Στατιστική ΙΙ

Γιώργος Τσιρογιάννης

Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων,

Πανεπιστήμιο Πατρών

Ακαδημαϊκό έτος: 2021-2022

Διάλεξη 2η

Συσχέτιση

6ο κεφάλαιο

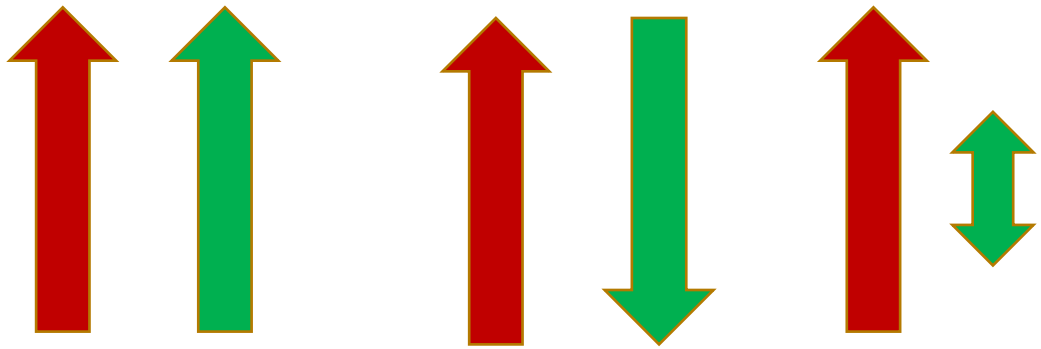
Συσχέτιση

- Υπάρχει σχέση μεταξύ του IQ των παιδιών και του εισοδήματος μιας οικογένειας;
- Κάποιος απόφοιτος με μεγαλύτερο βαθμό πτυχίου αμοίβεται καλύτερα τον 5ο χρόνο της δουλειάς του;
- Υπάρχει η έννοια του ζεύγους παρατηρήσεων: «IQ-εισόδημα», «Βαθμός, μισθός»

Συσχέτιση (παράδειγμα)

	Έστειλε	Έλαβε
Αντρι	5	10
Μαικ	7	12
Ντορις	13	14
Στιβ	9	18
Τζον	1	6

Ισχύει ότι όποιος στέλνει περισσότερες κάρτες λαμβάνει και περισσότερες;



Θετική σχέση

	Έστειλε	Έλαβε
Φ1	13	14
Φ2	9	18
Φ3	7	12
Φ4	5	10
Φ5	1	6

διατεταγμένο

Αρνητική σχέση

	Έστειλε	Έλαβε
Φ1	13	6
Φ2	9	8
Φ3	7	10
Φ4	5	11
Φ5	1	17

διατεταγμένο

Αντίθετα διατεταγμένο

Μικρή ή καμία σχέση

	Έστειλε	Έλαβε
Φ1	13	10
Φ2	9	18
Φ3	7	12
Φ4	5	16
Φ5	1	14

διατεταγμένο

Μη διατεταγμένο

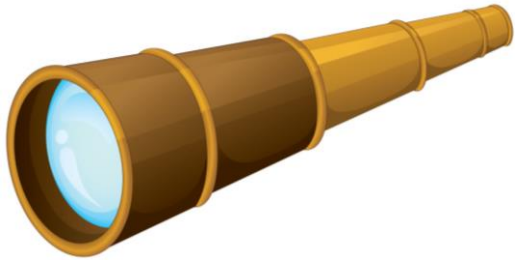
Θετική, αρνητική ή μη σχέση (γραμμική)

- Δύο μεταβλητές σχετίζονται θετικά μεταξύ τους αν τα ζέυγη των παρατηρήσεων/αποτελεσμάτων τείνουν να κατέχουν παρόμοιες σχετικές θέσεις (δηλαδή παρατηρούμε ζεύγη υψηλών, ζεύγη μεσαίων και χαμηλών) στις αντίστοιχες κατανομές τους.
- Αρνητικά όταν τα ζεύγη τείνουν κατέχουν αντίθετες θέσεις (δηλαδή παρατηρούμε υψηλές τιμές με χαμηλές, ζεύγη μεσαίων και χαμηλών με υψηλές) στις αντίστοιχες κατανομές τους.
- Διαφορετικά λέμε ότι δεν υπάρχει σχέση

Ερωτήσεις

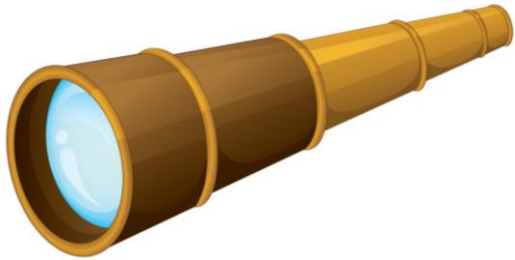
- Είναι αρνητική, θετική ή σχέση;
- Οι μαθητές που παρακολουθούν τηλεόραση, βαθμός τριμήνου
- Το βάρος των αυτοκινήτων, χιλιόμετρα που διανύει με 10 λίτρα καυσίμου
- Οι άνθρωποι με καλύτερη εκπαίδευση απολαμβάνουν καλύτερα εισοδήματα

Διαγράμματα διασποράς



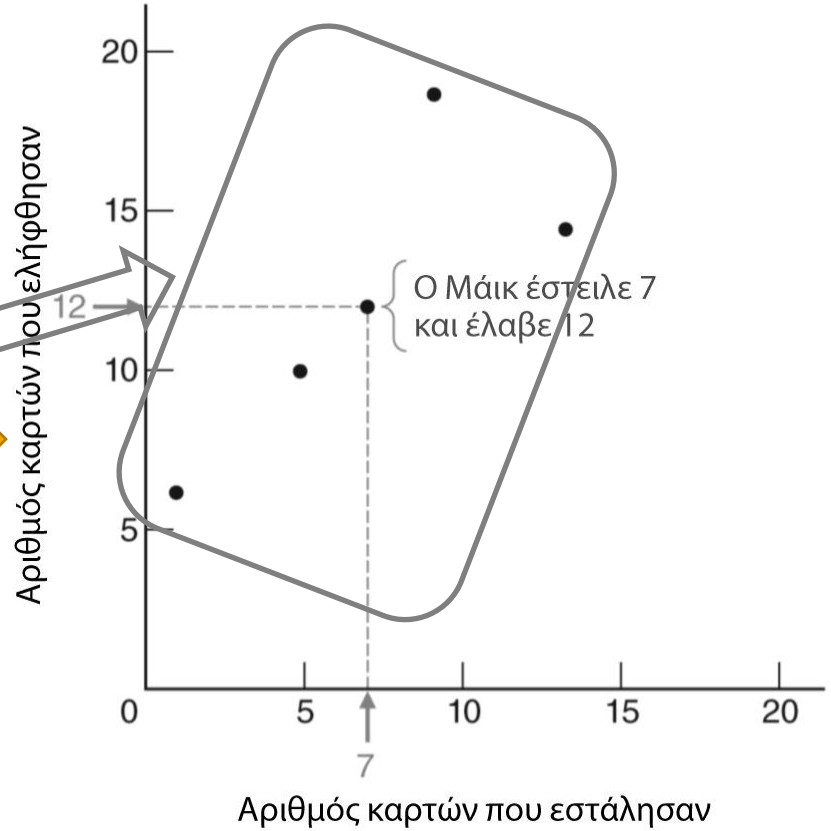
- Πρόκειται για σημειακή απεικόνιση των ζευγών παρατηρήσεων.
- Η κλίμακα των αξόνων μπορεί να διαφέρει αφού πολλές φορές έχουν πολύ διαφορετική φυσική σημασία

