היה 84 ושל מועמד 2 – 85. מכיוון שהמראיין לא רצה להחליט על חודה של נקודה, הוא פנה לאוניברסיטאות שהמועמדים למדו בהן וקיבל את המידע הזה:

– מועמד 1 למד באוניברסיטה A, שציון הגמר הממוצע בה היה 80, עם סטיית תקן של 7.5 נקודות.

– מועמד 2 למד באוניברסיטה B, שגם בה ציון הגמר הממוצע היה 80 אך סטיית התקן הייתה 5 נקודות.

א. תקננו את הציונים של שני המועמדים.

ב. מי משני המועמדים הצטיין יותר בלימודיו יחסית לממוצע הציונים באוניברסיטה שבה למד?

ג. לבוגרי אוניברסיטה B, כְּתבו מתחת לכל ציון מתוקנן את הציון המתאים ביחידות המקוריות. ערכו טבלה דומה לבוגרי אוניברסיטה A. מה גילית? נזכיר, יש להכפיל בסטיית התקן ולהוסיף את הממוצע.

| ביחידות מתוקננות | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| ביחידות מקוריות | | | | |

**משימה II: ציון פרויקט וציון תעודה**

א. האם מנתוני לוח 1 (עמ' 16) אפשר לדעת למי מהתלמידים שיפר הפרויקט את הציון הסופי? הסבירו. למי מהתלמידים הפרויקט לא עזר כלל?

ב. בלוח 3, חשבו את ציוני התקן של שני המשתנים והשלימו את שתי העמודות האחרונות.

**לוח 3.** חישוב ציוני התקן של המשתנים

| התלמיד | ציון תקן פרויקט $\frac{x-71}{22.78}$ | ציון תקן תעודה $\frac{y-82}{13.27}$ | כיוון ביחס לממוצע המתאים | איזה ציון חריג יותר ביחס לכיתה |
|---|---|---|---|---|
| 1 | -1.8 | -1.66 | שניהם מתחת לממוצע | פרויקט |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | -0.26 | -1.28 | | |

| | | | | |
|---|---|---|---|---|
| | | | | 9<br>10 |

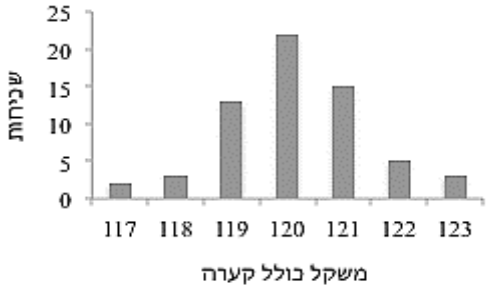
ג. חשבו בעזרת מחשבון את הממוצע ואת השונות של ציוני התקן.

ד. אם ציון הפרויקט היה מוצג בסקלה של 1 עד 10 (יש לחלק כל ציון ב-10), כיצד היו משתנים ממוצע הפרויקט, הסטיות מהממוצע וסטיית התקן של המשתנה? כיצד היו משתנים ציוני התקן המתאימים?

ה. לתלמיד מספר 8 ציון הפרויקט וציון התעודה זהים. השוו את הערכים המתוקננים של שני הציונים. איזה מהם חריג יותר בהשוואה לשאר הכיתה? מה גיליתם? הסבירו. השלימו את העמודה האחרונה.

**משימה III: תכונות הממוצע, סטיית התקן וציוני התקן**

בדיאגרמה הבאה מוצגת התפלגות המשקל (בגרמים) של 60 חפיסות שוקולד בטהובן כפי שנשקלו במכון התקנים. משקלן הנקוב הוא 100 גרם, אך להפתעת הבוחנים, המשקל שנמדד לכל חפיסה היה גבוה מאוד, בוודאי גבוה מ-100 גרם. בבדיקה התברר שהמשקל שנרשם כלל גם את משקל הקערית (20 גרם) שבה הונחו חפיסות השוקולד.




א. ממוצע המשקל שנמדד היה 120.1 גרם. סמנו נקודה זו בחץ בדיאגרמה – זוהי נקודת שיווי המשקל.

ב. מהו ממוצע המשקל נטו? מה הקשר בין שני הממוצעים הללו? נסו למצוא כלל בדבר **השינוי בממוצע הנובע מהוספה או מהחסרה של ערך קבוע** לערכי המשתנה.

ג. השונות מודדת פיזור סביב הממוצע. ענו ללא חישובים מה הקשר בין שונות המשקל ברוטו לשונות נטו. נסו למצוא כלל בדבר **השינוי בסטיית התקן הנובע מהוספה או מהחסרה של ערך קבוע**.

ד. באיור – הַציגו מתחת לסקלת המשקל ברוטו סקלה נוספת של המשקל המתאים נטו (החסירו את 20 הגרמים המיותרים). כך קיבלתם את הדיאגרמה של התפלגות המשקל נטו. מהי עתה נקודת שיווי המשקל?

ה. למבחני טעימות לוקחים מדי יום ביומו עשירית מהמשקל שנותר בכל חפיסה. מהו המשקל הממוצע של החפיסות לאחר יום הטעימות הראשון? לאחר היום השני? שימו לב שאחרי כל יום טעימות מוכפל משקל כל חפיסה ב-0.9. נסו למצוא כלל בדבר **השינוי בממוצע הנובע מהכפלה בקבוע**.

ו. באיור – הַציגו מתחת לסקלת המשקל נטו סקלה נוספת של המשקל לאחר יום הטעימות הראשון – הכפילו את המספרים ב-0.9. קיבלתם את דיאגרמת התפלגות המשקל בסוף אותו יום. מהי עתה נקודת שיווי המשקל?

ז. האם סטיית התקן של המשקל בסוף כל יום גדלה, קטנה או אינה משתנה? נסו לחשוב על כלל בדבר **השינוי בסטיית התקן הנובע מהכפלה בקבוע**.

ח. נסו להסיק מה קורה **לציוני התקן** של המשקל בכל השלבים הללו.

למשימה נוספת ראו שאלה I בשאלון הסיכום (עמ' 139 ).

# פרק 2
# ניתוח גרפי

לניתוח הקשר בין זוג משתנים כמותיים, הנתונים שבידינו הם $n$ תצפיות שבכל אחת מהן נמדדו ערכי שני המשתנים. כך מתקבלים $n$ זוגות של מספרים.

**דוגמאות**

– נמדדים גובה ומשקל של פעוטות. לפעוטה שי, לדוגמה, נמדד משקל 5.230 קילוגרם ואורך 54.5 ס״מ; כך נרשום את התצפית: (5.230, 54.5).

– בסקר בריאות, לכל משתתף נבדקה מידת העישון וכן צוין אם חלה או לא חלה בסרטן הריאות.

– נבדקו ציוני הבגרות וציוני שנה א של סטודנטים באוניברסיטה.

– נרשם השכר ומספר שנות הלימוד (לחלופין: הוותק בעבודה) של מועסקים.

– נבדק מספר השוטרים ומספר מקרים של שוד מזוין בערים שונות בארצות הברית.

מהדוגמאות שהצגנו עולה שהקשרים שנעסוק בהם אינם קשרים שבהם ערכו של אחד המשתנים נקבע לחלוטין על ידי המשתנה האחר, כמו למשל הקשר בין מהירות הנסיעה ובין המרחק שנעבור. לקשרים שבהם אנו עוסקים בסטטיסטיקה נכנס **גורם של אקראיות**, עקב הבדלים אינדיבידואליים בין הפרטים שתכונותיהם נבדקות. המטרה מוגבלת אפוא לניסיון לזהות בעזרת הנתונים **נטייה כללית של קשר** בין המשתנים ולתת לה ביטוי מספרי באמצעות **מדד לחוזק הקשר**. אם יש קשר, נשאל בהמשך גם כיצד אפשר להשתמש בקשר לצורכי **ניבוי** של אחד המשתנים אם מתבססים על ערכו הידוע של המשתנה האחר.

➢ **תיאור גרפי של זוג משתנים – דיון**

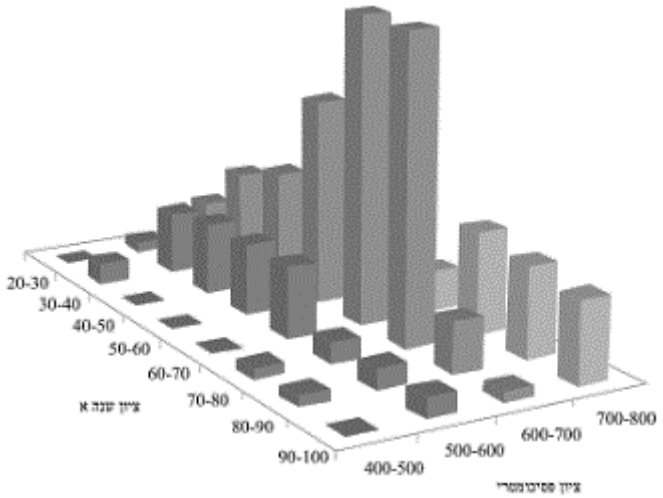
[אפשר להמשיך ישירות לסעיף 2.1 מבלי לפגוע בהמשך הקריאה.]

נזכיר שלמדנו לבנות **טבלת שכיחות של משתנה רציף**, שבה ערכי המשתנה הרציף קובצו לקטעים וליד כל קטע של ערכים רשמנו את השכיחות המתאימה. התיאור הגרפי שהצענו למשתנה רציף נקרא **היסטוגרמה**: מעל כל קטע של ערכים בציר האופקי בנינו מלבן, ששטחו מייצג את השכיחות בקטע. באופן זה התקבלה דיאגרמת מלבנים במישור.

אם כך, נצפה שההצגה הגרפית המתאימה לזוג משתנים תהיה דיאגרמת תיבות במרחב התלת-ממדי.

**דוגמה 3: ציונים פסיכומטריים וציוני שנה א.** הנתונים שבידינו עוסקים בקשר שבין ציון הבחינה הפסיכומטרית לבין הציון הממוצע השנתי של 198 תלמידי שנה א בחוג למתמטיקה באוניברסיטה בסוף שנת לימודים מסוימת. נציג כאן – ללא הנתונים עצמם וללא הסברים – את הדיאגרמה המרחבית המתאימה המתקבלת מ-198 הנתונים.

באיור 1, מעל כל מלבן במישור של ערכי זוג המשתנים, בנינו תיבה בגובה השכיחות שבטבלת הנתונים.



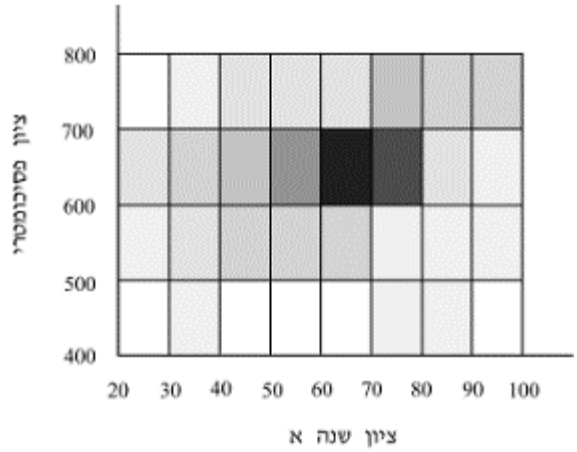

**איור 1.** היסטוגרמה תלת-ממדית של נתונים של תלמידי מתמטיקה; מעל כל מלבן בנינו תיבה שגובהה מייצג את השכיחות.

נשים לב,

- לא קל לצייר ידנית דיאגרמה כזאת; את ההיסטוגרמה שבאיור 1 יצרנו מטבלת הנתונים בעזרת תוכנת אקסל.
- באיור הקפדנו שלא להצמיד את תיבות השכיחויות, אך גם כך תיבות קדמיות מסתירות לעיתים תיבות אחוריות, ואי אפשר לראות בדיוק את כל מבנה הנתונים.

**תרגיל מחשבה: רעיון חלופי – הצגה מישורית**

באיור 2 מוצגת הצגה גרפית חלופית של הנתונים מדוגמה 3. חִשבו מה הרעיון שמאחורי השיטה המוצעת.




**איור 2.** מפת השכיחויות של נתונים של תלמידי מתמטיקה

גם שיטה זו אינה קלה לביצוע, וודאי שאינה מדויקת דיה. נציע עתה חלופה מישורית שתתברר כיעילה ושימושית וגם קלה יותר לביצוע, כאשר מספר הנתונים אינו גדול מדי.

## 2.1 דיאגרמות פיזור

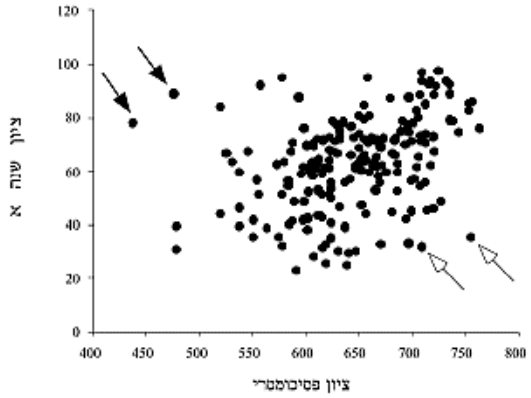
### התרשמות ראשונית מהנתונים

נסמן את זוג המשתנים ב-$X$ וב-$Y$, ואת $n$ זוגות הנתונים שהתקבלו ב- $(x_1, y_1),\ldots,(x_n, y_n)$.

את הנתונים שבידינו נהוג לתאר ב***דיאגרמת פיזור***, שמציינים בה במישור $(x, y)$ את מקומה של כל אחת מ- $n$ התצפיות (זוג מספרים) כנקודה במישור. נבחן תחילה דוגמה פשוטה.

**ציונים פסיכומטריים** (המשך דוגמה 3). באיור 3 מוצגת דיאגרמת הפיזור של נתוני דוגמה 3 (הנתונים עצמם לא הובאו כאן). כל אחד מ-198 התלמידים שנצפו מוצג באמצעות נקודה במישור.

בדיאגרמה רואים "ענן" של נקודות, שבו נמצאים, בין השאר, תלמידים עם ציונים פסיכומטריים נמוכים למדי שהצליחו יפה בלימודיהם באוניברסיטה (ראו חיצים שחורים), וגם כאלה שהציון הפסיכומטרי שלהם גבוה ועם זאת לא קיבלו ציונים גבוהים בלימודיהם (ראו חיצים לבנים).




**איור 3**. דיאגרמת פיזור של תלמידי מתמטיקה לפי הציון הפסיכומטרי וציון שנה א באוניברסיטה

מבט בוחן יותר בדיאגרמה מראה שהציונים הגבוהים באוניברסיטה שייכים ברובם לתלמידים בעלי ציון פסיכומטרי גבוה. במילים אחרות, נראה שזיהינו **מגמה של עלייה** של ציוני שנה א עם העלייה של הציון הפסיכומטרי, אם כי נראה שקשר זה אינו חזק כל כך (ניתוח מדויק יובא בהמשך).

בדיון הבא ניווכח שבעזרת דיאגרמת הפיזור אפשר לקבל מושג אם בכלל **יש קשר** בין המשתנים וגם להתרשם מהו **סוג הקשר**. הדיון בנושא זה ילווה בדוגמאות הממחישות מצבים אופייניים.

**דוגמה 4: מבחן פיז"ה.** נבחנו קשרים בין **המצב הכלכלי** של מדינות לבין **ציוני מתמטיקה** שלהן במבחני פיז"ה הבין-לאומיים. המצב הכלכלי של מדינה נמדד על פי התוצר המקומי הגולמי (תמ"ג) לנפש (בדולרים). לשם פשטות נציג וננתח את הנתונים של עשרים מדינות בלבד, אשר בחרנו מקרית מתוך כלל 31 המדינות שהשתתפו בשנה מסוימת במבחני פיז"ה. בבחירה כללנו את ישראל.

בלוח 4 מוצגים נתוני ממוצע ציוני מתמטיקה של תלמידי כיתות ו במבחני מחקר פיז"ה בעשרים המדינות שבחרנו וכן התמ"ג בכל אחת מהמדינות.

**לוח 4.** תמ״ג לנפש וציון ממוצע במתמטיקה בכיתות ו ב-20 מדינות

| המדינה | תמ״ג לנפש, $x$ (אלפי דולר) | ציון ממוצע במתמטיקה, $y$ |
|---|---|---|
| אלבניה | 8.9 | 394 |
| אסטוניה | 23.8 | 521 |
| הולנד | 41.3 | 523 |
| טוניסיה | 9.9 | 388 |
| טורקיה | 14.3 | 448 |
| יוון | 24.8 | 453 |
| **ישראל** | **35.1** | **466** |
| ליטא | 19.2 | 479 |
| מלזיה | 17.5 | 421 |
| ניו-זילנד | 30.9 | 500 |
| סלובקיה | 24.5 | 482 |
| **פולין** | **21.2** | **518** |
| פורטוגל | 22.5 | 487 |
| פינלנד | 37.1 | 519 |
| צ׳כיה | 26.9 | 499 |
| צרפת | 34.3 | 495 |
| קוסטה ריקה | 12.9 | 407 |
| קוריאה הדרומית | 34.0 | 554 |
| רומניה | 12.9 | 445 |
| תאילנד | 10.0 | 427 |

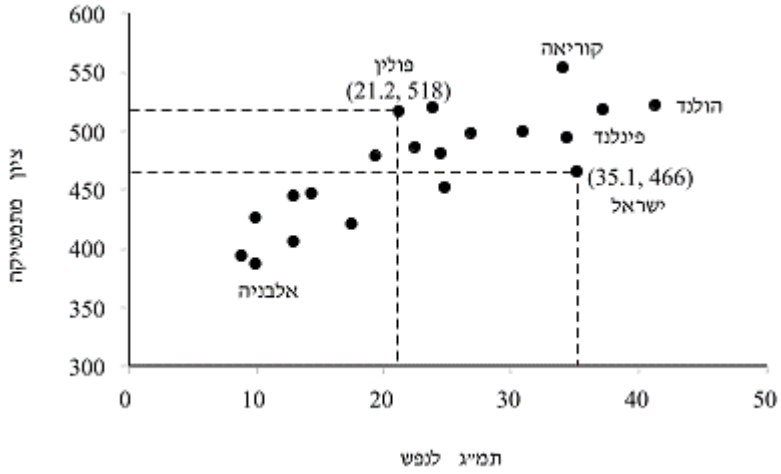
**שאלת חשיבה**: מה תוכלו לומר על מצבה של ישראל לעומת המדינות האחרות על פי לוח 4?

יש קושי לערוך השוואה כזאת מתוך הנתונים המוצגים בלוח. התיאור הגרפי שנציע מקל את ההשוואה.

**ניתוח גרפי**: בדיאגרמת הפיזור (איור 4), כל מדינה מיוצגת באמצעות נקודה, שבה הערך בציר ה-$x$ הוא התמ״ג המתאים, והערך בציר ה-$y$ הוא הציון הממוצע במתמטיקה. דיאגרמת הפיזור שהתקבלה נותנת תמונה יפה של הקשר בין שני המשתנים הללו, דבר שאינו מתאפשר מהלוח עצמו. מהדיאגרמה נראה שככל שהתמ״ג לנפש במדינה גבוה, כך ממוצע ציוני מתמטיקה בכיתות ו **נוטה לעלות** (קשר עולה).

**שאלת חשיבה**: מה תוכלו לומר על מצבה של ישראל על פי הדיאגרמה באיור 4?

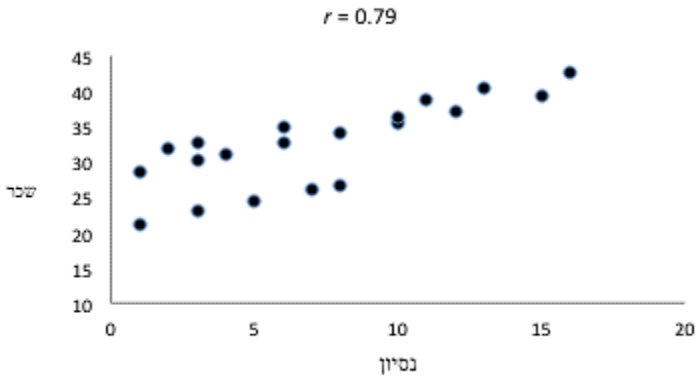
[בהמשך נציע ניתוח חלופי באמצעות ציוני התקן של ישראל על פי התמ״ג ועל פי הציון במתמטיקה.]




**איור 4.** דיאגרמת פיזור של עשרים מדינות לפי תמ"ג לנפש וממוצע ציוני מתמטיקה במבחן פיז"ה
(הערכים המתאימים לישראל ולפולין מפורטים)

במקבץ הדוגמאות המוצגות כאן (כולן אמיתיות) נמשיך ונזהה סוגים שונים של קשר בין משתנים בעזרת דיאגרמות הפיזור. בהמשך הספר, כשנציע מדד לקשר בין משתנים, נחזור לדוגמאות אלו ונבדוק אם הערך שיחושב שם תואם את הניתוח הגרפי הראשוני שנערוך כאן. בשלב זה נתעלם מהערך $r$ שרשום כבר עתה מעל כל איור. הדוגמאות מוצגות ללא פירוט של המחקר או הניסוי שנערך.

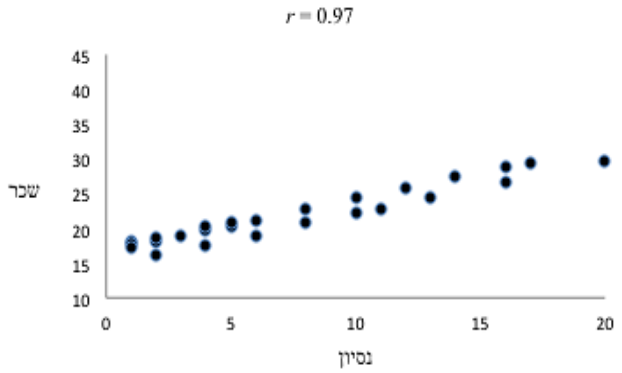
**א.** בדיאגרמת הפיזור באיור 5 מוצג השכר החודשי (באלפי שקלים) לפי שנות הניסיון (הוותק) בעבודה של עובדים **במשרות ניהול** בחברה גדולה למדי.




**איור 5.** דיאגרמת פיזור של שכר חודשי וניסיון בעבודה של עובדים במשרות ניהול

על פי הדיאגרמה נראה שאצל עובדים במשרות ניהול יש **קשר עולה** בין שנות הניסיון בעבודה לבין השכר: השכר נוטה להיות גבוה אצל עובדים עם ניסיון רב יותר.

**ב.** בדיאגרמת הפיזור באיור 6 מוצג השכר החודשי (באלפי שקלים) לפי שנות הניסיון בעבודה של העובדים **הזוטרים** בחברה זו.

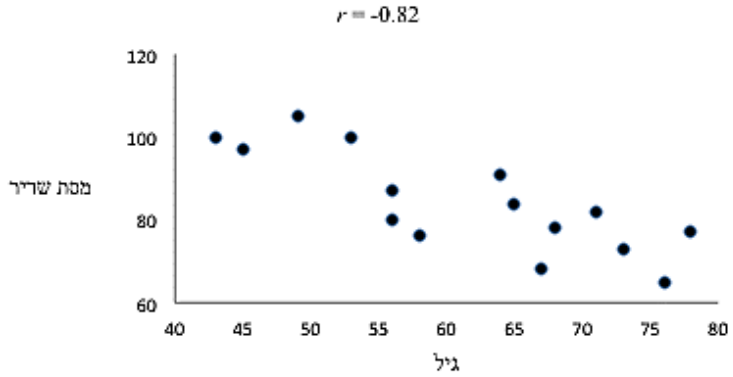



**איור 6.** דיאגרמת פיזור של שכר חודשי וניסיון בעבודה של עובדים זוטרים

על פי הדיאגרמה, מה תוכלו לומר על הקשר בין שנות הניסיון בעבודה לשכר העובדים הזוטרים? הביעו דעתכם על **חוזק הקשר** אצל עובדים זוטרים בהשוואה לזה שאצל עובדים בכירים.

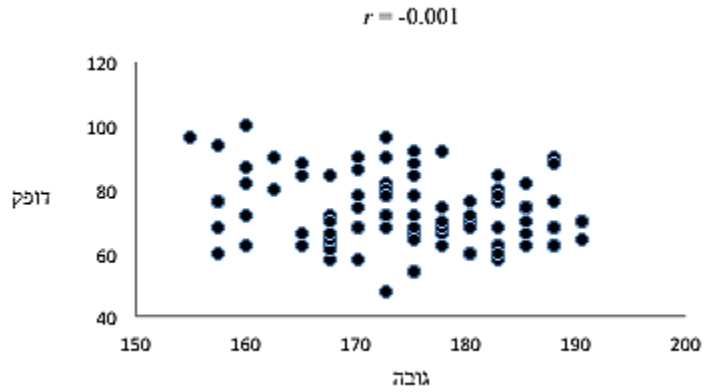
[בכל אחת משתי הדיאגרמות נסו לצייר אליפסה הדוקה סביב כל הנקודות או סביב מרביתן. השוו את רוחב האליפסות.]

**ג.** באיור 7 מוצגת דיאגרמת פיזור של מסת השריר של 16 נשים בוגרות בישראל לפי גילן. בדיאגרמה נראית **נטייה לירידה** של מסת השריר עם העלייה בגיל, ונראה גם שהקשר **חזק למדי**.




**איור 7.** דיאגרמת פיזור של נשים בישראל לפי גילן ומסת השריר שלהן

**ד.** באיור 8 מוצגת דיאגרמת פיזור של חיילים על פי גובהם וקצב הדופק שלהם.

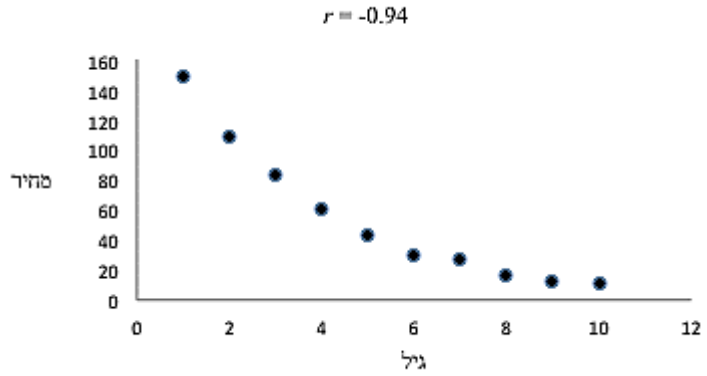



**איור 8.** דיאגרמת פיזור של חיילים לפי גובהם וקצב הדופק שלהם

ענן הנקודות באיור 8 מפוזר פחות או יותר באופן אחיד. מהדיאגרמה נראה שאין למעשה כל קשר בין שני המשתנים. כלומר, קצב הדופק של החיילים אינו קשור לגובהם.

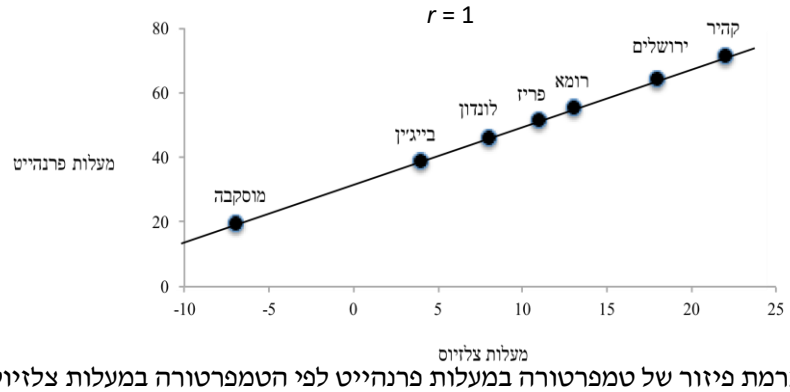
במצבים כאלה תם למעשה תפקידנו: אם בדיאגרמה לא התגלתה שום נטייה כללית המבטאת את הקשר, **אין טעם להמשיך בחישובים**.

**ה.** באיור 9 מוצגת דיאגרמת פיזור של מחיר **ממוצע** (באלפי שקלים) של מכוניות משומשות בשוק, לפי גיל המכוניות. הקשר בין המשתנים כאן הוא בבירור **קשר יורד**, והוא גם חזק מאוד: גיל המכונית קובע כמעט בוודאות את מחירה.




**איור 9.** דיאגרמת פיזור של מכוניות משומשות לפי גיל המכוניות ומחירן הממוצע

**ו.** באיור 10 מוצגת דיאגרמת פיזור של טמפרטורה במעלות פרנהייט לפי טמפרטורה במעלות צלזיוס כפי שנמדדה ב-12.12.2020 בכמה ערים בעולם. הסבירו מדוע הנקודות נמצאות כולן על קו ישר אחד.




**איור 10.** דיאגרמת פיזור של טמפרטורה במעלות פרנהייט לפי הטמפרטורה במעלות צלזיוס

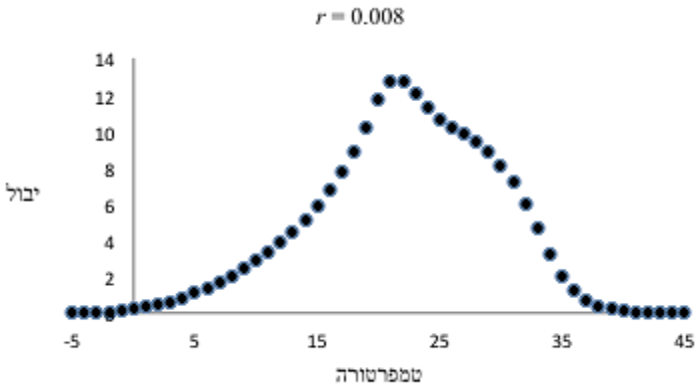
**סיכום הדוגמאות**

❤ חשוב לפתוח כל ניתוח העוסק במשתנה דו-ממדי בהצגה גרפית באמצעות דיאגרמת פיזור.

– אם הדיאגרמה מצביעה על חוסר קשר – פיזור אחיד – אין טעם להמשיך בניתוח.

– במקרים רבים (ראו למשל איור 3, 5, 7 וכן 9) הדיאגרמה מצביעה על **נטייה** של הנקודות להסתדר פחות או יותר סביב **קו ישר כלשהו** – **עולה או יורד**. רוב הפרק יוקדש למצבים כאלה.

לא נדון בספר זה במצבים שבהם עקומה אחרת מתאימה לתיאור הקשר בין המשתנים. ראו למשל את הדיאגרמה באיור 11, שמוצג בה הקשר בין יבול כותנה לטמפרטורה בארצות הברית:




**איור 11.** דיאגרמת פיזור של יבול הכותנה לפי הטמפרטורה

על פי הדיאגרמה, בערכים נמוכים של הטמפרטורה (עד כ-21 מעלות) היבול עולה ככל שהטמפרטורה גבוהה יותר. אבל מעבר לסף מסוים טמפרטורות גבוהות יותר מזיקות לכותנה, והיבול יורד עם העלייה בטמפרטורה. כפי שרואים, יש קשר ברור בין הטמפרטורה והיבול, אולם הקשר אינו קווי.

## 2.2 ניתוח סימן

### ➢ דיאגרמת פיזור מתוקננת

אחת המטרות העיקריות של יחידה זו היא להציע מדד לכיוון ולעוצמת הקשר הקווי בין שני משתנים $X$ ו- $Y$. כמובן, נרצה שהמדד יהיה מוחלט, שערכו לא יהיה תלוי ביחידות המדידה של המשתנים, ושהוא יאפשר גם להשוות את עוצמת הקשר בין זוגות שונים של משתנים שנמדדו ביחידות שונות.

נזכיר:

✓ הדרך להשתחרר מהתלות ביחידות המדידה היא למדוד את המשתנים ביחידות מתוקננות.

**סימון**: **ציוני התקן** של המשתנים יסומנו $\underset{\sim}{x}$ ו- $\underset{\sim}{y}$ בהתאמה, והם מתקבלים באמצעות החסרת הממוצע וחלוקה בסטיית התקן:

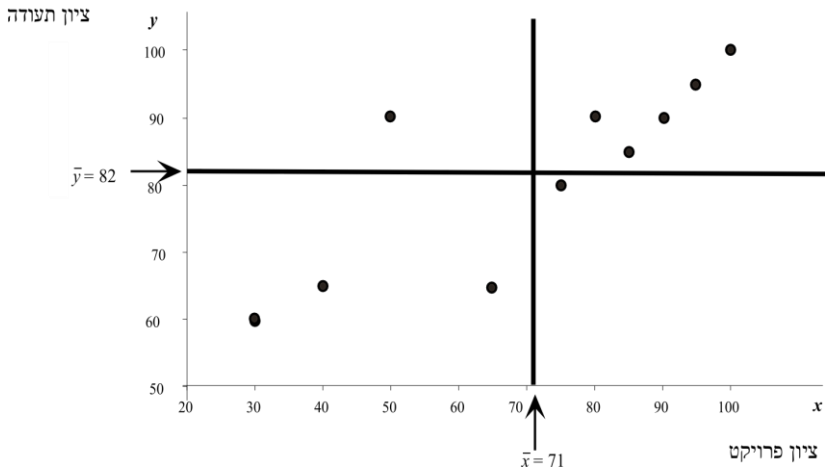
$$x \to \frac{x-\bar{x}}{\sigma_X} = \underset{\sim}{x} \qquad \text{ו-} \qquad y \to \frac{y-\bar{y}}{\sigma_Y} = \underset{\sim}{y} .$$

**שאלת חשיבה**: כיצד תשתנה דיאגרמת הפיזור כתוצאה מהמעבר ליחידות מתוקננות?

**ציון פרויקט וציון תעודה** (המשך דוגמה 1). מועצת התלמידים מעוניינת לבדוק את הקשר בין ציון הפרויקט (שאמור להיות ציון מגן) לבין הציון בתעודה.

הציונים מוצגים בלוח 1 (עמ' 16). נזכיר, ממוצעי הציונים שהתקבלו הם: $\bar{x} = 71$ ו- $\bar{y} = 82$.

באיור 12 מוצגת דיאגרמת הפיזור של הנתונים. באיור גם ציירנו קווים מקבילים לצירים העוברים כל אחד דרך הממוצע המתאים המצוין על הציר – אלו הם **קווי הממוצעים**.




**איור 12.** דיאגרמת פיזור של ציון פרויקט וציון תעודה עם קווי הממוצעים

בלוח 5 מוצגים ציוני התקן של המשתנים (החסרנו את הממוצעים המתאימים וחילקנו בסטיות התקן).

**לוח 5.** ציוני תקן של פרויקט ותעודה

| התלמיד | ציון תקן פרויקט $\tilde{x} = \frac{x-71}{22.78}$ | ציון תקן תעודה $\tilde{y} = \frac{y-82}{13.27}$ | הסימן של מכפלת ציוני התקן $\tilde{x} \cdot \tilde{y}$ | |
|---|---|---|---|---|
| 1 | -1.80 | -1.66 | + | |
| 2 | -0.92 | 0.60 | − | ← |
| 3 | -1.36 | -1.28 | + | ← |
| 4 | 1.05 | 0.98 | + | ← |
| 5 | 1.27 | 1.36 | + | |
| 6 | 0.83 | 0.60 | + | |
| 7 | 0.18 | -0.15 | − | |
| 8 | -0.26 | -1.28 | + | |
| 9 | 0.61 | 0.23 | + | |
| 10 | 0.40 | 0.60 | + | |

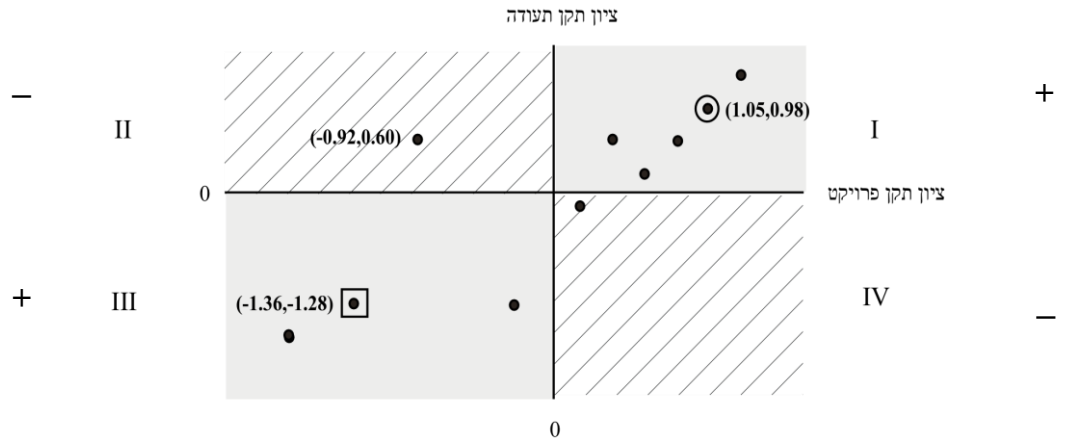
באיור 13 מוצגת **דיאגרמת הפיזור המתוקננת** שהתקבלה מהנתונים שבלוח 5. הדיאגרמה מתארת את ציוני התקן $\tilde{x}, \tilde{y}$ של זוג המשתנים. בדיאגרמה רשמנו את ציוני התקן של תלמידים 2, 3 ו-4 (בלוח בחיצים).

נזכיר: בדיקה $\tilde{x}$

✓ ממוצע הציונים המתוקננים של כל אחד מהמשתנים הוא 0 וסטיית התקן היא 1.

**מסקנה**: בדיאגרמת הפיזור המתוקננת, קווי הממוצעים מתלכדים עם הצירים החדשים.

נשים לב שבאיור 13 הצירים מחלקים את המישור לארבעה רביעים I–IV, שניים אפורים ושניים מקווקווים. בהמשך הדיון נסביר את סימני הרביעים באיור (+ או –) וכן את משמעות הסימנים בטור השמאלי בלוח 5.




**איור 13.** ציון פרויקט וציון תעודה: דיאגרמת פיזור מתוקננת וחלוקה לארבעה רביעים
(מפורטים הערכים המתאימים לתלמידים 2, 3, 4)

**שאלת חשיבה**: האם יש דרך לקבל את דיאגרמת הפיזור המתוקננת ישירות מהדיאגרמה המקורית גם ללא החישובים שערכנו בלוח 5?

כדי לענות נבדוק תחילה את ההשפעה של החסרת הממוצעים המתאימים על הדיאגרמה המקורית:

– בציר האופקי יש לחסר 71 מכל הערכים. המשמעות: הזזה שמאלה של כל הנקודות ב-71 יחידות.

– בציר האנכי יש לחסר 82 מכל הערכים. המשמעות: הזזה כלפי מטה של כל הנקודות ב-82 יחידות.

בכך הפכנו את קווי הממוצעים לצירים החדשים.

[חלוקה בסטיות התקן רק משנה את סקלת המספרים על הצירים. מבחינה גרפית זוהי פעולה של כיווץ או של מתיחה של הצירים באופן שיחידת המדידה בשניהם היא 1.]

### ➢רביעים חיוביים ורביעים שליליים

– בדיאגרמת הפיזור המתוקננת של זוג משתנים קווי הממוצעים מתלכדים עם הצירים.

– הצירים מחלקים את המישור לארבעה רביעים I–IV. על פי איור 13, נסמן שניים מהם ב- (+) ושניים ב- (–).

לניתוח משמעות הסימנים נחזור לדוגמה 1.

**ציון פרויקט וציון תעודה** (המשך דוגמה 1). נבחן נקודות מייצגות ברביעים השונים בדיאגרמה המתוקננת. נפתח בנקודות בשני הרביעים האפורים.

– הנקודה $(1.05, 0.98)$ (מוקפת בעיגול) מייצגת את תלמיד 4, שהוא תלמיד טוב ושני ציוני התקן שלו **חיוביים**, כלומר שניהם מעל לממוצע. על כן, לתלמיד זה מכפלת ציוני התקן חיובית.

– הנקודה $(-1.36,-1.28)$ (מוקפת בריבוע) מייצגת את תלמיד 3, שמתקשה בלימודים ושני ציוני התקן שלו **שליליים**, כלומר שני הציונים שלו מתחת לממוצע. על כן גם לתלמיד זה מכפלת ציוני התקן חיובית.

לכל הנקודות ברביעים האפורים, הערכים של שני הציונים $x$ ו- $y$ הם **באותו כיוון ביחס לממוצע** המתאים, כלומר שניהם מעל לממוצע או שניהם מתחת לממוצע. על כן שני ציוני התקן הם בעלי אותו סימן – שניהם חיוביים או שניהם שליליים. ולכן:

מכפלת ציוני התקן של כל נקודה ברביעים האפורים תהיה תמיד **חיובית**. בדיאגרמת הפיזור סימנו את הרביעים הללו בסימן פלוס: (+).

נבחן עתה נקודה ברביעים המקווקוים בדיאגרמת הפיזור המתוקננת.

למשל, הנקודה $(-0.92,0.60)$, שברביע המקווקו השמאלי, מייצגת את תלמיד 2, שלא הצליח בפרויקט – ה-$x$ נמוך מהממוצע ולכן ציון התקן הוא שלילי. אך הוא קיבל ציון תעודה טוב $y$ גבוה מהממוצע וציון התקן חיובי.

לכל הנקודות ברביעים המקווקוים, הערכים $x$ ו- $y$ הם **בכיוונים הפוכים ביחס לממוצע** המתאים – האחד מעל הממוצע והאחר מתחת לממוצע, ובהתאם לכך אחד מציוני התקן הוא חיובי והאחר שלילי.

מכאן, מכפלת ציוני התקן של כל נקודה ברביעים המקווקוים תהיה תמיד **שלילית**. בדיאגרמה סימנו את הרביעים הללו בסימן מינוס: (–).

בעמודה האחרונה בלוח 5 רשמנו את **הסימן** של מכפלת ציוני התקן לכל אחד התלמידים.

### ➢ניתוח סימן, קשר עולה וקשר יורד

בדיאגרמת הפיזור המתוקננת של זוג משתנים, הצירים מחלקים את המישור לארבעה רביעים I–IV. כמו באיור 13, נסמן שניים מהם ב- (+) ושניים ב- (–). כמו בדוגמה 1 שניתחנו,

✋ הסימן שייחסנו לכל רביע (+ או –) נקבע על פי הסימן של מכפלת ציוני התקן של הנקודות באותו רביע.

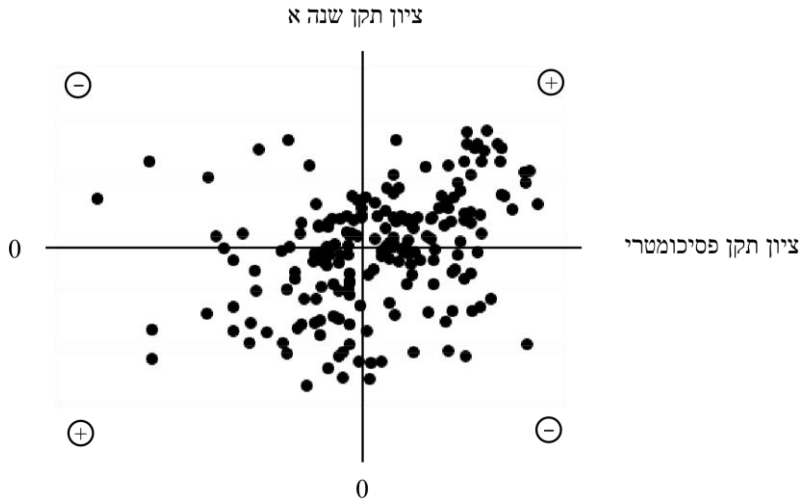
**משמעות**: הסימן של רביע הוא (+) אם לכל הנקודות ברביע ערכי המשתנים שניהם מעל הממוצע או שניהם מתחת לממוצע המתאים; הסימן של רביע הוא (–) אם אחד המשתנים הוא מעל לממוצע המתאים והשני מתחת לממוצע.

מה אנו למדים מכך על הקשר בין המשתנים? נחזור ונתבונן בכמה דוגמאות מייצגות.

**ציון פרויקט וציון תעודה** (המשך דוגמה 1). נשים לב שרוב הנקודות בדיאגרמת הפיזור המתוקננת שבאיור 13 נמצאות ברביעים האפורים, שהם הרביעים החיוביים. הדבר משקף את **הקשר העולה** הברור בין שני הציונים – ככל שערכו של ציון הפרויקט עולה כך ערכו של ציון התעודה נוטה לעלות. נאמר שיש ***קשר חיובי*** בין המשתנים.

