סטטיסטיקה 3- שיעור 4

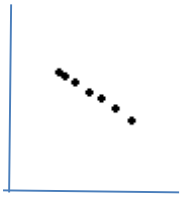
קשר לינארי

מטרות:

1.זיהוי הקשר

2. איתור ערכים קיצוניים

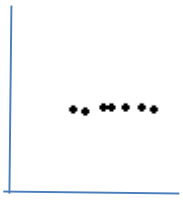
קשר מושלם חיובי



קשר מושלם שלילי

קשר חיובי חלקי

קשר שלילי חלקי



אין קשר

ערכים קיצוניים- במסגרת הקורס לא נוציא ערכים אלו.

נוסחאת פרסון- ההגדרה:

$$r_{(x,y)} = \frac{\mathrm{cov}(x,y)}{\hat{s}_x \hat{s}_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\hat{s}_x \hat{s}_y}$$

Rp יכול להיות בין (-1) ל 1 : (בערך מוחלט)

קשר חלש	עד 0.3
בינוני	0.3-0.6
קשר גבוה	0.6-1

$$cov(x,y) = \frac{\sum(xi - \bar{x}) * (yi - \bar{y})}{n-1}$$

Cov(x,x)= שונות של x

נעדיף לבחון את הקשר על פי R של פירסון, כיוון בה COV אנו יכולים לבחון האם יש קשר, אך אנו לא יודעים לתרגם את גודל הקשר למידת העוצמה (נקבל אותה במספרים ולא באחוזים) לעומת R של פירסון.

טרנספורמציות:

1. הוספה/ החסרה של קבוע לx/y או לשניהם- אינה משנה את מקדם המתאם rp
2. הכפלה/ חילוק פי קבוע של x/y או של שניהם- אינה משנה את מקדם המתאים rp

- טרנפורמציות לינאריות של משנות את rp, טרנספורמציות לא לינאריות משנות את rp (כמו למשל העלאה בריבוע)

Rp מבטא קשר סימטרי במובן ש xעל y ו yעל x ייתן תוצאה זהה.

הסקה לאוכלוסיה: אוכלוסיה

סוגי השערות: **ρ**- אוכ' , r- מדגם

חד כיוונית חיובית:

$$H_0 : \rho = 0$$

$$H_1 : \rho > 0$$

חד כיוונית שלילית:

$$H_0 : \rho = 0$$

$$H_1 : \rho < 0$$

דו כיוונית:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

הנחות:

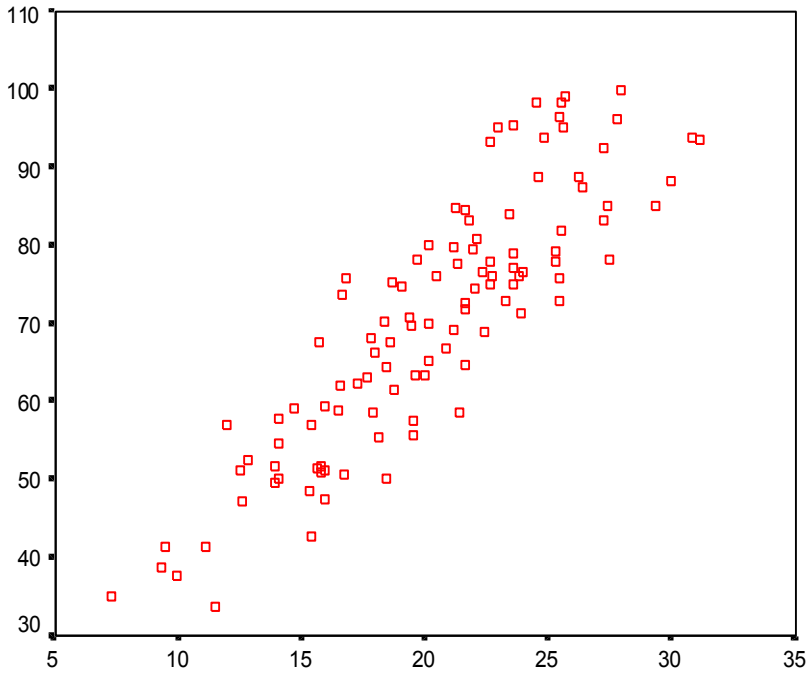
1. דגימה מקרית של תצפיות
2. התפלגות דו נורמלית- x מתפלג נורמלית עבור כל ערך של y
3. y מתפלג נורמלית עבור כל ערך של x

התפלגות דו נורמלית נראית כצורה של קובע על מערכת הצירים.

