

דיאגרמות פיזור ומדד בסיסי של קורלציה*

מאת: Derek W. Haylock

תרגום: נעמי רובינזון

דיאגרמות פיזור הן מקור להתנסויות מתמטיות חשובות עבור ילדים החל מגיל 11 בערך. השימוש בנתונים שנאספו מתוך כיתתם (גודל הכיתה הממוצעת מהווה אוכלוסיה נוחה מאוד לפעילות סטטיסטית בסיסית) יוצר מוטיבציה גבוהה אצל הילדים לגילוי קשר בין שני משתנים.

להלן דוגמאות לזוגות של משתנים שנחקרו לאחרונה באופן מוצלח בכיתות של תלמידים בני 11-12:

א. גובה במ"מ.

שטח כף הרגל בסמ"ר,

האם לאדם הגבוה ביותר, כף הרגל הגדולה ביותר?

דיאגרמת הפיזור מציגה קורלציה חיובית (ראה שרטוט 1).

רוב הנקודות מקובצות ברצועה אלכסונית עולה.

ב. אורך רגל במ"מ,

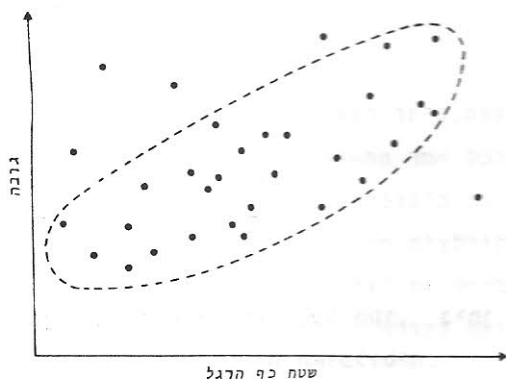
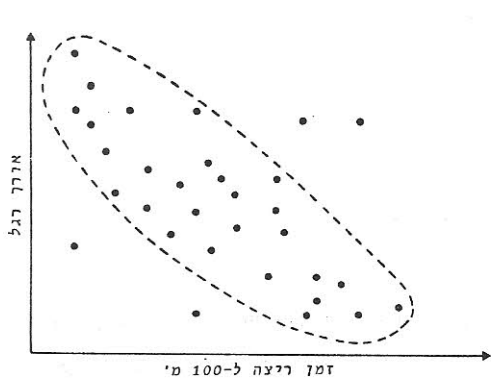
זמן הריצה למרחק של 100 מ' (הזמן נמדד בשניות בדיוק של ספרה אחת

לאחר הנקודה העשרונית),

האם האנשים בעלי הרגליים הארוכות ביותר, ירוצו בזמן הקצר ביותר?

דיאגרמת הפיזור מציגה קורלציה שלילית (ראה שרטוט 2).

רוב הנקודות מקובצות ברצועה אלכסונית יורדת.



*This article is taken with permission from Mathematics in School, Vol. 12 No. 1, January 1983.

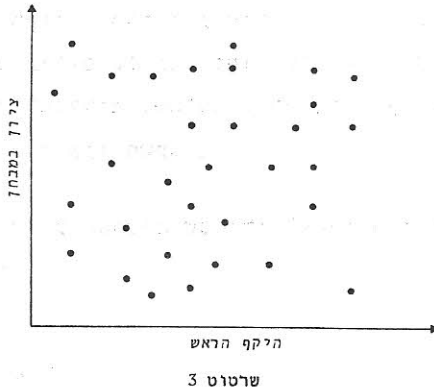
ג, גודל הראש (ההיקף נמדד במ"מ).

כמות השכל (נמדדת לפי ציון במבחן הבנה בחשבון).

האם האנשים בעלי הראשים הגדולים ביותר, יהיו חכמים יותר?

צריך לטפל בכך מאוד בזהירות. בשרטוט 3 אין קורלציה נראית לעין.

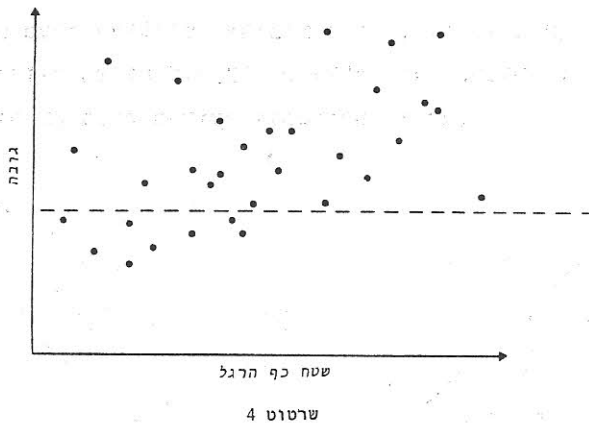
הנקודות מפוזרות ואין ריכוז ברצועה כלשהי.