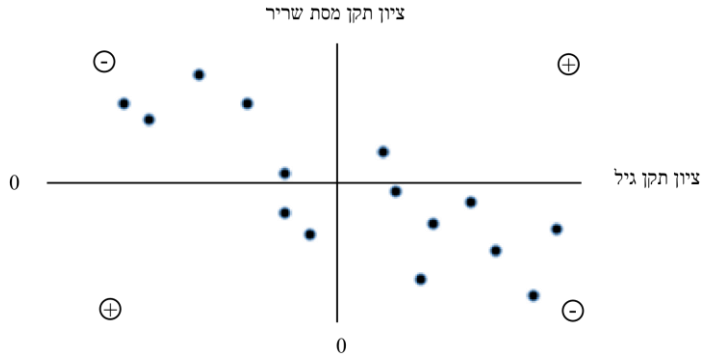
**ציונים פסיכומטריים** (המשך דוגמה 3). באיור 14 מוצג אותו ענן נקודות שבאיור 3, אך הפעם ביחידות מתוקננות. כפי שרואים, יש יותר נקודות ברביעים החיוביים מאשר בשליליים. הדבר מצביע על **קשר עולה**

בין שני המשתנים – ככל שערכו של הציון הפסיכומטרי עולה כך ערכו של ציון שנה א **נוטה לעלות**. גם במקרה זה נאמר שיש ***קשר חיובי*** בין המשתנים.




**איור 14.** חזרה לאיור 3 (דוגמה 3), והפעם ביחידות מתוקננות וסימן של כל רביע

**הגיל ומסת השריר**. נחזור לדיאגרמת הפיזור שבאיור 7, הפעם ביחידות מתוקננות. כפי שרואים באיור 15, רוב הנקודות נמצאות ברביעים השליליים. הדיאגרמה שבאיור מתארת אפוא **קשר יורד** (למרבה הצער) בין הגיל ומסת השריר אצל נשים – ככל שערכו של הגיל עולה כך מסת השריר **נוטה לרדת**. נאמר שיש **קשר שלילי** (קשר יורד) בין המשתנים.




**איור 15.** חזרה לאיור 7, הפעם ביחידות מתוקננות וסימן של כל רביע

בהמשך נדגים גם מצבים שבהם חלוקת הנקודות בדיאגרמת הפיזור שווה פחות או יותר בין הרביעים החיוביים והרביעים השליליים. במצבים כאלו מתקבל איזון בין המכפלות החיוביות והשליליות.

**מבחן פיז"ה** (המשך דוגמה 4). שימוש במחשבון נותן: $\bar{x} = 23.1$ , $\bar{y} = 471.3$ .

**מטלה:** על גבי דיאגרמת הפיזור המקורית שבאיור 4 (עמ' 27) ציירו את שני קווי הממוצעים וסמנו רביעים חיוביים ושליליים. נתחו בהתאם את כיוון הקשר שבין המשתנים.

## אז מה היה לנו? ניתוח גרפי

### מושגים חדשים

- דיאגרמת פיזור (Scatter diagram)
- קווי הממוצעים
- קשר קווי, עולה או יורד
- דיאגרמת פיזור מתוקננת
- מכפלת ציוני התקן $\tilde{x} \cdot \tilde{y}$
- רביעים חיוביים ורביעים שליליים

### תובנות שרכשנו

- לפני שפונים לניתוח הקשר בין שני משתנים, חשוב לצייר דיאגרמת פיזור של הנתונים. הדיאגרמה עוזרת לזהות מגמות ולהימנע ממסקנות חפוזות.
- כדי שתוצאות הניתוח שנערוך לא יהיו תלויות ביחידות המדידה נעדיף לעבור מייד לערכים מתוקננים ; כלומר הניתוח ייערך על ציוני התקן של שני המשתנים.
- בדיאגרמת הפיזור המתוקננת, יחידת המדידה זהה בשני הצירים והצירים נפגשים בנקודה (0,0).
- בדיאגרמה המתוקננת קווי הממוצעים מתלכדים עם הצירים.
- הצירים מחלקים את המישור לארבעה רביעים. הסימן (+) או (–) שייחסנו לכל רביע נקבע על פי **הסימן של מכפלת ציוני התקן** של הנקודות ברביע זה.
  - בתצפית שבה ערכי שני המשתנים באותו כיוון ביחס לממוצע שלהם – שניהם גבוהים מהממוצע או שניהם נמוכים מהממוצע – הסימן של מכפלת ציוני התקן $\tilde{x} \cdot \tilde{y}$ הוא **חיובי**.
  - בתצפית שבה ערכי שני המשתנים הפוכים בכיוונם – אחד המשתנים גבוה מהממוצע והאחר נמוך מהממוצע – הסימן של מכפלת ציוני התקן הוא **שלילי**.
- אם רוב הנקודות הן ברביעים **החיוביים** הקשר בין המשתנים הוא קשר **עולה**, אם רובן ברביעים **השליליים** הקשר הוא קשר **יורד**.

כלים גרפיים ממוחשבים – הכנה של דיאגרמת פיזור בעזרת תוכנת אקסל

בבעיות אמת עם נתונים רבים אפשר להכין דיאגרמת פיזור בקלות רבה יותר בעזרת תוכנת אקסל. בדף העבודה רושמים את ערכי שני המשתנים בשתי עמודות. בוחרים את שתי העמודות הללו. מתוך תפריט **הוספה** (insert) בדף העבודה בוחרים **תרשים** (chart). מקבלים אפשרויות שונות של דיאגרמות. אם בוחרים את האפשרות **פיזור** (scatter) מקבלים דיאגרמת פיזור, ואותה אפשר להתאים לצרכינו : לשנות את הסקלה של הצירים, להוסיף שמות של המשתנים על הצירים, לשנות את הפורמט של הנקודות בדיאגרמה ועוד ועוד.

## משימות חישוב וחשיבה – ניתוח גרפי

(פתרונות בעמ' 111)

משלב זה ואילך מומלץ להיעזר במחשבונים לכל החישובים.

**משימה I. עברית שפה קשה**

לקבוצת תלמידים נערכו שני מבחנים: מבחן A – הוקראו עשר מילים, ונרשם מספר המילים שכל תלמיד איית **נכון**; מבחן B – התלמידים התבקשו למצוא את המילים שהאיות שלהן **שגוי** ברשימה של עשרים מילים, ונרשם מספר המילים השגויות שמצא כל תלמיד.

א. הציגו את הנתונים שבטבלה הבאה בדיאגרמת פיזור. האם על פי הדיאגרמה נראה שיש קשר בין מספר המילים הנכונות בהכתבה לבין היכולת לאתר מילים שגויות ברשימה?

| התלמיד | מבחן A, $x$ | מבחן B, $y$ |
|---|---|---|
| א | 2 | 5 |
| ב | 5 | 8 |
| ג | 9 | 15 |
| ד | 6 | 12 |
| ה | 7 | 11 |
| ו | 4 | 9 |
| סה"כ | | |
| ממוצע | | |

ב. חשבו ממוצע וסטיית תקן של שני המשתנים.

ג. הוסיפו לטבלה שתי עמודות, ורשמו בהן את הציונים המתוקננים המתקבלים. בדקו שהממוצע בהן הוא 0 וסטיית התקן היא 1.

ד. ציירו דיאגרמת פיזור של המשתנים המתוקננים. חלקו את המישור לארבעה רביעים, רשמו את הסימן של כל רביע, והסבירו מה משמעות הסימן.

**משימה II. ציון פרויקט וציון תעודה** (המשך משימה II מעמ' 20)

א. ציירו את דיאגרמת הפיזור המתוקננת, וסמנו גם את קנה המידה הנכון על הצירים. מה הם קווי הממוצעים? סמנו רביעים חיוביים ורביעים שליליים. מה משמעות הסימן?

ב. מהי צורת הקשר בין המשתנים ומה כיוונו? ציירו בדיאגרמה אליפסה מהודקת סביב כל הנקודות או סביב מרביתן. מה כיוון האליפסה? האם האליפסה צרה או רחבה? מה תסיקו מכך על כיוון הקשר ועל עוצמתו?

**משימה III. חמודי הסבות מקצועיות**

מנכ"ל חברת "חמודי" להסבות מקצועיות עורך סדנאות אימון למלצרים מתחילים. כדי לברר אם שיטת האימון שלו יעילה, הוא ניסה אותה על עשרה מועמדים מקריים ורשם לכל אחד מהם ניקוד על פי ביצועיו בתום ימי הסדנה האישית, שנמשכה בין יום אחד לארבעה ימים. התקבלה טבלה זו:

| המועמד | ימי אימונים, $x$ | ניקוד, $y$ | $\underset{\sim}{x}=\frac{x-\bar{x}}{\sigma_X}$ | $\underset{\sim}{y}=\frac{y-\bar{y}}{\sigma_Y}$ |
|---|---|---|---|---|
| א | 1 | 4 | | |
| ב | 1 | 5 | | |
| ג | 1 | 5.5 | | |
| ד | 2 | 6 | | |
| ה | 2 | 7.5 | | |
| ו | 2 | 8 | | |
| ז | 3 | 7.8 | | |
| ח | 3 | 7.3 | | |
| ט | 4 | 6.5 | | |
| י | 4 | 6 | | |

א. הציגו דיאגרמת פיזור של שני המשתנים על פי הנתונים בעמודות הלבנות: $X$ – מספר ימי האימונים, $Y$ – הניקוד המתאים של עשרת המועמדים. חשבו ממוצעים ושרטטו בדיאגרמה את קווי הממוצעים.

ב. מה תוכלו לומר מהדיאגרמה בלבד על צורת הקשר בין שני המשתנים? התאימו ביד חופשית קו ישר שנראה לכם מתאים ביותר לתיאור הקשר. עתה ציירו עקומה שעשויה להתאים יותר לתיאור הקשר.

ג. חשבו את **סטיות התקן** של המשתנים, ובעזרתן חשבו את ציוני התקן (רשמו בעמודות האפורות).

ד. ציירו גם דיאגרמת פיזור מתוקננת. מהם קווי הממוצעים בדיאגרמה המתוקננת? ציינו מהם הרביעים החיוביים ומהם הרביעים השליליים. באילו רביעים נמצאות רוב הנקודות?

ה. באחת הדיאגרמות ציירו אליפסה מהודקת סביב נקודות רבות ככל האפשר. האם האליפסה שהתקבלה היא צרה או רחבה? מה תסיקו מכך על עוצמת הקשר?

**משימה IV. מחירי מחשב ביתי**

כדי לבדוק את הקשר בין מחיר דגם חדש של מחשב ביתי לבין הנכונות לקנות אותו, נבחנו 10 מחירים שונים (במאות דולרים), ולכל מחיר $x$ הוקצו אקראית 100 בתי אב. נסמן ב-$y$ את מספר בתי האב (מתוך 100) שאמרו שכנראה יקנו את המחשב במחיר הנקוב $x$. הנתונים מוצגים בטבלה זו:

| **מחיר** $x$ | 1 | 2.5 | 5 | 10 | 20 | 30 | 40 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| **קונים** $y$ | 90 | 80 | 65 | 46 | 34 | 26 | 17 | 15 | 6 | 4 |

א. שרטטו דיאגרמת פיזור. חשבו את הממוצעים של שני המשתנים ושרטטו בדיאגרמה את קווי הממוצעים.

ב. שרטטו ביד חופשית קו ישר שנראה לכם מתאים ביותר לצורכי ניבוי.

ג. עתה שרטטו ביד חופשית עקומה חלקה שנראית לכם מתאימה יותר לנתוני הדיאגרמה.

ד. שרטטו דיאגרמת פיזור מתוקננת. נסו לאתר נקודה אחת בכל רביע. באילו רביעים נמצאות מרבית הנקודות? מה תסיקו מכך?

ה. שרטטו אליפסה מהודקת סביב הנקודות בדיאגרמה. מהו כיוון האליפסה? האם האליפסה שהתקבלה היא צרה או רחבה? מה תסיקו מכך על כיוון הקשר ועל עוצמתו?

ו. הסבירו כיצד אפשר לקבל את הדיאגרמה המתוקננת (למעט הסקלות על הצירים) גם מבלי לחשב ערכים מתוקננים.

**תרגילים: ניתוח גרפי**

(פתרונות מקוצרים לתרגילים נבחרים בעמ' 134)

**תרגיל 1**. עברו על כל דיאגרמות הפיזור בפרק זה שבהן נראה שיש קשר קווי בין המשתנים. ציירו בכל אחת אליפסה מהודקת שתכיל את נקודות הדיאגרמה, פרט אולי לכמה חריגות מאוד. באילו נראה שמתקבלת אליפסה צרה ובאילו אליפסה רחבה? מהו כיוון האליפסה? דונו במשמעות הדברים.

**תרגיל 2.** לכל זוג משתנים להלן ציינו אם יש קשר ביניהם ואם הקשר עולה או יורד. מיינו את הזוגות לפי חוזק הקשר – חזק מאוד, חזק, בינוני, חלש, חלש מאוד.

א. מחיר מכונית משומשת וגיל המכונית;

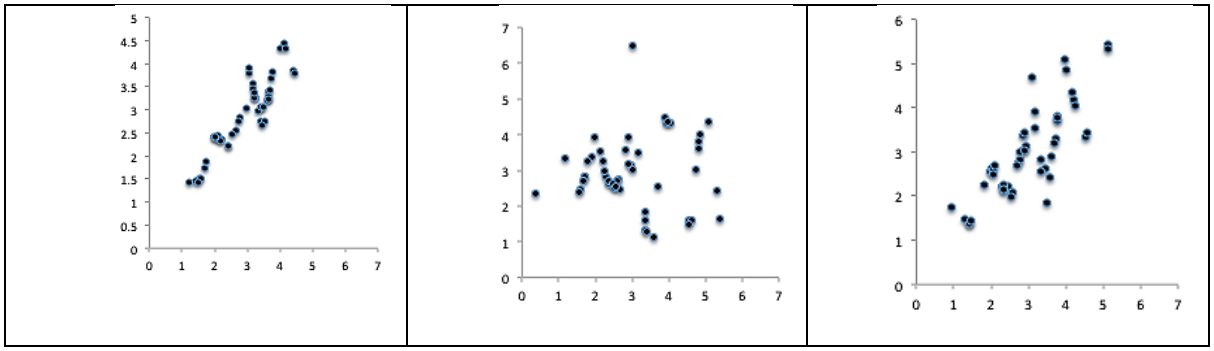
ב. בקרב זוגות נשואים, גיל הבעל וגיל האישה;

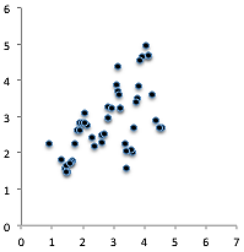
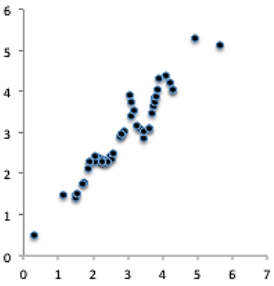
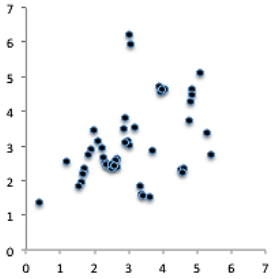
ג. גיל של תלמיד בבית הספר ושנת הלידה שלו;

ד. גיל של תלמיד בכיתה והגיל של אימו;

ה. מספר הילדים במשפחה וההשכלה של האם.

**תרגיל 3.** לפניכם שש דיאגרמות פיזור המתארות קשר עולה. ציירו בכל אחת 'לפי העין' את קווי הממוצעים וסמנו רביעים חיוביים ורביעים שליליים. דרגו את הדיאגרמות לפי חוזק הקשר הקווי.



| | | |
|---|---|---|
|  |  |  |

**תרגיל 4.** בחרו לפחות עשרה נבדקים (לא מהכיתה שלכם). הכינו לכל אחד עותק מוגדל של קוד ה- QR.

**בשלב א** בקשו מכל נבדק להעריך בזמן קצר קצוב כמה ריבועים **שחורים בודדים** יש בקוד. **בשלב ב** בקשו מכל נבדק להעריך את מספר הריבועים **הלבנים הבודדים** בקוד.

רשמו את זוגות המספרים והציגו אותם בדיאגרמת פיזור. על פי הדיאגרמה נתחו את הקשר בין שתי ההערכות, למשל: האם יש קשר קווי? האם הקשר עולה או יורד? מהי עוצמתו?

# פרק 3

# מתאם בין משתנים

בפרק זה נציג מדד מקובל לקשר קווי בין שני משתנים – מקדם המִתְאם (correlation coefficient) – ונבחן את ערכו בדוגמאות שהכרנו בפרק הקודם.

נתבונן תחילה במשתנים המתוקננים $\underset{\sim}{X},\underset{\sim}{Y}$ ובדיאגרמת הפיזור המתוקננת.

נזכיר:

- ✓ מתקבלת מערכת צירים רגילה: אותן יחידות בשני הצירים ונקודת המפגש של הצירים היא (0,0).
- ✓ קווי הממוצעים מתלכדים עם הצירים. הנקודה (0,0) היא מרכז הדיאגרמה.
- ✓ הצירים מחלקים את המישור לארבעה רביעים: שניים חיוביים (סומנו +) ושניים שליליים (סומנו –).
  - לנקודות ברביעים החיוביים מכפלת ציוני התקן היא חיובית;
  - לנקודות ברביעים השליליים מכפלת ציוני התקן היא שלילית;
- ✓ אם רוב נקודות הדיאגרמה מרוכזות ברביעים החיוביים, הקשר בין המשתנים הוא קווי עולה (קשר חיובי).
- ✓ אם רוב נקודות הדיאגרמה מרוכזות ברביעים השליליים, הקשר בין המשתנים הוא קווי יורד (קשר שלילי).

נראה טבעי אפוא לבסס מדד לקשר קווי בין המשתנים על סכום המכפלות של $n$ ציוני התקן.

## 3.1 מקדם המתאם

### מקדם המתאם (של פירסון) – הגדרה ראשונית

***מקדם המתאם*** $r$ הוא ממוצע של $n$ מכפלות ציוני התקן:

$$r = r(X,Y) = \frac{1}{n}\cdot\left[\underset{\sim}{x}_1 \cdot \underset{\sim}{y}_1 + ... + \underset{\sim}{x}_n \cdot \underset{\sim}{y}_n\right] \qquad (1)$$

**מכפלות ציוני התקן**

### הדגמות

– בדיאגרמה באיור 15 (עמ' 34) זיהינו **קשר יורד**. נשים לב שהמכפלות החיוביות של הציונים המתוקננים כאן הן קטנות ומועטות, ולעומתן המכפלות השליליות גדולות ורבות. לכן סכום המכפלות **שלילי** ומקדם המתאם (הממוצע) **שלילי** אף הוא.

– בדיאגרמות של **חוסר קשר**, כמו באיור 8 (עמ' 28), המכפלות החיוביות מקזזות את המכפלות השליליות, ומתקבל מקדם מתאם **קרוב ל-0**.

❤ למשתנים בעלי **קשר עולה** סימנו של מקדם המתאם **חיובי**, ולמשתנים בעלי **קשר יורד** סימנו **שלילי**.

בהמשך נדון במשמעות של ערך מקדם מתאם קרוב ל-0, וכמו כן נבדוק אם הערך המספרי עצמו הוא מדד לחוזק הקשר.

➢ **חישוב ידני של מקדם המתאם** – על פי נוסחה (1)

נערוך בטבלה את $n$ זוגות הערכים המתוקננים.

– בעמודה נוספת נחשב לכל זוג את מכפלת הערכים, ובתחתית העמודה נסכם את כל המכפלות.

– נחשב את הממוצע של $n$ המכפלות ונקבל את מקדם המתאם $r$.

בשתי הדוגמאות הבאות נעזרנו במחשבון כדי לערוך את כל החישובים הנדרשים. בבעיות מעשיות, שהנתונים בהן רבים, משתמשים בתוכנות ייעודיות לחישוב ישיר של מקדם המתאם. בנספח ב מופיע הסבר קצר על שימוש נוח בתוכנת אקסל.

**ציון פרויקט וציון תעודה** (המשך דוגמה 1). בלוח 5 מופיעים ציוני התקן (עמ' 32). בעזרתם הוספנו בלוח 6 עמודה של מכפלות ציוני התקן:

**לוח 6**. ציון פרויקט וציון תעודה של עשרה תלמידים – חישוב מקדם המתאם

| **התלמיד** | **מכפלת ציוני התקן** $\tilde{x}\cdot\tilde{y}$ |
|---|---|
| 1 | 2.9839 |
| 2 | -0.5558 |
| 3 | 1.7434 |
| 4 | 1.0321 |
| 5 | 1.7268 |
| 6 | 0.5028 |
| 7 | -0.0265 |
| 8 | 0.3374 |
| 9 | 0.1389 |
| 10 | 0.2382 |
| סה"כ | 8.1213 |
| **ממוצע** | $r = 0.8121$ |

מקדם המתאם, שהוא ממוצע המכפלות, רשום בתחתית העמודה השמאלית (בתא האפור): $r = 0.8121$. מייד נמשיך ונברר אם הקשר החיובי שהתגלה בין שני המשתנים נחשב חזק.

## 3.2 מקדם המתאם – תכונות

מקדם המתאם הוא **מספר נטול יחידות**, והוא אינו תלוי בקנה המידה של המשתנים.

❤ **סימן שלילי של מקדם המתאם מצביע על קשר יורד בין שני המשתנים, וסימן חיובי מצביע על קשר עולה.**

נקדים את המאוחר ונציג תכונות חשובות נוספות של מקדם המתאם, שנרחיב עליהן בהמשך (נספח ד, עמ' 107):

❤ מקדם המתאם מקבל ערכים בין $(-1)$ ל- $(+1)$ בלבד.

❤ ערכי הקיצון $(+1)$ או $(-1)$ מתקבלים כאשר $X$ קובע לחלוטין את ערכו של $Y$ באמצעות הקשר הקווי, כלומר כאשר הנקודות בדיאגרמת הפיזור נמצאות כולן על קו ישר אחד. $Y = a + bX$

❤ ערכי מקדם המתאם הקרובים ל- $(-1)$ או ל- $(+1)$ מצביעים על קשר קווי חזק – יורד או עולה, בהתאמה – בין המשתנים.

**ציוני פרויקט וציוני תעודה** (המשך דוגמה 1 מעמ' 41). מקדם המתאם הוא כאמור $r = 0.812$, כלומר מספר חיובי גבוה. לפנינו אפוא קשר **עולה חזק** בין שני הציונים – תלמיד שציון הפרויקט שלו גבוה יותר נוטה לקבל גם ציון תעודה גבוה יותר.

ננתח עתה מקרה שבו נופתע לגלות שמקדם המתאם אינו גבוה.

**ציונים פסיכומטריים וציוני שנה א** (המשך דוגמה 3 מעמ' 34). נחזור ונתבונן בדיאגרמת הפיזור (איור 14) של כלל תלמידי המתמטיקה לפי הציון הפסיכומטרי וציון שנה א באוניברסיטה. נוכחנו שהקשר כאן הוא חיובי. האם הקשר חזק? האם הוא חלש?

מניתוח הנתונים המקוריים בתוכנת אקסל התקבל מקדם מתאם של $r = 0.337$, שהוא אומנם חיובי אך נחשב חלש למדי. באיור 14 רואים שאכן הנקודות בדיאגרמת הפיזור יוצרות ענן מפוזר, שכיוונו הכללי הוא עלייה.

מאינפורמציה נוספת שבידינו הנוגעת לאותם תלמידים, נמצא שמקדם המתאם בין **ציון הבגרות** שלהם לבין ציון שנה א אף הוא **נמוך למדי**: $r = 0.393$. שילוב המשתנים – ממוצע בין ציון הבגרות והציון הפסיכומטרי – נותן את **ציון ההתאמה** הידוע, ומקדם המתאם בינו לבין ציוני שנה א עולה ל- $r = 0.427$, שאף הוא אינו כה גבוה.

תוצאות אלה מעידות שציון ההתאמה המקובל איננו קריטריון טוב כל כך לקבלת סטודנטים ללימודי מתמטיקה וגם לחוגים אחרים, ואולם ייתכן שהוא עדיף על הגרלה פשוטה.

**➢ מקדם המתאם – ערכי קיצון**

**הדגמה פעילה**. בלוח הבא רשמנו ערכים של משתנה פשוט X, ולידו משתנה נוסף $Y=2X+3$.

| $x$ | $y$ | **ציון תקן,** $\tilde{x}$ | **ציון תקן,** $\tilde{y}$ | **מכפלת ציוני התקן,** $\tilde{x}\cdot\tilde{y}$ |
|---|---|---|---|---|
| -1 | 1 | | | |
| 0 | 3 | | | |
| $\bar{x}=1/2$ | $\bar{y}$ = | | | $r=1$ |

**ממוצעים**

א. השלימו את הערכים בטבלה.

ב. חישוב ישיר מורה שהשונות של $X$ היא $1/4$. **הסיקו** מהי השונות של $Y$. הסיקו גם מהן סטיות התקן.

ג. חשבו את ציוני התקן של שני המשתנים. מה גיליתם? חשבו את מכפלת ציוני התקן ואת ממוצע המכפלות. בדקו שהערך של מקדם המתאם בין שני המשתנים הוא אכן 1.

**באופן כללי**: מתברר שערך הקיצון 1 מתקבל בדיוק כאשר **קובע לחלוטין** את ערכו של $Y$ באמצעות **קשר קווי** $Y=a+bX$, $b>0$, כלומר כאשר שני המשתנים בודקים את אותה תכונה אך קנה המידה שונה. דוגמאות למצב כזה הן, למשל, הקשר בין זמן בשעות לזמן בדקות או בין טמפרטורה במעלות צלזיוס לטמפרטורה במעלות פרנהייט.

**הדגמה פעילה**: במשחק סכום אפס של שני שחקנים, הרווח של שחקן אחד הוא ההפסד של האחר. יהיו $X$ – הרווח של שחקן א, $Y$ – הרווח של שחקן ב. היעזרו בקשר $Y=-X$ כדי להשלים את הטבלה.

| $x$ | $y$ | **ציון תקן** $\tilde{x}$ | **ציון תקן** $\tilde{y}$ | **מכפלת ציוני התקן** $\tilde{x}\cdot\tilde{y}$ |
|---|---|---|---|---|
| 0 | 1 | | | |
| 1 | -1 | | | |
| 2 | -2 | | | |
| $\bar{x}=1$ | $\bar{y}$ = | | | $r=-1$ |

**ממוצעים**

א. חישוב ישיר מורה שהשונות של $X$ היא $2/3$. **הסיקו** מהו הממוצע ומהי השונות של $Y$. הסיקו מהן סטיות התקן.

ב. חשבו את ציוני התקן של שני המשתנים. מה גיליתם? חשבו את מכפלת ציוני התקן ואת ממוצע המכפלות. בדקו שהערך של מקדם המתאם בין שני המשתנים הוא אכן $(-1)$. נסו להכליל.

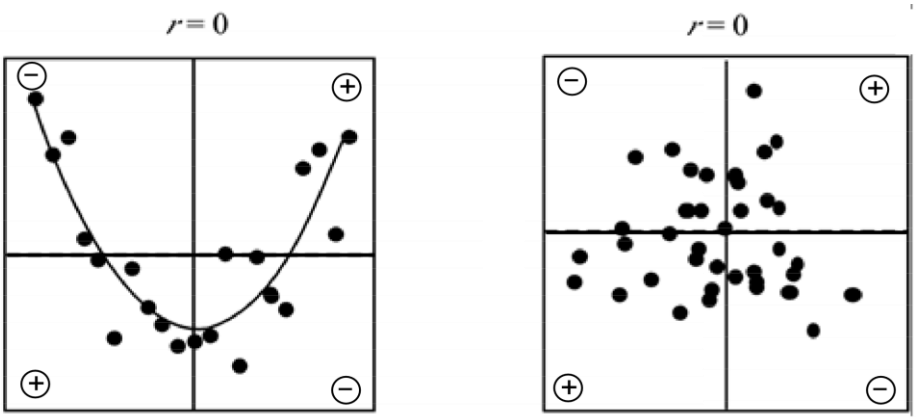
**שאלת חשיבה**

מה המשמעות של **מקדם מתאם 0** (או קרוב ל-0)? האם אפשר להסיק מכך על **חוסר קשר** בין המשתנים?

∴ **לא ולא!**

**מינוח**: במצב שבו מקדם המתאם בין $X$ ו-$Y$ הוא אפס, אנו אומרים שהמשתנים הם ***בלתי מתואמים***.

**ניתוח**: מקדם המתאם מתאפס כאשר סך המכפלות החיוביות של ציוני התקן מקזז את סך המכפלות השליליות. דבר זה יקרה, למשל, כאשר דיאגרמת הפיזור נראית כמו ענן שהוא פחות או יותר סימטרי ביחס לצירים, כמו בדיאגרמות שבאיור 16(א).




**16(ב)** **16(א)**

**איור 16. (א)**: אין קשר בין המשתנים; **(ב)**: קשר פרבולי בין המשתנים

עם זאת, מבט בדיאגרמה באיור 16(ב) מורה בבירור:

**אין לקפוץ מערך 0 של מקדם המתאם למסקנות גורפות בדבר חוסר קשר בין המשתנים!**

באיור 16(ב) הנקודות אכן מפוזרות באופן פחות או יותר שווה בין הרביעים, המכפלות החיוביות והשליליות מקזזות אלה את אלה ומקדם המתאם הוא לפיכך 0. אך כפי שרואים, הנקודות באיור 16(ב) למעשה מפוזרות בצורה טובה סביב **פרבולה** (המסומנת באיור), דבר המצביע דווקא על **קשר טוב** – שאינו קווי – **בין המשתנים**. וזאת אף שמקדם המתאם הוא 0 (דיון מפורט יותר על קשר שאינו קווי בהמשך).

עוד נשוב ונתח את שתי הדיאגרמות הללו בעזרת כלים שנרכוש בהמשך.

## 3.3 מקדם המתאם – נוסחאות חישוב מקובלות

בדוגמה 1 השתמשנו בנוסחה (1) המבוססת על ציוני התקן של המשתנים. עם זאת, מקובל יותר לחשב את מקדם המתאם ישירות מהנתונים המקוריים.

הצבת $\underset{\sim}{x}=\frac{x-\bar{x}}{\sigma_X}$ , $\underset{\sim}{y}=\frac{y-\bar{y}}{\sigma_Y}$ בנוסחה (1) נותנת $\underset{\sim}{x}=\frac{x-\bar{x}}{\sigma_X}$ , $\underset{\sim}{y}=\frac{y-\bar{y}}{\sigma_Y}$ :

$$r=\frac{1}{\sigma_X\cdot\sigma_Y}\cdot\frac{1}{n}\cdot\left\{\underset{\downarrow}{(x_1-\bar{x})\cdot(y_1-\bar{y})}+\cdots+\underset{\downarrow}{(x_n-\bar{x})\cdot(y_n-\bar{y})}\right\} \qquad (2)$$

**מכפלת הסטיות מהממוצע המתאים**

לאחר כמה פעולות אלגבריות פשוטות מתקבלת נוסחת החישוב המקובלת הבאה ל**מקדם המתאם** :

$$r=\frac{1}{\sigma_X\cdot\sigma_Y}\left\{\underbrace{\frac{1}{n}\cdot\left[x_1\cdot y_1+\ldots+x_n\cdot y_n\right]}_{\downarrow}-\underset{\downarrow}{\bar{x}\cdot\bar{y}}\right\} \qquad (3)$$

**מכפלת הממוצעים ממוצע המכפלות**

– נוסחה (1) קלה יותר להבנה, אולם נוסחה (3), שלכאורה נראית מסורבלת יותר, נוחה יותר לחישובים ידניים.

– כדי לקבל את מקדם המתאם על פי נוסחה (3) יש לחשב מהנתונים חמישה גדלים : הממוצעים של שני המשתנים, סטיות תקן של שני המשתנים וכן ממוצע המכפלות.

– שימוש קל ופשוט בתוכנת אקסל מייתר את כל החישובים הללו (ראו נספח ב).

**דוגמה 5.** בלוח 7 מוצגים נתוני הטמפרטורה המקסימלית (במעלות צלסיוס) והלחות היחסית (באחוזים) כפי שנמדדו בשמונה ישובים בישראל ב-10 באפריל 2009. בעזרת מחשבון קיבלנו את הממוצעים ואת סטיות התקן $\sigma_X=3.71$ ו- $\sigma_Y=11.39$ , וכן את מכפלת הציונים $x\cdot y$ .

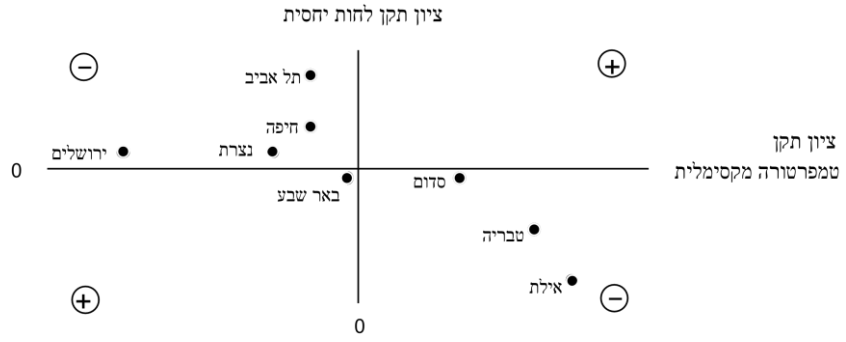
**לוח 7.** טמפרטורה מקסימלית ולחות יחסית בשמונה ישובים

| היישוב | טמפרטורה מקסימלית, $x$ | לחות יחסית, $y$ | $x\cdot y$ |
|---|---|---|---|
| אילת | 28 | 25 | 700 |
| באר שבע | 22 | 45 | 990 |
| חיפה | 21 | 55 | 1,155 |
| טבריה | 25 | 45 | 1,125 |
| ירושלים | 16 | 50 | 800 |
| נצרת | 20 | 50 | 1,000 |
| סדום | 27 | 35 | 945 |

| | | | |
|---|---|---|---|
| תל אביב | 21 | 65 | 1,365 |
| **ממוצעים** | $\bar{x} = 22.5$ | $\bar{y} = 46.25$ | 1,010 |

באיור 17 מוצגת דיאגרמת הפיזור של הנתונים מלוח 7. בדיאגרמה מתקבלת תמונה יפה של הקשר בין שני המשתנים, שקשה להבחין בו כאשר בוחנים רק את נתוני הלוח עצמו.

מהנתונים שבלוח 7 ציירנו את דיאגרמת הפיזור וסימנו בה את קווי הממוצעים. באמצעות הזזה מתאימה כלפי מטה ושמאלה (פרטו) מתקבלת דיאגרמת הפיזור המתוקננת שבאיור 17. בדיאגרמה מתקבלת תמונה של קשר טוב בין שני המשתנים, שקשה להבחין בו כאשר בוחנים את נתוני הלוח עצמו.




**איור 17.** דיאגרמת פיזור מתוקננת של שמונה יישובים לפי הטמפרטורה המקסימלית והלחות היחסית

כפי שרואים, רוב הנקודות בדיאגרמה המתוקננת נמצאות ברביעים השליליים ורק אחת (באר שבע) ברביע חיובי. נראה שככל שהטמפרטורה המקסימלית ביישוב גבוהה, כך הלחות היחסית בה **נוטה לרדת** – זהו **קשר יורד** בין המשתנים.

שימוש במחשבון או באקסל נותן את סטיות התקן $\sigma_X = 3.71$ ו- $\sigma_Y = 11.39$. מכל אלו נחשב את מקדם המתאם על פי נוסחה (3).

$$r = \frac{1}{3.71 \times 11.39}[1,010 - 22.5 \times 46.25] = -0.725 .$$

נשים לב שערכו של המתאם קרוב למדי ל- $(-1)$; ערך כגון זה מצביע על קשר **שלילי חזק למדי** בין המשתנים.

**שאלת חשיבה**. מה ישתנה בדיאגרמת הפיזור המתוקננת אם נמדוד את הטמפרטורות במעלות פרנהייט? מה יקרה למקדם המתאם?

## אז מה היה לנו? מתאם בין משתנים

### מושגים חדשים

- מקדם המתאם (correlation coefficient)
- קשר עולה, קשר יורד
- קשר חזק, קשר חלש

**תובנות חדשות**

- מקדם המתאם הוא מדד לטיב ולכיוון של הקשר **הקווי** בין שני משתנים. ערכו אינו תלוי ביחידות המדידה. שינוי ליניארי (קווי) בקנה המידה של המשתנים אינו משנה את מקדם המתאם.
- מתאם **חיובי** מעיד על קשר **עולה** בין המשתנים; מתאם **שלילי** מעיד על קשר **יורד** בין המשתנים.
- מקדם המתאם מקבל ערכים בין (1–) ל-(1+).
- מקדם המתאם מקבל את הערך 1 רק כש**כל** הערכים בדיאגרמת הפיזור נמצאים על קו ישר עולה, ואת הערך (1–) רק ש**כל** הערכים בדיאגרמת הפיזור נמצאים על קו ישר יורד. מצבים כאלה מתקבלים רק כאשר ערכי המשתנה $Y$ הם תוצאה של שינוי בקנה המידה של המשתנה $X$.
- מקדם מתאם קרוב לאפס אינו מעיד בהכרח על חוסר קשר בין שני המשתנים, אלא על חוסר קשר קווי.

**כלים חדשים**

- מקדם המתאם של פירסון

$$r = r(X,Y) = \frac{1}{n} \cdot \left[ \underset{\downarrow}{x_1 \cdot y_1} + \ldots + x_n \cdot y_n \right] \qquad (1)$$

**סכום מכפלות ציוני התקן**

- מקדם המתאם – נוסחת חישוב מקובלת (ראו גם נוסחה 2 בעמ' 45):

$$r = \frac{1}{\sigma_X \cdot \sigma_Y} \left\{ \underbrace{\frac{1}{n} \cdot \left[ x_1 \cdot y_1 + \ldots + x_n \cdot y_n \right]}_{\downarrow} - \underset{\downarrow}{\bar{x} \cdot \bar{y}} \right\} \qquad (3)$$

**מכפלת הממוצעים** **ממוצע המכפלות**

בחישוב ידני של מקדם המתאם על פי נוסחה (3), יש לחשב תחילה ממוצעים וסטיות תקן של המשתנים וכן את ממוצע המכפלות (חמישה גדלים). את הממוצעים וסטיות התקן קל לקבל באמצעות מחשבון ואולי גם את מקדם המתאם $r$.

לחישוב קל ומהיר של מקדם המתאם (של פירסון) בעזרת אקסל יש להקליד את הנתונים בדף העבודה בשתי עמודות, ולרשום בפקודה הבאה את טווח התאים המתאים לכל משתנה:

| | הסימון | פקודת אקסל |
|---|---|---|
| מקדם המתאם | $r$ | $=correl(\_:\_,\_:\_)$ |

## משימות חישוב וחשיבה – מתאם בין משתנים

(פתרונות בעמ' 115)

המטרה במשימות הבאות היא לחשב את מקדם המתאם **חישוב ידני**, על פי ההנחיות. מומלץ מאוד לצייר תחילה דיאגרמות פיזור, אפשר באקסל.

**משימה I. ערכי קיצון** – שימוש בנוסחה (2)

א. $X$ הוא המרחק היומי הכולל, בקילומטרים, שצעד יואב עם כלבו האהוב סטאר במהלך ארבעה ימים. $Y$ הוא המרחק כפי שנמדד במיילים (1 מייל = 1.6 קילומטר). רשמו את הנוסחה לחישוב ערכי $Y$, והשלימו את עמודת ערכי $Y$.

ב. חשבו את ממוצע ערכי $X$, **והסיקו** ממנו את ממוצע ערכי $Y$.

| **מרחק בק"מ,** $x$ | **מרחק במייל,** $y$ | **ציון תקן,** $\tilde{x}$ | **ציון תקן,** $\tilde{y}$ | **מכפלת ציוני התקן,** $\tilde{x}\cdot\tilde{y}$ |
|---|---|---|---|---|
| 2.2 | | | | |
| 1.6 | 1 | | | |
| 1 | | | | |
| 0.8 | | | | |
| 3.2 | | | | |
| $\bar{x}=$ | $\bar{y}=$ | | | $r=$ |

**ממוצעים**

ג. חשבו שונות וסטיית תקן של $X$, **הסיקו מכך** את סטיית התקן של $Y$. נזכיר, הוספה של קבוע למשתנה $X$ אינה משנה את סטיית התקן; הכפלה בקבוע חיובי מכפילה את סטיית התקן באותו קבוע.

ד. חשבו את ציוני התקן של שני המשתנים. מה גיליתם? חשבו את מכפלת ציוני התקן.

ה. חשבו את **מקדם המתאם** בין שני המשתנים. הסבירו את התוצאה והכלילו אותה.

**משימה II. ציון פרויקט וציון תעודה** (המשך דוגמה 1)

א. בלוח 1 (זהו אותו לוח 1 שבעמ' 16) מִצאו (בדרך שתבחרו) את ממוצע המכפלות $x\cdot y$ (בעמודה השמאלית למטה).

| **ציון פרויקט,** $x$ | **ציון תעודה,** $y$ | $x\cdot y$ |
|---|---|---|
| 30 | 60 | 1800 |
| 50 | 90 | |
| 40 | 65 | |

| | | |
|---|---|---|
| | 95 | 95 |
| | 100 | 100 |
| | 90 | 90 |
| | 80 | 75 |
| | 65 | 65 |
| | 85 | 85 |
| | 90 | 80 |
| $\frac{1}{10}\cdot\left[x_1\cdot y_1+\ldots+x_{10}\cdot y_{10}\right]=$ | $\bar{y}\ =\ 82$ | $\bar{x}=71$ |

**ממוצעים**

ב. עתה היעזרו בנוסחה (3) כדי לחשב את מקדם המתאם. השוו לתוצאת החישוב בעמ' 42.

**משימה III. חמודי הסבות מקצועיות** (המשך משימה III מעמ' 38)

מנכ"ל חברת חמודי להסבות מקצועיות עורך סדנאות אימון למלצרים מתחילים. כדי לברר אם שיטת האימון שלו יעילה, הוא רשם לכל אחד מהמועמדים ניקוד על פי ביצועיו בתום ימי הסדנה האישית, שנמשכה בין יום אחד לארבעה ימים. התקבלה טבלה זו:

| **ימי אימונים,** $x$ | **ניקוד,** $y$ | $x\cdot y$ | **מכפלת ציוני התקן** $\tilde{x}\cdot\tilde{y}$ |
|---|---|---|---|
| 1 | 4 | | |
| 1 | 5 | | |
| 1 | 5.5 | | |
| 2 | 6 | | |
| 2 | 7.5 | | |
| 2 | 8 | | |
| 3 | 7.8 | | |
| 3 | 7.3 | | |
| 4 | 6.5 | | |
| 4 | 6 | | |
| | | | |
| $\bar{x}=$ | $\bar{y}=$ | | $r=$ |

**סה"כ**

**ממוצע**

א. השתמשו בציוני התקן שחישבתם בפרק 2 בעמ׳ 37 כדי להשלים את העמודה השמאלית – מכפלת ציוני התקן. קבלו באמצעותם את מקדם המתאם על פי נוסחה (1).

ב. עתה השלימו גם את עמודת המכפלות $x \cdot y$ בטבלה וחשבו בעזרתה מחדש את **מקדם המתאם** בין שני המשתנים על פי נוסחה (3). השוו את התוצאות.

ג. כל יום אימונים נמשך 8 שעות. איך ישתנה מקדם המתאם אם במקום ימי אימונים נמדוד שעות אימונים?

**משימה IV. האם שכר עובדים ממריא עם גובה העובד?**

לאחד העובדים בחברת היי טק הייתה תחושה שהעובדים הגבוהים מקבלים שכר גבוה יותר מהעובדים הנמוכים. הנתונים בלוח שלהלן הם $x$ – הגובה (בסנטימטרים) ו-$y$ – השכר החודשי (באלפי שקלים) של 14 עובדים בדרג בינוני-נמוך בחברה זו.