Διαγράμματα διασποράς

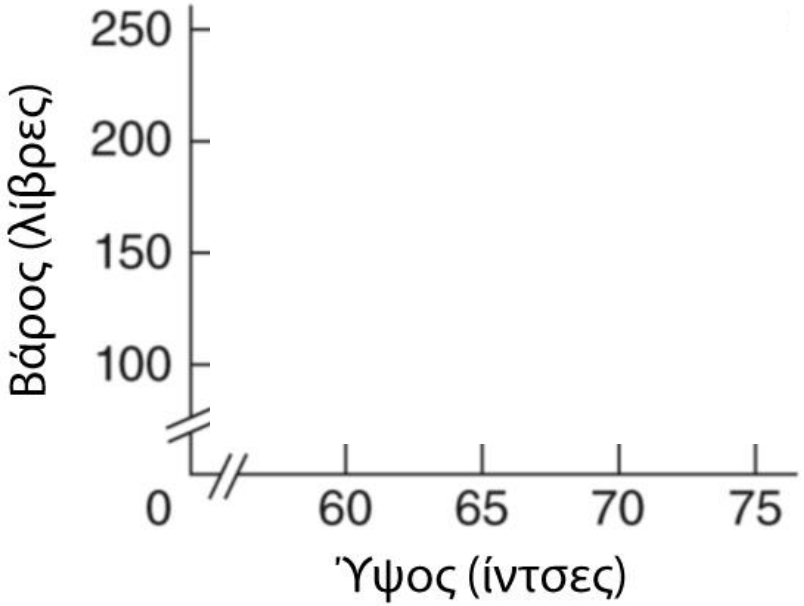


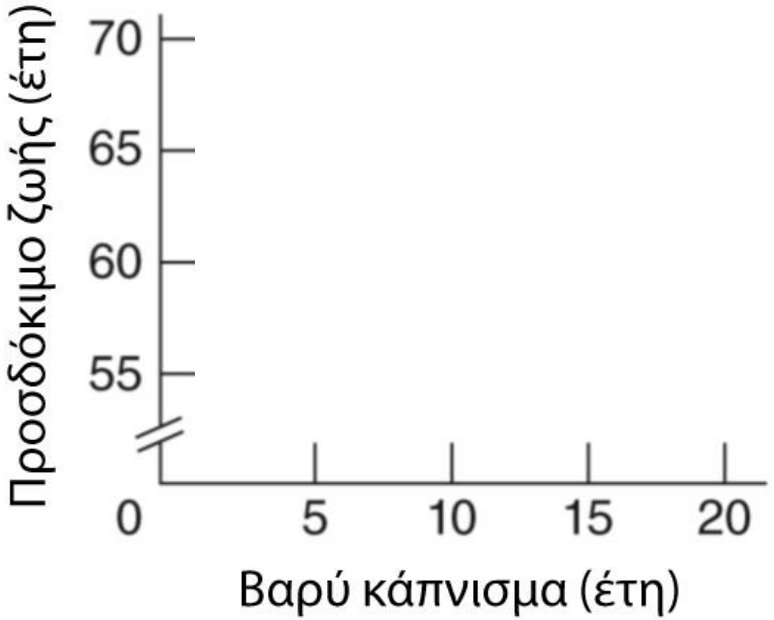
Αριθμός Καρτών

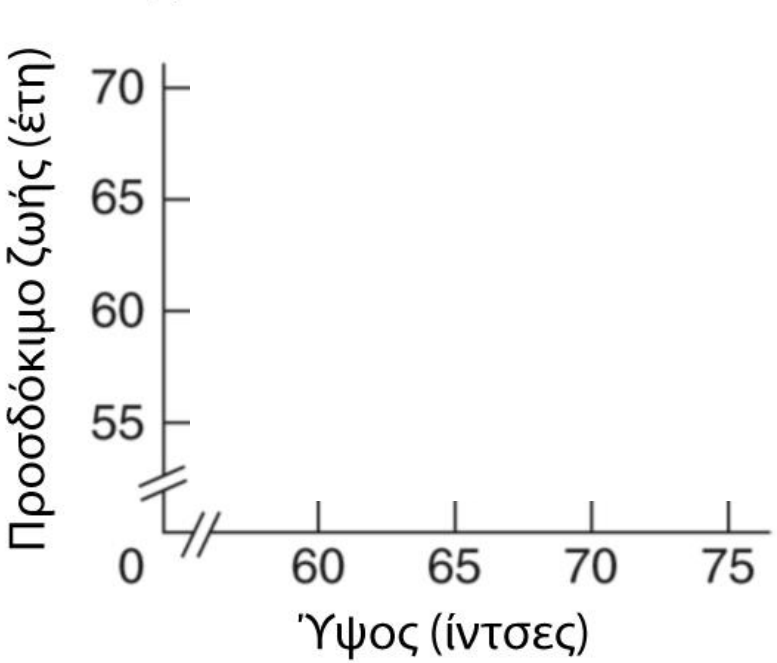
	Έστειλε	Έλαβε
Αντρι	5	10
Μαικ	7	12
Ντορις	13	14
Στιβ	9	18
Τζον	1	6