דרך פשוטה אך יעילה לניתוח דיאגרמות אלו היא כדלהלן:

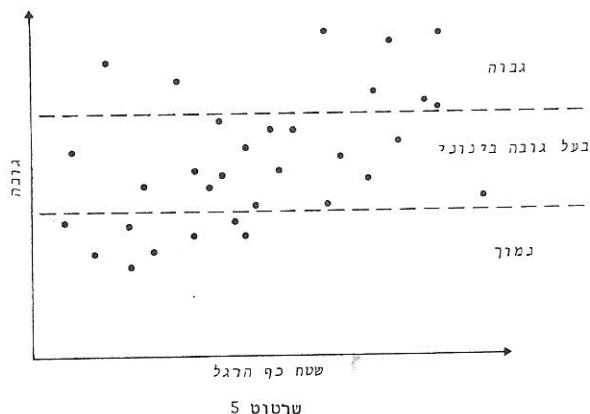
ראשית, מניחים סרגל במאוזן לרוחב הגרף ומזיזים אותו עד אשר רבע מהנקודות של הדיאגרמה נמצאות מתחתיו.

משרטטים קו כמו בשרטוט 4, לא תמיד ניתן לשרטט קו כך שבדיוק רבע מהנקודות יתרכזו מתחתיו, אך נשתדל להתקרב לכך ככל שנוכל.

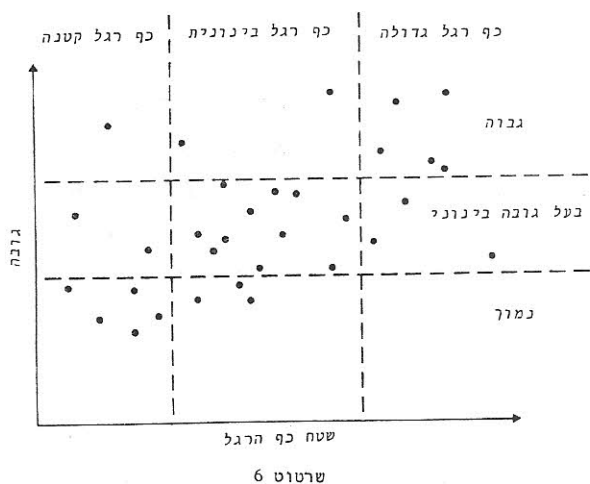


משרטטים קו נוסף באופן דומה, כך שכרבע מהנקודות יימצאו מעל הקו. ניתן לערוך כאן דיון עם הילדים, כיצד מחלקים הקווים הללו את האוכלוסיה.

בדוגמה שבשרטוט 5, האוכלוסיה מתחלקת באופן ברור ל-3 קבוצות: "גבוה", "בעל גובה בינוני", "נמוך".

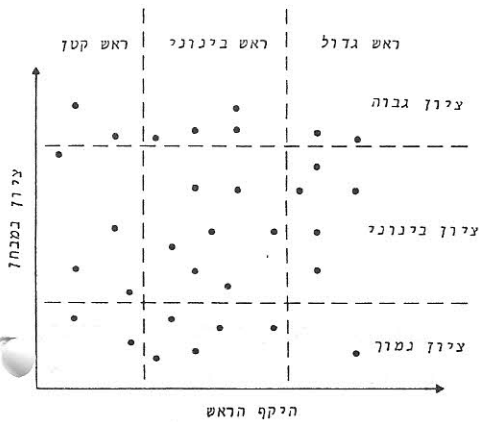


בנקודה זו, אם רצונך בכך, ניתן לדון גם ברבעונים, לאחר מכן נחזור על כך במשתנה האחר. ניצור שוב שלוש קבוצות, שבדוגמא שלנו הן: "כף רגל גדולה", "כף רגל בינונית", "כף רגל קטנה". באופן כזה, חילקנו את הדיאגרמה ל-9 איזורים (שרטוט 6).