דיאגרמות פיזור ומתאמים

א. קשר חיובי חזק

Graph



Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
HOURS No. of learning hours before exam	20.1754	5.03946	109
MARK Test mark	70.1530	16.22455	109

Correlations[b]

		No. of learning hours before exam	Test mark
No. of learning hours before exam	Pearson Correlation	1	.880**
	Sig. (2-tailed)		.000
Test mark	Pearson Correlation	.880**	1
	Sig. (2-tailed)	.000	

0.88 קשר חיובי חזק

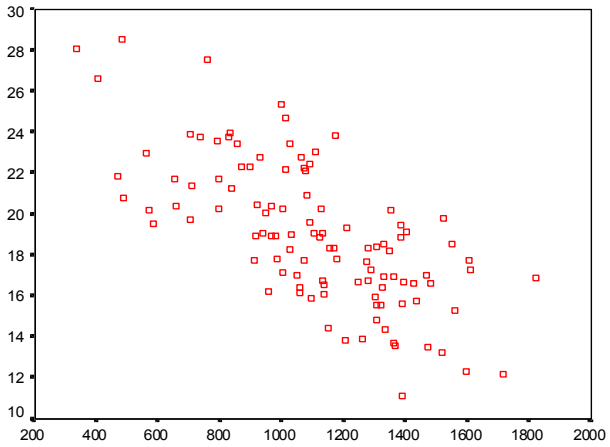
**. Correlation is significant at the 0.01 level (2-tailed).

b. Listwise N=109

Sig<0.05

ולכן נדחה H0- כלומר מתקיים קשר לינארי באוכלוסייה.

ב. קשר שלילי חזק



Graph

ככול שנפח המנוע גדול יותר, נסע פחות ק"מ לליטר (צורכים יותר דלק)

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
ENGINE Volume (cubic CM)	1101.791	300.050	109
FUEL Consumption (KM per Litter)	19.028	3.499	109

Correlations[b]

		Volume (cubic CM)	Consumption (KM per Litter)
Volume (cubic CM)	Pearson Correlation	1	-.704**
	Sig. (2-tailed)		.000
Consumption (KM per Litter)	Pearson Correlation	-.704**	1
	Sig. (2-tailed)	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

b. Listwise N=109

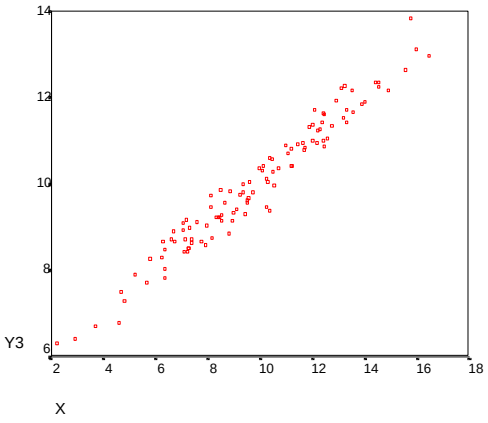
קשר שלילי חזק

נקודה חשובה: לא יודעים מהי העוצמה באוכ' , אלא רק במדגם.

ניתן לראות כי מתקיים קשר שלילי באוכ' בין נפח המנוע ולק"מ לליטר.

ג. מתאמים בעוצמות שונות וכיצד הם מתבטאים בדיאגרמת הפיזור

Graph



Correlations[a]

		X	Y3
X	Pearson Correlation	1	.979**
	Sig. (2-tailed)	.	.000
Y3	Pearson Correlation	.979**	1
	Sig. (2-tailed)	.000	.

**. Correlation is significant at the 0.01 level

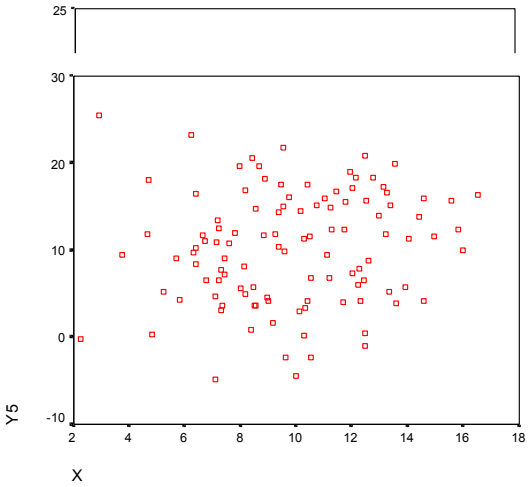
a. Listwise N=109

Correlations[a]

		X	Y4
X	Pearson Correlation	1	.518**
	Sig. (2-tailed)	.	.000
Y4	Pearson Correlation	.518**	1
	Sig. (2-tailed)	.000	.