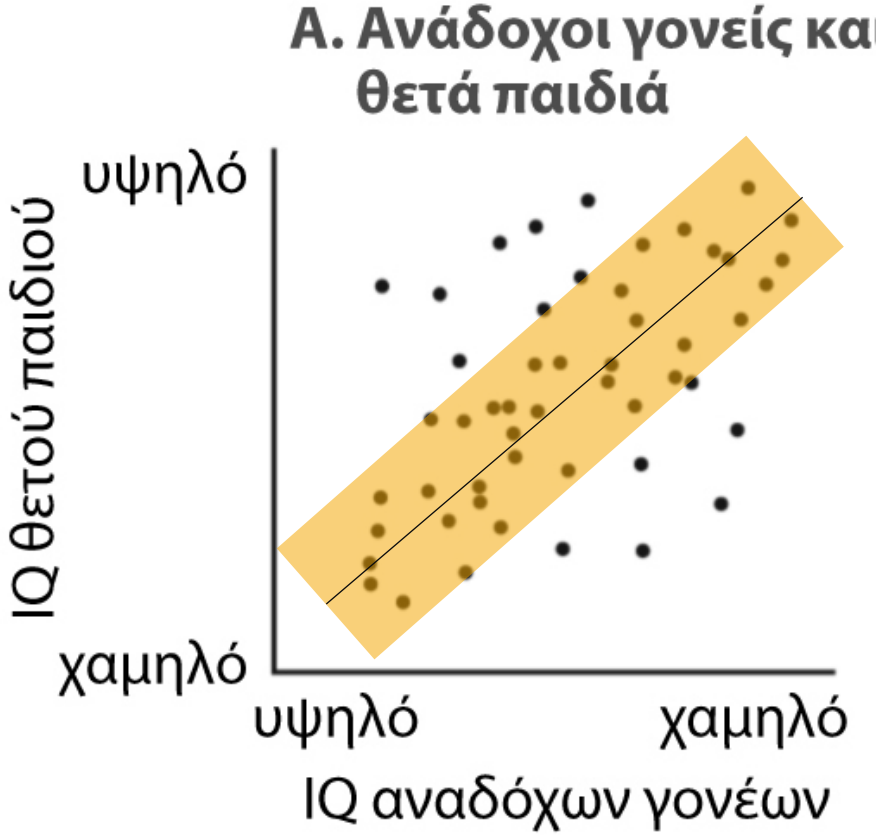
Διαγράμματα διασποράς
Θετική, αρνητική ή μη σχέση