א. ציירו דיאגרמת פיזור של השכר לפי הגובה. ללא שימוש באקסל אפשר להסתפק בניתוח עובדים 1 עד 7.

| העובד | **גובה** | **שכר** |
|---|---|---|
| 1 | 196 | 15.7 |
| 2 | 188 | 15.1 |
| 3 | 185 | 15.4 |
| 4 | 183 | 15.4 |
| 5 | 178 | 14.2 |
| 6 | 179 | 12.8 |
| 7 | 175 | 15.1 |

| העובד | **גובה** | **שכר** |
|---|---|---|
| 8 | 173 | 12.1 |
| 9 | 170 | 14.0 |
| 10 | 169 | 12.4 |
| 11 | 167 | 11.4 |
| 12 | 163 | 11.7 |
| 13 | 160 | 11.1 |
| 14 | 152 | 11.0 |

ב. חשבו את מקדם המתאם בין הגובה לשכר.

ג. מה יהיה ערכו של מקדם המתאם אם הגובה יימדד באינצ׳ים (1 אינץ׳ = 2.54 ס״מ) והשכר בדולרים?

**משימה V. סרטן עור וקרינה אולטרה סגולה**

אחד הגורמים לסרטן עור מסוג מלנומה הוא הקרינה האולטרה סגולה מהשמש. כמות הקרינה תלויה בעובי שכבת האוזון, ועובְיָה שונה בקווי רוחב שונים. בעמודה השמאלית מוצג שיעור מקרי המלנומה ל-100,000 תושבים, כפי שנמדד בתשעה אזורים בקווי רוחב שונים בארצות הברית במהלך שלוש שנים.

| **קו רוחב,** $x$ | **שיעור מקרי מלנומה,** $y$ |
|---|---|
| 32.8 | 9 |
| 33.9 | 5.9 |

| | |
|---|---|
| 34.1 | 6.6 |
| 37.9 | 5.8 |
| 40.0 | 5.5 |
| 40.8 | 3.0 |
| 41.7 | 3.4 |
| 42.2 | 3.1 |
| 45.0 | 3.8 |
| $\bar{x}=38.71$ | $\bar{y}=5.12$ |

**ממוצעים**

א. ציירו **דיאגרמת פיזור** של שיעור מקרי המלנומה לפי קו רוחב (מומלץ להיעזר באקסל). האם הקשר בין $X$ ו-$Y$ נראה קווי? חִשבו על עקומה אחרת שנראית לכם מתאימה יותר ושרטטו אותה ביד חופשית.

ב. הממוצעים רשומים בטבלה, סכומי הריבועים הם: $x_1^2+\ldots+x_9^2=13633.64$, $y_1^2+\ldots+y_9^2=267.87$. חשבו את ממוצעי המכפלות.

היעזרו בכל אלו כדי לחשב ידנית את סטיות התקן ואת **מקדם המתאם** בין המשתנים.

**משימה VI. עברית שפה קשה** (המשך משימה I מעמ' 37)

בטבלה שלהלן רשומים הנתונים מעמ' 36 וחישובים נוספים שערכנו. נזכיר, $\sigma_X=2.22$, $\sigma_Y=3.16$.

א. השלימו את הטבלה והיעזרו בה כדי לחשב ידנית את מקדם המתאם על פי נוסחה (2) (עמ' 45).

| **מבחן A**, $x$ | **מבחן B**, $y$ | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})(y-\bar{y})$ |
|---|---|---|---|---|
| 2 | 5 | -3.5 | -5 | 17.5 |
| 5 | 8 | -0.5 | -2 | |
| 9 | 15 | 3.5 | 5 | |
| 6 | 12 | 0.5 | 2 | |
| 7 | 11 | 1.5 | 1 | |
| 4 | 9 | -1.5 | -1 | |
| $\bar{x}=5.5$ | $\bar{y}=10$ | | | |

**ממוצעים**

ב. עתה היעזרו בחישוב ציוני התקן שערכתם בפרק 2 כדי לקבל את מקדם המתאם על פי נוסחה (1).

ג. נתבונן במשתנה $X^*$ – מספר השגיאות במבחן $A$. מהו הקשר בין X ל-$X^*$ ? הסיקו את הממוצע וסטיית התקן של $X^*$ .

ד. מהו מקדם המתאם בין $X^*$ ל-$Y$ ? הסבירו כיצד אפשר להסיק זאת ללא חישובים באמצעות שינוי מתאים בערכים שחישבתם בסעיף א. לחלופין, בנו עבור $X^*$ ו-$Y$ טבלה מקבילה לזו.

- לחישובים הנדרשים במשימות הבאות מומלץ מאוד להשתמש בתוכנת אקסל.

**משימה VII. פשיעה בדטרויט**[3]

בלוח שלהלן מוצגים נתוני שיעור מקרי הרצח (מספר המקרים ל-100,000 תושבים) בשנים 1961–1973 בעיר דטרויט בארצות הברית.

| **שיעור מקרי הרצח,** $y$ | **שיעור השוטרים,** $x_1$ | **שכר ממוצע לשעה,** $x_2$ |
|---|---|---|
| 8.60 | 260.35 | 2.98 |
| 8.90 | 269.80 | 3.09 |
| 8.52 | 272.04 | 3.23 |
| 8.89 | 272.96 | 3.33 |
| 13.07 | 272.51 | 3.46 |
| 14.57 | 261.34 | 3.60 |
| 21.36 | 268.89 | 3.73 |
| 28.03 | 295.99 | 2.91 |
| 31.49 | 319.87 | 4.25 |
| 37.39 | 341.43 | 4.47 |
| 46.26 | 356.59 | 5.04 |
| 47.24 | 376.69 | 5.47 |
| 52.33 | 390.19 | 5.76 |

[3] ראו מאמר : Fisher, J. C. (1976). Homicide in Detroit: The Role of Firears. *Criminology*. *vol.14*, 387–400.

על סמך 13 התצפיות הללו רוצים לנבא את שיעור מקרי הרצח בעיר בשנה. מתוך כלל 13 המשתנים המנבאים שהיו במחקר המקורי התמקדנו כאן בשני **משתנים מנבאים**: $X_1$ הוא מספר השוטרים ל-100,000 תושבים, ו- $X_2$ הוא השכר הממוצע לשעה.

א. מהנתונים הללו חשבו בעזרת **פקודת אקסל** (ראו בעמ' 48) את מקדמי המתאם בין שיעור מקרי הרצח לבין כל אחד משני המשתנים המנבאים.

ב. לפי דעתכם, האם כל אחד מהמשתנים הללו בנפרד יעיל לניבוי שיעור מקרי הרצח בעיר? [בספר זה לא נכנסנו לנושא **רגרסיה מרובה** – ניבוי ערך של משתנה על בסיס כמה משתנים מנבאים.]

ג. האם לדעתכם אפשר להשתמש בנתונים אלו לניבוי מספר מקרי הרצח בדטרויט בשנה הבאה?

### משימה VIII. מבחני פיז"ה

א. מנתוני לוח 4 (בעמ' 26) חשבו את מקדם המתאם בין התמ"ג לציון במתמטיקה. מה משמעות התוצאה?

ב. מצאו גם את ציוני התקן של שני המשתנים עבור ישראל והשוו ביניהם. מה גיליתם?

### משימה IX. חורף במיניאפוליס

בטבלה מוצגים נתוני הטמפרטורה הממוצעת בכל אחד מחודשי השנה בפרוור של מיניאפוליס בארצות הברית וכמות השוקולד (בקילוגרמים) שנמכרה באותו חודש (הנתונים פיקטיביים).

| | **כמות השוקולד,** $y$ | **טמפרטורה ממוצעת,** $x$ |
|---|---|---|
| ינואר | 98 | -5 |
| פברואר | 100 | -7 |
| מרץ | 75 | 5 |
| אפריל | 67 | 10 |
| מאי | 24 | 18 |
| יוני | 26 | 22 |
| יולי | 25 | 28 |
| אוגוסט | 27 | 25 |
| ספטמבר | 40 | 16 |
| אוקטובר | 55 | 10 |
| נובמבר | 88 | 2 |
| דצמבר | 95 | -3 |

א. ציירו דיאגרמת פיזור ידנית או באמצעות אקסל (ראו הוראות בעמ' 36), ונתחו את הקשר שהתקבל.

ב. הקלידו את הנתונים בדף עבודה של אקסל וחשבו את מקדם המתאם (ראו פקודה בעמ' 48).

### משימה X. הגיל והטלפון

בניסיון לבדוק את הקשר בין גיל המשתמש לבין מספר הכניסות לטלפון הסלולרי במהלך יום לימודים נבדקו עשרים נערים ונערות בגילים שונים, והתוצאות (פיקטיביות) נרשמו בטבלה.

א. ציירו את הנתונים בדיאגרמת פיזור ונתחו את הקשר המסתמן.

ב. מצאו את מקדם המתאם בין הגיל לבין מספר שיחות הטלפון היומיות. הסבירו את משמעות התוצאה.

| שם | גיל | מספר כניסות |
|---|---|---|
| א | 7 | 1 |
| ב | 14 | 8 |
| ג | 9 | 4 |
| ד | 10 | 3 |
| ה | 8 | 3 |
| ו | 17 | 7 |
| ז | 11 | 5 |
| ח | 13 | 6 |
| ט | 16 | 8 |
| י | 17 | 4 |

| שם | גיל | מספר כניסות |
|---|---|---|
| כ | 8 | 3 |
| ל | 18 | 5 |
| מ | 14 | 3 |
| נ | 9 | 3 |
| ס | 10 | 2 |
| ע | 8 | 1 |
| פ | 12 | 5 |
| צ | 9 | 1 |
| ק | 15 | 6 |
| ר | 8 | 2 |

### תרגילים: מתאם בין משתנים

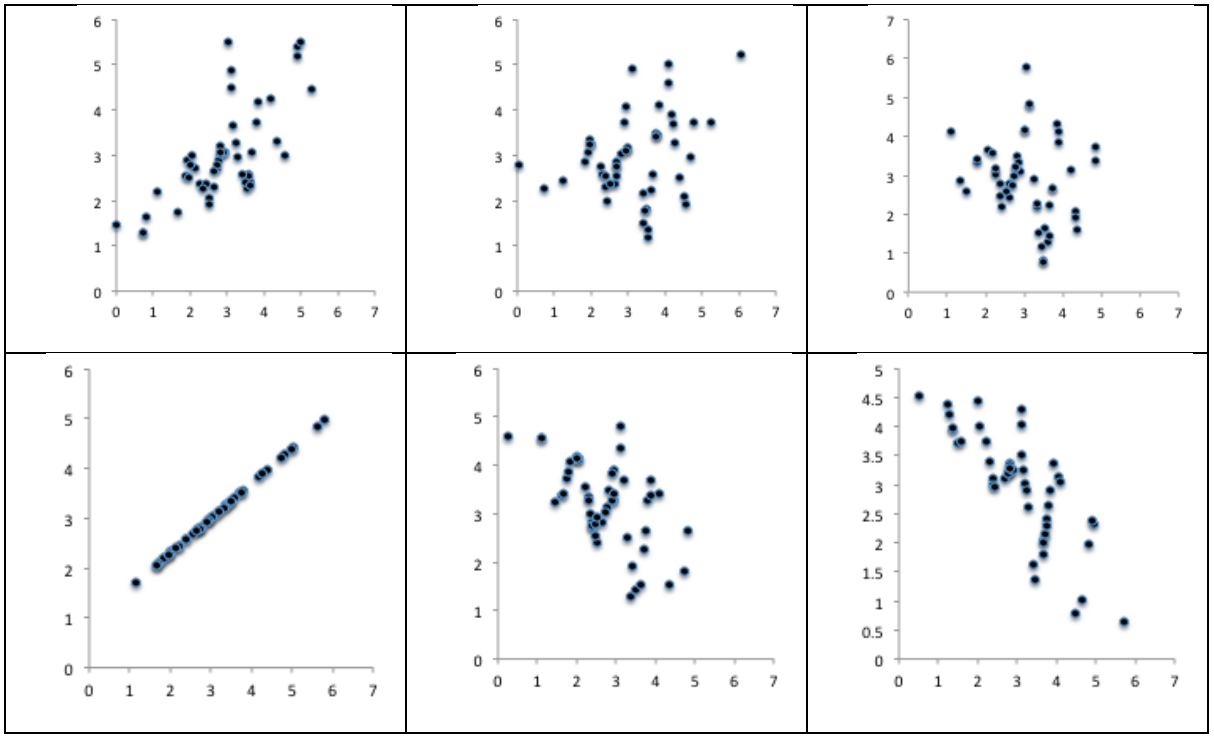
(פתרונות מקוצרים בעמ' 134)

**תרגיל 1** (המשך שאלה 3 מעמ' 38). אלו הם מקדמי המתאם של שש הדיאגרמות, לפי סדר עולה: 0, 0.4, 0.6, 0.8, 0.9, 0.95. התאימו לכל אחת מהדיאגרמות את מקדם המתאם המתאים לה.

**תרגיל 2.** לפניכם שש דיאגרמות פיזור, ורשימה של ערכי מקדמי המתאם שלהן:

0.3, 1.0, -0.20, 0.71, -0.93, -0.75.

התאימו לכל דיאגרמה את מקדם המתאם שלה מהרשימה. הסבירו את בחירתכם.



**תרגיל 3.** להלן דוגמאות של זוגות של משתנים. רשמו מהי הערכתכם באשר למקדם המתאם שיתקבל מאוסף מדידות של כל אחד מזוגות המשתנים הללו. האם המתאם יהיה חיובי או שלילי? האם יהיה קרוב לאפס או גדול בערכו המוחלט?

א. מספר המכוניות בכביש ומספר תאונות הדרכים;

ב. משקל המכונית וצריכת הדלק שלה (נמדדת במספר קילומטרים לליטר);

ג. הטמפרטורה המקסימלית ביום וכמות הגלידה שנמכרה ביום זה;

ד. מספר שנות הלימוד של שכיר והכנסתו;

ה. שטחו של מעגל והיקפו;

ו. ציון במתמטיקה וציון בלשון של תלמיד;

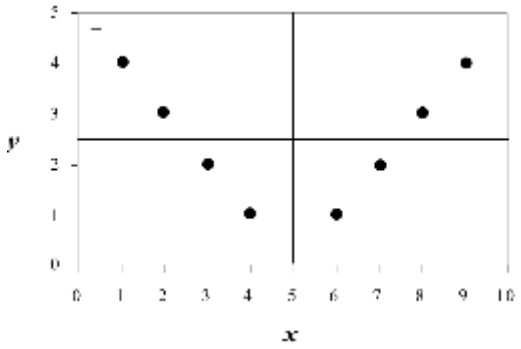
ז. מספר גני הילדים בערים בישראל והכנסות כל עירייה מקנסות על דוחות חניה בעיר.

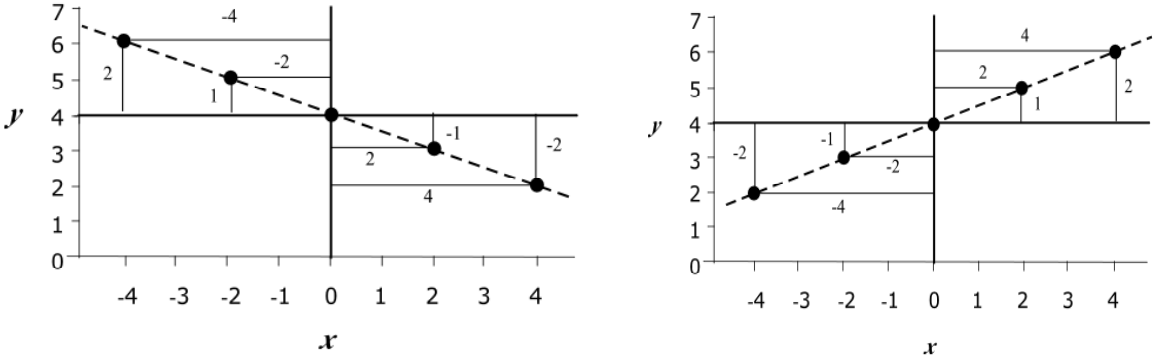
**תרגיל 4.** לפניכם שלוש דיאגרמות פיזור. חשבו את מקדם המתאם בין שני המשתנים בכל דיאגרמה.

א. דיאגרמת הפיזור המתוקננת העליונה היא סימטרית: הראו שמקדם המתאם כאן הוא $r = 0$. הסבירו.

ב. בדיאגרמה הימנית למטה הראו שמתקבל $r = 1$. הסבירו.

ג. בדיאגרמה השמאלית למטה הראו שמתקבל $r = -1$. הסבירו.





**תרגיל 5.** נבדקו משקליהם של שניים-עשר זוגות של אבות ובנים צעירים, ונמצא שהמתאם בין משקל האב למשקל הבן הוא $r = 0.28$.

א. אחרי שנה התברר שהאבות שמרו על משקלם הקודם, ואילו לכל אחד מהבנים נוספו 3 קילוגרם למשקל. מהו מקדם המתאם החדש?

ב. מהו מקדם המתאם החדש, אם תוספת המשקל של כל אחד מהבנים היא 20% ממשקלו הקודם.

**תרגיל 6.** בדוגמה 5 בעמ' 45, הטמפרטורות המקסימליות $x$ נתונות במעלות צלזיוס. מה ישתנה בלוח 6 אם הטמפרטורות יימדדו במעלות פרנהייט ($x^*$)? מה יקרה למקדם המתאם? [נוסחת המעבר בין צלזיוס לפרנהייט היא $x^* = 32 + (9/5)\,x$.]

# פרק 4

# בעיות ניבוי

לעיתים יש צורך לבטא את הקשר בין משתנים בצורה מפורשת, באופן שיאפשר **ניבוי** ערך עתידי של אחד המשתנים על סמך ערכו הידוע של המשתנה האחר. למשל:

כמות הגשם מחר על פי כמות הגשם היום; אחוז האבטלה בחודש אפריל על סמך אחוז האבטלה בחודש מרץ; מידת המעורבות בתאונות דרכים על פי גיל הנהג; מידת ההצלחה בתפקיד של מועמד על סמך נתונים שונים של המועמדים; הצלחה בלימודים גבוהים על סמך ציוני הבגרות של תלמידים.

אנו מחפשים אפוא **נוסחה פשוטה** שתאפשר **ניבוי** כזה בקירוב שיהיה טוב דיו.

## 4.1 שימוש בממוצעים לצורכי ניבוי

בידינו אוסף נתונים של משתנה כמותי יחיד – משך הזמן שחיכינו לאוטובוס במהלך עשרה ימים, גיל הנישואים של נשים בישראל וכדומה – ואנו עומדים לערוך מדידה נוספת של אותו משתנה. מהו הניבוי הטוב ביותר לתוצאת הניסוי?

**ניבוי ללא ידע נוסף**

♥ הניבוי המקובל בסטטיסטיקה לערכו של משתנה בניסוי עתידי כלשהו הוא **הממוצע של ערכי המשתנה שבידינו**.

בבעיות ניבוי רבות אין קושי לאסוף שפע של **נתונים סטטיסטיים** שהניבוי יתבסס עליהם, למשל:

– לניבוי המשקל $y$ של תינוק בלידתו נשתמש בממוצע הכללי $\bar{y} = 3.500$ קילוגרם.

– לניבוי גיל של כלה בישראל ביום נישואיה, נשתמש בממוצע גיל כלות בישראל, שהוא 25 (נתונים מ-2017). [ראו גם נספח ג (עמ' 105) – אתר הלשכה המרכזית לסטטיסטיקה.]

**ציונים פסיכומטריים** (המשך דוגמה 3): הניבוי הטוב ביותר לציון סוף שנה א של תלמיד מקרי במתמטיקה הוא הציון הממוצע של 198 התלמידים שבדקנו: $\bar{y} = 62.8$.

**ניבוי המבוסס על ידע נוסף**

בניבוי משקל של תינוק בלידתו, ייתכן שנוכל להגיע לידע יותר ספציפי שאפשר להתבסס עליו בניבוי, למשל מין התינוק, ארץ הלידה, באיזה שבוע של ההיריון נולד התינוק. כיצד משתמשים בידע כזה לצורכי הניבוי?

בחישוב ממוצע של ערכי המשתנה נצטמצם לנתונים הרלוונטיים לידע שבידינו. למשל:

– לניבוי המשקל של תינוק שנולד בישראל נשתמש בממוצע הישראלי $\bar{y} = 3.400$. הניבוי לתינוק בדרום-מזרח אסיה הוא $\bar{y} = 3.260$ ק״ג וכן הלאה.

– לניבוי גיל של כלה בישראל ביום נישואיה, עשוי להיות בידינו ידע נוסף כמו גיל החתן. מנתוני לוח 8 (גיל חתנים וכלות בישראל) בנספח ג של הלשכה המרכזית לסטטיסטיקה בישראל (הלמ״ס), חישבנו ומצאנו שלחתנים בני 18 ממוצע גיל הכלות הוא 19.

ברשות הלמ״ס יש נתונים רבים ומגוונים על אוכלוסיית ישראל. מומלץ מאוד לחפש באתר המתאים או אפילו לפנות בכתב באתר: www.cbs.gov.il.

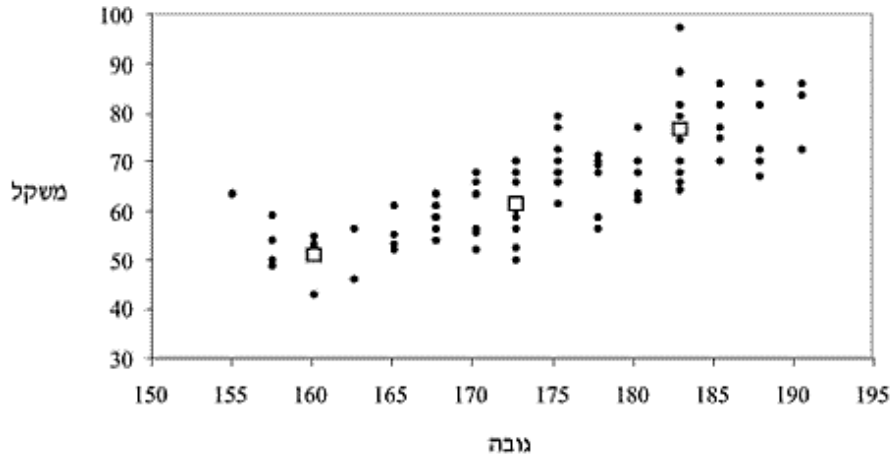
**באופן כללי**: בידינו אוסף נתונים של משתנה כמותי $Y$ ושל משתנה נוסף $X$ העשוי להיות רלוונטי לניבוי. אם ידוע לנו שבניסוי חדש המשתנה $X$ קיבל את הערך $x$, מהו הניבוי הטוב ביותר לערך הבלתי ידוע של המשתנה $Y$?

**מינוח**: המשתנה שעליו ננסה לבסס את הניבוי ייקָרא **משתנה מנבא** או משתנה מסביר.

❤ הניבוי הטוב ביותר ל-$Y$ בהינתן שהמשתנה המנבא $X$ קיבל את הערך $x$, הוא **הערך הממוצע של המשתנה $Y$, שיחושב רק לאותם נתונים שבהם $X$ קיבל את הערך $x$**.

**ציונים פסיכומטריים** (המשך דוגמה 3). הניבוי הטוב ביותר לציון סוף שנה א של תלמיד מתמטיקה כלשהו הוא כאמור הציון הממוצע של 198 התלמידים שבדקנו $(\bar{y} = 62.8)$. ואולם אם יתברר לנו שהציון הפסיכומטרי של התלמיד הוא 650, הרי הניבוי הטוב ביותר בעבורו הוא הציון הממוצע בשנה א שיחושב רק לאותם תלמידים שהציון הפסיכומטרי שלהם הוא 650. בדקנו ומצאנו שהניבוי המתקבל במקרה זה גבוה יותר – 69.1.

**דוגמה 6** (גובה ומשקל). באיור 18 מוצגת דיאגרמת פיזור של הגובה והמשקל של 92 סטודנטים בקורס מבוא לפסיכולוגיה. המשקל הממוצע של הסטודנטים הוא 66 קילוגרם. זהו גם הניבוי שלנו למשקלו של סטודנט נוסף המצטרף לקורס.




**איור 18.** דיאגרמת פיזור של סטודנטים לפי גובהם ומשקלם

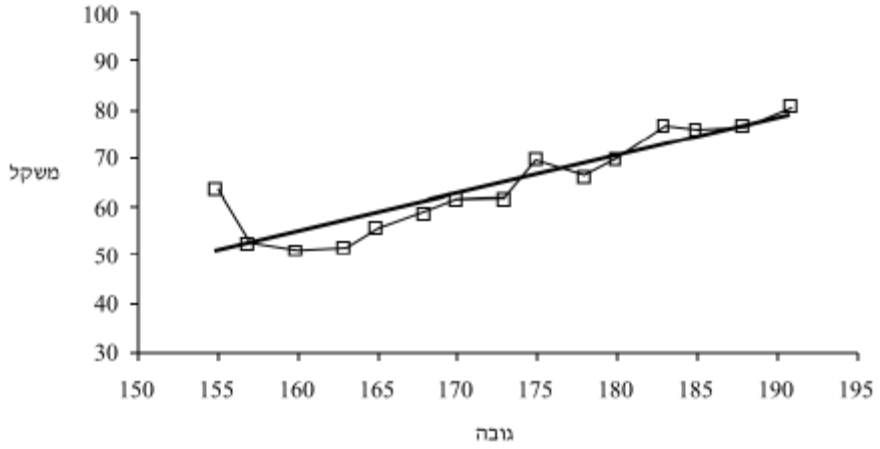
נשים לב שריבוי הנתונים שבידינו – שילוב של גובה ומשקל – מאפשר לחשב לכל גובה בנפרד את ממוצע המשקל של קבוצת הסטודנטים שזה גובהם. ממוצעים אלו מאפשרים לנבא ניבוי טוב את משקלו של סטודנט נוסף שיצטרף לקורס מתוך שימוש בידע שלנו על גובהו.
לדוגמה, נתבונן בסטודנט נוסף, שגובהו 173 ס״מ: הניבוי הטוב ביותר למשקלו יהיה המשקל הממוצע של הסטודנטים בגובה זה (מסומן בריבוע באיור 18). בדרך דומה נוכל לנבא את המשקל של כל סטודנט נוסף על בסיס גובהו הנתון, על פי המשקל הממוצע של כל הסטודנטים בעלי אותו גובה.

באיור 18 סימנו בריבועים את המשקל הממוצע לשלוש קבוצות של סטודנטים שווי גובה: קבוצת הסטודנטים שגובהם 160 ס״מ, סטודנטים שגובהם 173 ס״מ וסטודנטים שגובהם 183 ס״מ.

## 4.2 עקום הממוצעים

**המשך דוגמה 6**. באיור 19 מצוינים בריבועים (ללא פרטי דיאגרמת הפיזור) ממוצעי המשקל שחושבו לכל אחד מהגבהים בנפרד. אם מחברים את הממוצעים הללו מתקבל מעין קו עקום שנקרא **עקום הממוצעים**, המבטא את הניבוי הטוב ביותר לערכו של המשתנה $Y$ (המשקל) בהינתן ערכו של $X$ (הגובה).

[משמעות הקו הישר העבה שבאיור, שהוא קו הרגרסיה, תוסבר בהמשך.]




**איור 19.** עקום הממוצעים – המשקל הממוצע של סטודנטים בגבהים שונים

(באיור מופיע גם קו הרגרסיה המתאים, כפי שיוסבר בהמשך)

**קשיים**

- כדי לאפשר ניבוי של המשקל על פי הגובה בעזרת עקום הממוצעים, עלינו לחשב את המשקלים הממוצעים **לכל אחת** מקבוצות הגובה, ומובן שזוהי טרחה מרובה.
- אם יגיע סטודנט חדש שגובהו 172 או 152 ס״מ, לא נוכל לחשב ממוצע ולהציע ניבוי מתאים למשקלו מכיוון שבין הנתונים שבידינו לא היו סטודנטים בגובה זה בדיוק.

משום כך אנו מעוניינים להגיע ל**נוסחת ניבוי פשוטה** לערכו של $Y$ בהינתן ערכו של $X$, נוסחה שבה נוכל להציב כל ערך אפשרי $x$ ולקבל עבורו ניבוי לערכו המתאים של $Y$. הנוסחה יכולה לתאר קו ישר, פרבולה או כל צורה אחרת של קשר שזיהינו בדיאגרמת הפיזור.

❤ בספר זה נעסוק בניבוי באמצעות קו ישר (ניבוי ליניארי) בלבד.

אנו מחפשים אפוא קו ישר שינבא 'בצורה הטובה ביותר' את ערכו של $Y$ עבור כל ערך נתון של $X$.

## 4.3 ניבוי באמצעות קו ישר

לפנינו דיאגרמת פיזור של $n$ זוגות נתונים $(x_1, y_1), \ldots, (x_n, y_n)$.

נזכיר:

יש להתבונן תחילה בדיאגרמת הפיזור כדי להתרשם שאכן לענן הנקודות יש נטייה כללית להסתדר סביב קו ישר. אם הענן מצביע על חוסר קשר או על קשר שאינו קווי, למשל פרבולה, אין טעם להמשיך ולהפעיל את הטכניקות שבסעיף זה.

אנו מחפשים קו ישר שיהיה קרוב ככל האפשר לנקודות בדיאגרמת הפיזור, כדי שנוכל להשתמש בו לצורכי ניבוי. נזכיר שמשוואת הקו הישר היא $y = a + bx$.

**סימון**: נסמן ב- $\hat{y}_x = a + bx$ את ערכו של **הניבוי ל-$Y$** עבור הערך $x$ על פי הקו הישר המוצע.

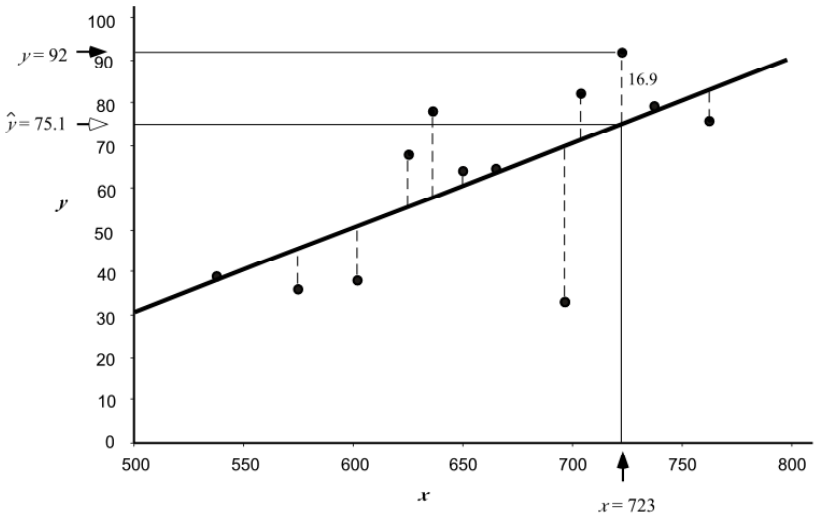
***הסטייה*** בין הערך שהתקבל בפועל עבור $x$ מסוים ובין הניבוי על פי הקו המוצע היא $y-\hat{y}_x$. טיבו של קו ניבוי נקבע על בסיס סטיות אלו, ואנו נרצה כמובן להקטין אותן ככל האפשר.

**דוגמה 7 (ציונים פסיכומטריים – נתונים חלקיים)**. בדיאגרמת הפיזור באיור 20 שבהמשך מוצגים נתוני הציון הפסיכומטרי ונתוני סוף שנה א של **מקצת** תלמידי המתמטיקה. באיור מוצג גם קו ניבוי מוצע: $\hat{y}_x = 0.2x - 69.5$. בהמשך נברר מדוע הצגנו כאן דווקא קו זה.

כדי לנתח את המשמעות של קו זה כקו ניבוי נחשב כמה סטיות טיפוסיות:

– לתלמיד שהציון הפסיכומטרי שלו הוא $x=723$, הניבוי לציון שנה א על פי קו הניבוי הוא $\hat{y}_{723} = 0.2\times 723 - 69.5 = 75.1$ (ראו חץ לבן על ציר ה-$y$). אך בפועל התלמיד השיג ציון גבוה יותר: $y=92$ (חץ שחור). הסטייה בין הערך שהתקבל לבין הניבוי היא: $y-\hat{y}_{723} = 92-75.1 = 16.9 > 0$.

– לתלמיד שהציון הפסיכומטרי שלו הוא $x=700$, הניבוי לציון שנה א על פי הקו הוא $\hat{y}_{700} = 0.2\times 700 - 69.5 = 70.5$, והוא גבוה מהערך שהתקבל בפועל: $y=32$. במקרה זה הסטייה היא $32-70.5 = -38.5 < 0$. בדרך זו אפשר לחשב את הסטיות של כל שאר הנקודות בדיאגרמה.




**איור 20.** דיאגרמת פיזור וקו ניבוי מוצע ל-12 מתלמידי המתמטיקה. $x$ – ציון פסיכומטרי, $y$ – ציון שנה א המתאים, $\hat{y}$ – הניבוי המתאים על פי הקו. הסטיות בין הניבוי לבין הערך שהתקבל בפועל מקווקוות.

באופן כללי, ***סטיות מקו הניבוי***: לכל אחת מ-$n$ הנקודות $(x,y)$ בדיאגרמת הפיזור אנו בוחנים את הסטייה בין $y$, שהוא הערך שהתקבל בפועל עבור $x$, לבין ערך הניבוי עבור $x$ על פי קו הניבוי המוצע $\hat{y}_x = a+bx$. בדרך זו ימצאו ערכי $y$ הגבוהים מערך הניבוי $\hat{y}_x$ שעל הקו, וערכים אחרים הנמוכים מערך

הניבוי (ראו איור 20). כלומר, סביב הקו יש סטיות חיוביות וסטיות שליליות. כדי לנטרל את הסימן, אנו בוחנים את ריבועי הסטיות:

לכל נקודה $(x, y)$ אנו מסתכלים על **ריבוע הסטייה** $\left(y-\hat{y}_x\right)^2$.

**חזרה לדוגמה 7**: ריבוע הסטייה בין הערך שהתקבל לבין הניבוי לתלמיד שהציון הפסיכומטרי שלו הוא $x=723$, הוא $\left(y-\hat{y}_{723}\right)^2=\left(92-75.1\right)^2=285.61$.

❤ **מידת הקִרבה** של קו ישר כלשהו $y=a+bx$ לנקודות בדיאגרמת הפיזור נמדדת באמצעות **סכום ריבועי הסטיות** בין ערכי $y$ שהתקבלו בפועל לבין ערכי הניבוי שלהם $\hat{y}$ על פי אותו קו.

**מינוח**: ***עקרון הריבועים הפחותים*** מורה לבחור את הקו שסכום ריבועי הסטיות ממנו הוא **מינימלי**.

הקו הישר שסכום ריבועי הסטיות ממנו הוא **מינימלי** נקרא ***קו הריבועים הפחותים***.

כל הפרק הבא מוקדש למציאת נוסחה כללית לקו הריבועים הפחותים.

✋ אפשר להפעיל את עקרון הריבועים הפחותים גם עבור עקומות נוספות, למשל פרבולה, כאשר הן המתאימות ביותר לנתונים, אך הנוסחאות אינן פשוטות.

**אז מה היה לנו? בעיות ניבוי**

**מושגים חדשים**

- עקום הממוצעים
- משתנה מנבא
- ניבוי ליניארי (קווי)
- סטיות מקו הניבוי
- עקרון הריבועים הפחותים

**תובנות חדשות**

- הניבוי המקובל בסטטיסטיקה לערכו של משתנה $Y$ בניסוי עתידי כלשהו הוא **הממוצע של ערכי המשתנה שבידינו**.
- הניבוי הטוב ביותר ל-$Y$ בהינתן שמשתנה מנבא $X$ קיבל את הערך $x$, הוא **הערך הממוצע של המשתנה $Y$, שיחושב עתה רק עבור אותם נתונים שבהם $X$ קיבל את הערך $x$**.
- למציאת הניבוי **הקווי** הטוב ביותר ל-$Y$ המתבסס על ערכי משתנה מנבא $X$, נפעל על פי עקרון הריבועים הפחותים: על פי עיקרון זה, המטרה היא לצמצם למינימום את סכום ריבועי הסטיות בין ערכי $y$ שהתקבלו בפועל לבין ערכי הניבוי שלהם $\hat{y}$ על פי אותו קו.

## משימות חישוב וחשיבה – בעיות ניבוי

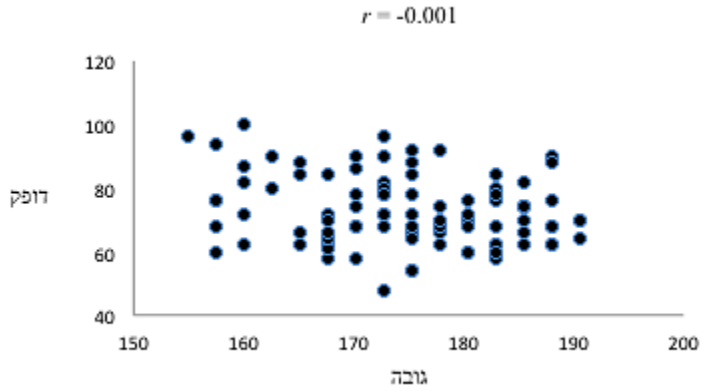
(פתרונות בעמ' 121)

**משימה I.**

עברו על דיאגרמות הפיזור באיורים 3–8 שבפרק 2 וציירו עליהן 'לפי העין' קו ניבוי שנראה לכם המתאים ביותר לנתונים. האם ערכו של מקדם המתאם (הרשום בחלק העליון של כל איור) משקף את מידת הקרבה של נקודות הדיאגרמה לקו?

**משימה II. קצב הדופק והגובה של חיילים**

א. בדיאגרמת הפיזור של קצב הדופק והגובה של חיילים (זהו איור 8 מעמ' 28) סמנו 'לפי העין' את עקום הממוצעים. סמנו את הממוצעים המתאימים בריבועים לבנים. האם התקבל קו ישר?




ב. עתה נסו לצייר קו ישר שנראה לכם המתאים ביותר לצורכי ניבוי קצב הדופק על בסיס הגובה.

**משימה III. שימוש בעקרון הריבועים הפחותים להשוואת טיבם של קווי ניבוי**

בשתי העמודות הימניות בטבלה שלהלן מוצגים 12 הנתונים המתאימים לדוגמה 7 – **ציונים פסיכומטריים – נתונים חלקיים** (עמ' 63).

| $(y-62.4)^2$ | $(y-\hat{y}_x)^2$ | $\hat{y}_x = 0.2x - 69.5$ | $y$ | $x$ |
|---|---|---|---|---|
| 547.6 | 0.8 | 38.1 | 39 | 538 |
| | | | 36 | 575 |
| | | | 38 | 602 |
| | | | 68 | 626 |
| | | | 78 | 636 |

| | | | | |
|---|---|---|---|---|
| 650 | 64 | | | |
| 665 | 64 | | | |
| 697 | 33 | | | |
| 704 | 82 | | | |
| 723 | 92 | | | |
| 738 | 79 | | | |
| 763 | 76 | | | |
| **סה"כ** | | | | |

א. בעמודה השלישית, רשמו לכל תלמיד את ערך הניבוי המתאים $\hat{y}_x$ על פי נוסחת קו הניבוי הרשומה בכותרת העמודה.

ב. עתה השלימו בהתאם את העמודה הרביעית. מהו סכום ריבועי הסטיות מהקו המוצע?

ג. משימוש במחשבון התקבל $\bar{y} = 62.4$ . בהתאם לכך, באיור 20 (עמ' 63) שרטטו קו ניבוי נוסף מקביל לציר $x$ , שמשוואתו $y = 62.4$ , ועל פיו ננבא עבור $Y$ את ממוצע הנתונים $\bar{y}$ בלי קשר לערך $x$ . עתה השלימו את העמודה השמאלית של ריבועי הסטיות מ- $\bar{y}$ , כולל הסכום (שיופיע בתא האפור בתחתית העמודה).

ד. מהו ממוצע ריבועי הסטיות של ערכי Y מקו הניבוי $y = \bar{y}$ ? מה למעשה חישבתם?

ה. על בסיס חישוב סכום ריבועי הסטיות משני קווי הניבוי שבחנו (רשמו את הערכים בתאים האפורים שבטבלה), הביעו דעתכם איזה משני הקווים מתאים יותר לצורכי ניבוי.

**משימה IV. גיל חתנים וגיל כלות בישראל – שימוש בנתוני הלמ"ס**

בנספח ג בעמ' 105, מוצג לוח שהוצאנו מתוך פרסומי הלשכה המרכזית לסטטיסטיקה. בנספח תתבקשו להיעזר בנתונים שבלוח לחישוב עקום הממוצעים לניבוי גיל הכלה על בסיס גיל החתן, וכן חישובים נוספים (פתרונות בעמ' 130).

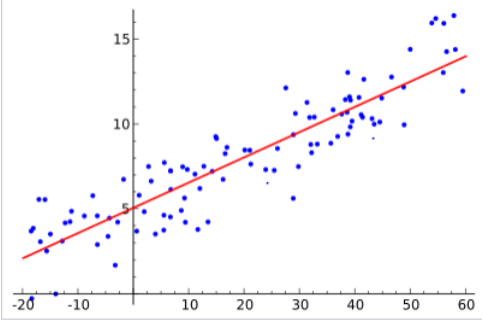
**תרגילים: בעיות ניבוי**

(פתרונות מקוצרים בעמ' 134)

**תרגיל 1 (גובה אבות וגובה בנים)**. חִזרו לעקום הממוצעים של פירסון (עמ' 11). באיור משורטט גם הקו הישר הטוב ביותר לניבוי גובה הבן על פי גובה האב. חוו דעתכם על מידת ההתאמה של הניבוי הקווי בהשוואה לניבוי על פי הממוצעים, שהוא הניבוי הטוב ביותר. מה הם התחומים שבהם השגיאות על פי הניבוי הקווי הן הגדולות ביותר?

[בפרק 6 נחזור לנקודה מעניינת זו.]

**תרגיל 2.** לפניכם דיאגרמת פיזור עם נתונים רבים מאוד ועם קו הניבוי הטוב ביותר (נוסחאות יפורטו בהמשך). ציירו 'לפי העין' נקודות אחדות על עקום הממוצעים. האם הן נופלות על קו הניבוי? הסבירו את משמעות הדבר.




**תרגיל 3** (אתגר). האם אפשר לדעתכם להתבסס על עקום הממוצעים בלבד (ולהתעלם מדיאגרמת הפיזור) כדי להתאים בעזרתו את קו הניבוי הטוב ביותר לכלל הנתונים? הציגו איור המדגים את הבעיה בקיצור דרך זה.

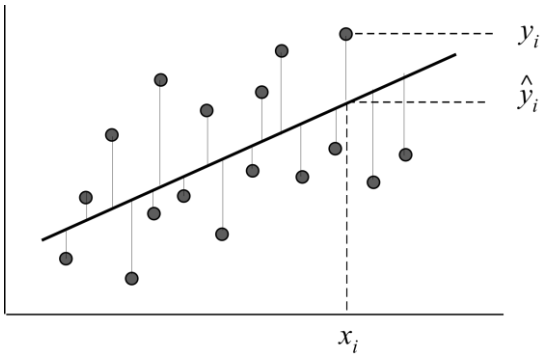
# פרק 5
# קו הרגרסיה

## 5.1 עקרון הריבועים הפחותים

אנו מחפשים **קו ישר** שיהיה **קרוב ככל האפשר** לנקודות $(x_1, y_1),...,(x_n, y_n)$ בדיאגרמת הפיזור, כדי שנוכל להשתמש בו לצורכי ניבוי.

**סימון**: נסמן ב- $\hat{y}_i$ את ערכו של **הניבוי** ל-$Y$ עבור $x_i$, על פי הקו הישר שנקבע.

בכל אחת מהנקודות בדיאגרמת הפיזור אנו מתבוננים ב***סטייה*** בין הערך של המשתנה $Y$ כפי שהתקבל בפועל לבין הערך שמתקבל על פי קו הניבוי – זוהי **הטעות** הנובעת מהניבוי שהצענו עבור $x_i$ (ראו איור 21):




**איור 21.** טעויות הניבוי: הסטיות $y_i - \hat{y}_i$ בין ערכי $Y$ לערכי הניבוי עבורם על פי קו הניבוי (בקווים אפורים)

❤ **מידת הקִרבה** של קו ישר כלשהו לנקודות בדיאגרמת הפיזור נמדדת באמצעות **סכום ריבועי הסטיות**[4] בין ערכי $y$ שהתקבלו בפועל לבין ערכי הניבוי שלהם $\hat{y}$ על פי קו זה.