**. Correlation is significant at the 0.01 level

a. Listwise N=109

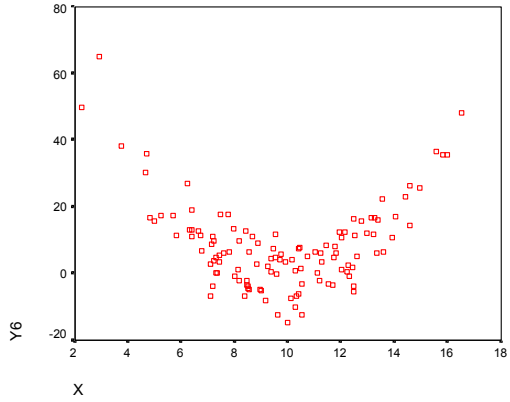


Correlations[a]

		X	Y5
X	Pearson Correlation	1	.122
	Sig. (2-tailed)	.	.207
Y5	Pearson Correlation	.122	1
	Sig. (2-tailed)	.207	.

a. Listwise N=109

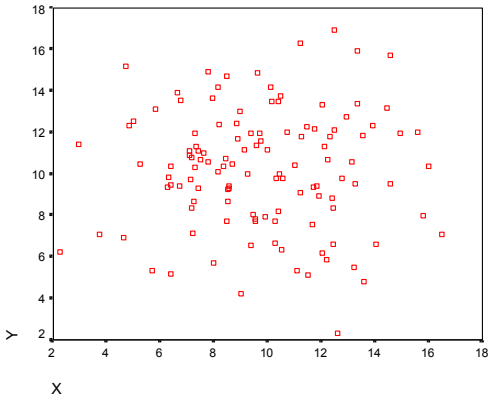
קשר שאינו ליניארי (פרבולי)



Correlations

		X	Y6
X	Pearson Correlation	1	-.043
	Sig. (2-tailed)	.	.639
	N	120	120
Y6	Pearson Correlation	-.043	1
	Sig. (2-tailed)	.639	.
	N	120	120

ד. חוסר קשר



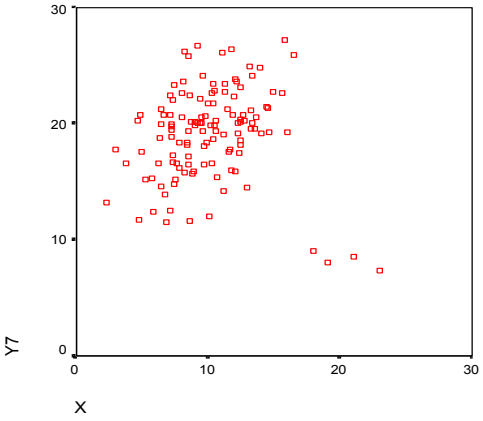
Correlations[a]

		X	Y
X	Pearson Correlation	1	.000
	Sig. (2-tailed)	.	.998
Y	Pearson Correlation	.000	1
	Sig. (2-tailed)	.998	.

a. Listwise N=120

ו. השפעות של ערכים קיצוניים

- **דוגמא לערכים קיצוניים המחלישים את הקשר**



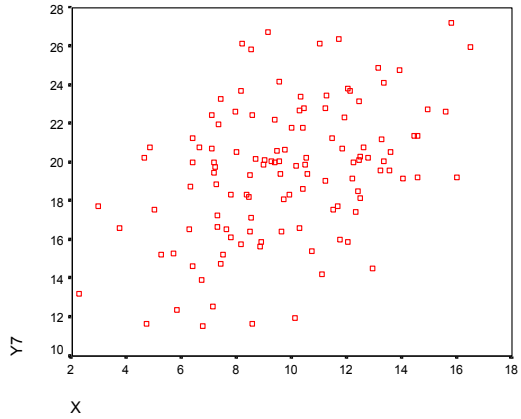
Correlations[a]