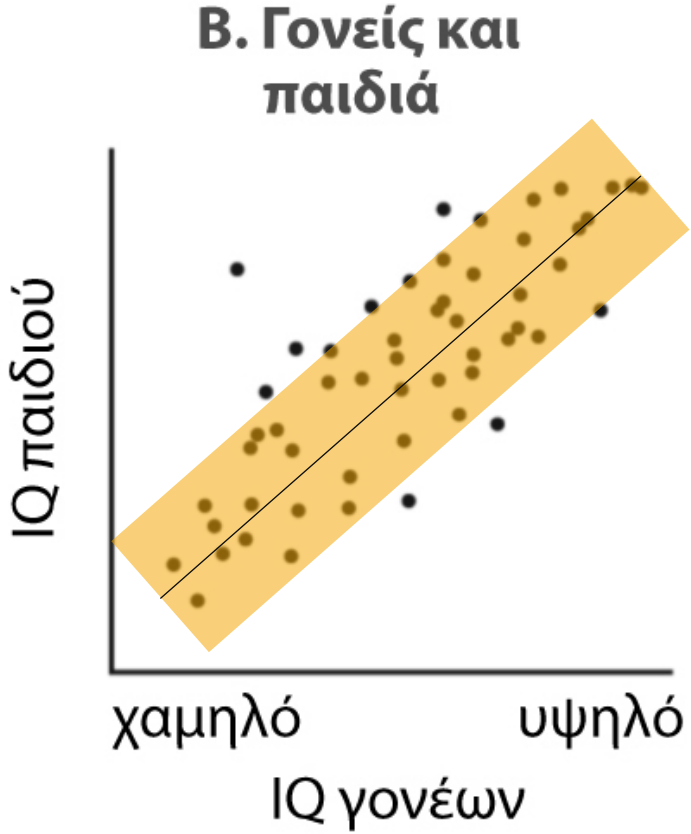




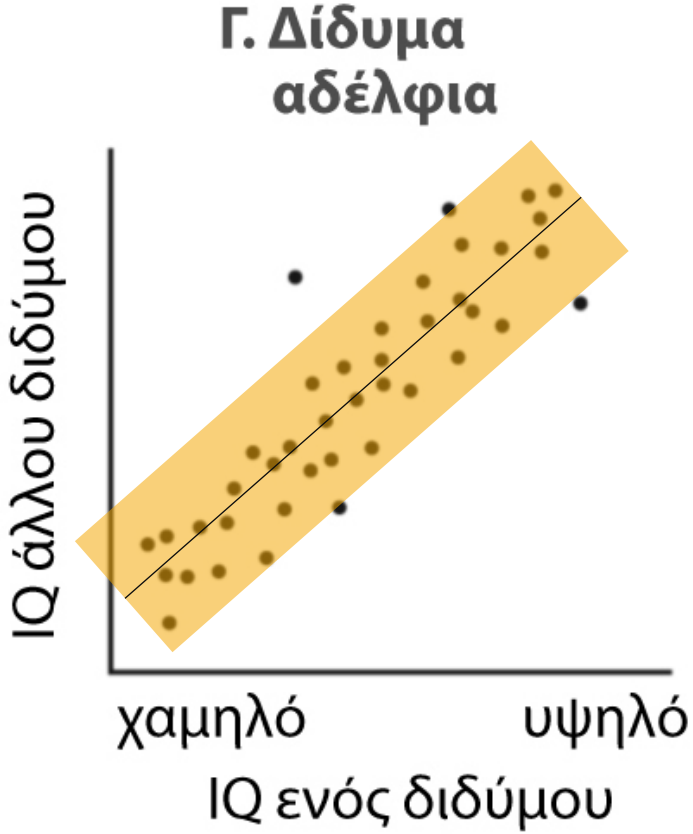
Ισχυρή vs ασθενής σχέση



Ασθενής

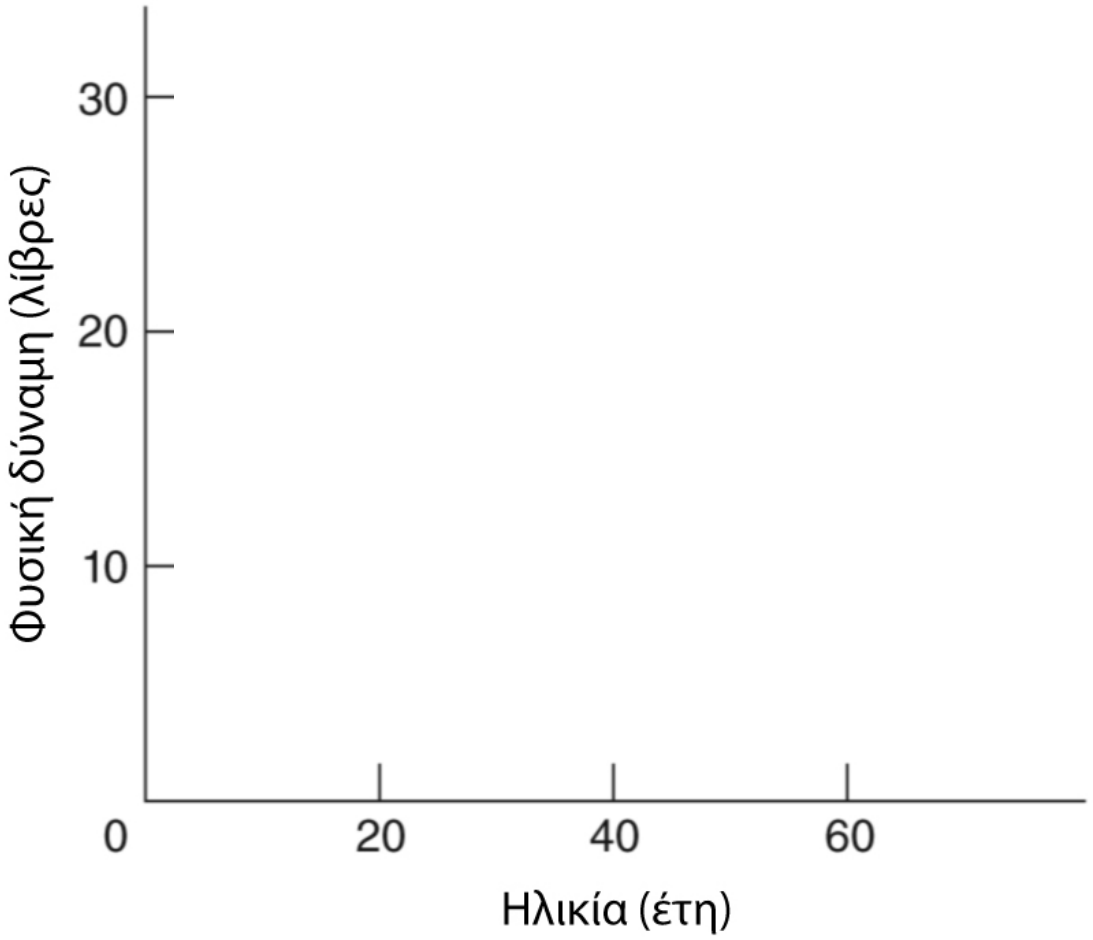


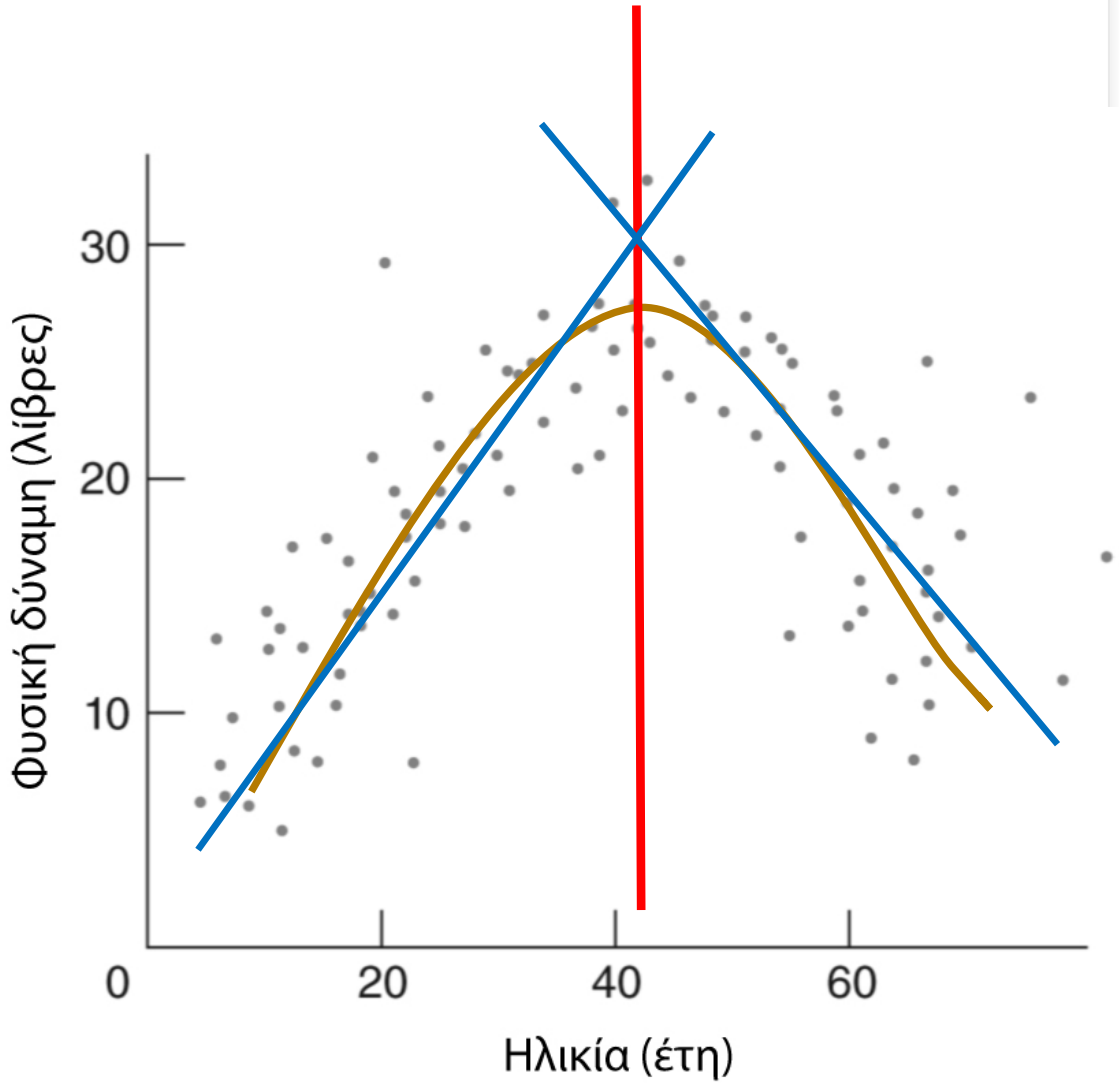
Μέτρια



Ισχυρή

Μη γραμμικές σχέσεις
Καμπυλόγραμμες σχέσεις





Συντελεστής συσχέτισης ποσοτικών δεδομένων

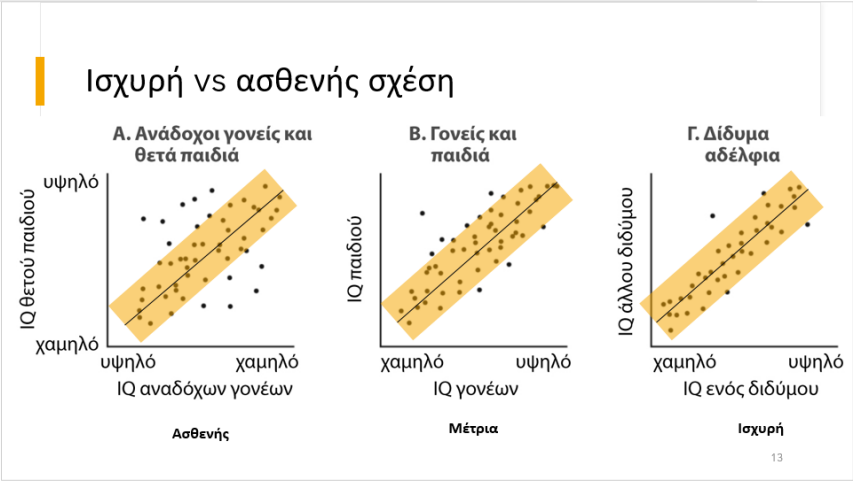
- Είναι ένας αριθμός στο διάστημα [-1,1] ο οποίος περιγράφει ποσοτικά την σχέση (γραμμική) των μεταβλητών
- Συνήθως συμβολίζεται με **r**
- Όταν αναφέρουμε απλά συσχέτιση θα εννοούμε την Pearson correlation coefficient
- Υπάρχουν και άλλα είδη συσχέτισης πχ
  - Spearman
  - Kendall