הקו הישר שסכום ריבועי הסטיות ממנו הוא מינימלי נקרא ***קו הריבועים הפחותים***.

***קו הרגרסיה*** שישמש אותנו לצורכי ניבוי הוא קו הריבועים הפחותים שתיארנו לעיל. מכאן שזוהי תכונתו:

❤ **סכום ריבועי הסטיות של ערכי $y$ מקו הרגרסיה הוא מינימלי.**

[4] נזכיר שגם בהגדרת השונות אנו מחשבים את ריבועי הסטיות מהממוצע $\bar{x}$ כמדד לפיזור של משתנה $X$, ולא את הערכים המוחלטים של הסטיות.

**הדגמה**: הקו הישר המתואר באיור 20 (עמ' 62) הוא למעשה קו הרגרסיה לניבוי ציוני שנה א על פי הציון הפסיכומטרי של שניים-עשר התלמידים (ראו משימה I בפרק זה, בעמ' 77). באיור מסומנות (בקווקו) כל הסטיות מקו זה.
כל קו ישר אחר שנשרטט – סכום ריבועי הסטיות ממנו יהיה גדול יותר.

## 5.2 קו הרגרסיה

קבועי הקו $a,b$ שעבורם מתקבל **המינימום** מחושבים מתוך גזירה לפי $a$ ולפי $b$ והשוואה ל-0. הפרטים הטכניים הם מעבר לרמה של חוברת זו. למעשה גם נוסחאות קו הרגרסיה שעבורו מתקבל המינימום הן מסורבלות. נציין שחישוב ידני על פי הנוסחאות סביר רק אם מספר הנתונים קטן, אך לנתוני אמת יהיה עלינו להיעזר בכלים חישוביים.

את **קו הרגרסיה** $\hat{y}_x = a + bx$ **לניבוי ערכו של** $y$ **על פי** $x$ מקבלים באמצעות נוסחאות אלה:

השיפוע $b$:

$$b = r \cdot \frac{\sigma_y}{\sigma_x} \qquad (4)$$

$$a = \bar{y} - b\bar{x} \qquad (5)$$

ל**ערך הניבוי** עבור $x$ נתון, יש להציב את $x$ במשוואת הקו $\hat{y}_x = a + bx$ המתקבלת.

לחישוב ידני של קו הרגרסיה נדרשים חמישה גדלים:

– הממוצעים $\bar{y},\bar{x}$ וסטיות התקן $\sigma_Y,\sigma_X$ של המשתנים (מתקבלים בקלות בעזרת מחשבונים).

– מקדם המתאם $r$ (בדקו אפשרות של התקנה במחשבון שלכם).

כל הגדלים הללו מתקבלים מיידית באמצעות תוכנות פשוטות (למשל אקסל). הדוגמאות הבאות מבהירות את השימוש בהם למציאת קו הרגרסיה וערך הניבוי.

**ציון פרויקט וציון תעודה** (המשך דוגמה 1). עבור נתוני לוח 1 (עמ' 16), כבר חישבנו ומצאנו את חמשת הגדלים: $\bar{x} = 71$, $\bar{y} = 82$, $\sigma_X = 22.78$, $\sigma_Y = 13.27$ (עמ' 19) וכן $r = 0.812$ (עמ' 41).

הצבה בנוסחה (4) נותנת את שיפוע הקו $b = r \cdot \frac{\sigma_y}{\sigma_x} = 0.812 \times \frac{13.27}{22.78} = 0.47$,

ומנוסחה (5) נקבל: $a = \bar{y} - b\,\bar{x} = 82 - 0.47 \times 71 = 48.63$.

משוואת קו הרגרסיה היא אפוא: $\hat{y}_x = 0.47 \cdot x + 48.63$.

תלמיד שקיבל ציון פרויקט $x=65$, הניבוי לציון התעודה הוא $\hat{y}_{65}=0.47\times65+48.63=79.18$.

תלמיד שקיבל ציון פרויקט $x=85$, הניבוי לציון התעודה הוא $\hat{y}_{85}=0.47\times85+48.63=88.58$.

למעשה, די בשתי הנקודות (65, 79.18) ו- (85, 88.58), שהקו עובר דרכן, כדי לשרטט את הקו (שרטטו).

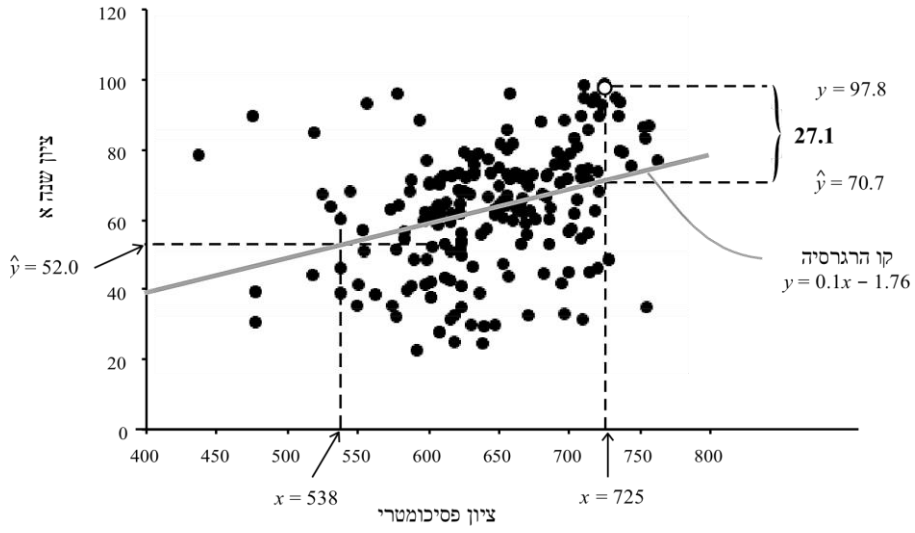
**ציונים פסיכומטריים** (המשך דוגמה 3). לעיבוד נתוני כלל 198 התלמידים, נעזרנו בפקודות אקסל (ראו נספח ב בעמ' 103) וקיבלנו: $\bar{x}=643.0$, $\bar{y}=62.8$, $\sigma_X=58.99$, $\sigma_Y=16.94$, כמו כן $r=0.337$. מכאן,

$b=0.337\times\frac{16.94}{58.99}=0.1$, ו- $a=62.8-0.1\times64.3=-1.5$.

משוואת קו הרגרסיה היא אפוא: $\hat{y}_x=0.1\cdot x-1.5$.

נתבונן בתלמיד שקיבל ציון פסיכומטרי נמוך מהממוצע: $x=600$. הצבה בנוסחה שהתקבלה נותנת ניבוי לציון שנה א $\hat{y}_{600}=0.1\times600-1.5=58.5$, וגם הניבוי נמוך מהממוצע.

תלמיד שהציון הפסיכומטרי שלו הוא $x=725$, הניבוי לציון שנה א על פי קו הרגרסיה הוא $\hat{y}_{725}=0.1\times725-1.76=70.74\cong70.7$, אך בפועל התלמיד השיג ציון גבוה יותר: $y=97.8$. הסטייה בין הערך שהתקבל לבין הניבוי היא $y-\hat{y}_{725}=97.8-70.7=27.1$ (ראו איור 22).




**איור 22.** קו הרגרסיה לניבוי ציון שנה א על פי הציון הפסיכומטרי (באפור). סכום ריבועי הסטיות מהקו הוא מינימלי.

באופן דומה, הניבוי הליניארי הטוב ביותר לתלמיד שקיבל ציון פסיכומטרי גבוה מעט מהממוצע $x=650$ הוא $\hat{y}_{650}=0.1\times650-1.5=63.5$, אף הוא גבוה במקצת מהממוצע. נזכיר שעל פי **עקום הממוצעים** (ראו בעמ' 60), הניבוי לציון שנה א לתלמיד שציון הפסיכומטרי שלו 650 הוא גבוה יותר – 69.1.

עקום הממוצעים נותן את הניבוי הטוב ביותר, ואילו קו הרגרסיה נותן את הניבוי הטוב ביותר רק מבין הניבויים הקוויים. מקובל להתמקד בניבוי באמצעות קווי רגרסיה בשל הקשיים הנלווים לשימוש בעקום הממוצעים. ההעדפה היא להסתייע בנוסחת ניבוי פשוטה שקל לחשב ולפרש את המקדמים שלה.

## 5.3 תיאור גרפי של קו הרגרסיה

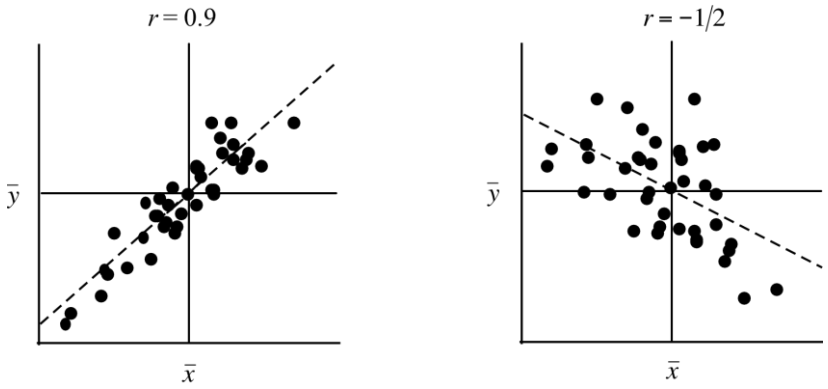
**ציונים פסיכומטריים** (דוגמה 3). מהו הניבוי הקווי הטוב ביותר לתלמיד שקיבל בפסיכומטרי **בדיוק** את הציון הממוצע $\bar{x}$?

כדי לקבל תשובה לשאלה זו יש להציב בנוסחת קו הרגרסיה $\hat{y}_x=0.1\cdot x-1.5$ את הערך $\bar{x}=643.0$. נקבל את הניבוי: $\hat{y}_{643}=0.1\times643-1.5=62.8=\bar{y}$. כלומר, הניבוי הוא הציון הממוצע של שנה א.

**באופן כללי**: הצבת הערך $\bar{x}$ במשוואת קו הרגרסיה נותנת את הניבוי $\hat{y}_{\bar{x}}=a+b\cdot\bar{x}=\left[\bar{y}-b\,\bar{x}\right]+b\cdot\bar{x}=\bar{y}$.

קו הרגרסיה עובר דרך **נקודת הממוצעים** $(\bar{x},\bar{y})$.

**הדגמה**: באיור 23 מוצגות שתי דיאגרמות פיזור עם קווי הממוצעים וקו הרגרסיה המתאים. קו הרגרסיה מקווקו ועובר דרך נקודת הממוצעים.




**איור 23.** דיאגרמות פיזור של משתנים עם קווי הרגרסיה ומקדמי המתאם ביניהם

באיור השמאלי מוצגת דיאגרמת פיזור של שני משתנים עם מקדם מתאם חיובי גבוה: $r=0.9$. על כן, קו הרגרסיה **עולה** והנקודות בדיאגרמה קרובות אליו, והניבוי על פי קו זה הוא טוב מאוד.

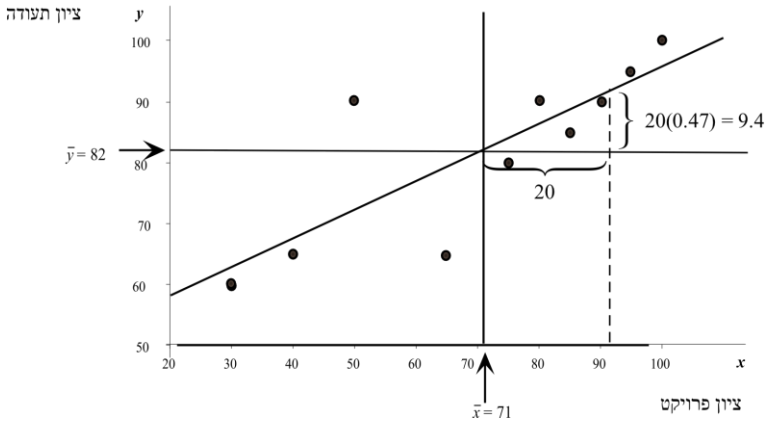
באיור הימני מקדם המתאם **שלילי** בינוני: $r=-1/2$. על כן, קו הרגרסיה **יורד**, והניבוי על פיו אינו טוב דיו.

נזכיר שכדי לצייר קו ישר די ב**נקודה** אחת וב**שיפוע** הקו.

כזכור:

- ✓ קו הרגרסיה עובר דרך נקודת הממוצעים $(\bar{x},\bar{y})$.
- ✓ השיפוע של קו הרגרסיה הוא $b = r \cdot \frac{\sigma_Y}{\sigma_X}$.

**ציון פרויקט וציון תעודה** (דוגמה 1). קו הרגרסיה עובר דרך נקודת הממוצעים $(\bar{x},\bar{y}) = (71,\ 82)$ והשיפוע שלו הוא $b = 0.47$. באיור 24 מודגם שימוש בנקודת הממוצעים ובשיפוע כדי לצייר את קו הרגרסיה.




**איור 24.** מציאת קו הרגרסיה על פי נקודת הממוצעים והשיפוע

**טיב הניבוי**: נזכיר שמקדם המתאם $r = 0.812$ קרוב ל-1, והניבוי באמצעות קו הרגרסיה יעיל.

### ➢ משוואת קו הרגרסיה – נוסחה מועילה

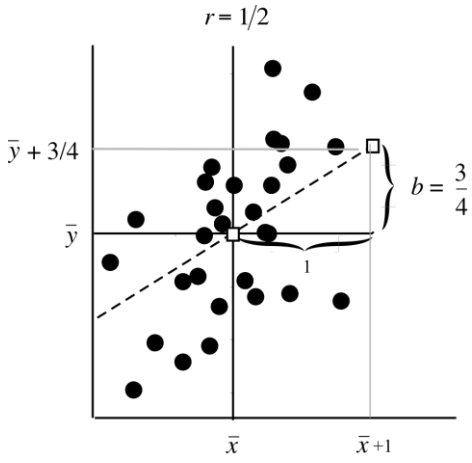
נתבונן בנקודה כלשהי $(x, y_x)$ על קו הרגרסיה. מהעובדה שהקו עובר גם דרך הנקודה $(\bar{x},\bar{y})$ ומהגדרה של שיפוע מתקיים $b = \frac{\hat{y}_x - \bar{y}}{x - \bar{x}}$. העברת אגפים נותנת:

$$\underset{\text{השינוי המתקבל בערך הניבוי } y}{\hat{y}_x} - \bar{y} = b \cdot (\underset{\text{השינוי בערכו של } x}{x} - \bar{x})$$

ומכאן מתקבלת **נוסחה מועילה למשוואת קו הרגרסיה**, נוחה מאוד לשימוש ידני:

$$\underset{\text{הניבוי עבור } x}{\hat{y}_x} = \bar{y} + b \cdot (x - \bar{x}) \qquad (6)$$

**הדגמה**: באיור 25 מוצגת דיאגרמת פיזור של ציוני מתמטיקה וציוני לשון בכיתה. במרכז האיור מצוינת נקודת הממוצעים $(\bar{x},\bar{y})$, שדרכה עובר קו הרגרסיה. כדי לשרטט את קו הרגרסיה על פי נוסחה (7) שבעמוד הבא, די לחשב את שיפוע הקו. מנתוני הבעיה חישבנו ומצאנו $\sigma_X=2, \sigma_Y=3$, ומקדם המתאם הוא $r=1/2$. על פי נוסחה (4) השיפוע הוא $b=\frac{1}{2}\cdot\frac{3}{2}=\frac{3}{4}$. באיור 25 מודגם אופן שרטוט של קו בעל שיפוע $\frac{3}{4}$ דרך נקודת הממוצעים.




**איור 25.** שרטוט קו הרגרסיה בעזרת נקודת הממוצעים והשיפוע

**הבחנה**: מנוסחת השיפוע $b=r\cdot\frac{\sigma_y}{\sigma_x}$, הסימן של השיפוע $b$ זהה לסימן של $r$.

מכאן,

❤ כשמקדם המתאם **חיובי** השיפוע הוא חיובי ו**הקו עולה**, וכשמקדם המתאם **שלילי** השיפוע הוא **שלילי** והקו **יורד**.

## 5.4 טיב הניבוי

מדד מקובל לטיב עקומת ניבוי כלשהי הוא ממוצע ריבועי הסטיות בין ערכי $Y$ שהתקבלו בפועל לבין ערכי הניבוי שהתקבלו על פי העקומה – זוהי **הטעות הריבועית הממוצעת של הניבוי**.

נזכיר שהניבוי הטוב ביותר שאינו מתבסס על משתנה מנבא הוא הקבוע $\bar{y}$. מהגדרת השונות של Y נסיק:

✋ הטעות הריבועית הממוצעת של הניבוי ללא משתנה מנבא היא $\sigma_y^2$.

הנוסחה הפשוטה הבאה – שלא נוכיח כאן – מייתרת חישובים ידניים מסורבלים של ריבועי הסטיות מקו הרגרסיה.

✋ הטעות הריבועית הממוצעת של קו הרגרסיה $= \sigma_y^2\cdot(1-r^2)$ (7)

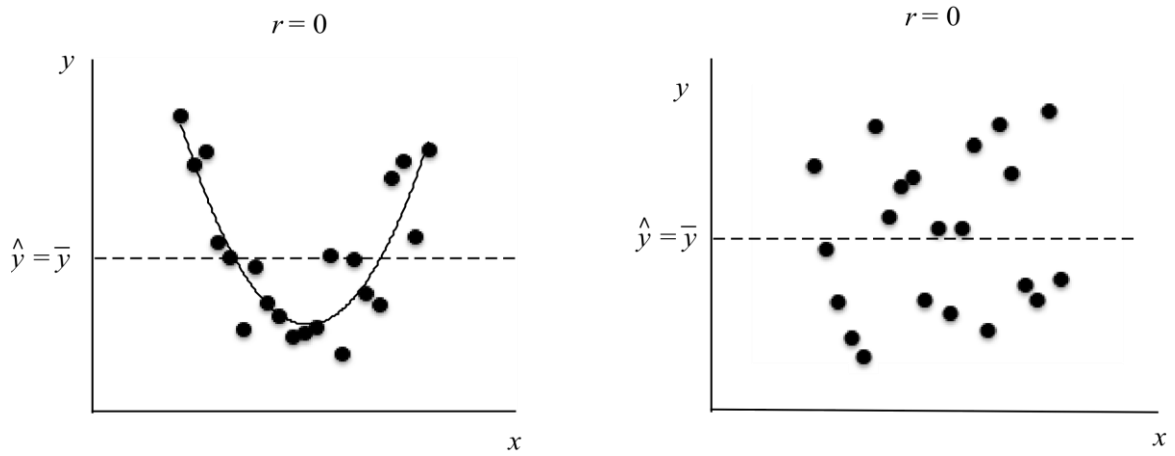
בנספח ד לספר (עמ' 107) נרחיב מעט בנושא טיב הניבוי וגם נציג דוגמה מעשית לחישוב השיפור של ניבוי באמצעות קו רגרסיה בהשוואה לניבוי ללא משתנה מנבא.

**➢ ניתוח ערכי קיצון**

– מנוסחה (7), הטעות הריבועית הממוצעת של קו הרגרסיה היא 0 כאשר $r^2 = 1$. כצפוי, כשכל הנקודות בדיאגרמת הפיזור נמצאות על קו ישר אחד אין כל טעות בניבוי על פי הקו.

– מנוסחה (7), טעות הניבוי של קו הרגרסיה היא מקסימלית כאשר מקדם המתאם הוא 0. ננסה לברר מה משמעות הדבר.

באיור 26 מוצגים שוב איורים 16(א) ו-16(ב) מעמ' 44, בשניהם מקדם המתאם בין המשתנים הוא $r = 0$.




**איור 26.** קו הרגרסיה (מקווקו) הוא $\hat{y}_x = \bar{y}$ (השיפוע 0, מקביל לציר $x$)

על פי נוסחת השיפוע של קו רגרסיה, בשני המקרים השיפוע של קו הרגרסיה – הקו המקווקו באיור – הוא 0. כלומר, קו הרגרסיה מקביל לציר ה-$x$ וערך הניבוי הוא תמיד $\bar{y}$. בשני המקרים אותו ניבוי מתקבל גם ללא ידיעת ערכו של $x$, ולפיכך הניבוי באמצעות קו הרגרסיה אינו יעיל כלל.

– באיור הימני המצב הוא שאכן **אין קשר** בין שני המשתנים.

– באיור השמאלי יש קשר ברור בין המשתנים, אולם הנקודות מפוזרות סביב ה**פרבולה** המסומנת באיור ולא סביב קו ישר כלשהו. לא ייפלא אפוא שהניבוי באמצעות קו ישר אינו יעיל כלל במקרה זה.

כאשר $r = 0$, אין הדבר מורה על חוסר קשר בין המשתנים, אלא על **חוסר קשר קווי**.

התובנה שרכשנו מניתוח האיורים היא:

בטרם עוברים לשלב הניבוי, חשוב לבחון את דיאגרמת הפיזור ולנסות לזהות בה מגמה כללית אשר תכריע **אם מתאים ניבוי קווי**, או אולי יש להפעיל מלכתחילה שיטות ניבוי מתוחכמות יותר.

## אז מה היה לנו? קו הרגרסיה

### מושגים חדשים

- עקרון הריבועים הפחותים
- קו הרגרסיה – קו ריבועים פחותים, קו הניבוי
- קשר ליניארי (קווי)
- ערך הניבוי

### תובנות חדשות

- הניבוי הקווי הטוב ביותר ל-$Y$ על בסיס ערכו של **משתנה מנבא** $X$ נקבע על פי קריטריון הריבועים הפחותים:
  זהו הקו הקרוב ביותר לנקודות בדיאגרמת הפיזור. מידת הקרבה נקבעת על פי סכום ריבועי הסטיות בין ערכי $y$ שהתקבלו בפועל לבין ערכי הניבוי שלהם $\hat{y}$ על פי קו זה.
- קו הרגרסיה עובר דרך נקודת הממוצעים $(\bar{x},\bar{y})$.
- הסימן של שיפוע קו הרגרסיה זהה לסימן של מקדם המתאם. כשמקדם המתאם חיובי הקו עולה, כשמקדם המתאם שלילי הקו יורד.
- מקדם מתאם קרוב לאפס אינו מעיד בהכרח על חוסר קשר בין שני המשתנים, אלא על חוסר קשר קווי.
- **משמעות השיפוע**: $\hat{y}_x - \bar{y} = b \cdot \underbrace{(x-\bar{x})}_{\text{סטייה מהממוצע בערך } x}$
  $\downarrow$
  **הסטייה המתאימה בערך הניבוי**
- מדד מקובל לטיב של עקומת ניבוי **כלשהי** הוא ממוצע ריבועי הסטיות בין ערכי $Y$ שהתקבלו בפועל לבין ערכי הניבוי המתקבלים על פי העקומה.

### כלים חדשים

- **קו הרגרסיה** לניבוי ערך של המשתנה $Y$ על סמך הערך $x$ של המשתנה $X$ הוא: $\boxed{\hat{y}_x = a + bx}$.
- בקו זה $\boxed{a = \bar{y} - b\bar{x}}$ וה**שיפוע** $b$ הוא $\boxed{b = r \cdot \frac{\sigma_y}{\sigma_x}}$.
- לקבלת **ערך הניבוי של Y** ל- $x$ נתון, יש להציב את $x$ במשוואה המתקבלת.
- נוסחה מקובלת, נוחה יותר, ל**משוואת קו הרגרסיה**: $\boxed{\hat{y}_x = \bar{y} + b \cdot (x - \bar{x})}$.
- המדד ל**טיב הניבוי** על פי קו הרגרסיה $\sigma_y^2 \cdot (1 - r^2)$:.
- **טיב הניבוי** ללא משתנה מנבא $\sigma_y^2$.

## משימות חישוב וחשיבה – קו הרגרסיה

(פתרונות בעמ' 124)

**משימה I. תלמידי מתמטיקה** (המשך דוגמה 7 מעמ' 61, 62)

מתוך הנתונים של שניים-עשר תלמידי מתמטיקה (ראו בעמ' 62) נקבל את הערכים האלה בעזרת מחשבון או תוכנת אקסל: $\bar{x}=659.7$ , $\sigma_x=65.4$ , $\bar{y}=62.4$ , $\sigma_y=19.8$ , $r=0.66$ .

א. מהו השיפוע של קו הרגרסיה? היעזרו בנקודת הממוצעים ובשיפוע כדי לשרטט את הקו בדיאגרמת הפיזור. שימו לב שקו הניבוי המוצג באיור 20 (עמ' 63) הוא למעשה קו הרגרסיה.

ב. מקדם המתאם שהתקבל, $r=0.66$ , נחשב גבוה. האם אפשר להסיק מכך שגם באוכלוסיית תלמידי מתמטיקה כולה יש קשר טוב בין המשתנים, והניבוי של הצלחה באוניברסיטה בעזרת ציון הפסיכומטרי הוא מספק?

**משימה II. חמודי הסבות מקצועיות** (המשך משימה III מעמ' 51)

היעזרו בערכים שהתקבלו בחישובים: $\bar{x}=2.3$ , $\sigma_x=1.10$ , $\bar{y}=6.36$ , $\sigma_y=1.24$ , $r=0.478$ , כדי לענות על שאלות הנוגעות לקו הרגרסיה לניבוי הניקוד בסדנה על בסיס מספר ימי האימון.

א. מהו הסימן של שיפוע הקו? מה משמעות הדבר?

ב. מהי משוואת קו הרגרסיה לניבוי הניקוד על בסיס מספר ימי האימון? שרטטו את קו הרגרסיה בעזרת נקודה אחת ושיפוע. האם הקו קרוב לנקודות בדיאגרמה? חוו דעתכם על טיב הניבוי.

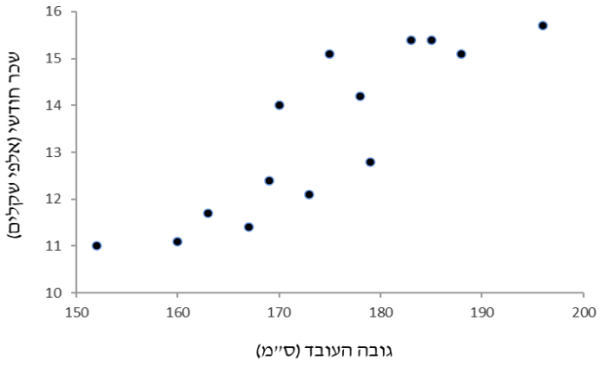
ג. מהו הצפי לניקוד של מועמד חדש שבחר בסדנה של שלושה ימי אימון ולמועמד שבחר ביום אימון יחיד.

ד. האם לדעתכם סביר להשתמש בקו הרגרסיה עבור מועמד שבוחר שישה ימי אימון?

**משימה III. הקשר בין שכר עובדים לגובהם** (המשך משימה IV מעמ' 52)

השאלה הנבדקת היא: האם העובדים הגבוהים נוטים לקבל שכר גבוה יותר מהעובדים הנמוכים?

א. מצאו את קו הרגרסיה לניבוי השכר $Y$ על פי גובה העובד $X$. מהו שיפוע הקו? מה משמעות הסימן? שרטטו את הקו על דיאגרמת פיזור זו:




ב. מה ניבוי השכר עבור עובד שגובהו 178 ס״מ? עבור עובד שגובהו 196 ס״מ?

**היפוך תפקידי המשתנים**: עובד סקרן במדור משכורת רוצה לחזות את הגובה $X$ של עובד על פי שכרו $Y$.

ג. מהו עתה **מקדם המתאם**?

ד. ענו **ללא חישובים** – מהו הסימן של שיפוע קו הרגרסיה לניבוי הגובה $X$ על בסיס השכר $Y$?

ה. חשבו את השיפוע של קו הרגרסיה לניבוי הגובה $X$ על בסיס השכר $Y$.

ו. מהו הניבוי לגובה של עובד ששכרו 16,200? השוו לתשובה של סעיף ב.

**משימה IV. סרטן עור וקרינה אולטרה סגולה** (המשך משימה V מעמ׳ 52)

א. מצאו את משוואת **קו הרגרסיה** לניבוי שיעור מקרי המלנומה $Y$ על פי קו הרוחב $X$. מה השיפוע של הקו? שרטטו את הקו בדיאגרמת הפיזור שערכתם בסעיף א (עמ׳ 52).

ב. מהי ההערכה (הניבוי) לשיעור מקרי המלנומה $y$ בקו רוחב 35? חוו דעתכם על טיב הניבוי.

ג. האם לדעתכם אפשר להשתמש בקו שמצאתם לניבוי שיעור מקרי המלנומה בתל אביב (בדקו מהו קו הרוחב)?

**משימה V. שירת הצרצר**

צרצרים משמיעים קול צרצור במעין פעולת ניסור, כשהם מעבירים במהירות קדימה ואחורה כנף אחת שלהם על פני הכנף האחרת. ידוע שיש קשר ליניארי בין הטמפרטורה לבין תדירות הצרצורים. רשמנו בטבלה את נתוני תדירות הצרצורים ואת הטמפרטורה (במעלות צלזיוס), שנאספו עבור חמישה-עשר צרצרים מזן מסוים (Striped Ground).

א. ציירו דיאגרמת פיזור. האם הקשר נראה אכן קווי?

| **תדירות לדקה,** $x$ | **טמפרטורה,** $y$ | $x \cdot y$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 20.0 | 31.4 | | | |

| | | | | | |
|---|---|---|---|---|---|
| | | | 22 | 16.0 | |
| | | | 34 | 19.8 | |
| | | | 29 | 18.4 | |
| | | | 27 | 17.1 | |
| | | | 24 | 15.5 | |
| | | | 14.7 | 14.7 | |
| | | | 27.8 | 17.1 | |
| | | | 20.8 | 15.4 | |
| | | | 28.5 | 16.2 | |
| | | | 26.4 | 15.0 | |
| | | | 28.1 | 17.2 | |
| | | | 27 | 16.0 | |
| | | | 28.6 | 17.0 | |
| | | | 24.6 | 14.4 | |
| 10644.47 | 4200.56 | 6645.75 | 393.9 | 249.8 | **סה"כ** |
| | | | | | **ממוצעים** |

ב. השלימו את שורת הממוצעים וחשבו בעזרתה את סטיות התקן ואת מקדם המתאם בין תדירות הצרצור לבין הטמפרטורה.

ג. האם מקדם המתאם ישתנה אם הטמפרטורה תימדד במעלות פרנהייט?

ד. מצאו את קו הריבועים הפחותים לניבוי הטמפרטורה על פי תדירות הצרצור.

ה. נמדד צרצור בתדירות של 18 פעמים בדקה. מהי ההערכה (הניבוי) המתקבלת לטמפרטורה?

### משימה VI. הקשר בין פסק הזמן מהעבודה לבין המשך הקריירה המקצועית

מחפשי עבודה שלקחו פסק זמן בקריירה המקצועית מאבדים ממידת האטרקטיביות שלהם בשוק העבודה, הן בגלל שחיקה בכישוריהם הן בגלל שאינם מעודכנים בידע המקצועי. בטבלה שלהלן מוצג אחוז בתי החולים $Y$ שהביעו נכונות לקלוט טכנאי רפואה שנעדרו $X$ שנים משוק העבודה. השלימו את שורת הממוצעים בטבלה.

| $y^2$ | $x^2$ | $x\cdot y$ | **אחוז בתי החולים,** $y$ | **שנות היעדרות,** $x$ |
|---|---|---|---|---|
| | | | 100 | 0.5 |
| | | | 94 | 1.5 |

| | | | | |
|---|---|---|---|---|
| 4 | 75 | | | |
| 8 | 44 | | | |
| 13 | 28 | | | |
| 18 | 17 | | | |
| 45 | 358 | 1513 | 575.5 | 27470 |
| | | | | |

**סה"כ**

**ממוצעים**

א. חשבו את מקדם המתאם והסבירו את סימנו.

ב. מצאו את משוואת קו הריבועים הפחותים לניבוי אחוז הנכונות $Y$ על בסיס שנות ההיעדרות $X$.

ג. עקב מחסור בכוח אדם מקצועי בבתי החולים, 100% מבתי החולים יהיו מוכנים לשכור עובדים שלא לקחו פסק זמן כלל. מהו ערך הניבוי שקיבלתם עבור $x=0$ ? מהי הסטייה של הניבוי מהערך האמיתי? הסבירו.

**משימה VII. חשיפה לחומרים רדיואקטיביים**

חוקרים מעוניינים למדוד את הקשר בין רמת החשיפה לחומרים רדיואקטיביים מתחנת כוח גרעינית ליד הנהר קולומביה באורגון לבין התמותה מסרטן בקרב תושבי הסביבה. לשם כך נאספו נתונים מתשעה אזורים לאורך הנהר. בכל אזור נמדדה רמת החשיפה ונקבע אינדקס רמת החשיפה $X$, וכן מקרי המוות מסרטן ל-100,000 תושבים $Y$.

א. על בסיס טבלת הנתונים הבאה ציירו דיאגרמת פיזור. האם סביר להניח שהקשר הוא קווי?

ב. חשבו את שורת הממוצעים והיעזרו בה לחישוב מקדם המתאם בין רמת החשיפה לבין התמותה מסרטן לאורך הנהר.

ג. חשבו קו רגרסיה לניבוי התמותה מסרטן על בסיס אינדקס החשיפה.

ד. באזור נוסף לאורך הנהר נמדד אינדקס חשיפה ברמה 5. מהי ההערכה המתאימה לתמותה מסרטן?

| **אינדקס החשיפה,** $x$ | **תמותה מסרטן,** $y$ | $x\cdot y$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 2.49 | 147.1 | | | |
| 2.57 | 130.1 | | | |
| 3.41 | 129.9 | | | |
| 1.25 | 113.5 | | | |
| 1.62 | 137.5 | | | |
| 3.83 | 162.3 | | | |

| | | | | | |
|---|---|---|---|---|---|
| | 11.64 | 207.5 | | | |
| | 6.41 | 177.9 | | | |
| | 8.34 | 210.3 | | | |
| **סה"כ** | 41.56 | 1416.1 | 7439.37 | 289.4222 | 232498.97 |
| **ממוצעים** | | | | | |

**משימה VIII. מחירי מכוניות משומשות בארצות הברית**

בטבלה הבאה מוצגים מחירים (בדולרים) של טויוטה קורולה משומשת על פי הגיל (בשנים), כפי שערכו סיטונאי רכב בארצות הברית. הוסיפו בתחתית הטבלה שורת ממוצעים.

| | **גיל**, $x$ | **מחיר סיטונאי**, $y$ | $x \cdot y$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| | 1 | 14,680 | | | |
| | 2 | 12,150 | | | |
| | 3 | 11,215 | | | |
| | 4 | 10,180 | | | |
| | 5 | 9,230 | | | |
| | 6 | 8,455 | | | |
| | 7 | 7,730 | | | |
| | 8 | 6,825 | | | |
| | 9 | 6,135 | | | |
| | 10 | 5,620 | | | |
| **סה"כ** | **55** | **92220** | **430350** | **385** | **924769600** |

א. שרטטו דיאגרמת פיזור. מהו כיוון הקשר בין המשתנים? האם הקשר נראה קווי?

ב. היעזרו בשורת הסכומים שבתחתית הטבלה כדי לחשב את מקדם המתאם בין המשתנים.

ג. מצאו קו רגרסיה לניבוי המחיר על פי גיל המכונית.

ד. מחיר מכונית חדשה באותה שנה היה 16,200. האם הקו עובר דרך הנקודה $(0,16200)$? הסבירו את הממצא.

**משימה IX. עברית שפה קשה** (המשך משימה I מעמ' 37)

לקבוצת תלמידים נערכו שני מבחני שליטה בשפה: מבחן A (כתיבת מילים **ללא** שגיאות) ומבחן B (זיהוי שגיאות). מחישובים בעזרת במחשבון התקבלו ערכים אלה: $\bar{x}=5.5$ , $\bar{y}=10$ , $\sigma_X=2.22$ , $\sigma_Y=3.16$.
בעזרת אקסל התקבל גם מקדם המתאם $r=0.95$.

א. מצאו את משוואת קו הרגרסיה לניבוי ציון מבחן B על סמך ציון מבחן A. היעזרו בה כדי לרשום (בעמודה השלישית) את ערכי הניבוי במבחן B לערכים המתאימים של מבחן A. חשבו גם (בעמודה הרביעית) את ריבועי הסטיות של ערכי הניבוי מהערכים שהתקבלו בפועל.

| מבחן A, $x$ | מבחן B, $y$ | ערך הניבוי, $\hat{y}_x$ | ריבועי הסטיות $(y-\hat{y}_x)^2$ |
|---|---|---|---|
| 2 | 5 | | |
| 5 | 8 | | |
| 9 | 15 | | |
| 6 | 12 | | |
| 7 | 11 | | |
| 4 | 9 | | |
| **ממוצע** | | | |

ב. חשבו מכל אלו את **הטעות הריבועית הממוצעת** בניבוי ציון מבחן B על סמך ציון מבחן A על פי קו הרגרסיה. השוו את התוצאה שקיבלתם לערך שמתקבל על פי נוסחה (7).

ג. המשיכו והיעזרו בנוסחה (7) לחישוב הטעות הריבועית הממוצעת בניבוי ציון מבחן B על סמך מספר **מילים שגויות** במבחן A $\left(X^*=10-X\right)$. הסבירו את התוצאה.

**משימה X. המשך עברית שפה קשה – היפוך תפקידים**

א. מצאו את משוואת קו הרגרסיה לניבוי ציון מבחן A על סמך ציון מבחן B.

ב. מהי **הטעות הריבועית הממוצעת** בניבוי ציון מבחן A על סמך ציון מבחן B על פי קו הרגרסיה? האם התקבל אותו ממוצע כמו בסעיף ב בשאלה IX?

**תרגילים: קו רגרסיה**

(פתרונות מקוצרים בעמ' 131 ; תרגילים נוספים ראו בחוברת תרגילים נלווית)

**תרגיל 1.** נבדק הקשר בין ההכנסות וההוצאות החודשיות של 200 משפחות באזור תל אביב. אלה התוצאות שהתקבלו (באלפי ש"ח):

– ממוצע הכנסות 16.5 וסטיית תקן 2.7 ;

– ממוצע הוצאות 14.3 וסטיית תקן 2.3, מקדם המתאם $r = 0.83$.

א. מצאו את קו הרגרסיה לניבוי ההוצאה המשפחתית על פי הכנסת המשפחה.

ב. מהו הניבוי להוצאות של משפחה שהכנסתה החודשית היא 16.5 אלף ש״ח.

ג. העריכו את ההוצאות של משפחה שהכנסתה החודשית 18 אלף ש״ח ושל משפחה שהכנסתה החודשית 11 אלף ש״ח.

ד. מהו **ממוצע ריבוע הטעות** של קו הרגרסיה לניבוי ההוצאות על פי ההכנסות?

ה. השוו את ממוצע ריבוע הטעות של קו הרגרסיה לממוצע ריבוע הטעות של הניבוי ללא משתנה מנבא.

**תרגיל 2**. בטבלה רשומים נתוני הרווח של חברה מסחרית (במאות אלפי דולרים) בכל אחת משנות קיומה. באמצעות קו רגרסיה, העריכו את הרווחים הצפויים לחברה בשנה ה-11. הסבירו עד כמה אפשר לסמוך על הערכה זו ובאילו תנאים.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **השנה**: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **הרווח**: | 2 | 2 | 4 | 5 | 7 | 6 | 8 | 8 | 10 | 11 |

**תרגיל 3.** (המשך דוגמה 5 מעמ׳ 45). הקשר בין טמפרטורה מקסימלית $X$ ולחות יחסית $Y$ בישראל. נזכיר: $\bar{x} = 22.5$, $\bar{y} = 46.25$, $\sigma_X = 3.71$ ו- $\sigma_Y = 11.39$ ומקדם המתאם הוא $r = 0.71$.

א. היעזרו בכל אלו כדי להעריך את הלחות היחסית ביישוב נוסף שבו הטמפרטורה המקסימלית היא 23 מעלות.

ב. על פי קו הרגרסיה, מהו הניבוי ללחות היחסית בטבריה, שהטמפרטורה בה היא 25 מעלות? מהו ריבוע טעות הניבוי המתקבלת?

היעזרו בנוסחה (7) לחישוב הטעות הריבועית הממוצעת בניבוי הלחות היחסית על בסיס הטמפרטורה על פי קו הרגרסיה. השוו לטעות הריבועית הממוצעת בניבוי ללא משתנה מנבא.

**תרגיל 4**. במחקר על גורמים למחלות לב מעוניינים לבדוק את הקשר בין המשתנים המנבאים עישון וצריכת קפה. לשם כך התבקשו עשרים וחמישה חולי לב לדווח על ערכי שני המשתנים: $X$ – מספר הסיגריות שעישנו ביום, $Y$ – מספר ספלי הקפה ששתו ביום. דיאגרמת פיזור הצביעה על קשר קווי עולה. הערכים שהתקבלו עובדו בעזרת תוכנה חישובית והתקבלו תוצאות אלה: $\bar{x} = 18$, $\bar{y} = 3.52$, $\sigma_X^2 = 175$, $\sigma_Y^2 = 2.5$. מקדם המתאם שהתקבל הוא $r = 0.71$.

א. מהי ההערכה (הניבוי) שלכם למספר ספלי הקפה ששותה חולה לב אשר מעשן 15 סיגריות ביום?

ב. מהי ההערכה (הניבוי) שלכם למספר הסיגריות שמעשן חולה לב אשר שותה 4 ספלי קפה ביום?

ג. היעזרו בנוסחה (7) כדי לחשב את ממוצע ריבוע הטעות של קו הרגרסיה לניבוי מספר הסיגריות על פי מספר ספלי הקפה. השוו לטעות בניבוי ללא משתנה מנבא.

ד. מהו ממוצע ריבוע הטעות של קו הרגרסיה לניבוי מספר ספלי הקפה על פי מספר הסיגריות? השוו לניבוי ללא משתנה מנבא.

**תרגיל 5**. במרכז האקדמי "ידע לעם" נבדק הקשר בין ממוצע הציונים בתעודת הבגרות ובין ממוצע ציוני הלימודים בסוף שנה א במרכז האקדמי. נמצא:

– ציוני בגרות: ממוצע 7.9, סטיית תקן 1.3;

– ציוני לימודים: ממוצע 7.5, סטיית תקן 2.1;

– מקדם המתאם: $r = 0.5$.

א. מהו הניבוי הקווי הטוב ביותר שתוכלו להציע לציון הלימודים של תלמיד חדש במרכז האקדמי, אם אין בידיכם שום אינפורמציה נוספת לגביו? על פי איזה עיקרון פעלתם? הסבירו. מהו ממוצע ריבוע הטעות בשיטת ניבוי זו?

ב. נניח שנמסר לכם כי לתלמיד הזה ציון בגרות של 8.5. מהי שיטת הניבוי שבה תשתמשו:

ב1. אם יש בידיכם קובץ הנתונים המקורי.

ב2. אם אלו כל הנתונים שבידיכם, מהו ערך הניבוי שתציעו? הציגו את משוואת הניבוי הקווי הטוב ביותר, וחשבו את טיב הניבוי.