		X	Y7
X	Pearson Correlation	1	.023
	Sig. (2-tailed)	.	.799
Y7	Pearson Correlation	.023	1
	Sig. (2-tailed)	.799	.

a. Listwise N=124

כמעט ואין קשר

<u>אותם נתונים לאחר השמטת 4 תצפיות שבניגוד למגמה</u>



Correlations[a]

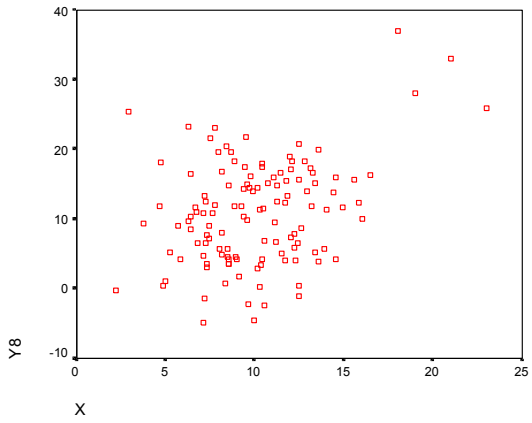
		X	Y7
X	Pearson Correlation	1	.413**
	Sig. (2-tailed)	.	.000
Y7	Pearson Correlation	.413**	1
	Sig. (2-tailed)	.000	.

**. Correlation is significant at the 0.01 level

a. Listwise N=120

קשר בינוני

- **דוגמא לערכים קיצוניים המחזקים את הקשר**

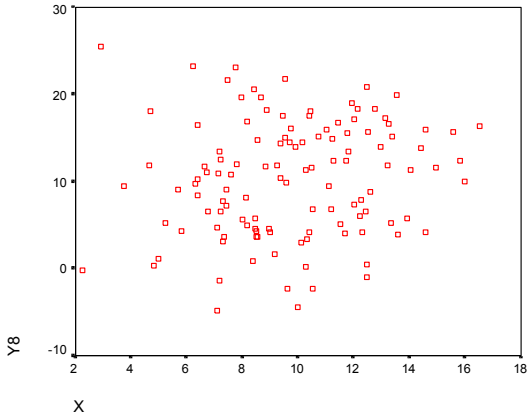


Correlations[a]

		X	Y8
X	Pearson Correlation	1	.350**
	Sig. (2-tailed)	.	.000
Y8	Pearson Correlation	.350**	1
	Sig. (2-tailed)	.000	.

**. Correlation is significant at the 0.01 level

a. Listwise N=124



<u>אותם נתונים לאחר השמטת 4 תצפיות שגורמות למגמה</u>

Correlations[a]

		X	Y8
X	Pearson Correlation	1	.128
	Sig. (2-tailed)	.	.165
Y8	Pearson Correlation	.128	1
	Sig. (2-tailed)	.165	.

a. Listwise N=120

<u>רגרסיה לינארית פשוטה</u>

משתנה ב"ת- אינסוף קטגוריות- משתנה כמותי

משתנה תלוי- כמותי

עקום הרגרסיה- קו באוכלוסייה העובר דרך התוחלות של Y המחושבות לכל ערך קבוע של Xi
קיימים אינסוף עקומי רגרסיה אפשריים- אנו נלמד על עקום רגרסיה 1- קו ישר.

$E(y/x=xi)=\mu yi$

מהו הקו הטוב ביותר? עקרון ריבועים פחותים LSE

$\min \sum \widehat{e^2}$ סטייה/טעות.

הקו באוכלוסיה- $\mu yi=\alpha+\beta xi$

הקו במדגם- $\hat{y}i = \hat{\alpha} + \hat{\beta}xi$

ערך אמיתי באוכ'- $y_i = \alpha + \beta x_i + \varepsilon_i$

ערך אמיתי במדגם- $\hat{y}i = \hat{\alpha} + \hat{\beta}xi + \hat{e}i$

$$\hat{\beta} = \frac{\text{cov}(x,y)}{\hat{s}_x^2} = r \cdot \frac{\hat{s}_y}{\hat{s}_x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

אנו נעבוד עם הקו של y, לא של x.

קשר לינארי, לרגרסיה פשוטה, המתאם הלינארי של פירסון, הסקה לאוכלוסייה, מתאם קווי מדגם, מדגם, ניתוח פלט, עקום הרגרסיה