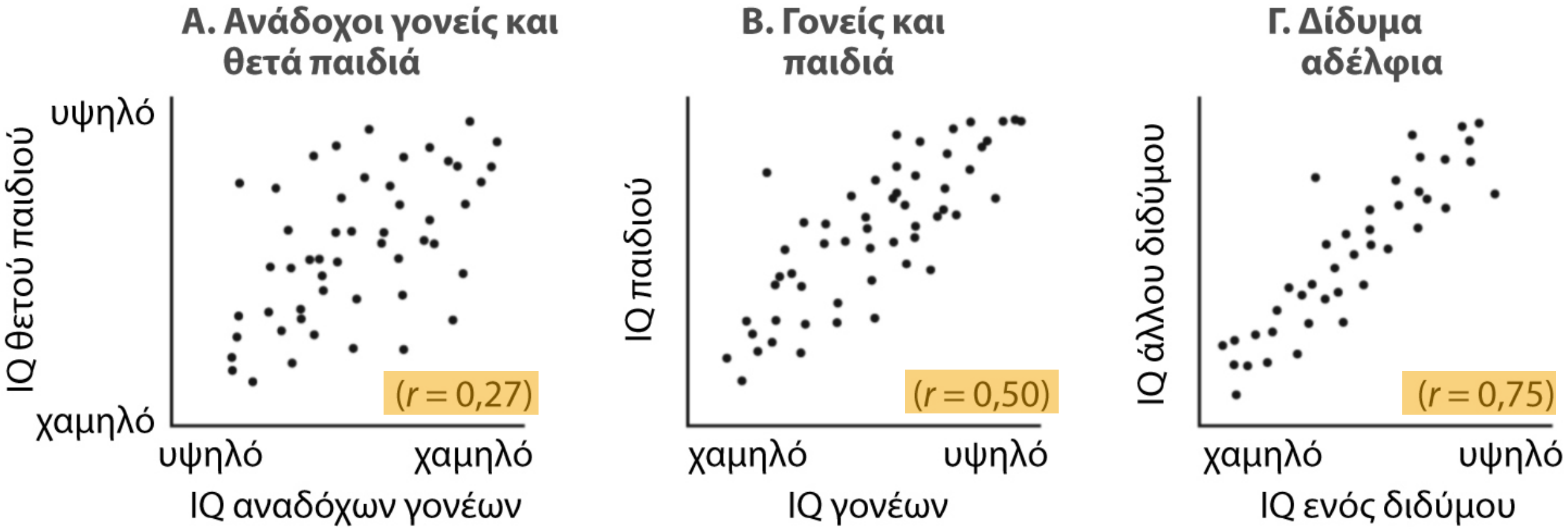
Carl Pearson (27 March 1857 – 27 April 1936)

Ιδιότητες του r

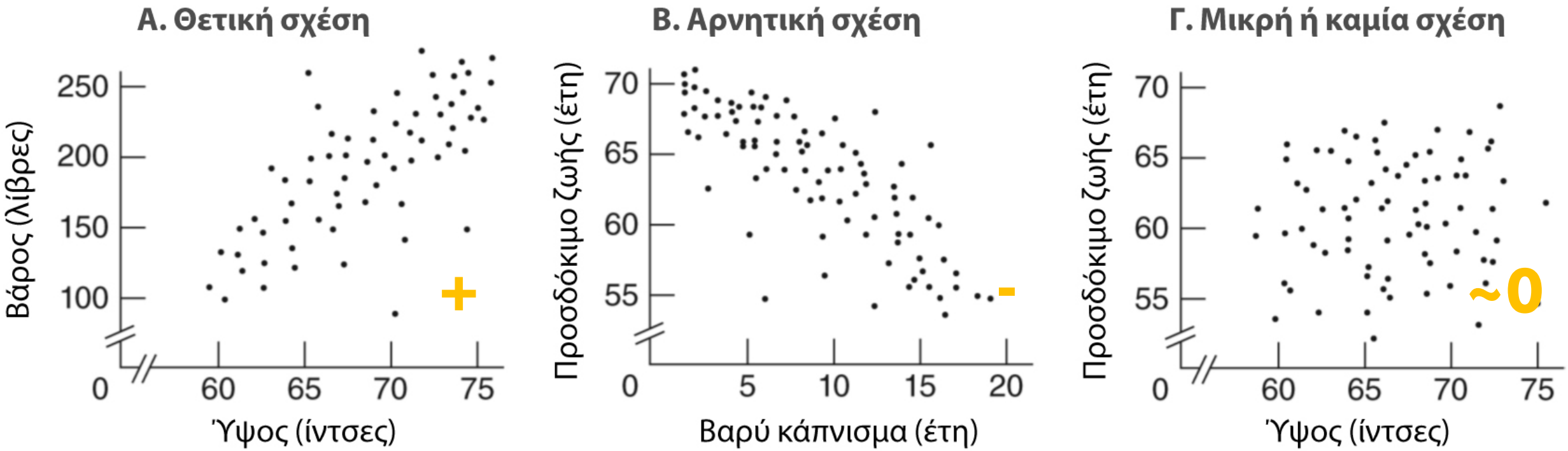
- Παίρνει τιμές στο [-1,1]
- Το πρόσημο δηλώνει αρνητική ή θετική σχέση
- Η απόλυτη τιμή υποδηλώνει ισχυρή ή ασθενή σχέση

Παράδειγμα


Ισχυρή vs ασθενής σχέση
A. Ανάδοχοι γονείς και θετά παιδιά
υψηλό
χαμηλό
IQ θετού παιδιού
υψηλό
χαμηλό
IQ αναδόχων γονέων
Ασθενής
B. Γονείς και παιδιά
IQ παιδιού
χαμηλό
υψηλό
IQ γονέων
Μέτρια
Γ. Δίδυμα αδέλφια
IQ άλλου διδύμου
χαμηλό
υψηλό
IQ ενός διδύμου
Ισχυρή
13