**תרגיל 6. נכונות לקנית מחשב ביתי** (המשך משימה IV מעמ' 38)

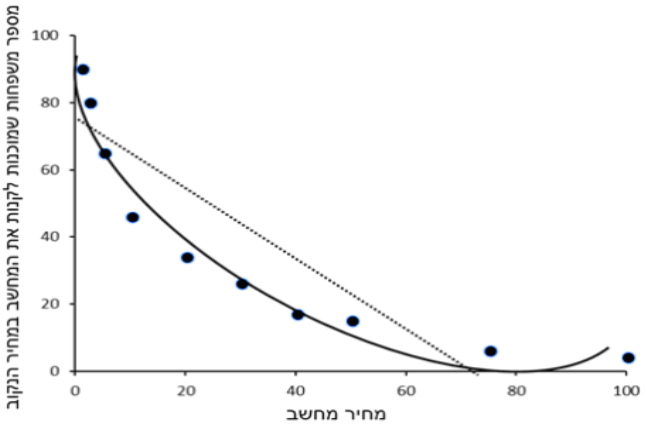
לבדיקת הקשר בין מחיר מחשב ביתי לבין הנכונות לקנות אותו, נבחנו 10 מחירים שונים (במאות דולרים); לכל מחיר הוקצו אקראית 100 בתי אב. בעזרת תוכנת אקסל חישבנו ומצאנו:

$\bar{x} = 33.35$, $\bar{y} = 38.30$, $\sigma_X = 31.68$, $\sigma_Y = 29.27$; כמו כן מקדם המתאם הוא $r = -0.87$.

א. מהי טעות הניבוי הריבועית הממוצעת ללא משתנה מנבא, כלומר מהי טעות הניבוי הקבוע $\overline{y}$?

ב. מהי טעות הניבוי הממוצעת של קו הרגרסיה? במה תלוי השיפור בהשוואה לניבוי ללא משתנה מנבא?

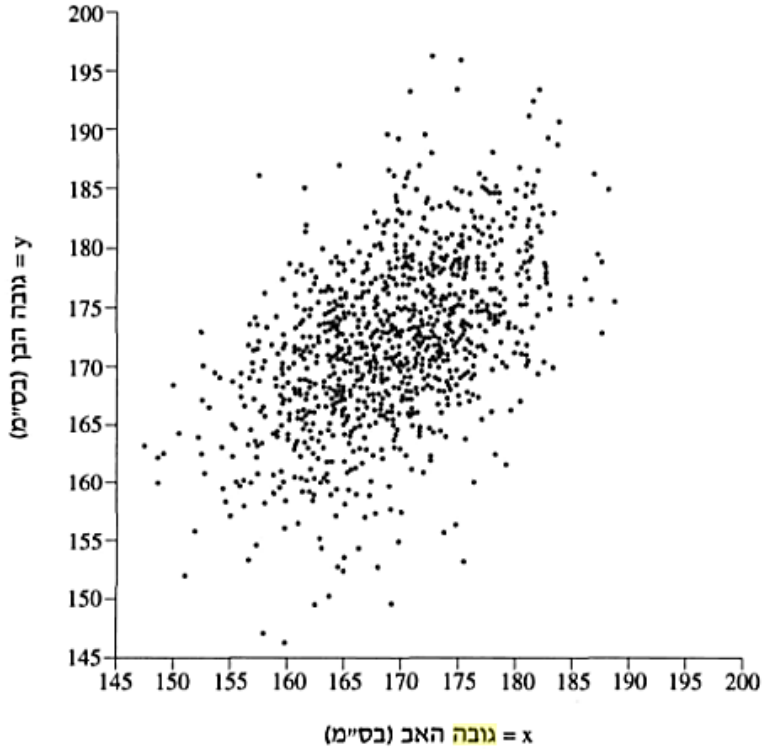
ג. מדיאגרמת הפיזור שלהלן ברור שניבוי טוב יותר היה מתקבל, למשל, על פי ה**פרבולה המודגשת**. חישבנו ומצאנו שממוצע ריבועי הסטיות מהקו הפרבולי המקווקו שבאיור הוא: 57.7. דונו בשיפור בטיב הניבוי בעזרת הפרבולה.


מספר משפחות שמוכנות לקנות את המחשב במחיר הנקוב
100
80
60
40
20
0
0
20
40
60
80
100
מחיר מחשב

# פרק 6
# כשלי חשיבה

## 6.1 נסיגה אל הממוצע

**דוגמה 8** (דוגמה מאלפת לסיום). נחזור ונציג כאן את העבודה שעשה הסטטיסטיקאי קרל פירסון, שהתפרסמה בשנת 1903. פירסון אסף נתונים של 1,078 משפחות על פי גובה האב וגובה בנו הבכור. דיאגרמת הפיזור המקורית של פירסון הוצגה בעמ' 11. בדיאגרמה שבאיור 27 מוצגים אותם נתונים מומרים לסנטימטרים.




**איור 27**. גובה של אבות וגובה של בנים

הנה סיכום התוצאות (המעוגלות) בסנטימטרים מתוך הנתונים:

– הגובה הממוצע של האבות הוא $\bar{x}=173$ וסטיית התקן היא $\sigma_X=7$.

– הגובה הממוצע של הבנים הוא $\bar{y}=175$ וסטיית התקן היא $\sigma_Y=7$.

נשים לב שבממוצע הבנים קצת גבוהים מאבותיהם, ואילו סטיות התקן זהות.

פירסון חישב ומצא שמקדם המתאם בין המשתנים היה $r=0.5$; מקדם המתאם אינו תלוי ביחידות המדידה.

מהנתונים שלפנינו, קו הרגרסיה לניבוי גובה הבן על פי גובה אביו הוא עולה ובעל שיפוע $b = 0.5 \times \frac{7}{7} = \frac{1}{2}$.

גם מבלי לרשום את משוואת הקו נוכל לומר:

אב שגובהו שווה לַגובה הממוצע של האבות – 173 ס״מ, הניבוי לגובה בנו הוא ממוצע גובהי הבנים – 175 ס״מ. נזכיר, קו הרגרסיה עובר דרך נקודת הממוצעים.

**שאלת חשיבה**: מכיוון שהמתאם חיובי, נצפה שלאבות גבוהים מהממוצע יהיו בנים גבוהים מהממוצע ולאבות נמוכים מהממוצע יהיו בנים נמוכים מהממוצע. יתר על כן, מכיוון שהבנים גבוהים בממוצע מהאבות וסטיות התקן בקרב האבות ובקרב הבנים זהות, נצפה שלאבות גבוהים הסטייה של גובה הבן כלפי מעלה תהיה גדולה לפחות כמו הסטייה של גובה האב כלפי מעלה. **האומנם**?

על פי נוסחה (6) בעמ׳ 71, אפשר לרשום את נוסחת הניבוי כך: $\hat{y}_x - \bar{y} = b \cdot (x - \bar{x})$.

ובמקרה שלנו:

$$\underbrace{\hat{y}_x - 175}_{\downarrow} = \frac{1}{2} \cdot \underbrace{(x - 173)}_{\downarrow}$$

**סטייה מהממוצע בערכו של $x$ — הסטייה המתאימה בערך הניבוי**

נתבונן עתה ב**אב גבוה מאוד**, שגובהו 185 ס״מ. אב זה גבוה מממוצע האבות (173) ב-12 ס״מ. היה אפשר לצפות מראש שהבן יהיה גבוה לפחות ב-12 ס״מ מממוצע הגובה של הבנים. ואולם, על פי קו הרגרסיה הניבוי לסטייה של גובה הבן מהממוצע הוא $0.5 \cdot (12) = 6$ בלבד.

לאב גבוה ב-12 ס״מ מהממוצע, הניבוי הוא שגם הבן יהיה גבוה מהממוצע, אבל רק ב-6 ס״מ.

בדומה לכך, נתבונן ב**אב נמוך מאוד**, שגובהו 163 ס״מ: $x - \bar{x} = -10$, כלומר האב נמוך ב-10 ס״מ מהגובה הממוצע של האבות. היינו מצפים שהבן של אב זה יהיה גם הוא נמוך מאוד. ובכן, הניבוי לסטייה של גובה הבן מהממוצע הוא: $0.5 \cdot (-10) = 5$.

לאב נמוך ב-10 ס״מ מהממוצע, הבן צפוי להיות אומנם נמוך מממוצע הבנים, אבל רק ב-5 ס״מ.

### ➢ נסיגה (רגרסיה) אל הממוצע

בדוגמה האחרונה נוכחנו כי לאבות גבוהים ננבא אומנם בנים גבוהים, אבל יחסית גבוהים פחות מהאבות; ובאופן דומה לאבות נמוכים ננבא בנים נמוכים, אבל יחסית פחות מהאבות. במילים אחרות, בגובה הבנים התגלתה תופעה של **נסיגה לכיוון ממוצע** הגובה של הבנים. פירסון בחן דוגמאות נוספות ובכולן מצא תופעה דומה.

**באופן כללי**: נתבונן בשתי תכונות שמקדם המתאם ביניהן חיובי וקטן מ-1, והן בעלות **שונויות שוות**. על פי נוסחה (6) נצפה שהסטיות מהממוצע של ערכי הניבוי עבור $y$ יהיו **קטנות** מהסטיות של ערכי $x$ מהממוצע (כמו בדוגמה 8 – גובה האב וגובה הבן).

מורהו של פירסון, סר פרנסיס גלטון, קרא לתופעה זו **רגרסיה (נסיגה) אל הממוצע** ( regression toward the mean). זוהי הסיבה שקו הניבוי הזה כונה "קו רגרסיה" (regression line).

האבחנה שתיארנו מסבירה תופעה שמתגלית כשעורכים שתי **מדידות חוזרות** של תכונה כלשהי באותם נבדקים. נשים לב שלפנינו שתי מדידות בעלות אותו פיזור, ומקדם המתאם ביניהן הוא חיובי אך אינו 1.

**דוגמאות**:

- כאשר מפיקים סרט המשך לסרט שהצליח כלכלית, יש סיכוי טוב שסרט ההמשך יצליח פחות, היות שהצלחת הסרט הראשון חרגה מאוד מהממוצע.
- בוואלה-ספורט תואר מצב טיפוסי שבו השבועות הראשונים לאחר החלפת מאמן הם "ירח דבש", שבמהלכו הקבוצה מצליחה לשפר את הישגיה. אולם בסיומם דועכות ההצלחות והאפקט מתפוגג. בעקבות זאת נשאלה שם השאלה מדוע בכל זאת מעדיפים בעלי הקבוצה פיטורים של המאמן על פני שמירה על יציבות הקבוצה.

התופעה המסתמנת במדידות חוזרות היא שנבדקים שהיו בקצה הערכים התחתון במדידה הראשונה מראים בממוצע שיפור במדידה השנייה, ונבדקים שהיו בקצה הערכים העליון במדידה הראשונה מראים בממוצע ירידה במדידה השנייה. בשפת יום-יום אפשר לקרוא לתופעה בשם "חזרה לבינוניות".

אחד ההסברים לתופעה זו נעוץ בגורם האקראיות שבמדידות, למשל בציונים: הישג גבוה הוא תולדה של כישרון, חריצות ועוד, אבל גם תלוי במזל (אם למשל פתרנו שאלה דומה לפני הבחינה). על גורם כזה אין שליטה בפעם הבאה.
הסבר נוסף הוא שהמדידה הראשונה **לא נבחרה מקרית**, בחנו מקרה קיצוני ושאלנו אם נצפה שגם המקרה הבא יהיה קיצוני.

בתרגילי החשיבה בסוף הפרק תמצאו דוגמאות אמת רבות נוספות.

**זהירות – על עונשים ופרסים**

חוסר מודעות לתופעת הרגרסיה אל הממוצע יכול להוביל לכשלי חשיבה: בסדרה של ניסויים חוזרים, ביצוע גרוע במיוחד עלול לגרור פעולה של ענישה. יש לזכור שגם אם יתרחש שיפור בעקבותיה, ייתכן בהחלט שהשיפור הוא תוצאה של חזרה לממוצע ולא של יעילות הענישה. זהירות דומה נדרשת במדיניות של תגמול על הצלחה יוצאת דופן.

## 6.2 קשר סטטיסטי וסיבתיות

**אין למהר ולהסיק מקשר סטטיסטי בין שני משתנים שיש ביניהם קשר סיבתי**. יש להביא תמיד בחשבון את האפשרות שיש **משתנה מתערב** המשפיע על שני המשתנים, והוא המסביר את הקשר ביניהם.

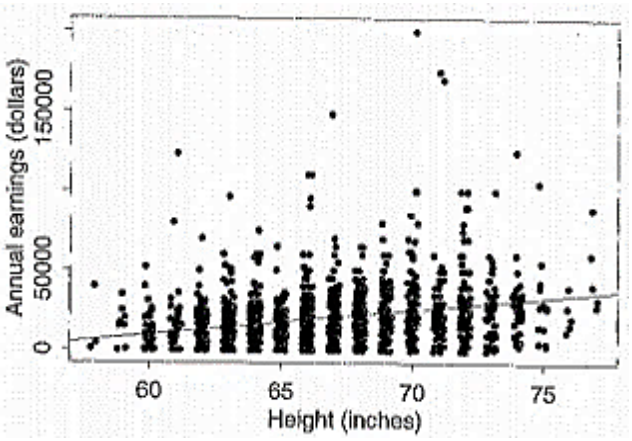
לדוגמה, יש קשר יורד בין גובהם של תלמידי בית ספר לבין מספר שגיאות הכתיב שלהם. האומנם לילדים נמוכים יש שגיאות כתיב רבות יותר? ובכן, אפשר לספק הסבר פשוט לקשר שהתגלה: הגורם המעורב הוא כמובן הגיל. עלייה בגיל התלמידים כרוכה בעלייה בגובה ובשיפור רמת הכתיבה.

סיפור משעשע בהקשר זה הוא על חייזר מהמאדים שחזר מסיור בכדור הארץ וסיפר בהתפעלות לאנשי המאדים שלבני האדם יש תכשיר מצוין המעודד השמנה, הנקרא סוכרזית. על פי תצפיותיו, שתיית משקה עם סוכרזית קשורה למשקל גבוה בקרב בני אדם.

לעיתים קרובות מדי אנו נתקלים בטעויות דומות אפילו אצל חוקרים מכובדים. על כן מומלץ תמיד לקרוא בביקורתיות תוצאות מדעיות. הנה שתי דוגמאות נוספות.

- מצאו שבקרב הרופאים המנתחים הנחשבים מומחים ביותר בתחומם יש נטייה לשיעור גבוה יותר של ניתוחים בלתי מוצלחים. האם זה אומר שלמעשה המומחים הללו אינם כה מומחים ועדיף לבחור רופא הנחשב פחות מומחה בתחומו?
חשיבה מעמיקה יותר מעלה שאין זה בהכרח כך. שהרי סביר יותר שהקשר השלילי שנמצא נובע למעשה מכך שהרופאים המומחים מנתחים בדרך כלל את החולים הקשים יותר, ולכן באופן טבעי שיעור ההצלחה בניתוחים שלהם נמוך יותר.

- בדיאגרמת הפיזור באיור 28 מוצגת הכנסה מול גובה במדגם מקרי גדול מאוד של אוכלוסייה בארצות הברית בשנת 1990. על פי הדיאגרמה, נראה שיש **קשר עולה** (אם כי לא חזק) בין שני המשתנים. באיור העברנו את **קו הניבוי הטוב ביותר**, והקו עולה.[5] כיצד אפשר להסביר את הקשר העולה?




**איור 28.** הכנסה שנתית מול גובה, ארצות הברית

הממצא אכן נשמע סנסציוני. מה עומד מאחורי הקשר שהתגלה? אפשר לתת לכך הסבר פשוט. המדגם כלל, כמובן, גברים ונשים. הנשים נוטות להיות נמוכות מהגברים, ולמרבה הצער המשכורות שלהן נוטות להיות נמוכות מאלו של הגברים. הקשר בין גובה להכנסה יכול להיות מוסבר אפוא באמצעות גורם המגדר.

[5] הגרף לקוח מהספר Teaching Statistics שכתבו Andrew Gelman & Deborah Nolan.

**מבחר ציטוטים עיתונאיים**

הציטוטים שלפנינו מדגימים כיצד עיתונאים מכובדים טועים ומטעים כשהם מצביעים על קשר **סיבתי** מתוך מידע על קשר בין המשתנים. נסו לנתח את הכשלים בקטעי כתבות אלו ולהיזהר מהם בעתיד.

1. "רוצים להיות טייסים? עדיף שתגורו במרכז הארץ, תיגשו לבגרות ריאלית מורחבת, ותהיו גם פעילים בתנועת נוער... כך עולה מנתונים שפרסם אמש חיל האוויר בנוגע לדמותם של בוגרי קורס הטיס שיסתיים השבוע..." (ישראל היום, 21.6.2009).

   האם נמליץ למעוניינים להיות טייסים לעבור מייד לתל אביב כדי לשפר את סיכוייהם?

2. "מחקר: השכלת ההורים מקפיצה את ציון הפסיכומטרי בעשרות נקודות; על פי הלמ"ס, הבחינה שמתיימרת לבדוק את היכולות הלימודיות מושפעת גם מהרקע החברתי של הנבחן" (כותרת מאמר, הארץ, 22.5.2009).

   בגוף המאמר מתברר שבלמ"ס מצאו קשר הדוק בין הציון הפסיכומטרי לבין רמת ההשכלה של ההורים: "כל שנת השכלה נוספת בהשכלת ההורים [...] מביאה לתוספת של כמעט 10 נקודות בבחינה עצמה". האם משמעות הממצא היא שאם תשלחו את הוריכם להמשך לימודים (מכללה או אוניברסיטה) של שלוש שנים, תצליחו טוב יותר בבחינה הפסיכומטרית והציון שלכם יעלה ב-30 נקודות?

3. "הדרך הטובה ביותר להצליח בבחינות הבגרות היא לעבור את מקצועות הבחינה ברמות המוגברות – ארבע וחמש יחידות לימוד. בדרך זאת התלמידים גם יכולים להגדיל את סיכוייהם להצטיין בתעודת הבגרות. לשם השוואה, מבין הניגשים לבחינת הבגרות במתמטיקה ברמה הבסיסית של שלוש יחידות לימוד, עברו את הבחינה 88% מהתלמידים. לעומת זאת, אלה שניגשו לבחינה במתמטיקה ברמה של ארבע יחידות לימוד, שיפרו את סיכויי ההצלחה שלהם ב-7%, ומי שעברו את הבחינה ברמה של חמש יחידות לימוד, שיפרו את סיכויי הצלחתם ב-9%..." (הארץ, 2001.7.5).

   ובכן, מאופן הדיווח עולה המלצה גורפת לכלל התלמידים – כדי להצליח בבחינת הבגרות במתמטיקה כדאי ללמוד ברמה של ארבע או חמש יחידות, כי אז מצליחים יותר מאשר ברמה של שלוש יחידות. האומנם?

חִשבו מהי הסיבה האמיתית לממצאים שהוצגו כאן. מה עומד למעשה מאחורי הממצאים?

❤ הסקת מסקנות על קיום קשר סיבתי בין משתנים בעקבות איתור מתאם ביניהם עלולה אף להוביל לאובדן חיי אדם. הדוגמה המפורסמת ביותר,[6] שהיא גם אחת המזיקות ביותר, היא הטענה שחיסונים גורמים לאוטיזם, שטען אנדרו וייקפילד במאמרו שפורסם בכתב העת החשוב Lancet בשנת 1998. וייקפילד הראה – במחקר גרוע מאוד – שיש מתאם חיובי בין מתן חיסון MMR לילדים ובין אבחון של אוטיזם אצל ילדים שחוסנו. אף שלא הודגם קשר סיבתי, די היה בכך כדי להצית תנועה רחבה של התנגדות לחיסונים, שפעילה עד היום. אי מתן חיסונים מוביל להתפרצות מגיפות ולמקרי מוות שהיו יכולים להימנע.

---

[6] תודה ליוסי לוי ולפוסט שלו 'סטטיסטיקה רעה: אי אבחנה בין מתאם לסיבתיות', מתוך האתר נסיכת המדעים.

**סיכום**: הבה ניזָהר שלא למהר ולהסיק מקשר סטטיסטי בין שני משתנים על הסיבתיות. עם זאת,

♥ **המסקנה ההפוכה נכונה**: קשר סיבתי בין שני משתנים הוא אינדיקציה למתאם (גבוה בדרך כלל) בין שני המשתנים, אם כי מתאם זה אינו בהכרח קווי.

## 6.3 על מקריות ומובהקות: 'הסקה' מהמדגם לאוכלוסייה

עד כה לא עסקנו כלל בשאלה מהו מקור הנתונים שבידינו: האם מדובר באוכלוסייה כולה – למשל אוכלוסיית תלמידי בית הספר, או רק במדגם מתוך האוכלוסייה – למשל בסקר בחירות. ואם מדובר במדגם – האם הוא מייצג את האוכלוסייה כולה?

יש לדעת:

☛ בדיווחים בתקשורת, הנתונים המשמשים לצורך החישובים הם בדרך כלל מדגם מהאוכלוסייה, ואילו מסקנות הדיווח משויכות בדרך כלל לאוכלוסייה כולה.

האם אפשר להסיק מערכו של מקדם המתאם שחושב מנתוני המדגם על הקשר בין המשתנים באוכלוסייה כולה?

בדיוק בכך עוסקת **ההסקה הסטטיסטית**. נציג כאן דוגמאות הבהרה קצרות.

**דוגמה 9 (הקשר בין גובה השכר למידת האטרקטיביות של עובדים)**. במחקר שנערך באוניברסיטת פלורידה בשנת 2009 בחנו החוקרים השערה שמעסיקים נוטים לשלם שכר גבוה יותר לאנשים יפים יותר. אוכלוסיית המחקר הייתה המועסקים כולם, ומתוכה נדגמו למחקר 191 מועסקים, נשים וגברים. לכל מועסק שנבחר נרשמו שכרו וכן הערכה של מידת האטרקטיביות שלו – מדד שקבע צוות שופטים על סמך תמונות המועסקים. **במדגם** נמצא מקדם מתאם חיובי $r = 0.24$ בין מדד האטרקטיביות לבין גובה השכר.

השאלה המעניינת בדוגמה זו נוגעת, כמובן, ל**אוכלוסיית המועסקים** כולה: האם המתאם החיובי שהתקבל במדגם מצביע במובהק על מתאם חיובי באוכלוסייה, דהיינו על העדפה בשכר של היפים יותר בקרב האוכלוסייה? שאלת המפתח היא אם ייתכן שהמתאם החיובי שהתקבל נובע פשוט מ**גורם האקראיות במדגם** ואינו מעיד על נטייה באוכלוסייה כולה.

ניתוח הנתונים בשיטות של הסקה סטטיסטית העלה שיש בנתוני המחקר **עדות מספקת** לנכונות הקביעה ששכרם של אנשים אטרקטיביים נוטה להיות גבוה משל אנשים שאינם כה אטרקטיביים.

**דוגמה 10 (משך הצפייה בטלוויזיה ומידת האגרסיביות)**. כדי לבחון אם צפייה מרובה בטלוויזיה מגבירה את האגרסיביות אצל ילדים, בחרו מדגם של עשרה ילדים ומצאו עבורם מתאם של $r = 0.32$ בין זמן צפייה יומי בטלוויזיה לציון אגרסיביות. גם במקרה זה נוגעת השאלה המעניינת למתרחש באוכלוסיית הילדים כולה: האם המתאם החיובי שהתקבל במדגם מצביע באופן מובהק על מתאם חיובי באוכלוסייה כולה?

בניתוח הנתונים בשיטות של הסקה סטטיסטית התברר ש**אין** בנתונים שהוצגו **משום הוכחה** לקשר בין משך הצפייה בטלוויזיה לרמת האגרסיביות אצל ילדים.

**הבחנה**: נשים לב שהמתאם שהתקבל בדוגמה 10 גבוה מזה שהתקבל בדוגמה 9. הסיבה שבדוגמה 10 לא נמצאה עדות מובהקת לקשר באוכלוסייה כולה נעוצה בגורם של גודל המדגם. במדגם קטן מאוד יש אפשרות סבירה שתוצאה גבוהה יחסית שהתקבלה בו היא **מקרית**.

❤ ההסקה הסטטיסטית מציעה שיטות מתוחכמות של הסקה ממדגם מקרי לאוכלוסייה כולה. מובן שחישוב הטעויות בשיטות אלו מביא בחשבון גם את גודל המדגם.

לסיום, נחזור להבדל בין **קשר סטטיסטי לקשר סיבתי**: בדוגמה 10, גם אם במדגם גדול יותר יימצא מתאם חיובי מובהק בין שני המשתנים, לא יהיה נכון להסיק שצפייה מרובה מגבירה את האגרסיביות, אלא רק שיש קשר (עולה) בין שני המשתנים. הרי ייתכן, למשל, שילדים אגרסיביים נוטים לצפות בטלוויזיה יותר מאחרים. בדוגמה 9 יש לברר אם הקשר הסטטיסטי שהתגלה הוא גם קשר סיבתי או שעשוי להיות גורם מתערב.

### אז מה היה לנו? כשלי חשיבה

**תובנות חדשות**

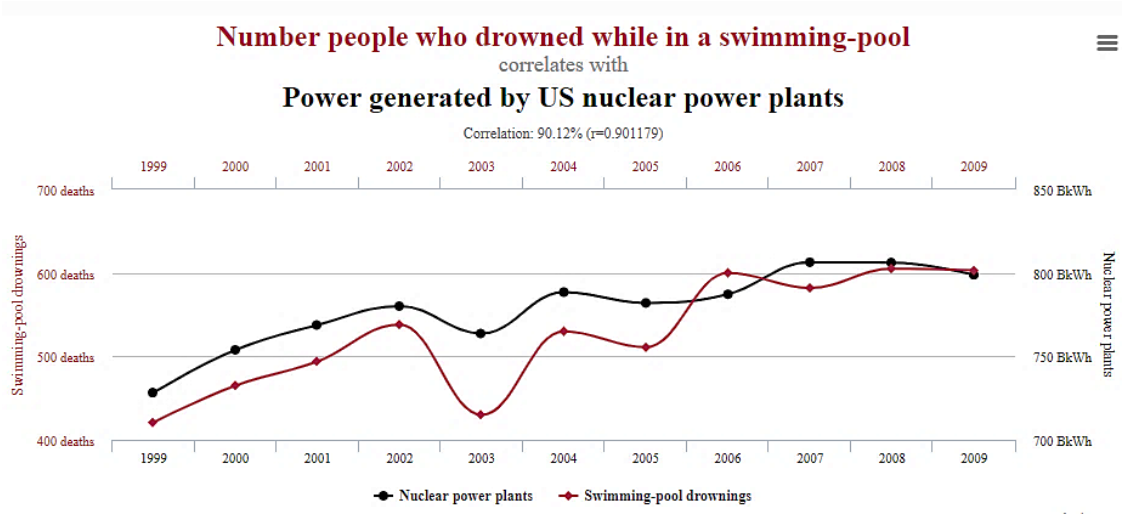
- חוסר מודעות לתופעת **הנסיגה אל הממוצע** יכול להוביל ל**כשלי חשיבה**: בסדרה של ניסויים חוזרים, יש לדעת שגם אם יתרחש שיפור בעקבות פעולה שננקוט, ייתכן בהחלט שההסבר הוא חזרה לממוצע ולא יעילות הפעולה. זהירות דומה נדרשת במדיניות של תגמול על הצלחה יוצאת דופן.
- אין למהר ולהסיק מ**קשר סטטיסטי** שמצאנו בין שני משתנים בדבר **קשר סיבתי** ביניהם. יש להביא תמיד בחשבון את האפשרות ש**משתנה מתערב** כלשהו משפיע על שני המשתנים והוא שמסביר את הקשר ביניהם.
- מסקנה חפוזה שיש קשר סיבתי בין $Y$ למשתנה מנבא $X$ בעקבות קבלה של מתאם גבוה יכולה להוביל ל**כשל חשיבתי**, שתוצאותיו עשויות להיות חמורות: אנו עלולים לנסות להשפיע על ערכו של $Y$ באמצעות שינויים יזומים בערכי המשתנה $X$, שפעמים רבות יש לנו אפשרות לשלוט בו.
  המלצות רפואיות פופולריות ברפואה ובתזונה הן פעמים רבות תוצאה של כשל כזה.
- בדיווחים בתקשורת, הנתונים שנותחו הם בדרך כלל **מדגם מאוכלוסייה רחבה**, ואילו המסקנות הגורפות של הדיווח נוגעות בדרך כלל לאוכלוסייה כולה. הסקה מהמדגם לאוכלוסייה דורשת שימוש בשיטות ניתוח מתוחכמות המביאות בחשבון את גורם האקראיות במדגם. נוכחנו שבמדגמים קטנים יש אפשרות סבירה שתוצאה של מקדם מתאם גבוה שהתקבלה במדגם היא מקרית.
  בזאת ועוד עוסקת **ההסקה הסטטיסטית**.

## תרגילי חשיבה

**תרגיל 1.** נתחו את הציטוט מדבריו של פרופסור דניאל ליברמן, מהרווארד: "כל הנושא של שמונה שעות שינה זה מיתוס. יש מחקרים המצביעים על כך שאנשים שישנים שמונה שעות בלילה או יותר, סובלים משיעורי תמותה גבוהים יותר מאלה של אנשים שישנים שבע שעות".

**תרגיל 2.** נסו לחשוב על דוגמאות נוספות שבהן נתקלתם בכשלים כגון אלה שדנו בהם בפרק זה. מעתה היו ערים לכך...

**תרגיל 3 (מתאמים מלאכותיים[7]).** האם לדעתכם יש קשר סיבתי המסביר את המתאם הגבוה ($r = 0.901$) בין מספר האנשים שטבעו בבריכות שחייה בארצות הברית לבין כמות האנרגיה שיוצרה בכורים הגרעיניים שבה באותן שנים?




**תרגיל 4 (מדידת לחץ דם).** בקבלה לבית חולים עורכים גם בדיקה של לחץ הדם. לנבדקים שלחץ הדם שלהם גבוה מאוד במדידה זו נוהגים לערוך בדיקה שנייה. מתברר שבבדיקה השנייה לחץ הדם בדרך כלל נמוך יותר. ההסבר של הרופאים הוא שבמדידה הראשונה הנבדקים עצבניים מעט ובשנייה הם נינוחים יותר. תופעה זו מכונה 'אפקט החלוק הלבן'.

כדי לבדוק את אמיתות המסקנה הוחלט להתמקד בנבדקים שהמדידה הראשונה שלהם הייתה נמוכה מאוד. התברר שבמדידה השנייה לחץ הדם היה גבוה יותר. האם נבדקים אלו עצבניים יותר בבדיקה השנייה?

א. עזרו לרופאים למצוא הסבר לתופעות אלה.

[7] ראו https://en.wikipedia.org/wiki/Spurious_relationship.

ב. נמצא שבבדיקה הראשונה לחץ הדם הממוצע היה 130 מילימטר כספית ובשנייה 120. בשני המקרים סטיית התקן הייתה 15 מילימטר. האם ממצאים אלה תומכים בתאוריית העצבנות והנינוחות של הנבדקים? נסו להציע הסבר אחר.

**תרגיל 5 (גובהי אבות ובנים).** במחקר של פירסון נמצא שממוצע הגובה של בנים לאבות שגובהם הוא 183 ס״מ היה 180.3 ס״מ בלבד. היעזרו בדיאגרמת הפיזור כדי לענות 'נכון' או 'לא נכון':

א. אם נתמקד בבנים שגובהם 180.3 ס״מ, הגובה הממוצע של האבות שלהם הוא כ-183 ס״מ.

ב. לאב גבוה מהממוצע, גם בנו גבוה מהממוצע המתאים.

**תרגיל 6 (הרמת משקולות).** מחקר בדק אפיונים פיזיים שונים של מרימי משקולות מקצועיים. במחקר נמצא, למשל, שמקדם המתאם בין משקל האתלט לבין המשקל שהוא מסוגל להרים הוא $r = 0.6$.

ענו 'נכון' או 'לא נכון' והסבירו:

א. בממוצע מרים משקולות מקצועי יכול להרים משקל שהוא 60% ממשקל גופו.

ב. אם משקל האתלט יעלה ב-5 קילוגרם, הוא יוכל לצפות להרים משקל נוסף של 3 קילוגרם.

ג. ככל שהאתלט שוקל יותר, כך הוא יכול לצפות להרים משקל גדול יותר.

ד. ככל שהאתלט יכול להרים משקל גדול יותר, כן בממוצע המשקל שלו גדול יותר.

ה. אפשר ליחס 60% מיכולת הרמת המשקלות של אתלט למשקל שלו.

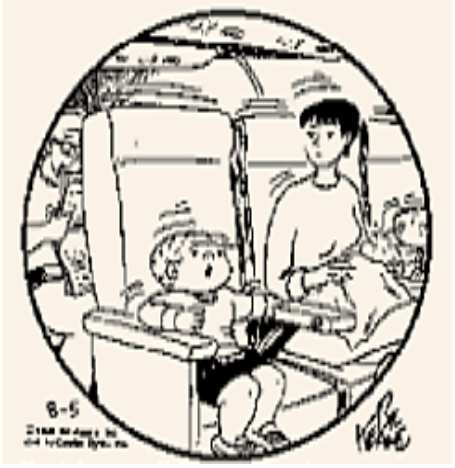
**תרגיל 7 ( IQ של ילדי הגן).** בתוכנית ניסיונית להעלאת ה-IQ של ילדי גן נבחנו הילדים במבחן IQ לפני ואחרי הרצת התוכנית. בשני המבחנים הממוצע היה כ-100 וסטיית התקן הייתה כ-15. במבט ראשון נראה אפוא שלתוכנית לא הייתה השפעה. במבט שני התגלתה תופעה מפתיעה. ילדים שהיו הרבה מתחת לממוצע במבחן המקדים עלו בממוצע ב-5 נקודות, ואילו ילדים שהיו הרבה מעל לממוצע ירדו בממוצע ב-5 נקודות.

האם השפעת התוכנית הייתה הקטנה של האי-שוויון ברמת האינטליגנציה של ילדי הגן בעקבות הפעילויות המשותפות של ילדים מבריקים עם ילדים מתקשים?

תנו הסבר אחר בעזרת התופעה של נסיגה אל הממוצע.

# סיכום

**קשר סטטיסטי וסיבתיות בראי הקריקטורה**



הלוואי שיפסיקו להדליק שוב ושוב את השלט 'להדק חגורות בטיחות'.

בכל פעם שהם מדליקים אותו המטוס מתחיל להיטלטל (נבדקות זכויות יוצרים)

**עיקרי הדברים**

ביחידה זו עסקנו בזוג משתנים כמותיים $X, Y$. הניתוחים שערכנו מתבססים על $n$ **מדידות** של ערכי שני המשתנים, כלומר $n$ זוגות של נתונים. השאלות ששאלנו את עצמנו הן:

– האם המשתנים הללו קשורים זה לזה?

– האם יש דרך **למדוד** את כיוון הקשר וכן את עוצמתו?

– האם נוכל לנבא את ערכו העתידי של אחד המשתנים על בסיס הידע על ערכו של המשתנה האחר?

הנה עיקרי הדברים:

- בשלב ראשון יש לצייר דיאגרמת פיזור של זוגות הנתונים שבידינו. מהדיאגרמה נלמד כיצד נראה הקשר (קווי, פרבולי, ענן סתמי וכן הלאה), נלמד על כיוון הקשר, וגם מעט על עוצמתו, ונבחן אם הנקודות מהודקות סביב עקומה כלשהי שתוכל לשמש בהמשך לצורכי ניבוי.
- התמקדנו בניבוי קווי. בדיאגרמת הפיזור יש לבחון אפוא אם יש לנקודות נטייה להסתדר סביב קו ישר עולה או יורד. אם לא זה המצב – אין טעם להמשיך בניתוח שנציע.

- השלב הבא הוא חישוב מקדם המתאם $r$ בין המשתנים באמצעות נוסחה (2) או (3). מקדם המתאם מודד את כיוון ואת עוצמת הקשר הקווי בין המשתנים.
- תכונות: ערכי מקדם המתאם נעים בין $(-1)$ ל-$1$. ערך חיובי מציין קשר עולה, וערך שלילי – קשר יורד, ערכי מקדם מתאם קרובים ל-$1$ או ל-$(-1)$ מצביעים על קשר קווי חזק בין המשתנים; ערכים קרובים ל-0 מצביעים על קשר קווי חלש.

❤ מקדם המתאם בודק אם הקשר הקווי בין המשתנים חזק דיו כדי **להציע ניבוי** (הערכה) קווי ל-$Y$ על בסיס ערכו של $X$ (ראו שאלות מעשיות בהמשך).

- שלב הניבוי: הצגנו נוסחאות – נוסחאות (4) ו-(5) – לחישוב ידני של הניבוי הקווי הטוב ביותר, הוא קו הרגרסיה. לפני השימוש בנוסחאות אלה יש לחשב את הממוצעים ואת סטיות התקן עבור שני המשתנים וכן את מקדם המתאם ביניהם.
- כאשר מספר הנתונים קטן מאוד, לא קשה לערוך ידנית את כל החישובים הדרושים או להיעזר במחשבון. אך בבעיות מעשיות שהנתונים בהן רבים, נזדקק לאמצעי חישוב מתקדמים יותר כדי למצוא את מקדם המתאם.
- תוכנת **אקסל**, המצויה בכל מחשב וקלה להפעלה, מספקת בקלות רבה את אמצעי החישוב שנדרשים להם (ראו נספח ב). מומלץ להימנע משימוש בנוסחאות ובחישובים ידניים.
- **טיב הניבוי**: נוסחה (7) מאפשרת השוואת טיב הניבוי באמצעות רגרסיה לניבוי ללא משתנה מנבא.

**משימה**: על בסיס עיקרי הדברים עִרכו תרשים זרימה של כל שלבי ניתוח קשרים בין משתנים שערכנו בספר זה [תשובה בעמ' 133].

### נוסחאות שימושיות לפתרון ידני של תרגילים

כל ההתייחסויות בפתרונות המשימות שבהמשך הן לנוסחאות אלו (מספור הנוסחאות הוא כפי שמופיע בפרקים עצמם). נזכיר שוב שמומלץ להיעזר ככל האפשר בתוכנה חישובית כגון אקסל (ראו נספח ב).

| | | |
|---|---|---|
| **שונות** | $\sigma^2 = \frac{1}{n}\left[x_1^{\,2} + \cdots + x_n^{\,2}\right] - \bar{x}^2$ | (0) |
| **פעולת תקנון** | $\underset{\sim}{x} = \frac{x - \bar{x}}{\sigma_X}$ | |
| **מקדם המתאם** | $r = \frac{1}{\sigma_X \cdot \sigma_Y} \cdot \frac{1}{n} \cdot \left\{ \underset{\downarrow}{(x_1 - \bar{x}) \cdot (y_1 - \bar{y})} + \cdots + \underset{\downarrow}{(x_n - \bar{x}) \cdot (y_n - \bar{y})} \right\}$ | (2) |

**מכפלת הסטיות מהממוצע המתאים**

| | | |
|---|---|---|
| | $r = \frac{1}{\sigma_X \cdot \sigma_Y} \left\{ \underbrace{\frac{1}{n} \cdot \left[ x_1 \cdot y_1 + \ldots + x_n \cdot y_n \right]}_{\downarrow} - \underset{\downarrow}{\bar{x} \cdot \bar{y}} \right\}$<br>ממוצע המכפלות — מכפלת הממוצעים | (3) |
| **קו הניבוי של $Y$ על פי $X$**<br>**(קו רגרסיה, קו ריבועים פחותים)** | $\hat{y}_x = a + bx$<br>$\downarrow$<br>**הניבוי עבור $x$** | |
| **השיפוע $b$** | $b = r \cdot \frac{\sigma_y}{\sigma_x}$ | (4) |
| | $a = \bar{y} - b\,\bar{x}$ | (5) |
| **נוסחה נוחה לשימוש** | $\hat{y}_x = \bar{y} + b \cdot (x - \bar{x})$ | (6) |
| **מדד לטיב הניבוי של קו הרגרסיה** | $\sigma_y^2 \cdot (1 - r^2)$ $b$ | (7) |

## מהעיתונות[8] – קשרים בין משתנים

# מדענים: שינה טובה מונעת עלייה במשקל

לרוב האנשים יש נטייה מאוד פשוטה לאלו שרוצים לעשות דיאטה: תשנו יותר. מחקר שנערך באחרונה גילה מיתאם חזק ומפתיע בין מספר שעות השינה של אנשים לבין הסיכוי שלהם לסבול ממשקל יתר.

הסיכוי להשמנה של נבדקים שישנו פחות מארבע שעות שינה בלילה, היה גבוה ב־73 אחוז מהסיכוי לה־שמנה של אלו שישנו את הכמות המומלצת – בין שבע לתשע שעות בלילה. וככל שישנים יותר, לפי ממצאי המחקר, יורד הסיכוי לעלייה במשקל: הסיכון של אלו שישנו חמש שעות בלילה גבוה ב־50 אחוז בהשוו־אה לסיכון של אלה ש־ישנו שבע שעות, ושל אלה שיש־נו שש שעות – ב־23 אחוז.

"ייתכן שיש כאן חלון הזדמנויות לתפוס שתי צי־פורים במכה – גם לעודד אנשים לישון יותר וגם לסייע להם לשמור על המשקל", אומר ד"ר סטיבן היימספילד מאוניברסיטת קולומביה. הוא ועמיתו ג'יימס גנגורוש הובילו את המחקר, שממצאיו מרג־גים השבוע בכנס שנערך בלאס־וגאס.

לכאורה הממצאים מפתיעים, אומר גנגורוש, כי מי שישן שורף פחות קלוריות. אולם מתברר כי המחסור בשינה מוריד את רמות ההלבון לפטין, שמרכא את התיאבון ומשפיע על הדרך שבה קובע המוח אם הגוף שבע. מחסור בשינה גם מעלה בגוף את רמת הגרלין, הזורר מעורר תיאבון.

מעריב, 18.11.2004

# מחקר: העישון עלול להזיק לאינטליגנציה

## מדליקים סיגריה כדי להתרכז? אתם למעשה פוגעים ביכולת המחשבה שלכם ■ חוקרים מסקוטלנד גילו כי מעשנים מקבלים ציונים נמוכים יותר במבחני רמת משכל

מאת מיכל שפירא

מחקרים רבים שנעשו בשנים האחרונות טע־נו כי למרות הסכנות, לסרטריות יש גם כמה ית־רונות: העישון מסחרת מחשובן לחלות באלצהיי־מר, ומחזק את הזיכרון וחקשב. אבל מחקר חדש חדשף כי לעישון עלולה לחיות דוקא השפעה שלי־לית על האינטליגנציה.

החוקרים, מאוניברסיט־את אברדין ואדינבורו שבסקוטלנד, בדקו האם ע־שון מהווה גורם סיכון לי־רידה באינטליגנציה (תוב־נה). הם נעזרו בכ־500 גשים וגברים, אשר נולדו ב־1936 ולקחו חלק באותו מבחן IQ (מדד לאינטליגנ־ציה) כשהיו בני 11. הנחק־רים השכימו להיבחן מחדש בגיל 64, כדי לבדוק האם חלה אצלם ירידה כ־IQ עם השנים.

התוצאות שפרסמו כעת זיהו את העישון כגר־רם משמעותי ובלתי תלוי, יחד עם גורמים נוס־פים כמו עישון, חינוך ותפקוד ריאות, המנבא את התפקוד המנטלי ומשפיע על מגת המשכל. כמו כן נמצא קשר כי מעשנים (בעבר או בהווה) הפגינו ביצועים נמוכים יו־תר במבחנים פסיכומטרי־ים מאשר אנשים ללא היסטוריה של עישון. מח־צא נוסף בדוח הראה כי למעשנים יש ירידה גדולה יותר ביכולת הקוגניטיבית (תפיסה מוחית־הכרתית).

"אנו מסיקים כי עישון ארוך טווח ארגו גורם ליתרד־גות קוגניטיבים, אלא להמשך", אומרי החוקרים. "עישון תורם תרומה קטנה, אך משמעו־תית, לירידה ביכולת הקוגנ־טיבית בין הגילים 11־64".

צוות המחקר הוסיף כי "לאורות שההשפעות הש־ליליות על התובנה הן יחסית קטנות במובנים של האדם הבודד, אלו השפעות השובות ברמת האוכלוסייה ובבחירת סגנון החיים".

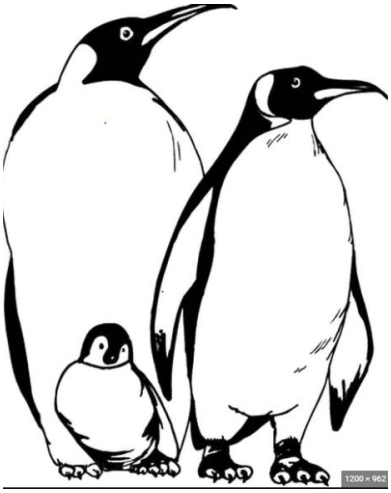

לעשן זה לא צעד חכם. עיבוד מחשב

מעריב, 7.12.2004

---

[8] תודה לסיגל לוי שמלקטת בחריצות פנינים מהעיתונות. תודה לעיתון מעריב שהסכים שנשתמש בכתבות. הקוראים מוזמנים למצוא כתבות עדכניות – בחיפוש פשוט בגוגל – בנושאים אלו וברבים אחרים.