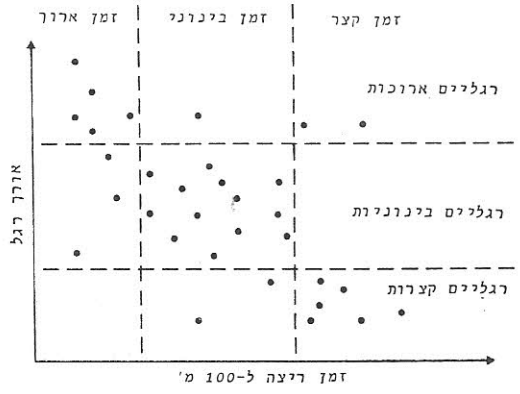
עתה, ניתן לבקש מהילדים לתאר את האנשים שהנקודות המתאימות להם מופיעות בכל אחד מהאיזורים: אנשים גבוהים בעלי רגל גדולה, אנשים גבוהים בעלי רגל בינונית, וכן הלאה. בדרך פשוטה זו, ניתן לראות באופן ברור שישנה קורלציה חיובית בין שני המשתנים, במקרה זה, שכן, יש מספר רב של נקודות במחלקות "גבוה-גבוה", "נמוך-נמוך", נקודה אחת בלבד במחלקה "נמוך-גבוה" ואף נקודה במחלקה "גבוה-נמוך".

מנסיוני נוכחתי כי ילדים מסוגלים ללמוד מהר מאוד כיצד לנתח דיאגרמות בדרך כזו וליישם את מימצאיהם בתבונה, שרטוטים 7 ו-8 מראים כיצד ניתן לחלק את הדיאגרמות בשרטוטים 2 ו-3 בדרך דומה.



היקף הראש

שרטוט 8



זמן ריצה ל-100 מ'

שרטוט 7

באחרונה, פיתחתי דרך פרימיטיבית למדידה של דרגת הקשר בין שני משתנים, הצומחת מן השיטה של ניתוח דיאגרמות אשר תוארה קודם לכן. שיטה זו יוצרת מדד הדומה מאוד למקדם הקורלציה הקונבנציונלי, מדד זה נראה מתאים לתחושה "האינטואיטיבית" לגבי דיאגרמת פיזור, הוא ממלא את הדרישה להיות בין (-1) (עבור קורלציה שלילית לחלוטין) ו-(+1) (עבור קורלציה חיובית לחלוטין), אך בו בזמן הוא קל לחישוב, כך שילדים צעירים מאוד יכולים להשתמש בו ולרכוש נסיון מוקדם במדידת קורלציה.

הצעד הראשון בפיתוח המדד הזה הוא רישום מספר הנקודות הנמצאות בכל אחד מתשעת האיזורים של דיאגרמת הפיזור המחולקת, אם נעשה זאת לגבי הדוגמאות שבשרטוטים 1, 2 ו-3, נקבל את התוצאות הבאות:

מקרה 1:	1	2	5	מקרה 2:	5	1	2	מקרה 3:	2	4	2
	2	11	3		3	13	0		4	7	5
	5	3	0		0	2	6		2	5	1

כאן המקום לשאול - אילו תוצאות נקבל במקרים המיוחדים, כלומר, במקרה של קורלציה חיובית מושלמת, קורלציה שלילית מושלמת או קורלציה אפס מושלמת?

כיון שבכל שורה ובכל טור חייבת להימצא רבע מהאוכלוסיה, אזי עבור אוכלוסיה של 32, נגיע לתוצאות הבאות:

מקרה 4:	0	0	8	מקרה 5:	0	0	8	מקרה 6:	2	4	2
	0	16	0		0	16	0		4	8	4
	8	0	0		0	0	8		2	4	2

(יש עוד אפשרויות)

נשאלת השאלה – כיצד ניתן להשתמש בניתוחים אלה כדי למדוד בדרך פשוטה את דרגת הקשר בין שני משתנים? ברור כי אנו רוצים שמקרה 4 יתן מדד השווה ל (+1), מקרה 5 יתן מדד של (-1) ומקרה 6 יתן מדד השווה ל 0. התכונה הבולטת ביותר של הקורלציה החיובית (מקרים 1 ו 4) היא כי ישנו מספר גדול של נקודות באיזורים "גבוה-גבוה" ו"נמוך-נמוך", ומעט נקודות או בכלל לא, באיזורים "נמוך-גבוה" ו"גבוה-נמוך", באופן דומה, בקורלציה השלילית (מקרים 2 ו 5), הנקודות מרוכזות באיזורים "נמוך-גבוה" ו"גבוה-נמוך", יותר מאשר באיזורים "גבוה-גבוה" ו"נמוך-נמוך". לכן, בשיטה זו ניתן (+1) לכל נקודה באיזורים "גבוה-גבוה" ו"נמוך-נמוך" ו (-1) לכל נקודה באיזורים "גבוה-נמוך" ו"נמוך-גבוה".