A. Ανάδοχοι γονείς και θετά παιδιά
υψηλό
χαμηλό
IQ θετού παιδιού
(r = 0,27)
υψηλό
χαμηλό
IQ αναδόχων γονέων
B. Γονείς και παιδιά
IQ παιδιού
(r = 0,50)
χαμηλό
υψηλό
IQ γονέων
Γ. Δίδυμα αδέλφια
IQ άλλου διδύμου
(r = 0,75)
χαμηλό
υψηλό
IQ ενός διδύμου

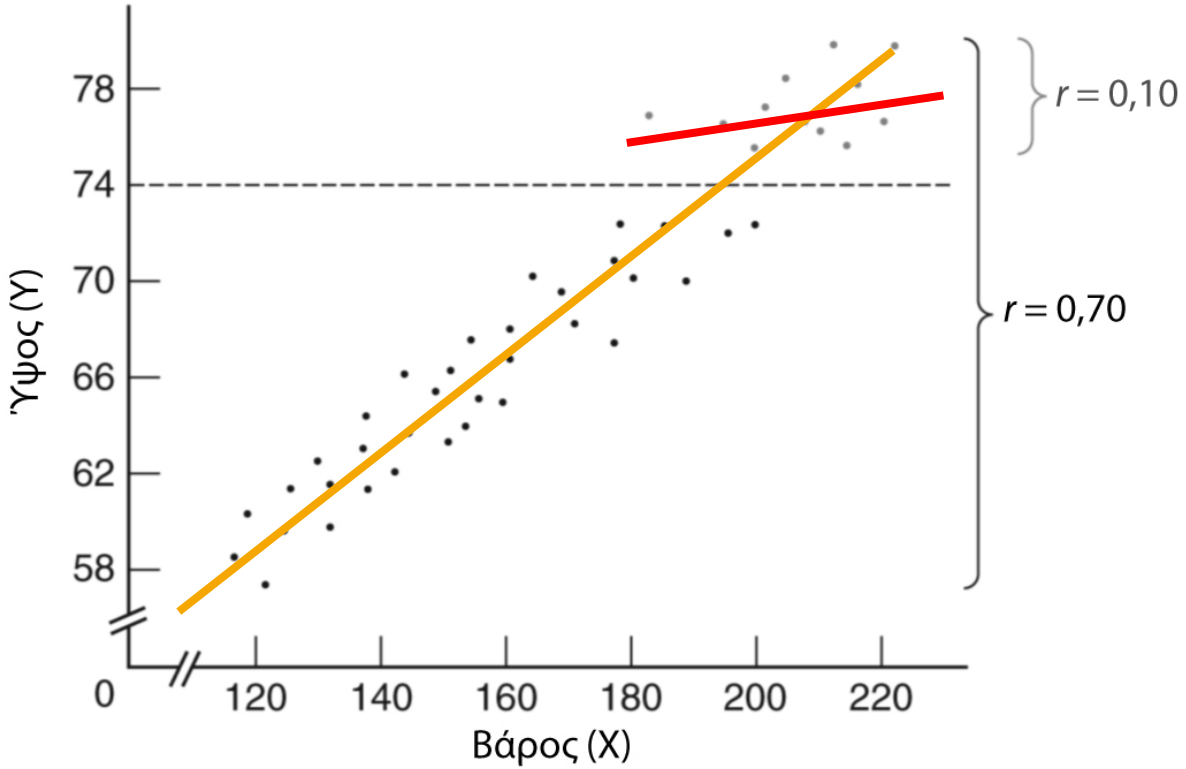
Παράδειγμα (πρόσημο)



Ερμηνεία της αριθμητικής τιμής του r

- Δηλώνει το είδος της σχέσης
- Δηλώνει την ισχύ της σχέσης
- Τι συμβαίνει όταν γενικεύουμε πέρα από το δείγμα;
  - Αν το θεωρήσουμε προϊόν τυχαίας μεταβλητότητας δειγματοληψίας, τότε η τιμή του r πρέπει να υπολογίζεται με μεθόδους επαγωγικής στατιστικής
  - Ο αριθμός των σημείων πρέπει να είναι αρκετά μεγάλος πχ εκατοντάδες
  - Ισχυρές τιμές δεν εγγυόνται πάντα πραγματικά ισχυρές σχέσεις (ο τρόπος οργάνωσης του πειράματος παίζει ρόλο, πχ ίσως πήραν την εξέταση 2 φορές και r=0.8 δεν είναι τόσο ισχυρό όσο δείχνει)

Περιορισμοί διαστημάτων



Τι δεν είναι το r

- Έστω ότι μια τιμή του r=0.7 που συνδέει ύψος και το βάρος των ανδρών
- Δεν πρέπει να ερμηνεύεται ότι πχ το βάρος κατά 70% ερμηνεύει την ισχύ της σχέσης με το ύψος
- Δηλαδή το r δεν δηλώνει ποσοστό ή ισχύ μιας τέλειας γραμμικής σχέσης
- Μια πιθανή προφορική ερμηνεία είναι: «οι ψηλότεροι άνδρες τείνουν να ζυγίζουν περισσότερο».
- Η συσχέτισης δεν είναι απαραίτητα σχέση αιτίου και αποτελέσματος

Ερωτήσεις

- Σχολιάστε:
  - r=-0.84, μεταξύ της τιμής και χιλιάδων χιλιομέτρων ενός αυτοκίνητου
  - r=-0.35, μεταξύ αριθμού ημερών απουσιών και επίδοσης σε διαγώνισμα
  - r=0.03, μεταξύ επιπέδου άγχους και βαθμού πτυχίου
  - r=0.56, μεταξύ ηλικία παιδιών πρωτοβάθμιας εκπαίδευσης και ικανότητας κατανόησης κειμένου

Πως υπολογίζουμε το r;

Συνδυασμός 2 μεταβλητών

$$SP_{xy} = \sum\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}$$

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

$$SS_x = \sum\left(X - \bar{X}\right)^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$$

Άθροισμα τετραγώνων απόστασης από τον μέσο

$$SS_y = \sum\left(Y - \bar{Y}\right)^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

Άθροισμα τετραγώνων απόστασης από τον μέσο

Πως υπολογίζουμε το r;

- Βρίσκουμε ποσό είναι το n (1)
- Υπολογισμός των αθροισμάτων για τις μεταβλητές X (2) και *Y* (3)
- Υπολογισμός των γινομένων κάθε ζεύγους (4) και το άθροισμά τους (5)
- Υπολογισμός των τετραγώνων των X (6), και το άθροισμά τους (7)
- Υπολογισμός των τετραγώνων των Y (8), και το άθροισμά τους (9)
- Κάνουμε χρήση των (1-9) για τον υπολογισμό των SS_x, SS_y, SP_{xy}
- Υπολογισμός του r