# נספחים



חושב שכל צוותי הצילום שבאים לכאן לאחרונה גורמים להתחממות הגלובלית?

## נספח א: שאלות ותשובות להרחבת הדעת

באמצעות סדרה של שאלות חקר ננסה ללבן שאלות המשך העולות בעקבות הדיונים ביחידה זו.

**האם דיאגרמת פיזור היא כלי יעיל לתיאור גרפי של בעיות אמת שהנתונים בהן רבים?**

כאשר הנתונים רבים, ייתכן בהחלט מצב של ריבוי נתונים זהים או קרובים מדי זה לזה. במצב כזה דיאגרמת פיזור אינה יעילה, ומתעורר צורך בכלים גרפיים מתוחכמים יותר. במקרים כאלו עדיף לחלק את תחום ערכי המשתנים לקטעים ולהוסיף לדיאגרמה ממד של גובה שיבטא שכיחות. מתקבלת דיאגרמה תלת-ממדית (ראו דיון בעמ' 25).

**בבעיות מעשיות, מה נחשב קשר קווי חלש? מה נחשב קשר חזק? ומה נחשב קשר בינוני?**

אין לשאלה זו תשובה חד-משמעית – התשובה תלויה מאוד בתחום המחקר הרלוונטי. במדעים המדויקים מקובל שערכים קרובים ל-1 או ל-(1–) מובילים למסקנה שהקשר חזק, אך במדעי החברה, למשל, מקובל שמקדם מתאם של 1/2 או של (1/2–) נחשב חזק.

נזכיר שמקדם המתאם בין הציון הפסיכומטרי לבין הצלחה בלימודי המתמטיקה באוניברסיטה, $r = 0.337$ , נחשב חזק דיו כדי למיין בעזרתו את המועמדים.

**האם מקדם מתאם הקרוב ל-0 בין שני משתנים מעיד שאין קשר בין המשתנים? נסו לחשוב על דוגמאות מחיי היום-יום כדי לבסס את תשובתכם.**

נתבונן, למשל, בקשר שבין כמות החטיפים שאוכלים בערב אחד מול הטלוויזיה לרמת ההנאה מאכילתם. למספר קטן של חטיפים, מן הסתם הקשר עולה; אך מעל כמות אינדיבידואלית מסוימת של חטיפים ההנאה הולכת ופוחתת ככל שמגזימים בכמות. מקדם המתאם שיתקבל במקרה זה קרוב ל-0, אף שיש קשר ברור בין המשתנים (כנראה שצורתו פרבולה; ראו איור בהמשך).

**מדוע, אם כך, נהוג להתמקד בניבוי קווי?**

הסיבה היא מעשית. הצגנו נוסחה נוחה לניבוי הקווי הטוב ביותר – זהו הקו הישר 'הקרוב ביותר' לנקודות בדיאגרמת הפיזור. מבחינה מתמטית טהורה, אין בעיה למצוא באופן דומה נוסחה לעקומת ניבוי פרבולית טובה ביותר וכן עקומות ניבוי המתאימות לקשרים נוספים, אך הנוסחאות המתקבלות מסורבלות מאוד לשימוש.

**האם יש דרך למדוד עד כמה שיפרנו את הניבוי המתבסס על משתנה מנבא $X$ על פני הניבוי בלעדיו? עד כמה אפשר לשפר עוד את הניבוי אם נאפשר עקומות ניבוי לא קוויות?**

בסעיף 5.4 הצגנו מדד כללי שנקרא **הטעות הריבועית הממוצעת של הניבוי**. המדד מאפשר להשוות את טיב הניבוי של עקומות ניבוי מסוגים שונים. התמקדנו בחישוב טיב הניבוי של קו הרגרסיה בהשוואה לטיב הניבוי ללא משתנה מנבא (ראו גם דוגמה חישובית במשימה III בעמ' 66).

**מצאנו, למשל, מתאם חיובי בין ההכנסה $Y$ לבין מספר שנות הלימוד. אבל לניבוי ההכנסה ייתכנו משתנים מנבאים נוספים העשויים להיות רלוונטיים: מגדר, גיל, ותק בעבודה ועוד. האם יש דרך להתחשב בכל המשתנים הללו יחד? האם יש דרך לבדוק אם אכן כך ישתפר הניבוי?**

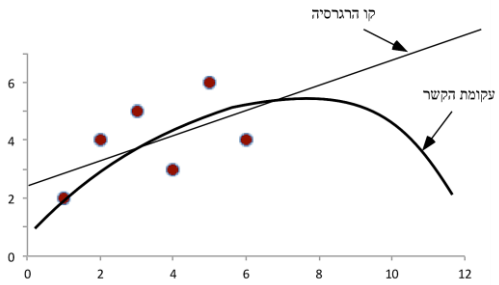
את ההגדרה של מקדם המתאם אפשר להרחיב למדידת קשר בין יותר משני משתנים – **מתאם מרובה**. יש גם נוסחאות ניבוי קווי המתבססות על כמה משתנים בעת ובעונה אחת – **רגרסיה מרובה**.

רמזנו כבר שיש מדד לטיב הניבוי המאפשר לבדוק אם הוספה של משתנים מנבאים חדשים אכן משפרת את הניבוי. נציין, לדוגמה, שהניבוי המקובל להצלחה באוניברסיטה מבוסס על הציון הפסיכומטרי ועל ממוצע ציוני הבגרות.

**האם נכון להשתמש בקו הרגרסיה לצורך ניבוי ערכו של $Y$ עבור ערכים של $X$ החורגים מטווח הערכים של הנתונים שבידינו?**

**בהחלט לא!!!** ייתכן, למשל, שהקשר בין המשתנים הוא למעשה פרבולי, אך הנתונים שאספנו כללו רק ערכים של $X$ בתחום העלייה או רק בתחום הירידה של העקומה (ראו איור). במצב כזה לא נצליח לגלות את הקשר הפרבולי בעזרת דיאגרמת הפיזור שקיבלנו, והניבוי באמצעות הקו הישר לערכי $X$ החורגים מתחום זה לא יהיה נכון.

בדוגמת החטיפים, למשל, בכל התצפיות שערכנו מספר החטיפים שנאכלו באותו ערב לא עלה על 5 (ראו איור). במצב כזה כל הנקודות בדיאגרמת הפיזור יסתדרו סביב קו ישר עולה. עם זאת, הניבוי על פי נוסחת קו זה עשוי שלא להתאים לניבוי עבור עשרה חטיפים בערב אחד.




**הנתונים שאספנו מהווים בדרך כלל מדגם מאוכלוסייה רחבה יותר. אם מקדם המתאם שחושב מתוך המדגם שונה מ-0, באיזו מידה אפשר להסיק מכך על קשר בין המשתנים באוכלוסייה כולה?**

בדיוק בכך עוסקת **ההסקה הסטטיסטית**. שאלת המפתח כאן היא אם ייתכן שהמתאם החיובי או השלילי שהתקבל בניתוח הנתונים שבידינו נובע פשוט מגורם האקראיות במדגם ואינו מעיד על נטייה באוכלוסייה כולה. ההסקה הסטטיסטית מציעה שיטות מתוחכמות לבחון שאלות מסוג זה (ראו דיון קצר בעמ' 92-91).

**היפוך תפקידים: האם ייתכנו מקרים שבהם נתעניין גם בניבוי $Y$ על פי $X$ וגם בניבוי $X$ על פי $Y$? נסו למצוא דוגמה פשוטה וגם להסביר את הקשר בין שני הניתוחים הנדרשים.**

המצב הרגיל הוא שברור לנו מהו משתנה המחקר $Y$ ואנו תרים אחר משתנים – בדרך כלל נוחים למדידה, שעשויים לתרום לניבוי ערכו העתידי של $Y$. עם זאת, למשל אפילו בבעיה הקלסית של הקשר בין גובה האבות וגובה הבנים, בהחלט ייתכן שנרצה גם להעריך את גובה האב על פי גובה בנו. הקשר בין שתי בעיות הניבוי הוא פשוט: מקדם המתאם כמובן זהה; דיאגרמת הפיזור של $X$ על פי $Y$ מתקבלת מהחלפת תפקידי הצירים (שיקוף ביחס לאלכסון); קו הניבוי עובר דרך נקודת הממוצעים $(\bar{y},\bar{x})$, ולא קשה למצוא את השיפוע מהחישובים שכבר נעשו.

**במחקרים מעשיים שנמצא בהם קשר טוב דיו בין המשתנים, פעמים רבות השלב הבא הוא לנסות להשפיע על ערכו העתידי של $Y$ באמצעות שינויים יזומים בערכם של המשתנים המנבאים. האם זהו ניסיון לגיטימי?**

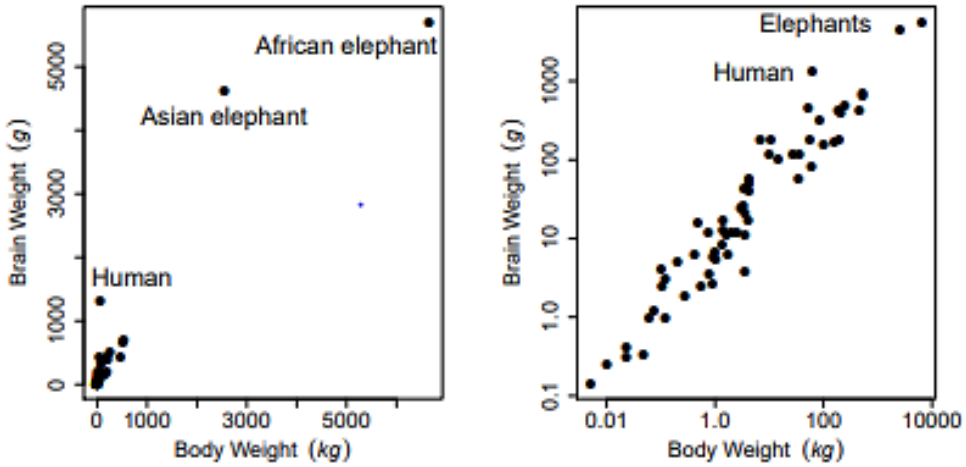
כאן יש לזכור את **כשל ההסקה החפוזה** מקשר סטטיסטי לקשר סיבתי. על המומחים בתחום הספציפי להמשיך לחקור אם ייתכן שהקשר הסטטיסטי שנמצא בין $X$ ל-$Y$ הוא תוצאה של גורם מתערב שלישי הקשור לשני המשתנים, וגורם זה הוא שאחראי לקשר הסטטיסטי. במצב כזה, שינוי בערכו של המשתנה המנבא $X$ לא יגרום לשינוי המיוחל בערכו של $Y$. זהו אחד המסרים החשובים בספר זה.

**איך החוקרים מזהים מהנתונים קשר שאינו קווי בין משתנים?**

בתשובה נציג מחקר שצוטט רבות בספרות המקצועית על הקשר בין משקל הגוף ומשקל המוח בקרב יונקים:[9]

[9] זוהי תמצית מחקר שצוטט רבות בספרות המקצועית. ראו למשל, עמ' 186–187 בספר Applied liner regression (S. Weisberg).

על בסיס מדידות שנערכו ב-62 סוגים של יונקים יבשתיים, ניסו החוקרים לבדוק קשרים בין משקל הגוף (בקילוגרמים) ובין משקל המוח (בגרמים) בקרב יונקים. דיאגרמת הפיזור שהתקבלה תחילה (השמאלית) הייתה חסרת משמעות.




במטרה לפענח את סוד הקשר, החוקרים בדקו אם יש קשר קווי בין פונקציות שונות של שני המשתנים, למשל בין הערכים הלוגריתמיים של המשתנים. ואכן, לאחר השינוי לסקלה לוגריתמית, התקבלה דיאגרמת הפיזור הימנית המצביעה על קשר קווי טוב בין הערכים הלוגריתמיים של המשתנים. נמצא אפוא **קשר טוב ומועיל בין שני משתני המחקר, אבל הקשר אינו קווי.**

## נספח ב: שימוש בתוכנת אקסל לחישובים

בבעיות מעשיות שהנתונים בהן רבים יש צורך להיעזר בתוכנות חישוביות ולהימנע ככל האפשר מחישובים ידניים באמצעות הנוסחאות שהצגנו בספר זה. אקסל (Excel) היא תוכנת גיליון עבודה נוחה, הנמצאת כמעט בכל מחשב אישי. בתוכנה יש אפשרויות רבות להצגת נתונים, לעריכתם ולחישוב של חישובים שונים בעזרתם. נצא כאן מתוך הנחה שלקוראים יש ידע כללי בתוכנת אקסל,[10] ורק נסקור בקצרה את אופן השימוש הרלוונטי לתכנים שבספר זה.

- אם הנתונים שאנו מתבססים עליהם הם רשימה של $n$ מספרים: $x_1,...,x_n$, נקליד אותם זה מתחת לזה בעמודה אחת של גיליון העבודה.

בטבלה שלהלן רשומות הפקודות המתאימות לעריכת החישובים הסטטיסטיים שנידרש להם לעיבוד הנתונים הללו (בסוגריים מצוין טווח התאים שבהם רשומים הנתונים):

| | | |
|---|---|---|
| ממוצע | $\bar{x}=\frac{1}{n}(x_1+...+x_n)$ | $=average(\_:\_)$ |
| שונות | $\sigma^2=\frac{1}{n}\left[(x_1-\bar{x})^2+...+(x_n-\bar{x})^2\right]$ | $=varp(\_:\_)$ |
| סטיית התקן | $\sigma=\sqrt{\sigma^2}$ | $=stdevp(\_:\_)$ |

שימוש בפונקציית אקסל $y=\frac{x-\bar{x}}{\sigma_X}$ מאפשר לחשב בקלות ובמהירות את **ציוני התקן** של כל הערכים ברשימה.

- אם בידינו רשימה של $n$ זוגות מספרים $(x_1,y_1),...,(x_n,y_n)$, נרשום בגיליון העבודה את ערכי $x$ זה מתחת לזה בעמודה אחת, ואת ערכי $y$ המתאימים בעמודה הסמוכה (כל זוג באותה שורה).

תוכנת אקסל מאפשרת לנו לקבל את **דיאגרמת הפיזור** של הנתונים, ביחידות המקוריות וביחידות המתוקננות. לשם כך, בוחרים את שתי עמודות הנתונים. מתוך תפריט **הוספה** (insert) בדף העבודה בוחרים **תרשים** (chart). מקבלים אפשרויות שונות של דיאגרמות. בוחרים את האפשרות **פיזור** (scatter) ומקבלים דיאגרמת פיזור שאפשר להתאים לצרכינו: לשנות את הסקלה של הצירים, להוסיף את שמות המשתנים, לשנות את תצוגת הנקודות בדיאגרמה ועוד ועוד.

הנה הפקודה לחישוב קל ומהיר של מקדם המתאם של פירסון; בסוגריים מצוין טווח התאים של ערכי שני המשתנים:

---

[10] למדריך בסיסי לשימוש באקסל ראו נספח אקסל בשערים "סטטיסטיקה תיאורית" ו"הסקה סטטיסטית" בסדרת הספרים הדיגיטלית **סטטיסטיקה בגובה העיניים**, בהוצאת כותר ספרי עיון (תלמה לויתן ואלונה רביב). השימוש בספר חופשי לתלמידים ולסטודנטים.

| | הסימון | פקודת אקסל |
|---|---|---|
| מקדם המתאם | $r$ | $=correl(\_:\_,\_:\_)$ |

## נספח ג: אתר הלשכה המרכזית לסטטיסטיקה

באתר הלמ"ס https://www.cbs.gov.il/he/pages/default.aspx ובפרסומי הלמ"ס יש שפע של נתונים מרתקים הנוגעים לישראל.

פנייה ישירה (על פי הכתובת info@cbs.gov.il) מניבה תגובה מהירה, מפורטת ומועילה.

שאלה על **גיל חתנים וכלות בישראל** ביום נישואיהם הניבה כמה לוחות רלוונטיים, ומהם בחרנו את לוח 8 – זוגות יהודים נישאים, לפי גיל החתן וגיל הכלה. דוגמה ממנו הבאנו כאן:

CBS, Marriage and Divorce in Israel 2018

למ"ס, נישואין וגירושין בישראל 2018

**TABLE 8. JEWISH COUPLES MARRYING, BY AGE OF THE GROOM AND THE BRIDE**

**לוח 8. זוגות יהודים נישאים, לפי גיל החתן וגיל הכלה**

Absolute numbers — **2018** — מספרים מוחלטים

| Age of groom | Age of bride: 24 | 23 | 22 | 21 | 20 | 19 | 18 | עד 17 Up to 17 | Grand total | גיל החתן |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | גיל הכלה | |
| **Total(1)** | **2,432** | **2,147** | **2,265** | **2,570** | **2,722** | **1,953** | **989** | **39** | **34,473** | **סך הכל(1)** |
| Up to 19 | 5 | *16* | *39* | *141* | *396* | *718* | *462* | 17 | *1,807* | *עד 19* |
| *Up to 17* | - | *1* | - | - | *1* | - | - | *1* | *3* | *עד 17* |
| 18 | - | 2 | 8 | 25 | 84 | 155 | 148 | 7 | 433 | 18 |
| 19 | 5 | 13 | 31 | 116 | 311 | 563 | 314 | 9 | 1,371 | 19 |
| 20 | 8 | 17 | 74 | 172 | 339 | 277 | 179 | 3 | 1,083 | 20 |
| 21 | 19 | 49 | 210 | 476 | 551 | 289 | 114 | 7 | 1,743 | 21 |

הקוראים מוזמנים להיכנס באתר ללוח 8 עצמו בקישור זה:

<https://www.cbs.gov.il/he/publications/doclib/2020/nisuim2017/t08-18.pdf>

**משימה אתגרית** (פתרון בעמ' 133)

בעמוד הבא מופיע תצלום של קובץ אקסל עם הנתונים המקוריים של לוח 8.

א. התבוננו במשתנים $X$ – גיל החתן, $Y$ – גיל הכלה. כמה זוגות של נתונים מוצגים בלוח זה? תנו דוגמה לזוג אחד של נתונים. האם נראה שדיאגרמת פיזור היא כלי יעיל לנתונים רבים כל כך? מהי הצעתכם?

ב. ציירו מערכת צירים מתאימה והציגו נקודות אחדות של דיאגרמת הפיזור.

ג. מבלי לחשב מקדם מתאם, חוו דעתכם על סוג הקשר – כיוונו ועוצמתו – בין המשתנים.

ד. התמקדו בחתנים בני 18 וחשבו את הגיל הממוצע של כלותיהם. מה משמעות תוצאות החישוב במונחי ניבוי?

ה. אם ממשיכים ומחשבים לכל אחד מגילי החתנים את ממוצע גיל הכלה המתאים ומציגים במערכת צירים את הזוגות (ממוצע גיל הכלה, גיל החתן), מה מתקבל?

ו. כמה ממוצעים כאלו יש לחשב כדי לדעת אם הנקודות בעקומה המתקבלת יוצרות קו ישר? חשבו ממוצעים אחדים.

ז. למומחים בתוכנת אקסל: ציירו את עקום הממוצעים בשלמותו, ואם תרצו ציירו גם את קו הרגרסיה.

**לוח 8. זוגות יהודים נישאים, לפי גיל החתן וגיל הכלה**

**TABLE 8. JEWISH COUPLES MARRYING, BY AGE OF THE GROOM AND THE BRIDE**

Absolute numbers **2018** מספרים מוחלטים

| Age of groom | Age of bride | | | | | | | | גיל הכלה | גיל החתן |
|---|---|---|---|---|---|---|---|---|---|---|
| | 24 | 23 | 22 | 21 | 20 | 19 | 18 | 17 עד Up to 17 | Grand total | |
| **Total(1)** | **2,432** | **2,147** | **2,265** | **2,570** | **2,722** | **1,953** | **989** | **39** | **34,473** | **סך הכל(1)** |
| Up to 19 | 5 | *16* | *39* | *141* | *396* | *718* | *462* | 17 | *1,807* | *עד 19* |
| *Up to 17* | - | *1* | - | - | *1* | - | - | *1* | *3* | *עד 17* |
| 18 | - | 2 | 8 | 25 | 84 | 155 | 148 | 7 | 433 | 18 |
| 19 | 5 | 13 | 31 | 116 | 311 | 563 | 314 | 9 | 1,371 | 19 |
| 20 | 8 | 17 | 74 | 172 | 339 | 277 | 179 | 3 | 1,083 | 20 |
| 21 | 19 | 49 | 210 | 476 | 551 | 289 | 114 | 7 | 1,743 | 21 |
| 22 | 59 | 180 | 449 | 604 | 577 | 259 | 75 | 2 | 2,256 | 22 |
| 23 | 152 | 316 | 383 | 413 | 325 | 147 | 37 | 4 | 1,914 | 23 |
| 24 | 343 | 379 | 302 | 239 | 177 | 72 | 30 | 3 | 1,851 | 24 |
| 25 | 405 | 294 | 266 | 168 | 88 | 49 | 14 | - | 2,032 | 25 |
| 26 | 404 | 271 | 138 | 109 | 54 | 44 | 16 | - | 2,317 | 26 |
| 27 | 310 | 202 | 127 | 67 | 38 | 14 | 10 | 1 | 2,566 | 27 |
| 28 | 228 | 125 | 67 | 39 | 27 | 14 | 8 | - | 2,483 | 28 |
| 29 | 136 | 108 | 41 | 26 | 25 | 6 | 5 | - | 2,335 | 29 |
| *30–34* | *293* | *143* | *103* | *45* | *28* | *10* | *8* | - | *6,702* | *34-30* |
| *35–39* | *41* | *18* | *16* | *9* | *8* | *3* | - | - | *2,242* | *39-35* |
| *40–44* | *10* | *5* | *1* | *3* | *1* | *1* | - | - | *1,051* | *44-40* |
| *45–49* | *1* | - | *1* | *1* | *2* | *2* | - | - | *538* | *49-45* |
| *50+* | *1* | - | - | *1* | - | - | - | - | *1,139* | *+50* |

**לוח 8. זוגות יהודים נישאים, לפי גיל החתן וגיל הכלה (המשך)**

**TABLE 8. JEWISH COUPLES MARRYING, BY AGE OF THE GROOM AND THE BRIDE (Cont.)**

Absolute numbers **2018** מספרים מוחלטים

| Age of groom | Age of Bride | | | | | | | | גיל הכלה | גיל החתן |
|---|---|---|---|---|---|---|---|---|---|---|
| | 50+ | 45–49 | 40–44 | 35–39 | 30–34 | 29 | 28 | 27 | 26 | 25 |
| **Total(1)** | **701** | **364** | **631** | **1,487** | **4,250** | **1,703** | **2,099** | **2,558** | **2,857** | **2,530** | **סך הכל(1)** |
| *Up to 19* | - | - | - | - | - | *1* | - | *3* | *2* | - | *עד 19* |
| *Up to 17* | - | - | - | - | - | - | - | - | - | - | *עד 17* |
| 18 | - | - | - | - | - | 1 | - | 2 | - | - | 18 |
| 19 | - | - | - | - | - | - | - | 1 | 2 | - | 19 |
| 20 | - | - | - | 1 | 1 | - | - | - | 1 | 2 | 20 |
| 21 | - | 1 | - | - | 2 | 2 | 1 | 4 | 2 | 9 | 21 |
| 22 | - | - | - | - | 1 | 2 | 2 | 2 | 10 | 20 | 22 |
| 23 | - | - | - | 2 | 10 | 6 | 4 | 15 | 24 | 65 | 23 |
| 24 | - | - | - | 2 | 16 | 12 | 13 | 25 | 60 | 172 | 24 |
| 25 | - | - | 2 | 3 | 26 | 17 | 39 | 76 | 203 | 376 | 25 |
| 26 | - | - | 2 | 1 | 54 | 40 | 72 | 173 | 490 | 444 | 26 |
| 27 | - | - | 1 | 8 | 86 | 80 | 173 | 464 | 543 | 437 | 27 |
| 28 | - | - | 1 | 15 | 155 | 157 | 383 | 493 | 467 | 296 | 28 |
| 29 | - | 1 | 1 | 23 | 280 | 267 | 395 | 447 | 343 | 223 | 29 |
| *30–34* | *2* | *4* | *14* | *270* | *2,280* | *885* | *853* | *731* | *602* | *396* | *34-30* |
| *35–39* | *1* | *8* | *77* | *555* | *944* | *180* | *125* | *92* | *83* | *58* | *39-35* |
| *40–44* | *10* | *39* | *217* | *389* | *272* | *32* | *22* | *17* | *14* | *11* | *44-40* |
| *45–49* | *24* | *87* | *174* | *142* | *69* | *12* | *5* | *5* | *3* | *4* | *49-45* |
| *50+* | *656* | *220* | *134* | *71* | *29* | *1* | *3* | *5* | *1* | - | +50 |

## נספח ד: תכונות מקדם המתאם – הוכחות

בסעיף העוסק בטיב הניבוי של קו הרגרסיה ( סעיף 5.4, עמ' 74) הגדרנו מדד לטיב הניבוי של עקומת ניבוי כלשהי – **הטעות הריבועית הממוצעת**, שהיא ממוצע ריבועי הסטיות בין הערכים בפועל לבין הערכים המתאימים שנובאו. נזכיר שהצבת ערכי הניבוי על פי קו הרגרסיה נתנה את הטעות הריבועית הממוצעת של קו הרגרסיה:

**הטעות הריבועית הממוצעת של קו הרגרסיה** היא

$$\sigma_y^2 \cdot (1-r^2) \qquad (7)$$

**מסקנות: תכונות מקדם המתאם**

מקֻשָּׁר (7) אפשר להסיק בנקל תכונות חשובות של מקדם המתאם שהזכרנו בפרק 3 ללא הוכחה.

- מכיוון שהטעות הריבועית – שהיא ממוצע של ריבועים – איננה יכולה להיות שלילית, הרי $1-r^2 \geq 0$, או $r^2 \leq 1$. כלומר, $-1 \leq r \leq 1$.

♥ מקדם המתאם מקבל רק ערכים שבין (1–) ל-1+.

מקֻשָּׁר (7) ברור גם שככל ש-$r^2$ קרוב יותר ל-1, כך **קטֵנה** הטעות בניבוי על פי הקו. נבחן ערכי קיצון:

הערך **המינימלי** של $\sigma_y^2 \cdot (1-r^2)$ מתקבל כשמקדם המתאם הוא 1+ או $(-1)$. נשים לב שהטעות היא מינימלית (0) כאשר יש אפס סטיות מהקו, כלומר **כאשר לכל** ערך $x_i$, הניבוי $\hat{y}_i$ שווה בדיוק לערך הנכון $y_i$. במילים אחרות, הטעות היא מינימלית **כאשר כל** הנקודות בדיאגרמת הפיזור נמצאות על קו ישר, הוא קו הרגרסיה.

♥ **מקדם המתאם הוא 1+ או (1–) אך ורק כאשר כל הנקודות בדיאגרמת הפיזור נמצאות על קו ישר.**

הערך **המקסימלי** של הטעות – כלומר המצב שבו הניבוי בעזרת קו הרגרסיה הוא הגרוע ביותר, מתקבל כאשר $r = 0$. נזכיר שבמצב כזה גם השיפוע הוא אפס וקו הניבוי הוא $\hat{y}_x = \bar{y}$ (הניבוי הוא קבוע ושווה לממוצע ערכי $y$), ולכן ידיעת הערך של $x$ כלל אינה רלוונטית לניבוי ואין תועלת במשתנה המנבא לניבוי הקווי.

מכל אלו הסקנו כי:

♥ מקדם המתאם $r$ מודד את טיב הקשר הקווי בין שני משתנים:

– ככל שמקדם המתאם קרוב יותר ל-1 או ל-1–, כך הניבוי באמצעות קו הרגרסיה טוב יותר.

– ככל שמקדם המתאם קרוב יותר ל-0, כך הניבוי באמצעות קו הרגרסיה גרוע יותר.

# פתרונות למשימות

(ההפניות הן לנוסחאות החישוביות בעמ' 97-96)

## פתרונות למשימות

### פרק 1. ציוני תקן

#### משימה I

**א.** תקנון הציונים של שני המועמדים:

- מועמד ראשון: $\frac{x_A-\bar{x}}{\sigma_A}=\frac{84-80}{7.5}=0.5$
- מועמד שני: $\frac{x_B-\bar{x}}{\sigma_B}=\frac{85-80}{5.0}=1.0$

**ב.** בהשוואה לציונים האחרים באוניברסיטה שלמד בה (B), המועמד השני הצטיין יותר.

**ג.** ציונים מתוקננים ומקוריים של בוגרי שתי האוניברסיטאות:

| | | | | |
|---|---|---|---|---|
| ביחידות מתוקננות | 0 | 1 | 2 | 3 |
| ביחידות מקוריות<br>אוניברסיטה A | 80 | 85 | 90 | 95 |
| ביחידות מקוריות<br>אוניברסיטה B | 80 | 87.5 | 95 | 102.5 |

נשים לב שהציונים המתוקננים כמעט שאינם חורגים מעל לערך 3.

מתברר שבתופעות רבות **ציוני התקן אינם חורגים מעל 3 ומתחת ל- (3–).**

#### משימה II. ציון פרויקט וציון תעודה

**א.** ייתכן שהפרויקט עזר לתלמידים 4, 5, 6, 8, אך אין דרך לדעת. לתלמידים 1, 2, 3, 7, 10 הפרויקט לא עזר כלל.

**ב.** חישוב ציוני התקן של המשתנים:

| התלמיד | ציון תקן פרויקט $\frac{x-71}{22.78}$ | ציון תקן תעודה $\frac{y-82}{13.27}$ | כיוון יחסית לממוצע | איזה ציון חריג יותר בהשוואה לשאר התלמידים |
|---|---|---|---|---|
| 1 | -1.80 | -1.66 | שניהם מתחת לממוצע | פרויקט |
| 2 | -0.92 | 0.60 | פרויקט מתחת תעודה מעל | פרויקט |

| | | | | |
|---|---|---|---|---|
| 3 | -1.36 | -1.28 | שניהם מתחת לממוצע | פרויקט |
| 4 | 1.05 | 0.98 | שניהם מעל לממוצע | פרויקט |
| 5 | 1.27 | 1.36 | שניהם מעל לממוצע | תעודה |
| 6 | 0.83 | 0.60 | שניהם מעל לממוצע | פרויקט |
| 7 | 0.18 | -0.15 | פרויקט מעל תעודה מתחת | פרויקט |
| 8 | -0.26 | -1.28 | שניהם מתחת לממוצע | תעודה |
| 9 | 0.61 | 0.23 | שניהם מעל לממוצע | פרויקט |
| 10 | 0.40 | 0.60 | שניהם מעל לממוצע | תעודה |

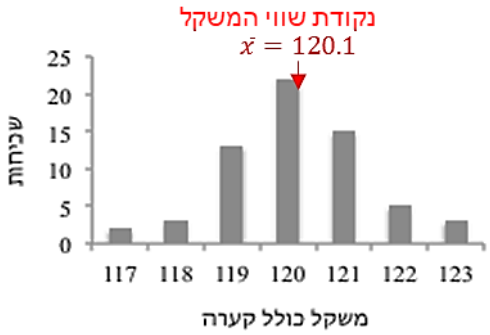
**ג.** כצפוי, ממוצע ציוני התקן הוא 0 והשונות היא 1.

**ד.** הכפלת הערכים בקבוע 1/10 תכפיל באותו קבוע את ממוצע הפרויקט, את הסטיות מהממוצע, וכן את סטיית התקן. ציוני התקן המתאימים יישארו זהים, מכיוון ששינוי ליניארי במשתנה אינו משנה את ציוני התקן.

**ה.** בחינת ציוני התקן של תלמיד מספר 8 מורה שאף שציון הפרויקט שלו זהה לציון התעודה, מצבו בהשוואה לכיתה גרוע יותר בתעודה מאשר בפרויקט.

**משימה III. הממוצע, סטיית התקן, ציוני התקן – תכונות**

**א.** ממוצע התפלגות המשקל (בגרמים) של 60 חפיסות שוקולד, כולל משקל הקערה:




**ב.** ממוצע המשקל נטו הוא 100.1 גרם (20 גרם פחות מממוצע המשקל עם הקערה).

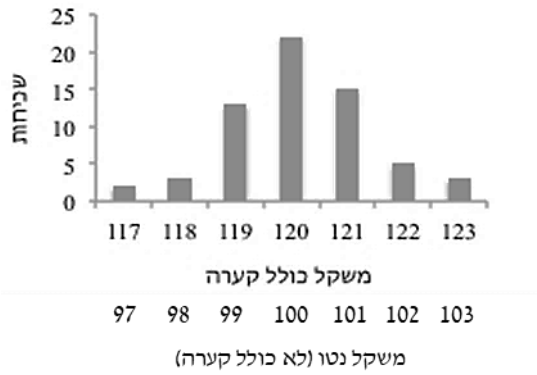
- הוספה או החסרה של קבוע לערכי משתנה תביא להוספה של אותו קבוע לממוצע או להחסרתו מהממוצע, בהתאמה.

**ג.** השונות ברוטו זהה לשונות נטו.

- הוספה או החסרה של קבוע למשתנים אינה משנה את השונות ולכן גם אינה משנה את סטיית התקן.

הסבר: ריבועי הסטיות מהממוצע של ערכי המשתנה לא השתנו, ולכן גם סך ריבועי הסטיות לא השתנה.

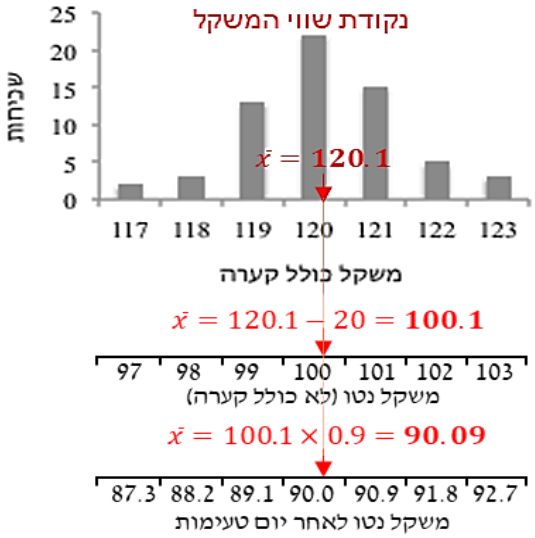
**ד.** התפלגות המשקל נטו בסקלה התחתונה באיור. נקודת שיווי המשקל היא 100.1 (סמנו).




**ה.** המשקל הממוצע של החפיסות לאחר יום הטעימות הראשון הוא 90.09 גרם, ולאחר יום הטעימות השני 81.08 גרם.

- כשמכפילים או מחלקים את ערכי המשתנה בקבוע, הממוצע יוכפל או יחולק (בהתאמה) באותו קבוע.

**ו.** דיאגרמת העמודות של המשקל נטו לאחר יום טעימות: זוהי אותה דיאגרמה, אלא שסקלת המשקל שבציר האופקי משתנה:




**סקלת המשקל נטו**

**סקלת המשקל לאחר יום אחד של טעימות**

**ז.** סטיית התקן בסוף כל יום קטנה פי 0.9.

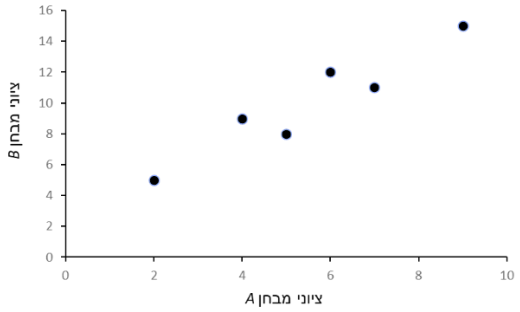
✋ בהכפלה או בחלוקה של ערכי המשתנה בקבוע, סטיית התקן תוכפל או תחולק באותו קבוע.

**ח.** ציוני התקן אינם משתנים בעקבות שינוי ליניארי בערכים – המיקום היחסי בהתפלגות נשמר.

## פרק 2. ניתוח גרפי

### משימה I. עברית שפה קשה

**א.** דיאגרמת פיזור של ציוני מבחן A (מספר מלים **נכונות** בהכתבה) ומבחן B (מספר מלים **שגויות** שאותרו ברשימה).

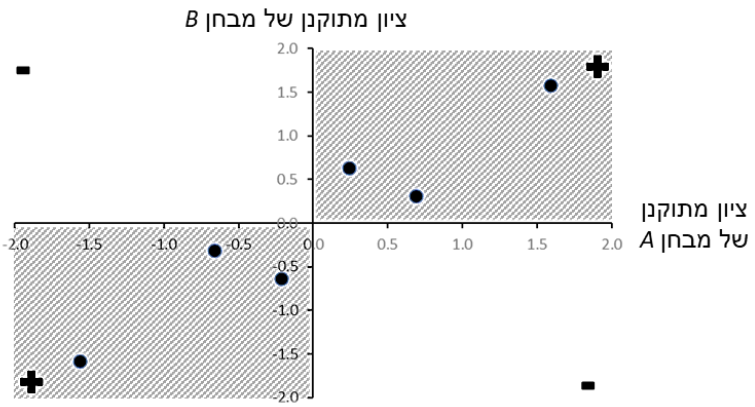



על פי הדיאגרמה, נראה שיש קשר **עולה** ברור בין ציון מבחן A לציון מבחן B.

**ב–ג.** ממוצע וסטיית התקן של ציוני מבחן A ומבחן B מודגשים בתחתית הטבלה להלן. הציונים המתוקננים למבחן A ולמבחן B מוצגים בעמודות האפורות.

| התלמיד | מבחן A, $x$ | מבחן B, $y$ | מבחן A, ציון תקן | מבחן B, ציון תקן | סימן המכפלה |
|---|---|---|---|---|---|
| א | 2 | 5 | -1.58 | -1.58 | + |
| ב | 5 | 8 | -0.23 | -0.63 | + |
| ג | 9 | 15 | 1.58 | 1.58 | + |
| ד | 6 | 12 | 0.23 | 0.63 | + |
| ה | 7 | 11 | 0.68 | 0.32 | + |
| ו | 4 | 9 | -0.68 | -0.32 | + |
| **ממוצע** | **5.5** | **10** | 0 | 0 | |
| **סטיית תקן** | **2.22** | **3.16** | 1.00 | 1.00 | |

**ד.** להלן דיאגרמת הפיזור של הציונים המתוקננים למבחן A ולמבחן B, כולל סימני הרביעים.



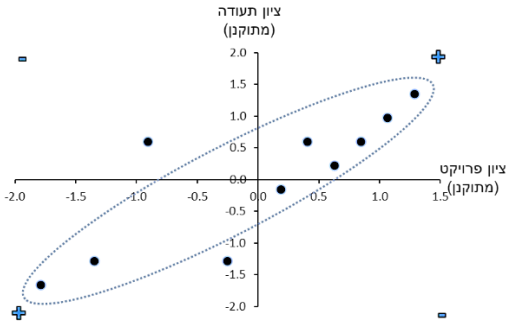

בשני הרביעים האפורים, הציונים של אותו תלמיד שניהם מעל לממוצע או שניהם מתחת לממוצע. לכן הסימנים של שני ציוני התקן זהים: שניהם חיוביים או שניהם שליליים. מכאן, בשני הרביעים האפורים מכפלת ציוני התקן חיובית – אלו הם **הרביעים החיוביים**. בשני הרביעים הלבנים המכפלה שלילית (הסבירו) – אלו הם **הרביעים השליליים**.

נשים לב שברביעים השליליים בדיאגרמת הפיזור אין נקודות, וכל הנקודות נמצאות ברביעים החיוביים.

**משימה II. ציון פרויקט וציון תעודה**

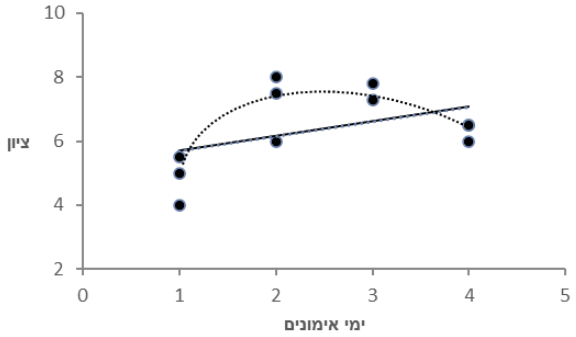
**א.** בדיאגרמת הפיזור של הציונים המתוקננים המוצגת בהמשך, קווי הממוצעים מתלכדים עם הצירים. ציון התקן של כל הערכים שהיו מעל לממוצע הוא חיובי, וציון התקן של כל הערכים שהיו מתחת לממוצע הוא שלילי. ברביעים החיוביים שני ציוני התקן הם בעלי אותו סימן, ולכן המכפלה שלהם חיובית. ברביעים השליליים ציון תקן אחד הוא חיובי והאחר שלילי, ולכן המכפלה שלהם שלילית.

**ב.** מדיאגרמת הפיזור נראה שהקשר בין הציונים הוא קשר עולה. הנקודות מסתדרות היטב לאורך קו ישר ונמצאות ברובן ברביעים החיוביים. לכל תלמיד, ככל שציון הפרויקט שלו גבוה יותר כך הציון הסופי בתעודה נוטה להיות גבוה. כיוונה של האליפסה שבאמצעותה הקפנו את מרבית הנקודות בדיאגרמה עולה משמאל לימין וכמו כן היא **צרה מאוד**. מכאן נסיק כי הקשר הוא **עולה וחזק.**




**משימה III. חמודי הסבות מקצועיות**

**א.** להלן דיאגרמת פיזור של שני המשתנים $X$ ו-$Y$ (הערכים רשומים בטורים האפורים בטבלה שבהמשך).



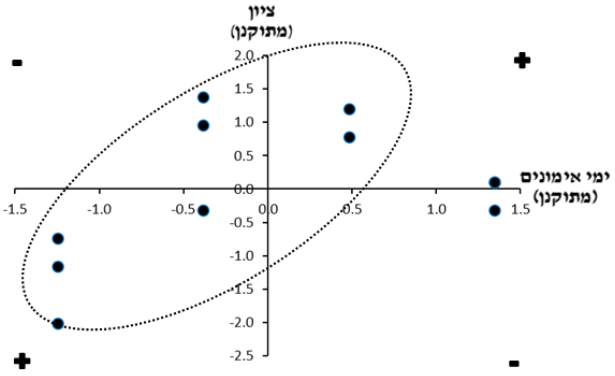

**ב.** נראה שהנקודות מסתדרות טוב יותר לאורך העקומה שתיארנו באיור מאשר לאורך הקו הישר, כלומר הקשר בין המשתנים **אינו קווי**.

**ג.** הממוצעים וסטיות התקן רשומים בתחתית הטבלה. ציוני התקן מופיעים בעמודות האפורות.

| $\tilde{y}=\frac{y-\bar{y}}{\sigma_Y}$ | $\tilde{x}=\frac{x-\bar{x}}{\sigma_X}$ | **ציון,** $y$ | **ימי אימונים,** $x$ | |
|---|---|---|---|---|
| -1.903 | -1.182 | 4 | 1 | |
| -1.096 | -1.182 | 5 | 1 | |
| -0.693 | -1.182 | 5.5 | 1 | |
| -0.290 | -0.273 | 6 | 2 | |
| 0.919 | -0.273 | 7.5 | 2 | |
| 1.322 | -0.273 | 8 | 2 | |
| 1.161 | 0.636 | 7.8 | 3 | |
| 0.758 | 0.636 | 7.3 | 3 | |
| 0.113 | 1.545 | 6.5 | 4 | |
| -0.290 | 1.545 | 6 | 4 | |
| **0.00** | **0.00** | **6.36** | **2.3** | **ממוצעים** |
| **1.00** | **1.00** | **1.24** | **1.10** | **סטיות תקן** |

**ד–ה.** בדיאגרמת הפיזור המתוקננת של ימי האימונים ושל הציונים, קווי הממוצעים מתלכדים עם הצירים. רוב הנקודות נמצאות ברביעים החיוביים.