להלן התוצאות שהתקבלו מניקוד זה, עבור ששת המקרים שעסקנו בהם:

$$\begin{aligned} \text{מקרה 4:} & \quad (8 + 8) - (0 + 0) = 16 \\ \text{מקרה 1:} & \quad (5 + 5) - (1 + 0) = 9 \\ \text{מקרה 3:} & \quad (2 + 2) - (2 + 1) = 1 \\ \text{מקרה 6:} & \quad (2 + 2) - (2 + 2) = 0 \\ \text{מקרה 2:} & \quad (2 + 0) - (5 + 6) = -9 \\ \text{מקרה 5:} & \quad (0 + 0) - (8 + 8) = -16 \end{aligned}$$

רואים, אם כן, כי שיטת ניקוד זו "מארגנת" את המקרים הנ"ל לפי מידת הקשר שחשנו בקיומו באופן אינטואיטיבי.

עתה, כדי להמיר את הניקוד למדד, יש לקחת בחשבון את גודל האוכלוסיה. כיון שבמקרה בו עסקנו, גודל האוכלוסיה הוא 32, הניקוד הגבוה ביותר האפשרי הוא (+16) והנמוך ביותר (-16). ברור כי אם נחלק תוצאות אלה ב 16, נקבל תוצאות בין (+1) ל (-1), כפי שנדרש. בדרך זו, נקבל את הערכים הבאים עבור "מדד הקורלציה" הבסיסי שלנו:

מקרה 4 :	$16/16 = 1$
מקרה 1 :	$9/21 \approx 0,4$
מקרה 3 :	$1/24 = 0,04$
מקרה 6 :	$0/16 = 0$
מקרה 2 :	$-9/9 \approx 0.5$
מקרה 5 :	$-16/16 = -1$

לסיכום, אם הדיאגרמה מחולקת לחלקים כפי שהוסבר לעיל, והמספרים של הנקודות בכל איזור מסומנים כדלהלן:

A	B	C
D	E	F
G	H	J

אזי, מדד הקורלציה יחושב כך:
$$\frac{(C + G) - (A + J)}{(N - E)}$$

(כאשר N הוא גודל האוכלוסיה).

ברוב המקרים שניסיתי את מדד הקורלציה הזה, במדגמים של 30 או יותר, ולגבי התפלגויות "נורמליות" עבור שני המשתנים, הוא נתגלה כהערכה טובה למדי (לעיתים קרובות עד כדי 0.1), של מקדם הקורלציה הקונבנציונלי. (לשם השוואה, מקדמי הקורלציה למקרים 1, 3 ו 2 לעיל הם: 0.49, 0.08 ו -0.53 בהתאמה). אך הנקודה החשובה ביותר היא שמדד זה "מתנהג" בדיוק כמו מקדם הקורלציה, שכן הוא מצוי בין (+1) ו (-1) ונותן תוצאות גבוהות יותר כאשר באופן אינטואיטיבי הקשר נראה חזק יותר, כמו כן, הילדים מבינים מה הם עושים עם המספרים: הניקוד של (+1) ו (-1) נראה הגיוני וכן גם החלוקה במספר מסוים כדי לקבל את המדד הסופי. כמו כן, כמובן, אין חישובים מסובכים לבצע.

מורים רבים לגיאוגרפיה ולמדע, לדוגמא, משתמשים עתה בדיאגרמות פיזור כדי להדגים קורלציות בין כמה מהמשתנים החשובים במקצועותיהם. חישוב מקדם הקורלציה של Spearman, אותו מתבקשים תלמידים לחשב לעיתים קרובות, הוא מסובך. מסיבה זו עוסקים בדרך כלל במדגמים שאינם גדולים מ 10. לדעתי, מדד הקורלציה שעסקנו בו במאמר זה, מתאים לצרכיהם של מורים במקצועות שונים, ונותן לתלמידים נסיון בעל ערך רב במדידת קורלציה, בלי חישובים מסובכים ומייגעים.

מאמר דומה מאת אותו מחבר, בשם A Simplified Approach to Correlation, הופיע בגליון מאי 83 של Mathematics Teacher.

במאמר זה, מדד הקורלציה מחושב באופן הבא: $\frac{(G + C) - (A + J)}{N/2}$

כאשר N הוא גודל האוכלוסיה,

בדרך זו מתקבל מדד קצת שונה מהמדד שהוצג לעיל במאמר, אך מאוד קרוב אליו.