Παράδειγμα

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} \qquad SP_{xy} = \sum\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}$$

$$SS_x = \sum\left(X - \bar{X}\right)^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n} \qquad SS_y = \sum\left(Y - \bar{Y}\right)^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

CARDS

FRIEND	SENT, X	RECEIVED, Y
Doris	13	14
Steve	9	18
Mike	7	12
Andrea	5	10
John	1	6

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

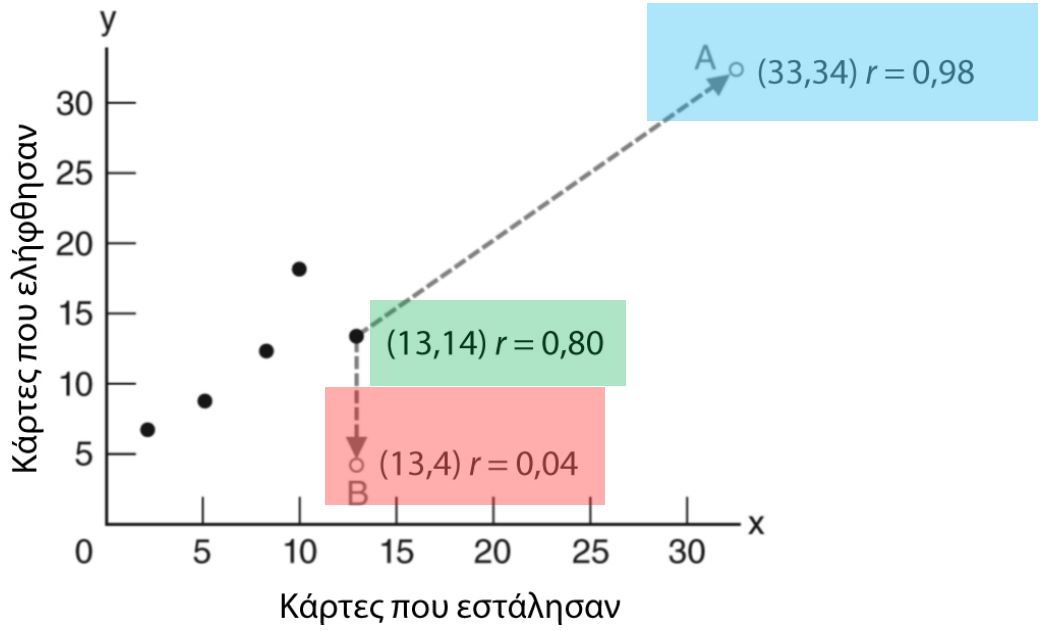
$$SS_x = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

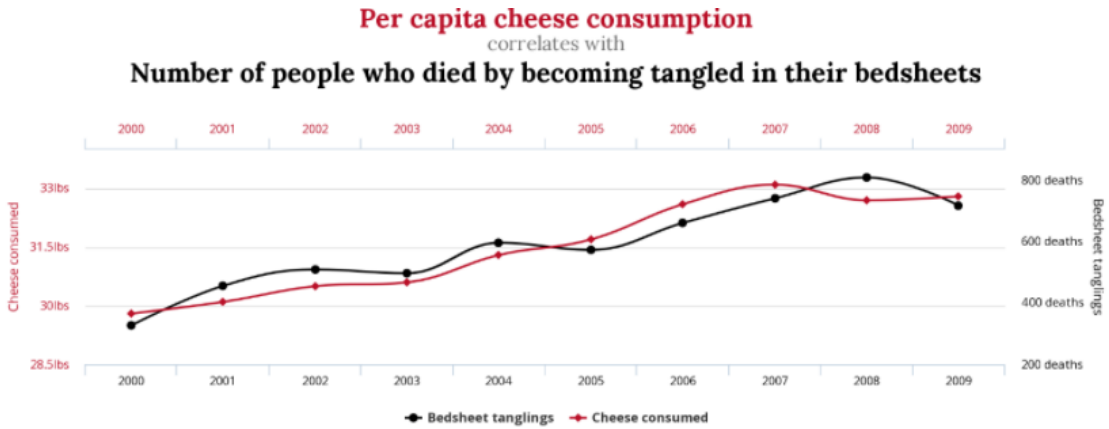
11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

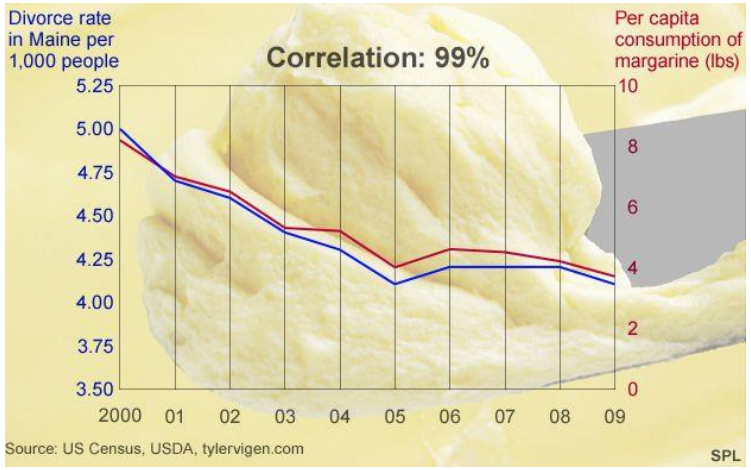
Ο ρόλος των ακραίων τιμών

- Ακραίες τιμές: τα πολύ ακραία αποτελέσματα/μετρήσεις τα οποία χρήζουν ειδικής προσοχής εξαιτίας της επίπτωσης που ενδεχομένως μπορούν να έχουν στην στατιστική ανάλυση.