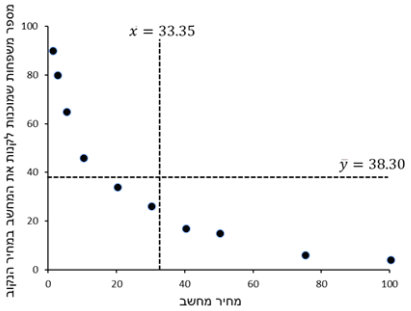
האליפסה המקיפה את הנקודות בדיאגרמה היא רחבה, והיא מקיפה את מרבית הנקודות אך לא את כולן. מכאן נסיק כי הקשר הוא בינוני-חלש.



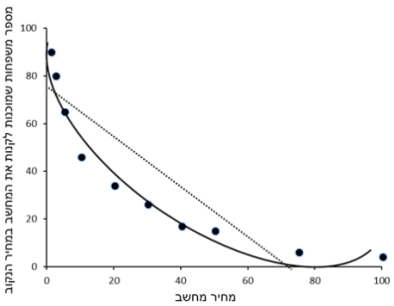

**משימה IV. מחירי מחשב ביתי**

**א–ב.** מחיר ממוצע של מחשב במאות דולרים: $\bar{x} = 33.35$. מתוך 100 בתי האב שהשתתפו, המספר הממוצע של בתי אב שאמרו שירכשו מחשב במחיר שהוצע להם: $\bar{y} = 38.30$. באיור מימין (א) דיאגרמת הפיזור עם קווי הממוצעים, משמאל (ב) קו ניבוי לעומת עקומת ניבוי מתאימה יותר.

א. דיאגרמת פיזור עם קווי ממוצעים




ב. דיאגרמת פיזור עם קווי ניבוי



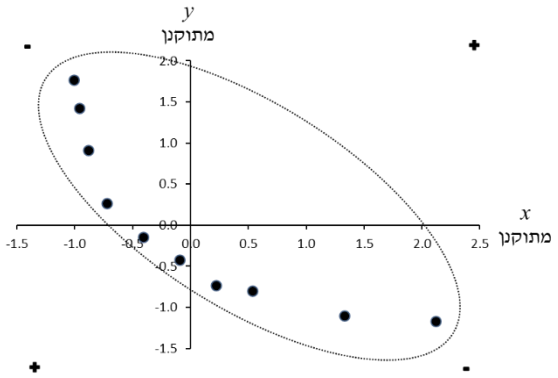

**ג.** סטיות התקן הן: $\sigma_X = 31.68$, $\sigma_Y = 29.27$.

ערכי המשתנים ביחידות מתוקננות $\tilde{x} = \frac{x - 33.35}{31.68}$, $\tilde{y} = \frac{y - 38.3}{29.27}$ רשומים בלוח זה:

| $x$ | 1 | 2.5 | 5 | 10 | 20 | 30 | 40 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 90 | 80 | 65 | 46 | 34 | 26 | 17 | 15 | 6 | 4 |
| $\tilde{x}$ | -1.02 | -0.97 | -0.89 | -0.74 | -0.42 | -0.11 | 0.21 | 0.53 | 1.31 | 2.10 |
| $\tilde{y}$ | 1.77 | 1.42 | 0.91 | 0.26 | -0.15 | -0.42 | -0.73 | -0.80 | -1.10 | -1.17 |

**ד.** בדיאגרמה המתוקננת שלהלן, קווי הממוצעים מתלכדים עם הצירים. כפי שאפשר לראות, רוב הנקודות נמצאות ברביעים השליליים. פירוש הדבר שהקשר יורד.

**ה.** האליפסה יורדת משמאל לימין – הקשר יורד. האליפסה איננה צרה, ולכן עוצמת הקשר **בינונית**.




**ו.** יש לבצע הזזה שמאלה ולמטה של הנקודות בדיאגרמה (א) הימנית, באופן שקווי הממוצעים יתלכדו עם הצירים החדשים.

## פרק 3. מתאם בין משתנים

### משימה I. ערכי קיצון

**א.** ערכי y (מרחק במייל) על פי הקשר $y = x/1.6$ רשומים בטבלה:

| מרחק בקילומטר, $x$ | מרחק במייל, $y$ | ציון תקן, $\tilde{x}$ | ציון תקן, $\tilde{y}$ | מכפלת ציוני התקן, $\tilde{x}\cdot\tilde{y}$ |
|---|---|---|---|---|
| 2 | 1.25 | -0.32 | -0.32 | 0.10 |
| 1.6 | 1 | -0.47 | -0.47 | 0.22 |
| 1 | 0.625 | -0.70 | -0.70 | 0.49 |
| 0.8 | 0.5 | -0.77 | -0.77 | 0.59 |
| 3.2 | 2 | 0.12 | 0.12 | 0.02 |
| $\bar{x} = 1.72$ | $\bar{y} = 1.08$ | | | $r = 1.00$ |

**ממוצעים**

**ב.** חישוב ידני נותן $\bar{x} = 1.72$ קילומטר. על כן, מתכונות הממוצע $\bar{y} = 1.72/1.6 = 1.08$ מייל.

**ג.** חישוב ידני נותן $\sigma_X = 0.85$ קילומטר. על כן, מתכונות סטיית התקן $\sigma_Y = 0.85/1.6 = 0.53$ מייל.

**ד.** ציוני התקן של שני המשתנים זהים מכיוון שמדובר בשינוי ליניארי: המשתנה $Y$ מתקבל מחלוקה של המשתנה $X$ בקבוע. כזכור, ציון התקן אינו תלוי ביחידת המדידה של המשתנה.

**ה.** כפי שאפשר לראות, מקדם המתאם בין שני המשתנים הוא 1. שני המשתנים מודדים בדיוק את אותה תכונה אך בקנה מידה אחר, ולכן ההתאמה הקווית ביניהם מושלמת.

**באופן כללי**: מקדם המתאם בין $X$ ל- $a+bX$ (שינוי ליניארי) הוא 1 עבור $b>0$.

**משימה II. ציון פרויקט וציון תעודה**

**א.** בתחתית עמודת מכפלת הציונים $x\cdot y$ בטבלה שלהלן הוספנו את ממוצע המכפלות. כזכור, סטיות התקן הן $\sigma_x = 22.78$ ו- $\sigma_y = 13.27$.

| ציון פרויקט, $x$ | ציון תעודה, $y$ | מכפלת הציונים, $x\cdot y$ |
|---|---|---|
| 30 | 60 | 1800 |
| 50 | 90 | 4500 |
| 40 | 65 | 2600 |
| 95 | 95 | 9025 |
| 100 | 100 | 10000 |
| 90 | 90 | 8100 |
| 75 | 80 | 6000 |
| 65 | 65 | 4225 |
| 85 | 85 | 7225 |
| 80 | 90 | 7200 |
| $\bar{x}=71$ | $\bar{y}=82$ | $\frac{1}{10}\cdot[x_1\cdot y_1+\ \ldots\ +x_{10}\cdot y_{10}]=6067.5$ |

**ממוצע**

**ב.** חישוב מקדם המתאם: מנוסחת החישוב (3) (עמ' 94) נקבל:

$$r=\frac{1}{\sigma_x\cdot\sigma_y}\left\{\frac{1}{10}\cdot[x_1\cdot y_1+\cdots+x_{10}\cdot y_{10}]-\bar{x}\cdot\bar{y}\right\}=\frac{6067.5-71\cdot 82}{22.78\cdot 13.27}=0.812$$

בעמ' 41 חישבנו את מקדם המתאם על פי הגדרה (1) (עמ' 40) וקיבלנו אכן $r=0.8121$.

**משימה III. חמודי הסבות מקצועיות**

**א–ב.** בטבלה הבאה השלמנו את העמודות: מכפלות המשתנים (עמודה שלישית) ומכפלות ציוני התקן (עמודה אחרונה), וכן חישבנו ממוצעים מתאימים (בתחתית כל עמודה).

| ימי אימונים, $x$ | ניקוד, $y$ | $x\cdot y$ | $\tilde{x}\cdot\tilde{y}$ |
|---|---|---|---|
| 1 | 4 | 4 | 2.25 |
| 1 | 5 | 5 | 1.30 |

| | | | |
|---|---|---|---|
| 0.82 | 5.5 | 5.5 | 1 |
| 0.08 | 12 | 6 | 2 |
| -0.25 | 15 | 7.5 | 2 |
| -0.36 | 16 | 8 | 2 |
| 0.74 | 23.4 | 7.8 | 3 |
| 0.48 | 21.9 | 7.3 | 3 |
| 0.17 | 26 | 6.5 | 4 |
| -0.45 | 24 | 6 | 4 |
| $r = 0.478$ | $\frac{1}{10} \cdot [x_1 \cdot y_1 + \ \dots \ + x_{10} \cdot y_{10}] = 15.28$ | $\bar{y} = 6.36$ | $\bar{x} = 2.3$ |

**ממוצעים**

חישוב מקדם המתאם על פי נוסחה (1) – ממוצע מכפלת ציוני התקן (עמ' 40):

$$r = \frac{1}{10} \cdot \left[ \underset{\sim}{x_1} \cdot \underset{\sim}{y_1} + \cdots + \underset{\sim}{x_{10}} \cdot \underset{\sim}{y_{10}} \right] = \frac{1}{10} \cdot 4.78 = 0.478$$

נזכיר, סטיות התקן הן $\sigma_x = 1.1$ ו- $\sigma_y = 1.24$. כמו כן, ממוצע מכפלות הנתונים המקוריים רשום בתחתית העמודה השלישית. על פי נוסחה (3) מקדם המתאם הוא:

$$r = \frac{1}{\sigma_x \cdot \sigma_y} \left\{ \frac{1}{10} \cdot [x_1 \cdot y_1 + \cdots + x_{10} \cdot y_{10}] - \bar{x} \cdot \bar{y} \right\} = \frac{15.28 - 2.3 \cdot 6.36}{1.10 \cdot 1.24} = 0.478$$

אכן, התקבלה אותה תוצאה באמצעות שתי הנוסחאות.

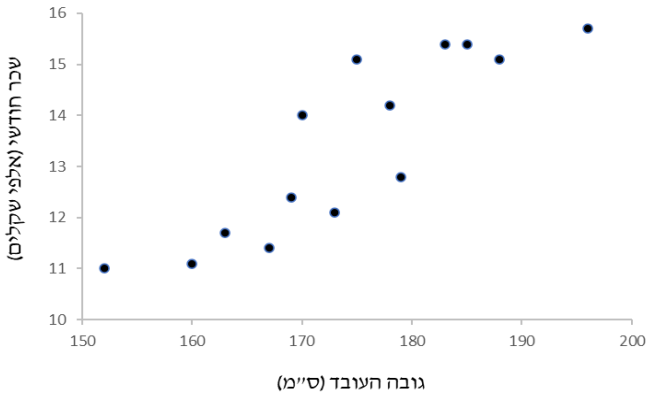
**ג.** הכפלה של מספר הימים בקבוע 8, שהוא מספר שעות האימון היומי, אינה משנה את מקדם המתאם. מקדם המתאם אינו מושפע מיחידות המדידה של המשתנים (ראו "אז מה היה לנו?" פרק 3, עמ' 47).

**משימה IV. האם שכר עובדים ממריא עם גובה העובד?**

**חישובים**: בעזרת מחשבון התקבלו הממוצעים $\bar{x} = 174.14$, $\bar{y} = 13.39$, וסטיות התקן $\sigma_x = 11.39$, $\sigma_y = 1.71$. מהטבלה חישבנו גם את מכפלות ערכי המשתנים ואת ממוצע המכפלות:

$\frac{1}{14} \cdot [x_1 \cdot y_1 + \cdots + x_{14} \cdot y_{14}] = 2348.02$.

**א.** להלן דיאגרמת פיזור של השכר לפי הגובה:



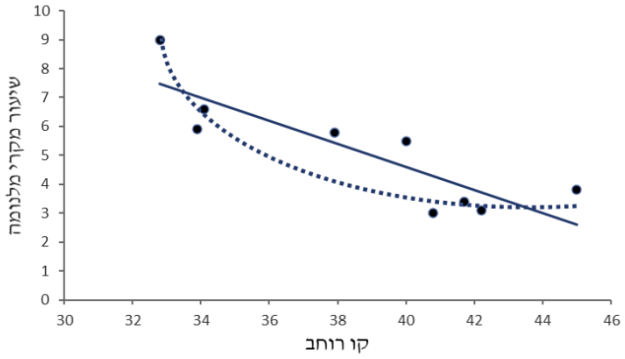

**ב.** מקדם המתאם בין הגובה לשכר:

$$r = \frac{1}{\sigma_x \cdot \sigma_y}\left\{\frac{1}{14}\cdot[x_1 \cdot y_1 + \cdots + x_{14} \cdot y_{14}] - \bar{x}\cdot\bar{y}\right\} = \frac{2348.02 - 174.14 \cdot 13.39}{11.39 \cdot 1.71} = 0.871$$

**ג.** מדידת הגובה באינצ׳ים (1 אינץ׳ = 2.54 ס״מ) והשכר בדולרים הם שינויים ליניאריים במשתנים, ולכן מקדם המתאם לא ישתנה ($r = 0.871$).

**משימה V. קרינה אולטרה סגולה וסרטן עור**

**א.** להלן דיאגרמת פיזור של שיעור מקרי המלנומה ($y$) לפי קו רוחב ($x$), שמצוינים בה קו ניבוי ישר ועקומת ניבוי שאינה קווית (העקומה המקווקות). על פניו לא ברור מי מהם מתאים יותר.




**ב.** בטבלה שבהמשך הוספנו עמודה של מכפלת המשתנים (עמודה שמאלית) ואת ממוצע המכפלות (בתחתית העמודה).

- מהנתונים שבידינו קיבלנו גם את השונויות על פי נוסחה (0), ומהן חישבנו את סטיות התקן של המשתנים (בתחתית הטבלה). לטבלה הוספנו גם עמודת מכפלות ערכי המשתנים.

| $x \cdot y$ | $y$ | $X$ |
|---|---|---|

| | | |
|---|---|---|
| 295.20 | 9.0 | 32.8 |
| 200.01 | 5.9 | 33.9 |
| 225.06 | 6.6 | 34.1 |
| 219.82 | 5.8 | 37.9 |
| 220.00 | 5.5 | 40.0 |
| 122.40 | 3.0 | 40.8 |
| 141.78 | 3.4 | 41.7 |
| 130.82 | 3.1 | 42.2 |
| 171.00 | 3.8 | 45.0 |
| **191.79** | **5.12** | **38.71** |
| | $\sigma_y$ = **1.88** | $\sigma_x$ = **4.04** |

**ממוצעים**

**סטיות התקן**

חישוב מקדם המתאם בין $X$ ל-$Y$ :

$$r = \frac{1}{\sigma_x \cdot \sigma_y}\left\{\frac{1}{9}\cdot[x_1\cdot y_1 + \cdots + x_9\cdot y_9] - \bar{x}\cdot\bar{y}\right\} = \frac{191.79 - 38.71\cdot 5.12}{4.04\cdot 1.88} = -\mathbf{0.857}$$

**משימה VI. עברית שפה קשה**

**א.** בעמודה הימנית בטבלה להלן השלמנו את העמודה החסרה בלוח שבעמ' 52. משמאלה נוסיף עמודות מתאימות למשתנה $X^*$ – מספר השגיאות, על פי $x^* = 10 - x$ .

| $(x-\bar{x})(y-\bar{y})$ | **שגיאות, $x^*$** | **מבחן B, $y$** | $x^*-\bar{x}^*$ | $y-\bar{y}$ | $(x^*-\bar{x}^*)(y-\bar{y})$ |
|---|---|---|---|---|---|
| 17.5 | 8 | 5 | 3.5 | -5 | -17.5 |
| 1.0 | 5 | 8 | 0.5 | -2 | -1.0 |
| 17.5 | 1 | 15 | -3.5 | 5 | -17.5 |
| 1.0 | 4 | 12 | -0.5 | 2 | -1.0 |
| 1.5 | 3 | 11 | -1.5 | 1 | -1.5 |
| 1.5 | 6 | 9 | 1.5 | -1 | -1.5 |
| 6.67 | 5 .5 | 10.0 | 0 | 0 | -6.67 |

**ממוצעים**

מהעמודה הימנית ומנוסחה (2) מתקבל מקדם המתאם בין $X$ ל-$Y$ : $r = r(X,Y) = \frac{6.67}{(2.22)\cdot(3.16)} = 0.95$ .

**ב.** בעזרת טבלת ציוני התקן בעמ' 107, חישבנו את ממוצע המכפלות שבעמודות האפורות וקיבלנו אכן $r = 0.95$ (בדקו בעזרת מחשבון).

**ג.** מתכונות הממוצע וסטיית התקן $\bar{x}^* = 10 - \bar{x} = 4.5$ , $\sigma_{X^*} = \sigma_X = 2.22$ .

**ד.** נשים לב שהערכים בעמודה $x^* - \bar{x}^*$ זהים לערכים בעמודה $x - \bar{X}$ , אלא ש**סימניהם הפוכים** (הסבירו למה). ומכאן – גם סימניהם של הערכים בעמודת המכפלות הפוכים מסימני הערכים של המכפלות בעמודה הימנית. שימוש בנוסחה (2) נותן אפוא $r^* = r(X^*, Y) = \frac{-6.67}{(2.22)(3.16)} = -0.95$ . ערך מקדם המתאם זהה לקודם, אך סימנו הפוך. את כל זאת היה אפשר לראות מלכתחילה ללא החישובים המדוקדקים. נמקו.

**משימה VII. פשיעה בדטרויט**

מקדמי המתאם חושבו בתוכנת אקסל. הנתונים הוקלדו בשלוש עמודות מקבילות: נתוני $Y$ בתאים A2 עד A14, נתוני $X_1$ בתאים B2 עד B14, **ונתוני** $X_2$ בתאים C2 עד C14.

פקודת אקסל: $=correl(A2:A14, B2:B14)$ נתנה את התוצאה: $r_{X_1,Y} = 0.964$ .

פקודת אקסל: $=correl(A2:A14, C2:C14)$ נתנה את התוצאה: $r_{X_2,Y} = 0.913$ .

שני המתאמים גבוהים מאוד, ולכן כל אחד מהמשתנים המסבירים יעיל לניבוי שיעור מקרי הרצח בעיר. הנתונים עשויים להיות לא רלוונטיים לצורכי ניבוי אחרי שנים רבות כל כך.

**משימה VIII. מבחני פיז"ה**

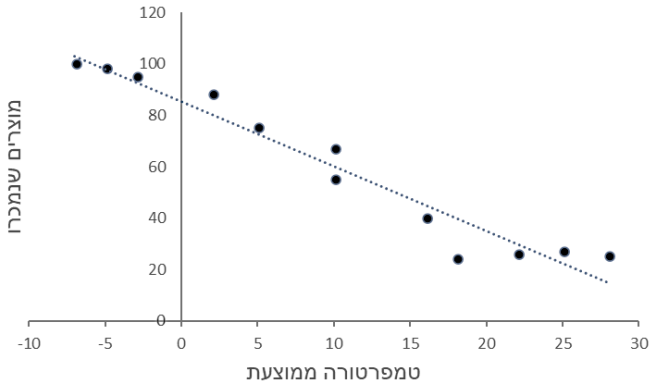
**א.** מקדמי המתאם חושבו בתוכנת אקסל. הנתונים הוקלדו בשתי עמודות מקבילות: נתוני $Y$ בתאים A2 עד A21, נתוני $X$ בתאים B2 עד B21. פקודת אקסל: $=correl(A2:A21, B2:B21)$ נתנה את התוצאה: $r_{X,Y} = 0.8$ . הקשר בין התמ"ג לציונים במבחן פיז"ה הוא חיובי וחזק.

**ב.** ציוני התקן עבור ישראל: תמ"ג $\frac{35.1 - 23.1}{9.69} = 1.238$ , ציון מתמטיקה $\frac{466 - 471.3}{45.91} = -0.115$ .

הערכים הללו מדברים בעד עצמם (נתחו).

**משימה IX. חורף במינאפוליס**

**א.**



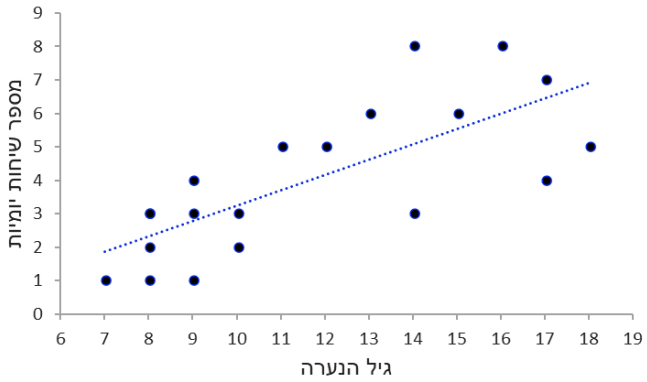

מדיאגרמת הפיזור רואים שהקשר הוא יורד וחזק מאוד, כצפוי. משמעות הקו האפור תובהר בפרק 5.

**ב.** מקדם המתאם חושב בתוכנת אקסל. הנתונים הוקלדו בשתי עמודות מקבילות: נתוני $Y$ בתאים A2 עד A13, נתוני $X$ בתאים B2 עד B13.

פקודת אקסל: $=correl(A2:A13,B2:B13)$ נתנה את התוצאה: $r_{X,Y}=-0.97$.

**משימה X. הגיל והטלפון**

**א.**




**ב.** משמעות הקו המנוקד תובהר בפרק 5. מקדמי המתאם חושבו בתוכנת אקסל. הנתונים הוקלדו בשתי עמודות מקבילות: נתוני $Y$ בתאים A2 עד A21, נתוני $X$ בתאים B2 עד B21.

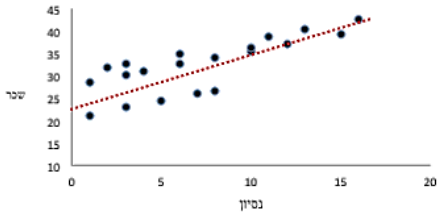
פקודת אקסל: $=correl(A2:A21,B2:B21)$ נתנה את התוצאה: $r_{X,Y}=0.964$.

## פרק 4. בעיות ניבוי

**משימה I.**

**א.** בדיאגרמות הפיזור שלהלן צוירו קווי ניבוי (באדום מקווקו). כמו כן צוינה גם מידת הקִרבה של הנקודות לקו הניבוי ביחס לערכו של מקדם המתאם.

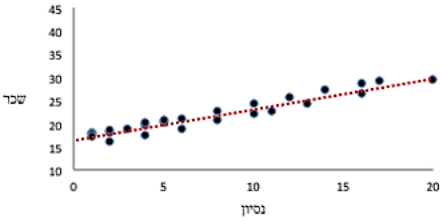
**א**




r=0.79

הנקודות קרובות לקו הניבוי.

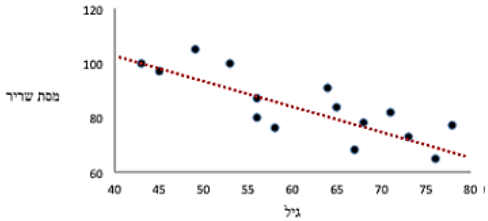
**ב**




r=0.97

הנקודות קרובות **מאוד** לקו הניבוי.

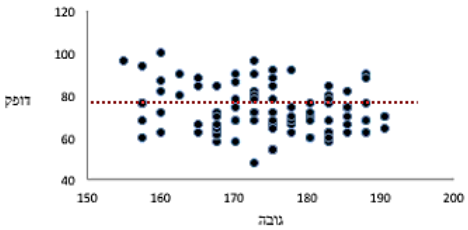
**ג**




r= –0.82

הנקודות קרובות לקו הניבוי.

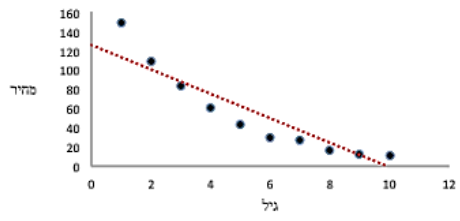
**ד**




r=–0.001

הנקודות רחוקות מקו הניבוי.
אין קשר בין הנקודות לקו.

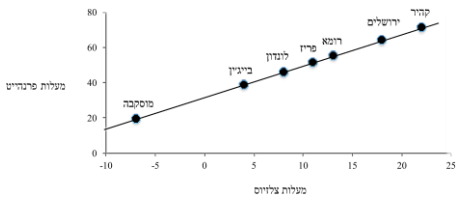
**ה**




r=–0.94

הנקודות קרובות לקו הניבוי.

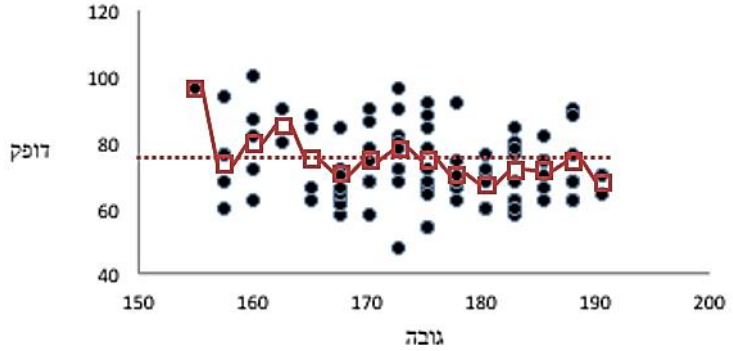
**ו**




r=1.00

הנקודות נמצאות על קו הניבוי.

**משימה II. קצב הדופק והגובה של חיילים**




**א.** באיור, הריבועים מסמנים את ממוצע הערכים לכל גובה ($x$) בנפרד. חיבור הריבועים יוצר את עקום הממוצעים (באדום) והוא אינו קו ישר.

**ב.** הקו הישר המתאים ביותר לצורכי ניבוי (מסומן כקו מקווקו) מקביל לציר ה-$x$.

**משימה III. השוואת טיב קווי הניבוי באמצעות עקרון הריבועים הפחותים**

**א.** בעמודה השלישית בטבלה רשומים ערכי הניבוי $\hat{y}_x$ על פי משוואת קו הניבוי $\hat{y}_x = 0.2x - 69.5$.

**ב.** בעמודה הרביעית רשומים ריבועי הסטיות של ערכי $y$ שהתקבלו בפועל – סטיות מהערכים המנובאים $\hat{y}_x$ על פי קו הניבוי. סכום ריבועי הסטיות הוא 2646.3.

**ג.** לאיור 18 מעמ' 59 הוספנו גם את קו הניבוי $\bar{y} = 62.4$ (הקו המקווקו שבהמשך התשובה).

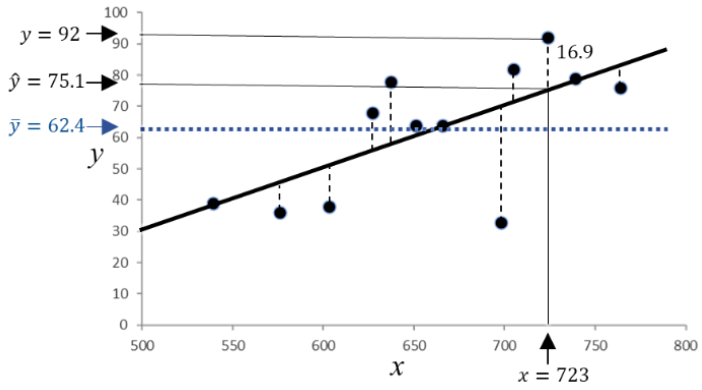
בעמודה האחרונה בטבלה רשמנו את ערכי ריבועי הסטיות מ-$\bar{y}$ וחושב סכומם בתחתית הטבלה: 4704.9.

| $x$ | $y$ | $\hat{y}_x = 0.2x - 69.5$ | $(y - \hat{y}_x)^2$ | $(y - 62.4)^2$ |
|---|---|---|---|---|
| 538 | 39 | 38.1 | 0.8 | 547.6 |
| 575 | 36 | 45.5 | 90.3 | 697.0 |
| 602 | 38 | 50.9 | 166.4 | 595.4 |
| 626 | 68 | 55.7 | 151.3 | 31.4 |
| 636 | 78 | 57.7 | 412.1 | 243.4 |
| 650 | 64 | 60.5 | 12.3 | 2.6 |
| 665 | 64 | 63.5 | 0.3 | 2.6 |
| 697 | 33 | 69.9 | 1361.6 | 864.4 |
| 704 | 82 | 71.3 | 114.5 | 384.2 |
| 723 | 92 | 75.1 | 285.6 | 876.2 |

| | | | | |
|---|---|---|---|---|
| 738 | 79 | 78.1 | 0.8 | 275.6 |
| 763 | 76 | 83.1 | 50.4 | 185.0 |
| **סה"כ** | | | **2646.3** | **4704.9** |

**ד**. חלוקת סכום ריבועי הסטיות מהקו $y=\bar{y}$ במספר התצפיות (12), נותנת 392.1. נשים לב שהמספר שחושב הוא למעשה השונות של המשתנה $Y$.

**ה**. מבין שני קווי הניבוי שבאיור, הקו $\hat{y}_x = 0.2x - 69.5$ עדיף על פני הקו $\bar{y} = 62.4$; סכום ריבועי הסטיות של הערכים המנובאים על פי קו זה מהערכים שהתקבלו בפועל קטן יותר.

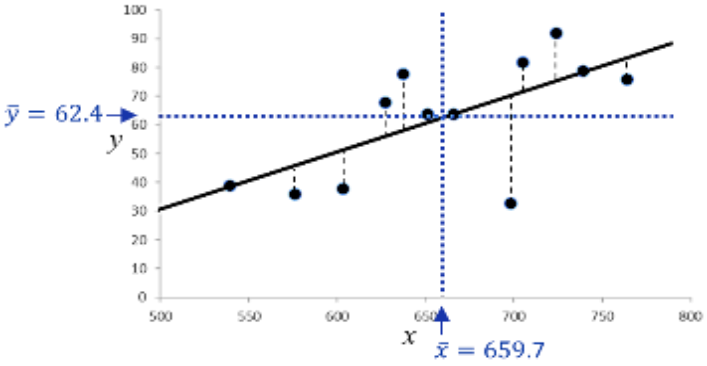



## פרק 5. קו רגרסיה

כדי לפתור ידנית חלק מהמשימות של פרק זה נדרשים חישובים מייגעים מאוד. כדאי להיעזר במחשבון.

### משימה I. תלמידי מתמטיקה – נתונים חלקיים

**א.** קו הרגרסיה עובר דרך נקודת הממוצעים (659.7 ,62.4) והשיפוע שלו הוא $b = 0.66 \cdot \frac{19.8}{65.4} = 0.2$ (ראו איור). קו הניבוי המוצג באיור 20 (בעמ' 63) הוא אכן קו הריבועים הפחותים.




**ב.** אי אפשר להסיק מ-12 תלמידים בלבד מהו המתאם באוכלוסיית תלמידי המתמטיקה כולה, מכיוון שלא ידוע לנו איך נבחרו תלמידים אלו ואם אכן הם מייצגים את כל אוכלוסיית התלמידים.

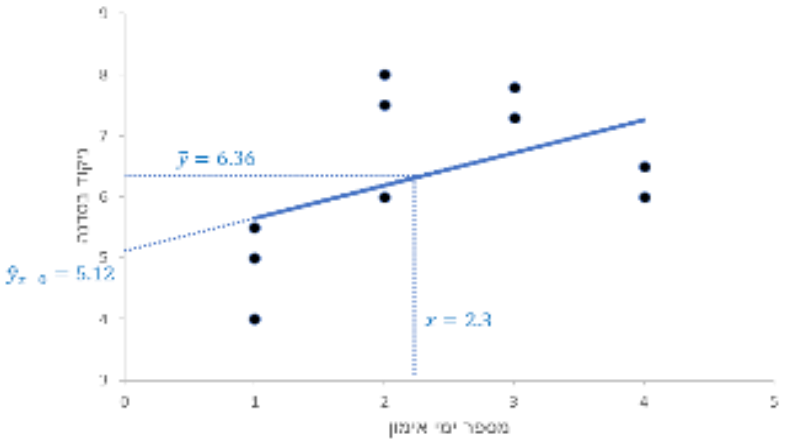
באופן כללי: בדרך כלל, מדגם שגודלו 12 פרטים אינו גדול דיו כדי להסיק ממנו על כלל האוכלוסייה. נזכיר כאן שבמדגם של 198 התלמידים שהצגנו בדוגמה 3 (בעמ' 69) מקדם המתאם היה 0.337 בלבד.

**משימה II. חמודי הסבות מקצועיות**

הנתונים: $\bar{x} = 2.30$, $\sigma_x = 1.10$, $\bar{y} = 6.36$, $\sigma_y = 1.24$, $r = 0.478$.

**א.** סימנו של השיפוע זהה לסימן של מקדם המתאם, כלומר השיפוע חיובי. המשמעות – הקו עולה, וככל שמספר ימי האימון גבוה יותר כך הניקוד בסדנה נוטה להיות גבוה יותר.

**ב.** שיפוע הקו הוא $b = 0.478 \cdot \frac{1.24}{1.10} = 0.54$ על פי נוסחה (6), ולכן משוואת קו הרגרסיה לניבוי הניקוד כשמתבססים על מספר ימי האימון היא: $\hat{y}_x = 6.36 + 0.54(x - 1.24) = 5.12 + 0.54\,x$.




זהו הקו הקרוב ביותר לנקודות בדיאגרמה (קו הריבועים הפחותים), אך כפי שרואים, הנקודות אינן קרובות מאוד לקו. נציין שמקדם מתאם שערכו בערך 0.5 נחשב בינוני, אך בתחומים רבים (למשל במדעי החברה) נהוג להתבסס על מתאם של 0.5 לצורכי ניבוי.

**ג.** לחישוב צפי הניקוד של מועמד שבחר בסדנה של שלושה ימי אימון נציב $x = 3$ במשוואת קו הרגרסיה, ונקבל: $\hat{y}_3 = 0.54 \cdot 3 + 5.12 = 6.74$.

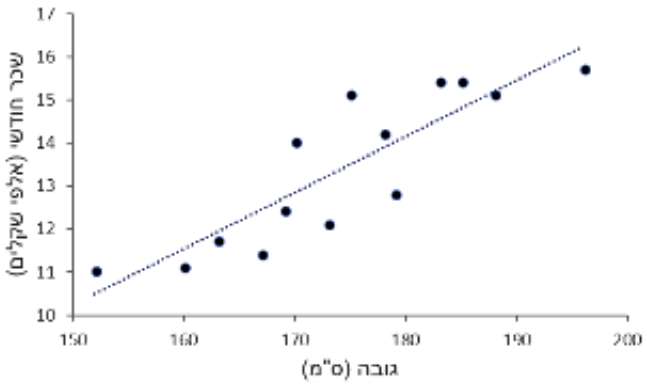
למועמד שבחר ביום אחד, הצפי הוא: $\hat{y}_1 = 0.54 \cdot 1 + 5.12 = 5.66$.

**ד.** לא סביר להשתמש בקו הרגרסיה לשישה ימי אימון. ערך זה הוא ערך חריג – אינו נמצא בטווח הנתונים שנדגם ואשר על פיו חושבה משוואת הקו, ולכן הטעות בניבוי עשויה להיות גדולה.

**משימה III. הקשר בין שכר העובדים לגובהם**

**א.** לצורך הפתרון חישבנו תחילה את חמשת הגדלים: $\bar{x} = 174.14$, $\sigma_x = 11.4$, $\bar{y} = 13.39$, $\sigma_y = 1.7$, וכן $r = 0.871$. מכאן, שיפוע הקו הוא $b = 0.871 \cdot \frac{1.7}{11.4} = 0.131$ וסימנו חיובי, כלומר הקו עולה. עובד גבוה יותר נוטה לקבל שכר גבוה יותר.

מנוסחה (5) $a = -9.43$ (חשבו), ומשוואת קו הרגרסיה לניבוי השכר $Y$ על פי גובה העובד $X$ היא אפוא: $\hat{y}_x = -9.43 + 0.131x$ (ראו איור בהמשך).




**ב.** ניבוי השכר (באלפי שקלים) לעובד שגובהו 178 ס״מ הוא $\hat{y}_{178} = 13.9$, ולעובד שגובהו 196 ס״מ הניבוי הוא $\hat{y}_{196} = 16.2$.

**ג.** בין אם ננבא את השכר באמצעות הגובה ובין אם ננבא את הגובה באמצעות השכר, מקדם המתאם בין המשתנים נותר זהה והוא $r = 0.871$.

**ד.** סימנו של שיפוע הקו לניבוי הגובה על פי השכר זהה לסימן של מקדם המתאם, כלומר גם הוא חיובי, וזהו קו עולה.

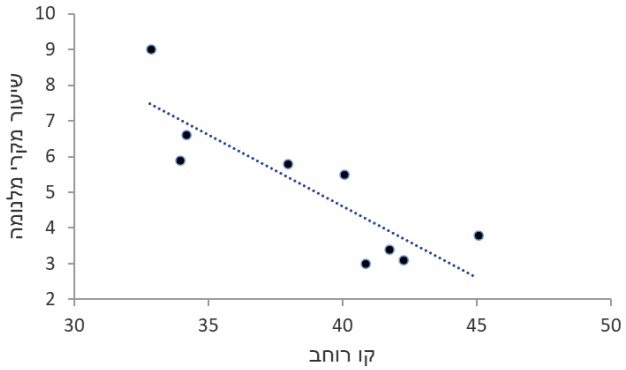
**ה.** שיפועו של קו הרגרסיה לניבוי הגובה ($X$) על פי השכר ($Y$) הוא $0.871 \cdot \frac{11.4}{1.7} = 5.8$.

כמו כן, לפי נוסחה (5): $a = 174.14 - 5.8 \times 13.39 = 96.6$. מכאן, משוואת הקו לניבוי הגובה על בסיס השכר היא: $\hat{x}_y = 96.6 + 5.8y$.

**ו.** הצבה של $y$=16.2 במשוואה נותנת ניבוי גובה (בסנטימטרים) של $\hat{x} = 96.6 + 5.8 \cdot 16.2 = 191$. נשים לב שהניבוי נמוך מהערך שהתקבל – 196 ס״מ. בסעיף 6.1 נסביר את התופעה.

**משימה IV. סרטן עור וקרינה אולטרה סגולה**

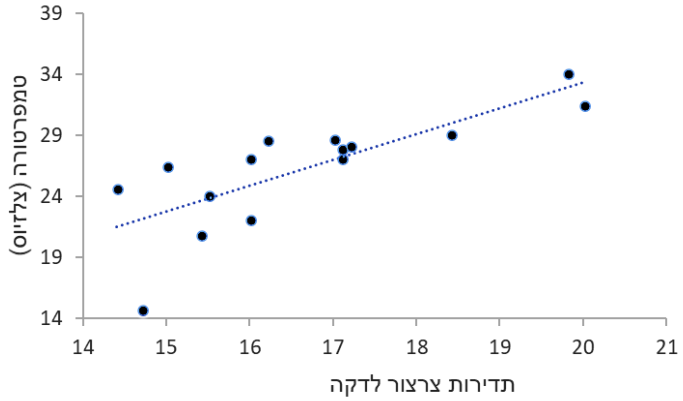
**א.** חישבנו ומצאנו $r = -0.857$, ומכאן שיפוע הקו הוא $b = -0.857 \cdot \frac{1.88}{4.04} = -0.4$. מנוסחה (5) מקבלים: $a = \bar{y} - b\bar{x} = 5.12 - (-0.4) \cdot 38.71 = 20.6$. מכאן, משוואת קו הרגרסיה לניבוי שיעור מקרי המלנומה על פי קו הרוחב היא: $\hat{y}_x = 20.6 - 0.4x$ (ראו איור).



**ב.** בקו רוחב $x = 35$, הניבוי לשיעור מקרי המלנומה הוא: $\hat{y}_{35} = 20.6 - 0.4 \cdot 35 = 6.6$. נציין שקו הרגרסיה מניב ניבויים סבירים ($r^2 = 0.735$).

**משימה V. שירת הצרצר**

**א.** על פי דיאגרמת הפיזור להלן, הקשר בין תדירות הצרצור לטמפרטורה נראה בהחלט קווי:




**ב.** מקדם המתאם מחושב באמצעות השורות הבאות של הטבלה הבאה:

| | תדירות לדקה, $x$ | טמפרטורה, $y$ | $x \cdot y$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| **סה"כ** | **249.8** | **393.9** | **6645.75** | **4200.56** | **10644.47** |
| **ממוצעים** | $\bar{x} = 16.65$ | $\bar{y} = 26.26$ | | | |

מנתונים אלו חושבו השונויות על פי נוסחה (0):

$\sigma_X^2 = \frac{1}{15}[4200.56] - 16.65^2 = 2.70$ , $\sigma_Y^2 = \frac{1}{15}[10644.47] - 26.26^2 = 20.04$.

לכן, $\sigma_x = 1.64$, $\sigma_y = 4.48$.

מקדם המתאם חושב על פי נוסחה (3):

$$r = \frac{1}{1.64 \cdot 4.48} \cdot \left\{\frac{1}{15} \cdot [6645.75] - 16.65 \cdot 26.26\right\} = 0.78.$$

**ג.** שינוי ביחידות המדידה של אחד המשתנים או שניהם אינו גורר שינוי במקדם המתאם. לפיכך, מקדם המתאם לא ישתנה אם הטמפרטורה תימדד בפרנהייט במקום בצלזיוס.

**ד.** מהחישובים שערכנו, שיפוע הקו הוא $b = 0.78 \cdot \frac{4.48}{1.64} = 2.13$, והקו עובר דרך נקודת הממוצעים (16.65, 26.26). מנוסחה (6), משוואת קו רגרסיה לניבוי הטמפרטורה על פי תדירות הצרצור היא:

$$\hat{y}_x = 26.26 + 2.12(x - 16.65) = -9.04 + 2.12x.$$

**ה.** הניבוי של הטמפרטורה במעלות צלזיוס לצרצור בתדירות של 18 פעמים לדקה הוא:

$\hat{y}_{x=18} = -9.04 + 2.12 \cdot 18 = 29$.

**משימה VI. פסק זמן מהעבודה והמשך הקריירה המקצועית**

**א.** לפנינו טבלת הנתונים של אחוז בתי החולים שהביעו נכונות לקלוט טכנאי רפואה שנעדרו משוק העבודה:

| $y^2$ | $x^2$ | $x \cdot y$ | **אחוז בתי החולים,** $y$ | **שנות היעדרות,** $x$ | |
|---|---|---|---|---|---|
| 10000 | 0.25 | 50 | 100 | 0.5 | |
| 8836 | 2.25 | 141 | 94 | 1.5 | |
| 5625 | 16 | 300 | 75 | 4 | |
| 1936 | 64 | 352 | 44 | 8 | |
| 784 | 169 | 364 | 28 | 13 | |
| 289 | 324 | 306 | 17 | 18 | |
| **27470** | **575.5** | **1513** | **358** | **45** | **סה"כ** |
| | | | $\bar{y} = 60$ | $\bar{x} = 7.5$ | **ממוצעים** |

חישוב השונויות על פי נוסחה (0): $\sigma_x^2 = \frac{1}{6}[575.5] - 7.5^2 = 39.7$,

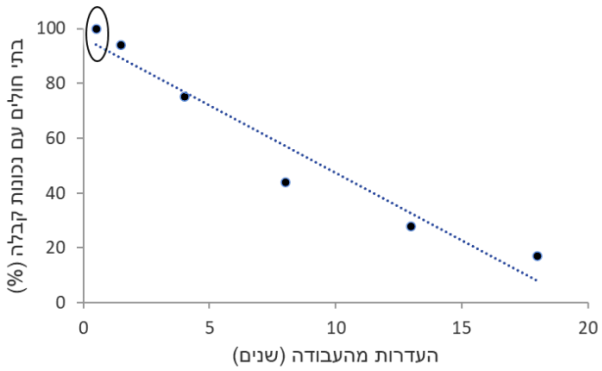
$\sigma_y^2 = \frac{1}{6}[27470] - 60^2 = 1018.2$. לכן, $\sigma_x = 6.3$, $\sigma_y = 31.9$.

מקדם המתאם חושב על פי נוסחה (3): $r = \frac{1}{6.3 \cdot 31.9} \cdot \left\{\frac{1}{6} \cdot [1513] - 7.5 \cdot 60\right\} = -0.972$.

הערך קרוב ל-$(-1)$, כלומר הקשר בין $X$ ל-$Y$ הוא קווי יורד וחזק. כצפוי, ככל שעולות שנות ההיעדרות של הטכנאים, כך יורד אחוז בתי החולים שהביעו נכונות לקלוט אותם לעבודה.

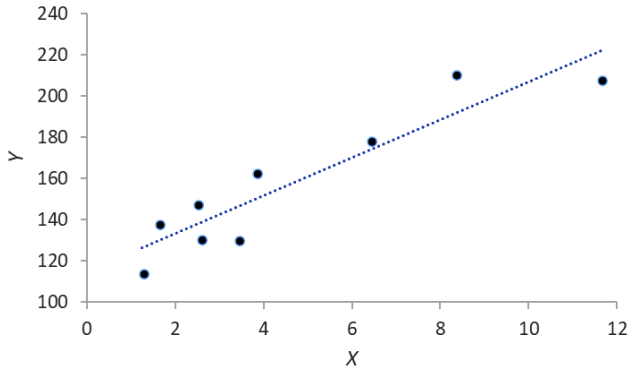
**ב.** באמצעות סטיות התקן ומקדם המתאם נחשב את שיפוע הקו: $b = -0.972 \cdot \frac{31.9}{6.3} = -4.92$. הקו עובר דרך נקודת הממוצעים (7.5, 60). על פי נוסחה (6), משוואת קו הרגרסיה לניבוי אחוז הנכונות $Y$ על בסיס שנות ההיעדרות $X$ היא: $\hat{y}_x = 60 - 4.92\ (x - 7.5) = 96.6 - 4.924x$.

**ג.** אפשר לראות בדיאגרמת הפיזור שקו הניבוי אינו עובר דרך הנקודה (100, 0). אכן, הניבוי של אחוז בתי החולים שהביעו נכונות כאשר שנות ההיעדרות הן $x = 0$ הוא $\hat{y}_0 = 96.6$, וזוהי גם נקודת החיתוך של קו הניבוי עם הציר האנכי. הסטייה של הניבוי מהערך 100%, שהוא הערך האמיתי, היא 3.4%. זוהי טעות הניבוי והיא צפויה.




**משימה VII. חשיפה לחומרים רדיואקטיביים**

**א.** בדיאגרמת הפיזור מוצג הקשר בין רמת החשיפה לחומרים רדיואקטיביים $X$ ובין מקרי המוות מסרטן ל-100,000 תושבים $Y$. נראה שהקשר הוא קווי וחזק:




כל החישובים הנדרשים בהמשך יתבססו על הנתונים המוצגים בשורות האפורות בטבלה זו:

| $y^2$ | $x^2$ | $x\cdot y$ | תמותה מסרטן, $y$ | אינדקס החשיפה, $x$ | |
|---|---|---|---|---|---|
| | | | 147.1 | 2.49 | |
| | | | 130.1 | 2.57 | |
| | | | 129.9 | 3.41 | |
| | | | 113.5 | 1.25 | |
| | | | 137.5 | 1.62 | |
| | | | 162.3 | 3.83 | |
| | | | 207.5 | 11.64 | |
| | | | 177.9 | 6.41 | |
| | | | 210.3 | 8.34 | |
| **232499** | **289.4** | **7439.4** | **1416.1** | **41.6** | **סה"כ** |

| | | |
|---|---|---|
| **ממוצעים** | $\bar{x} = 4.6$ | $\bar{y} = 157.3$ |

**ב.** חישוב השונויות על פי נוסחה (0): $\sigma_x^2 = \frac{1}{9}[289.4] - 4.6^2 = 10.8$,

$\sigma_y^2 = \frac{1}{9}[232499] - 157.3^2 = 1076$. ולכן, $\sigma_x = 3.29$, $\sigma_y = 32.8$.

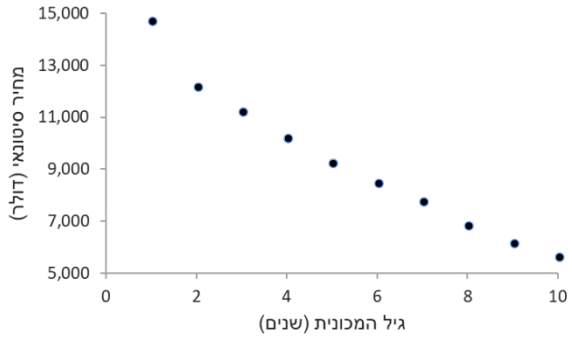
מקדם המתאם מחושב על פי נוסחה (3): $r = \frac{1}{3.29 \cdot 32.8} \cdot \left\{\frac{1}{9} \cdot [7439.4] - 4.6 \cdot 157.3\right\} = 0.926$.

**ג.** מכל אלו מתקבל ששיפוע הקו הוא $b = 0.926 \cdot \frac{32.8}{3.29} = 9.232$, והוא עובר דרך נקודת הממוצעים (4.62, 157.34). מנוסחה (6) נקבל את משוואת קו הרגרסיה לניבוי התמותה מסרטן על בסיס אינדקס החשיפה: $\hat{y}_x = 114.72 + 9.232x$ (ראו איור בסעיף א).

**ד.** ההערכה לתמותה מסרטן לאינדקס חשיפה ברמה 5 היא: $\hat{y}_5 = 114.72 + 9.232 \cdot 5 = 161$, כלומר 161 מקרי תמותה מתוך כל 100,000 איש.

### משימה VIII. מחירי מכוניות משומשות

**א.** דיאגרמת פיזור של מחירים של טויוטה קורולה משומשת (בדולרים) על פי גיל המכונית (בשנים):




הקשר בין המשתנים הוא קווי יורד וחזק.

**ב.** מקדם המתאם בין המשתנים מתקבל מתוצאות החישובים בשורות התחתונות בטבלה:

| $x$, **גיל** | $y$, **מחיר סיטונאי** | $x \cdot y$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 1 | 14,680 | | | |
| 2 | 12,150 | | | |
| 3 | 11,215 | | | |
| 4 | 10,180 | | | |
| 5 | 9,230 | | | |
| 6 | 8,455 | | | |

| | | | | |
|---|---|---|---|---|
| 7 | 7,730 | | | |
| 8 | 6,825 | | | |
| 9 | 6,135 | | | |
| 10 | 5,620 | | | |
| **55** | **92,220** | **430,350** | **385** | **924,769,600** |
| $\bar{x}=5.5$ | $\bar{y}=9222$ | | | |

**סה"כ**

**ממוצעים**

חישוב השונויות על פי נוסחה (0): $\sigma_x^2=\frac{1}{10}[385]-5.5^2=8.25$,

$\sigma_y^2=\frac{1}{10}[924769600]-9222^2=7431676$. ולכן, $\sigma_x=2.87$, $\sigma_y=2726$.

מקדם המתאם חושב על פי נוסחה (3): $r=\frac{1}{2.87\cdot 2726}\cdot\left\{\frac{1}{10}\cdot[430350]-5.5\cdot 9222\right\}=-0.982$.
הקשר אכן יורד וחזק, כצפוי.

**ג.** נחשב מכל אלו את שיפוע הקו: $b=-0.982\cdot\frac{2726}{2.87}=-931.64$. הקו עובר דרך נקודת הממוצעים (5.5, 9222). מהנקודה ומהשיפוע נקבל בעזרת נוסחה (6) את משוואת קו הרגרסיה לניבוי המחיר על פי גיל הרכב: $\hat{y}_x=14{,}346-931.64x$.

**ד.** כאשר מציבים במשוואת הקו $x=0$, מקבלים עבור $y$ את הערך המנובא 14,346 דולר, הנמוך ב-1,854 ממחירה של מכונית חדשה (16,200 דולר). התוצאה צפויה, מכיוון שלקנייה של מכונית חדשה יש ערך מוסף שאינו יכול להימדד מהנתונים הנוגעים למכוניות משומשות.

**משימה IX. עברית שפה קשה** (המשך משימה I מעמ' 37)

**א.** נתון: $\bar{x}=5.5$, $\bar{y}=10$, $\sigma_X=2.22$, $\sigma_Y=3.16$, ו-$r=0.95$.

מנוסחה (4), שיפוע קו הניבוי הוא $b=0.95\cdot\frac{3.16}{10}=1.356$. מנוסחה (5) מקבלים:

$a=\bar{y}-b\bar{x}=10-1.356\cdot 5.5=2.542$.

מכאן, משוואת קו הרגרסיה לניבוי ציון מבחן B על סמך ציון מבחן A היא $\hat{y}_x=2.542+1.356x$. ערכי הניבוי במבחן B לערכים המתאימים של מבחן A רשומים בעמודה השלישית בטבלה.

**ב.** בעמודה הרביעית בטבלה רשמנו את ריבועי הסטיות של ערכי הניבוי מהערכים שהתקבלו בפועל. בתא האפור בתחתית העמודה רשומה הטעות הריבועית הממוצעת בניבוי ציון מבחן B על סמך ציון מבחן A.

| **מבחן A**, $x$ | **מבחן B**, $y$ | **הניבוי**, $\hat{y}_x$ | **ריבועי הטעויות**, $(y-\hat{y}_x)^2$ |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 0.06 | 5.25 | 5 | 2 |
| 1.75 | 9.32 | 8 | 5 |
| 0.06 | 14.75 | 15 | 9 |
| 1.75 | 10.68 | 12 | 6 |
| 1.07 | 12.03 | 11 | 7 |
| 1.07 | 7.97 | 9 | 4 |
| **0.96** | | | **ממוצע** |

**ג.** נשים לב שידיעת מספר השגיאות שקולה לגמרי לידיעת מספר התשובות הנכונות. לכן גם טעויות הניבוי לא ישתנו, ולכן הטעות הריבועית הממוצעת על בסיס מספר השגיאות גם היא 0.96. לחלופין, בנינו עבור המשתנים $X^*$ ו-$Y$ טבלה מקבילה לזו שבסעיף ב. כפי שרואים, העמודה האחרונה זהה לזו שהתקבלה בסעיף ב, לכן גם הממוצע לא השתנה.

| $(y-\hat{y}_{x^*})^2$ | **ערך הניבוי,** $\hat{y}_{x^*}$ | **מבחן B,** $y$ | **מבחן A,** $x^*$ |
|---|---|---|---|
| 0.06 | 5.25 | 5 | 8 |
| 1.75 | 9.32 | 8 | 5 |
| 0.06 | 14.75 | 15 | 1 |
| 1.75 | 10.68 | 12 | 4 |
| 1.07 | 12.03 | 11 | 3 |
| 1.07 | 7.97 | 9 | 6 |
| **0.96** | | | **ממוצע** |

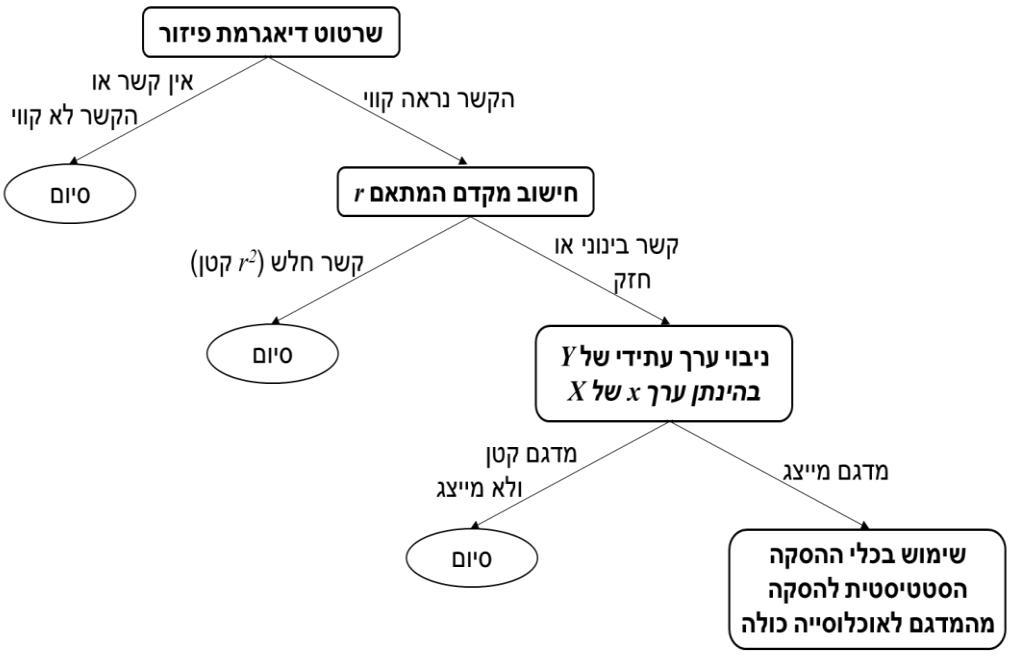
**משימה X. עברית שפה קשה (המשך משימה IX)**

שיפוע הקו לניבוי $X$ על פי $Y$ הוא $b = 0.95 \cdot \frac{10}{3.16} = 3$, והקו עובר דרך נקודת הממוצעים (10, 5.5). משוואת קו הניבוי היא אפוא: $\hat{x}_y = 5.5 + 3(y-10)$. במקום לערוך טבלה לחישוב ריבועי הטעויות, נחשב את הטעות הריבועית הממוצעת באמצעות נוסחה (7): $(2.22)^2(1-0.95^2) = 0.48$ (בדקו).

הפער בין טעויות הניבוי נובע מהפער בין השונויות המתאימות.

## סיכום

**משימת סיכום: תרשים זרימה**



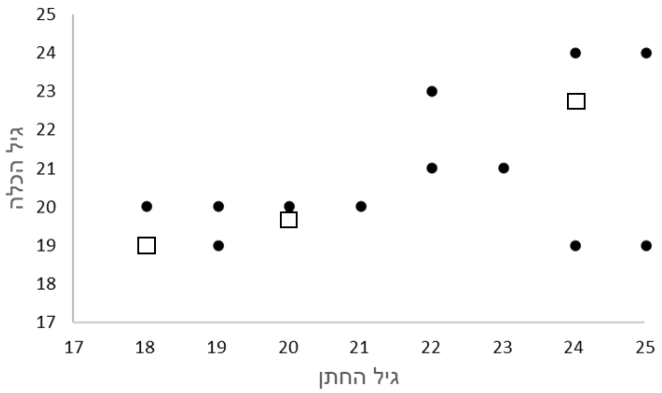

## נספח הלמ"ס

**משימה אתגרית: שימוש בנתוני הלמ"ס לכריית נתונים**

**א.** בלוח שהצגנו יש 34,473 זוגות נתונים מהסוג (גיל הכלה, גיל החתן), לדוגמה (18, 21).

**ב.** ריבוי נתונים זהים אינו מאפשר לשרטט דיאגרמת פיזור, ולכן יש להשתמש בכלים גרפיים מתוחכמים יותר. אין קושי לצייר מערכת צירים ולסמן בה את הזוג (18, 21) למשל, אלא שיש 114 זוגות נתונים זהים לזוג זה. לכן צריך להוסיף לדיאגרמה ממד של גובה, שיבטא שכיחות, כלומר להציג דיאגרמה תלת-ממדית (ראו בעמ' 23).

נסתפק אפוא בהצגת נקודות אחדות מתוך דיאגרמת הפיזור:




הנקודות השחורות באיור הן חלק מדיאגרמת הפיזור, ואילו הריבועים הלבנים נמצאים על עקום הממוצעים (ראו בהמשך בסעיף ו).

**ג.** נראה שהקשר הוא עולה ועוצמתו בינונית-חזקה.

**ד.** ממוצע גיל הכלה לחתנים בני 18:

$$\overline{y}_{x=18} = \frac{17\cdot7+18\cdot148+19\cdot155+20\cdot84+21\cdot25+22\cdot8+23\cdot2+27\cdot2+29\cdot1}{433} = 19$$

זהו הניבוי הטוב ביותר של גיל הכלה לחתנים בני 18.

**ה.** באופן זה מתקבל **עקום הממוצעים**, שהנקודות על פניו הן הניבויים הטובים ביותר של גיל הכלה על פי גיל החתן.

**ו.** לכל אחת מקבוצות גיל החתן (18 ערכים) יש לחשב את ממוצע גיל הכלה; אם רשומה קבוצת גילים ניקח את אמצע הקבוצה כערך המתאים. אנו חישבנו את הגיל הממוצע של הכלה לשלושה ערכים של גיל החתן:

$$\overline{y}_{x=18} = 19 \,,\, \overline{y}_{x=20} = 19.8 \,,\, \overline{y}_{x=24} = 22.7$$

באיור שבסעיף ב ציינו את הנקודות המתאימות בריבועים לבנים באותה מערכת צירים. כפי שרואים באיור, עקום הממוצעים אינו קו ישר, כלומר זה אינו קו הרגרסיה. מעניין לבצע חישובים נוספים.

## פתרונות מקוצרים לתרגילים נבחרים

### פרק 2. ניתוח גרפי

**תרגיל 3.**

| | | |
|---|---|---|
| בינוני | חלש | חזק מאוד |
| חלש | חזק מאוד | בינוני |

## פרק 3. מתאם בין משתנים

**תרגיל 1.**

| | | |
|---|---|---|
| 0.8 | 0.4 | 0.9 |
| 0 | 0.95 | 0.6 |

**תרגיל 2.**

| | | |
|---|---|---|
| –0.2 | 0.3 | 0.71 |
| –0.93 | –0.75 | 1 |

**תרגיל 4.** בדיאגרמה העליונה $r = 0$. ביחידות מתוקננות, לכל מכפלה חיובית יש מכפלה זהה עם סימן מינוס.

בדיאגרמה הימנית למטה $r = 1$. רמז: לאחר תקנון די לחשב מכפלה עבור שתי הנקודות העליונות, לכפול ב-2 ולחשב ממוצע. באופן דומה, בדיאגרמה השמאלית $r = -1$.

**תרגיל 5.** א–ב. על פי נוסחה (2), מקדם המתאם בין $(-X)$ ל-$Y$ הוא $(-r)$, מכיוון שלכל ערכי המכפלה כאן סימן הפוך. מאותה סיבה מקדם המתאם בין $(-X)$ ל-$(-Y)$ הוא $-(-r) = r$.

**תרגיל 6.** מקדם המתאם אינו משתנה עם שינוי ליניארי במשתנים.

## פרק 4. בעיות ניבוי

**תרגיל 1.** נשים לב שעבור ערכים קיצוניים (גבוהים מאוד ונמוכים מאוד) של גובהי האבות יש נסיגה – כלומר התרחקות מקיצוניות – של ממוצע גובהי הבנים. במילים אחרות, הניבוי הטוב ביותר לערכי קיצון אינו קיצוני כל כך. ראו דיון בפרק 6.

**תרגיל 3. התשובה שלילית.** אין קיצורי דרך.

סוג הבעייתיות (ציירו): נניח ש-X מקבל רק שני ערכים. עבור אחד מהם התקבלו ערכים רבים מתאימים של Y והניבוי הוא הממוצע שלהם. עבור השני התקבל ערך יחיד חריג מאוד – רחוק מאוד משאר הנקודות בדיאגרמה, אך הניבוי שווה לערך חריג זה. על עקום הממוצעים יתקבלו שתי נקודות בלבד וקו הניבוי יעבור דרכן. אילו היו לנו כל הנתונים, הרי השפעתה של הנקודה החריגה על קו הניבוי הייתה מזערית, ואילו במקרה שתיארנו כאן ההשפעה שלה זהה לזו של הממוצע של ערכים רבים.

## פרק 5. קו הרגרסיה

**תרגיל 1.**

**א.** קו הרגרסיה לניבוי ההוצאה על בסיס ההכנסה הוא: $\hat{y}_x = 2.634 + 0.707 \cdot x$.

**ב.** הניבוי המתאים לערך הממוצע של $Y$ הוא הערך הממוצע של $X$, במקרה זה 14.3.

**ג.** הניבוי עבור $x$=18 הוא 15.36, הניבוי עבור $x$=11 הוא 10.41.

**ד.** יש להציב את נתוני הבעיה בנוסחה (7) ולהשוות את התוצאה המתקבלת לשונות של $Y$.

**תרגיל 2.** קו הרגרסיה לניבוי הרווחים על בסיס מספר שנות הקיום הוא: $\hat{y}_x = 62.56 - 2.23 \cdot x$. מקדם המתאם גבוה מאוד – 0.979, על כן הניבוי אמין מאוד. הניבוי לשנה ה-11 הוא 11.8. אומנם זהו ניבוי מעבר לטווח הערכים, אולם מכיוון שמדובר בחריגה של שנה אחת בלבד נראה שהניבוי אמין.

**תרגיל 3.** קו הניבוי של הלחות על פי הטמפרטורה הוא: $\hat{y}_x = 2.634 + 0.707 \cdot x$.

**א.** ניבוי הלחות ליישוב שהטמפרטורה בו היא 23 מעלות הוא 45.89.

**ב.** לטבריה, שהטמפרטורה בה היא 25 מעלות, ניבוי הלחות הוא 44.44. הערך שהתקבל בפועל הוא 45, וריבוע טעות הניבוי הוא 0.316.

**ג.** ממוצע ריבוע טעות הניבוי הוא 61.54.

**תרגיל 4.**

**א.** שיפוע קו הניבוי של $Y$ על פי $X$ הוא 0.085, ומשוואת קו הניבוי היא $\hat{y}_x = 3.52 + 0.085(x - 18)$. מכאן, הניבוי לחולה המעשן 15 סיגריות ביום הוא: $\hat{y}_{15} = 3.26 \cong 3$ ספלי קפה ביום.

**ב.** שיפוע קו הניבוי של $X$ על פי $Y$ הוא 5.95, ומשוואת קו הניבוי היא $\hat{x}_y = 18 + 5.95(y - 3.52)$. מכאן, הניבוי לחולה לב השותה 4 ספלי קפה ביום הוא $\hat{x}_4 = 20.86 \cong 21$ סיגריות ביום.

**ג–ד.** ממוצע ריבוע הטעות של קו הרגרסיה בהשוואה לניבוי ללא משתנה מנבא:

ניבוי מספר הסיגריות על פי מספר ספלי הקפה: $2.5 \cdot [1 - (.71^2)] = 1.24$, וזהו שיפור רב בהשוואה ל-2.5.

ניבוי מספר ספלי הקפה על פי מספר הסיגריות: $175 \cdot [1 - (.71^2)] = 86.78$, וזהו שיפור רב בהשוואה ל-175.

**תרגיל 5.**

**א.** הניבוי הוא ממוצע ציוני הלימודים, כלומר 7.5. ממוצע ריבוע הטעות הוא השונות 4.41.

**ב.** ב1. נחשב לשם כך את הציון הממוצע בלימודים של אותם תלמידים שציון הבגרות שלהם הוא 8.5.

ב2. יש למצוא קו רגרסיה לניבוי ציון הלימודים $Y$ על בסיס ציון הבגרות $X$. מהנתונים מתקבל ששיפוע הקו הוא 0.81, והקו עובר דרך נקודת הממוצעים (7.5, 7.9). משוואת הקו היא אפוא: $\hat{y}_x = 7.5 + 0.81(x - 7.9)$. לניבוי ציון לימודים של מועמד שקיבל ציון בגרות 8.5 נציב את הערך 8.5

במשוואת הקו. נקבל שציון הלימודים הצפוי הוא $\hat{y}_{8.5}=8$. ממוצע ריבוע הטעות הוא $(1.3)^2\cdot(1-0.5^2)=3.31$. השיפור בהשוואה לניבוי ללא משתנה מנבא אינו גדול.

**תרגיל 6.**

**א.** $\sigma_Y^2 = 856.73$

**ב.** $856.73\cdot(1-0.86^2)=225.9$, וזהו שיפור ניכר. השיפור תלוי במקדם המתאם (נמקו).

**ג.** השימוש בפרבולה משפר מאוד את הניבוי.

תמיד האמנתי שממתאם חזק

אפשר להסיק על סיבתיות!

עד שקראתי את החוברת הזו

ואז הפסקתי להאמין

וואו, נשמע כאילו החוברת

הייתה מועילה

או להפך

# שאלון סיכום 🤔 (שאלות חשיבה)

**שאלה I. האם מלפפון הוא יותר ירוק או יותר ארוך?**

על שאלה זו לא נענה כאן, אך נענה על שאלות דומות – עם נתונים ממחקרי אמת!

א. ההורים המותשים של התינוקת לילית התלוננו שהיא מתעוררת 5 פעמים בלילה (מעל לממוצע). עד כמה הנתון אכן חריג? במחקר על איכות שינה של תינוקות התבררו נתונים אלה: משך ממוצע של שנת הלילה הוא $\bar{x} = 9.67$, עם סטיית תקן של $\sigma_X = 0.76$. כמו כן, ממוצע מספר ההתעוררויות בלילה הוא $\bar{y} = 4.87$, עם סטיית תקן של $\sigma_Y = 1.27$.

בירור נוסף עם ההורים העלה שהתינוקת ישנה 10.5 שעות בלילה. לנוכח כל זאת, מה תענו להורים של לילית? [חשבו והשוו את מידת החריגות של ציוני התקן ביחס לממוצע המתאים.]

ב. בדוגמה 4 – מבחני פיז"ה (עמ' 24), הציעו דרך לבחון את מצבה היחסי של ישראל במתמטיקה בהשוואה למצבה הכלכלי בזירה הבין-לאומית.

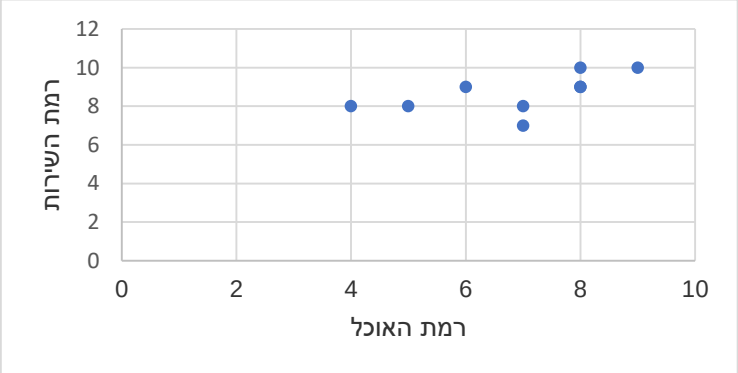
**שאלה II.**

עבור כל אחת מהדיאגרמות הבאות בדקו:

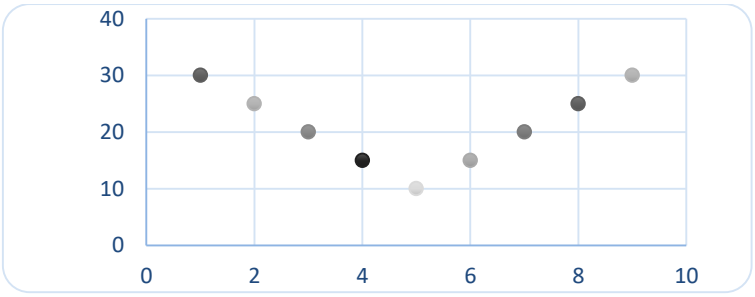
– האם על פי הדיאגרמה יש קשר קווי בין המשתנים? אם כן, האם הקשר עולה או יורד?

– מה הערכתכם, האם מקדם המתאם הוא: 0, חלש, בינוני, חזק, 1, 1-, גדול מ-1.

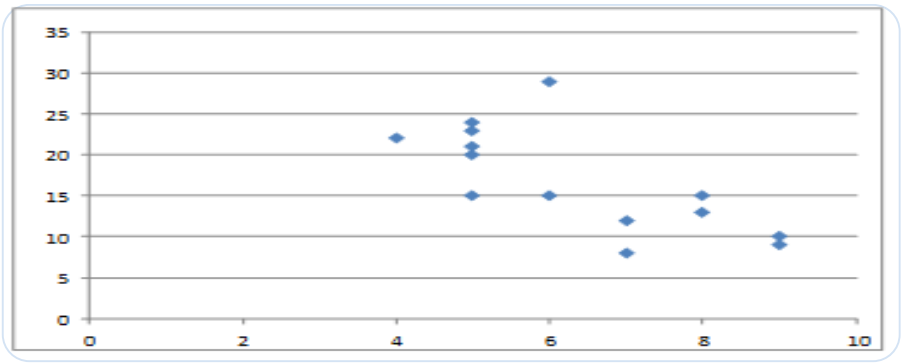
א. מידת שביעות הרצון של לקוחות בית הקפה "פת לחם" באשר לרמת האוכל ולרמת השירות.



ב.

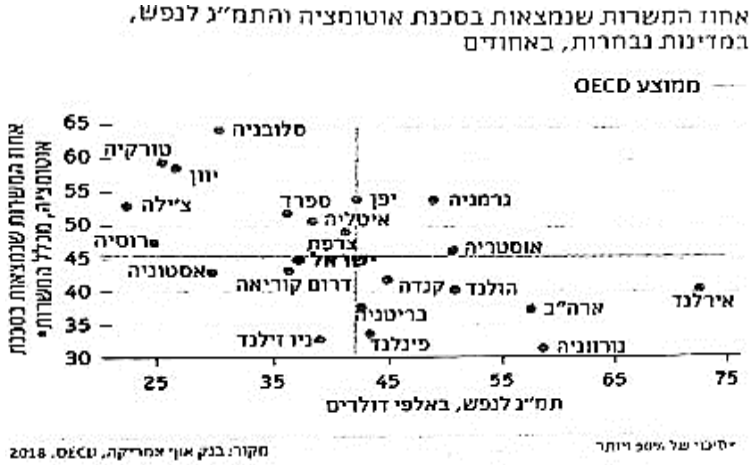



ג.




**שאלה III.**

דיאגרמת הפיזור הבאה עוסקת בקשר שבין המצב הכלכלי של מדינות שונות (נמדד על פי התמ"ג) ובין אחוז המשרות הנמצאות בסכנת היעלמות בעתיד הקרוב (מקור: בנק אוף אמריקה):




א. שרטטו אליפסה הדוקה סביב נקודות הדיאגרמה ונתחו בעזרתה את הקשר בין המשתנים (כיוון, חוזק). באילו רביעים נמצאות מרבית הנקודות? מה משמעות הדבר?

ב. השוו את מצבה היחסי של ישראל בזירה הבין-לאומית על פי התמ"ג ואחוז המשרות שייעלמו.

**שאלה IV.**

מורה אסף נתונים של 15 תלמידים בתיכון 'האלופים' כדי לבדוק את הקשר בין מספר השעות בשבוע שתלמידים מקדישים לבילויים לבין ציון התעודה הממוצע שלהם. על פי דיאגרמת הפיזור נראה שהקשר בין המשתנים הוא קווי ויורד. המורה חישב ומצא ש- $r^2 = 0.81$. חוו דעתכם כל על אחת המסקנות:

א. בקרב תלמידי תיכון בישראל, אלו המבלים יותר נוטים לקבל ציונים נמוכים יותר.

ב. בקרב הנבדקים, ככל שהם מקדישים פחות זמן לבילויים כך הציון שלהם נוטה לרדת.

ג. אם נועם (אחד הנבדקים) יצמצם את שעות הבילוי שלו, הציון שלו יעלה.

**שאלה V.**

עבור רשימת נתונים של שני משתנים $X, Y$ התקבלו תוצאות אלה: הממוצע של $X$ הוא 1.5, והשונויות של שני המשתנים הן 4. כמו כן, קו הרגסיה של $Y$ על בסיס $X$ הוא $\hat{y}_x = -0.2x + 0.5$.

א. הסיקו מהו מקדם המתאם בין $X$ ל-$Y$.

ב. נסו להסיק מנתונים אלו גם מהו הממוצע של $Y$.

**שאלה VI. שיפוע קו הרגרסיה**

ערכתם מחקר על קשר בין אורך מכונית (במטרים) לבין המשקל שלה (בקילוגרמים). בין השאר חישבתם את קו הרגרסיה. ללא חשיפה לנתונים עצמם, בחרו את התשובה שנראית לכם הסבירה ביותר. אם אף אחת איננה סבירה, מהי הערכתכם?

השיפוע של קו הרגרסיה הוא:

3, 30, 300, 3000. הסבירו מה משמעות השיפוע כאן על פי הנוסחה של משמעות השיפוע (עמ' 75).

**שאלה VII. תפיסה על-חושית**

שני משתתפים ש'הוכיחו' יכולת תפיסה על-חושית בתוכנית ריאליטי השתתפו במבחן שבחן את יכולותיהם. משתתף א' קיבל ציון שהוא 2.5 סטיות תקן מעל הממוצע ומשתתף ב' קיבל ציון שהוא 1 סטיית תקן מעל הממוצע. עורכי המחקר הציעו לכל משתתף לערוך מבחן חוזר.
האם כדאי למשתתף א' לנסות להיבחן שוב? הסבירו. מה באשר למשתתף ב'?

**שאלה VIII.**

קבעו לכל אמירה אם היא נכונה או אינה נכונה. נמקו היטב.
אם מקדם המתאם בין שני משתנים הוא 1, אזי:

א. עבור כל תצפית, הערך של משתנה אחד שווה לערך של המשתנה השני.

ב. יש קשר ליניארי חזק בין המשתנים.

ג. שני המשתנים מודדים את אותה תכונה.

**שאלה IX**.

קבעו לכל אמירה אם היא נכונה או אינה נכונה. נמקו היטב.

אם מקדם המתאם בין שני משתנים הוא 1-, אזי:

א. אין קשר בין המשתנים.

ב. יש קשר ליניארי חזק בין המשתנים.

ג. כשמשתנה אחד נוטה לרדת גם לשני יש נטייה לרדת.

ד. יש קשר בין שני המשתנים, אך אי אפשר לדעת באיזה סוג של קשר מדובר.

**שאלה X.**

קבעו לכל אמירה אם היא נכונה או אינה נכונה. נמקו היטב.

אם מקדם המתאם בין שני משתנים הוא 0, אזי:

א. אין קשר בין המשתנים.

ב. אין דרך להעריך את ערכו של המשתנה $Y$ על בסיס ערכו של המשתנה $X$.

ג. יש לשרטט דיאגרמת פיזור לפני שמנסים להסיק מסקנות.

**שאלה XI.**

במדגם רחב היקף שבדק ציוני בגרות של תלמידים, נמצא מתאם חיובי גבוה בין $X$ – הציון בלשון ובין $Y$ – הציון במתמטיקה. מכאן (נמקו):

א. קיימת נוסחה של קו ישר שאפשר לחשב באמצעותה את הציון בבגרות במתמטיקה על פי הציון בבגרות בלשון.

ב. תלמידי ישראל המצליחים יותר בלשון מצליחים יותר גם במתמטיקה.

ג. תלמיד שיתכונן היטב למבחן במתמטיקה, צפוי שהציון שלו בלשון יעלה.

ד. תלמיד ממוצע בלשון צפוי להיות תלמיד ממוצע במתמטיקה.

ה. תלמיד שהצטיין באופן מיוחד בלשון (מעל 95) צפוי להצטיין באופן מיוחד במתמטיקה.

**שאלה XII.**

קבעו לכל אמירה אם היא נכונה או אינה נכונה. נמקו היטב.

חוקר אקלים בחר באקראי 10 ימים בשנה, ומדד בכל יום את הטמפרטורות המקסימליות בטורונטו שבקנדה ובסידני שבאוסטרליה. החוקר חישב ומצא במדגם הימים מקדם מתאם שלילי בין הטמפרטורות בערים אלו. מכאן דווח בתקשורת:

א. פרסום א: אין קשר בין הטמפרטורה בטורונטו ובין הטמפרטורה בסידני.

ב. פרסום ב: עם עליית הטמפרטורות בסידני יש נטייה לירידה בטמפרטורה בטורונטו.

**שאלה XIII.**

במדגם זוגות של נתונים התברר שלשני המשתנים יש אותו ממוצע 0 ואותה שונות 1. הוכיחו או הפריכו את המסקנות הבאות:

א. $a = 0$ .

ב. מקדם המתאם בין המשתנים הוא 1.

**שאלה XIV.**

השלימו: לפי הפתגם "רחוק מהעין – רחוק מהלב", יש קשר ___ בין קִרבה פיזית לקִרבה נפשית.

א. חיובי

ב. שלילי

ג. אפסי

ד. אי אפשר לדעת.

**שאלה XV.**

מבחן אמי"ר הוא מבחן מיון באנגלית של המרכז הארצי לבחינות והערכה. הציון המינימלי בבחינה הוא 150 והמקסימלי 250. בקורס הכנה למבחן השתתפו 19 תלמידים. ממוצע הציונים שלהם היה 197.47 והשונות 536.25. התבקשתם להוסיף לטבלת הציונים של 19 התלמידים שתי עמודות: $X$ – כמה נקודות חסרות לכל תלמיד כדי להגיע לציון המקסימלי בבחינה; $Y$ – בכמה נקודות גבוה ציונו של כל תלמיד מהציון המינימלי בבחינה.

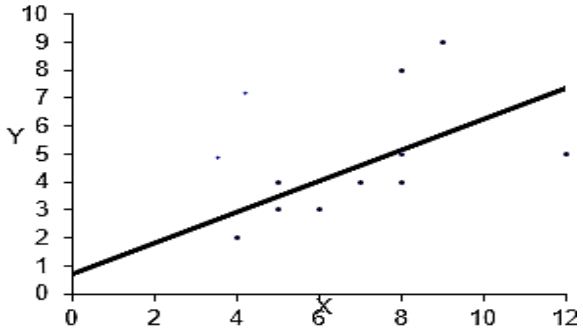
א. מהו הממוצע ומהי השונות של כל אחד מהמשתנים?

ב. מה תוכלו לומר על מקדם המתאם בין המשתנים $Y,X$ :

1, -0.5, 0.5, -1.

**שאלה XVI.**

בידיכם דיאגרמת פיזור המבוססת על 10 זוגות של תצפיות, בצירוף קו הרגרסיה המתאים.




$$\hat{y}_x = 0.69 + 0.56 \cdot x$$

א. שרטטו 'לפי העין' קווי ממוצעים. באילו רביעים נמצאות רוב הנקודות?

ב. מהו שיפוע הקו? מה השינוי הצפוי בערך הניבוי של $Y$ עם שינוי של שתי יחידות במשתנה $X$?

ג. בהנחה שהשונויות של המשתנים שוות פחות או יותר, העריכו מהו מקדם המתאם.

ד. אם למעשה השונות של $Y$ גדולה מעט יותר מזו של $X$, כיצד תשתנה ההערכה של מקדם המתאם בהשוואה להערכתם בסעיף הקודם?

**שאלה XVII.**

בכל סעיף ציינו נכון/לא נכון:

א. קו הריבועים הפחותים הוא הקו הישר שסכום ריבועי הסטיות (של ערכי $Y$) ממנו הוא מקסימלי.

ב. סכום הסטיות (של ערכי $Y$) מערכי קו הרגרסיה הוא 0.

ג. אם דיאגרמת פיזור מצביעה על חוסר קשר בין שני משתנים (פיזור אחיד), אין טעם להמשיך בניתוח.

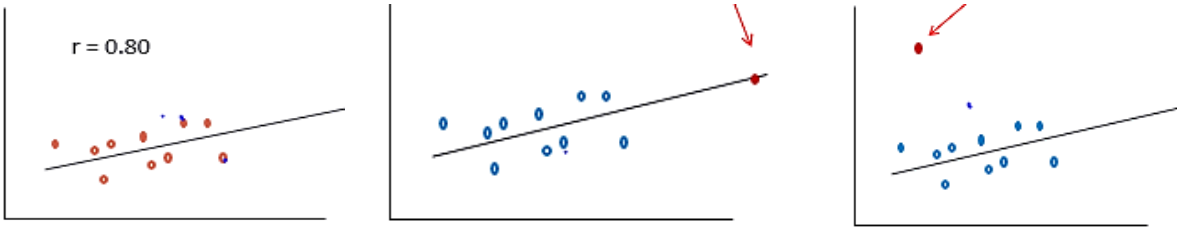
ד. אם מקדם המתאם בין המשתנים קרוב ל-0, פירוש הדבר שאין קשר בין המשתנים.

ה. אם מקדם המתאם בין המשתנים הוא 0, הקו הטוב ביותר לניבוי $Y$ על בסיס $X$ הוא $\hat{y}_x = \bar{y}$ (הממוצע).

ו. אם נמצא קשר קווי טוב בין $X$ ל-$Y$ כך שאפשר להשתמש בו לצורכי ניבוי ערכו העתידי של Y, אפשר גם להשפיע על הערך תוך כדי שינוי מתאים של ערכי $X$.

**שאלה XVIII.**

משמאל דיאגרמת פיזור עם מקדם מתאם 0.8. עם הוספת התצפית המסומנת בחיצים, האם המתאם יעלה, יישאר זהה או ירד?




**שאלה XIX.**

חוו דעתכם:

התבקשתם לחשב את מקדם המתאם בין גילו של המן הרשע בעת התרחשות מאורעות מגילת אסתר ובין הגילים של עשרת בניו באותה עת. מה תענו?

א. ברור שמקדם המתאם הוא שלילי שהרי מדובר בטיפוס שלילי.

ב. כדי לענות יש לדעת מה היו הגילים של בני המן באותה עת.

יש אפשרות אחרת?

**תשובות**

**I.** א. ציוני התקן עבור לילית הם: שעות השינה 1.09 (סטיית תקן אחת מעל לממוצע), מספר ההתעוררויות הוא 0.1 (קרוב לממוצע). השוואת ציוני התקן מורה שההורים עוד יצאו בחסד...

ב. הדרך להשוות מצב יחסי של פרט בשני המשתנים היא להשוות את ציוני התקן של אותו פרט.

במשימה VIII בפרק 3 חישבנו עבור ישראל: התמ"ג 1.23, ציון פיז"ה 0.13– . הפער גדול מאוד. לנוכח המתאם החזק בין המשתנים (0.8), ציון פיז"ה של ישראל במתמטיקה מאכזב מאוד.

**II.** מקדמי המתאם הם: א. חיובי חזק; ב. 0; ג. שלילי בינוני.

**III.**

א. רוב הנקודות נמצאות ברביעים השליליים, כלומר, הקשר יורד.

ב. מצב המשרות שבסכנה בישראל גרוע (45%), אבל אולי מעט פחות גרוע ממדינות אחרות ביחס לתמ"ג (התמ"ג של ישראל מתחת לממוצע ואילו אחוז המשרות שבסכנה קרוב לממוצע).

**IV.** א. ההסקה לאוכלוסייה איננה נכונה, נסחו טענה נכונה. למעשה אף אחת מהמסקנות אינה נכונה.

**V.** נתון, $b=-0.2$. על כן $r=-0.2$. על פי נוסחה (5) $0.5=\bar{y}+0.2\cdot 0.15$. לכן $\bar{y}=0.2$.

**VI.** 300. כל תוספת של 1 מטר לאורך מוסיפה בממוצע 300 קילוגרם למשקל.

**VII.** למשתתף א' לא כדאי להסתכן (אולי הצליח במזל...), למשתתף ב' כנראה כדאי לנסות.

**VIII.** ב ו-ג נכונים.

**IX.** ב נכון.

**X.** רק ג נכון.

**XI.** רק ד נכון.

**XII.** לא נכון. המדגם קטן מדי להסקה.

**XIII.** א. נכון. ב. דוגמה נגדית פשוטה: נבחר $X = -Y$.

**XIV.** חיובי.

**XV.** א. ממוצעים 52.53, 47.47, השונויות שתיהן 536.25; ב. 1-.

**XVI.** א. ילפי העיון הממוצע של $X$ הוא מעט פחות מ-7 ושל $Y$ הוא בערך 4.5. הקו עובר דרך נקודת הממוצעים. רוב הנקודות נמצאות ברביעים החיוביים; ב. 0.56, שינוי של 1.12 יחידות באותו כיוון; ג. 0.56; ד. הערכה של מקדם המתאם תקטן מעט על פי היחס.

**XVII.** א, ד, ו – לא נכון; ב, ג, ה – נכון.

**XVIII.** בדיאגרמה האמצעית המתאם יעלה, בימנית ירד.

**XIX.** א. חחח... הנתון אינו רלוונטי למסקנה; ב. למעשה אין צורך באינפורמציה זו, שהרי ממילא אי אפשר להפעיל את הנוסחה (נדרשים לפחות שני נתונים על כל אחד מהמשתנים כדי להפעיל את הנוסחאות).