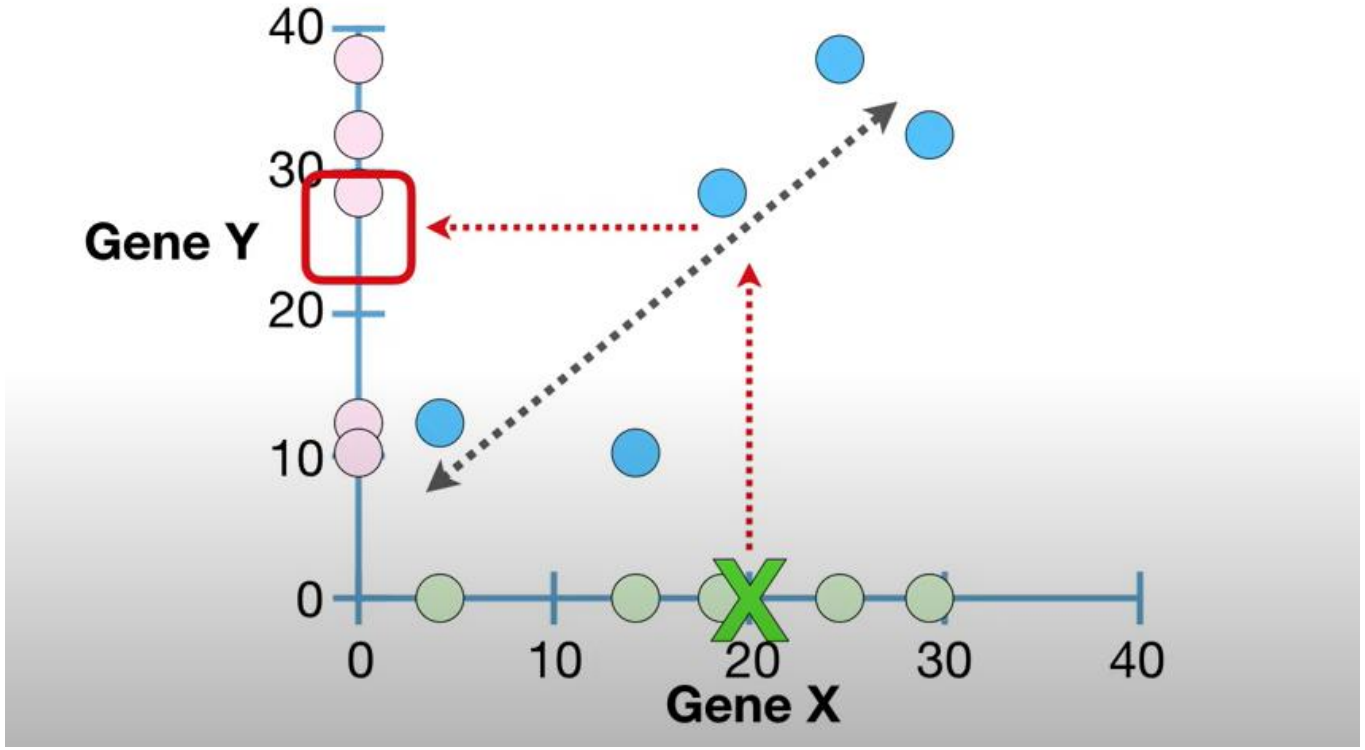


Ψευδοσυσχετίσεις
Spurious correlations





Online σχετικό υλικό



https://www.youtube.com/watch?v=xZ_z8KWkhXE

Πρακτικό κομμάτι στην γλώσσα R

Διαδικαστικό τμήμα: φόρτωση των δεδομένων

```
library(readxl)
FinalData <- read_excel("../FinalData.xlsx")
attach(FinalData)
View(FinalData)
str(FinalData)
```

Υπολογισμός των συσχετίσεων

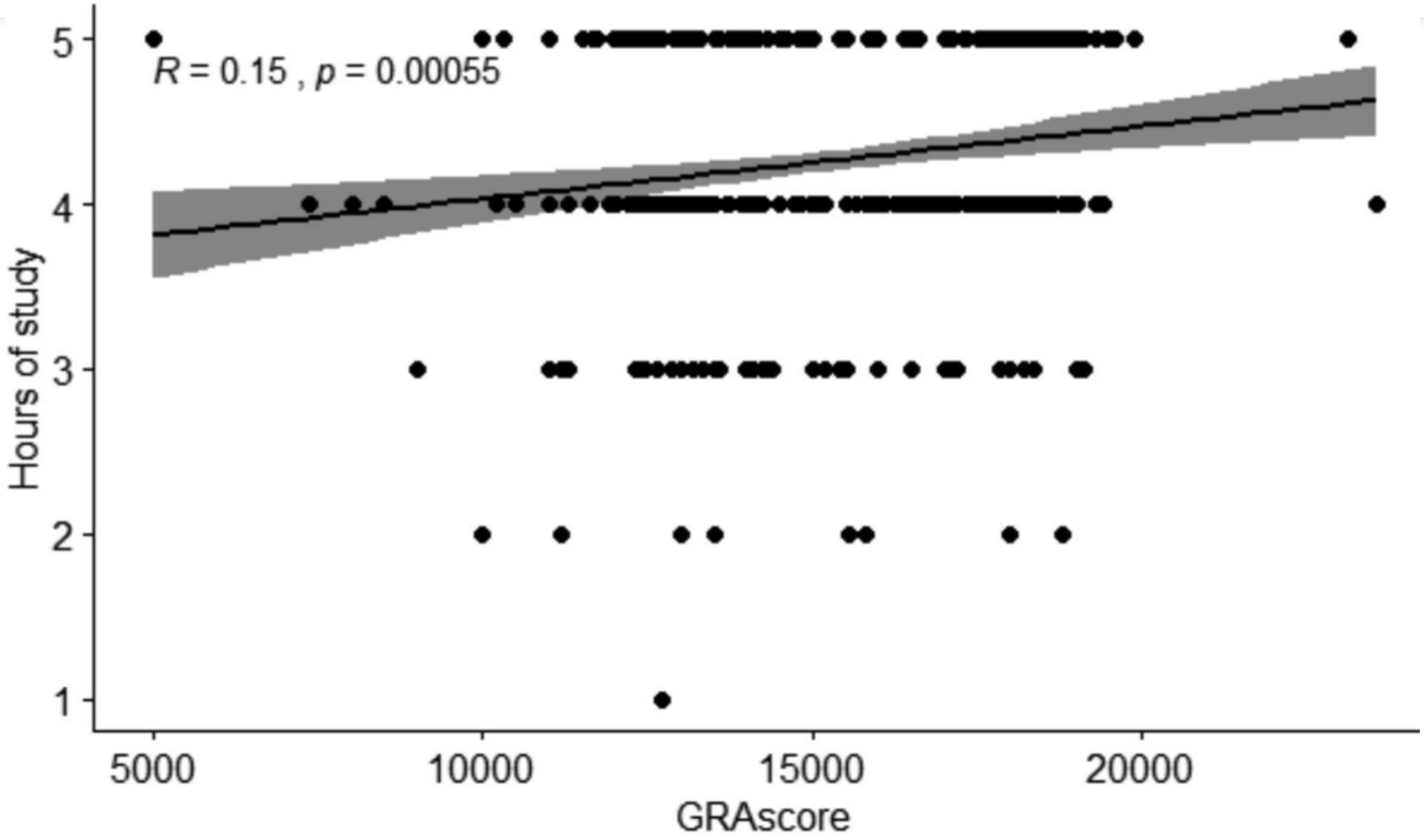
```
cor(FinalData$score, FinalData$workload, method = "pearson",use = "complete.obs")
cor(FinalData$score, FinalData$workload, method = "spearman",use = "complete.obs")
cor(FinalData$score, FinalData$workload, method = "kendal",use = "complete.obs")
```

Αποτέλεσμα:
0.1461
0.1311
0.1051

Στατιστική μελέτη των συντελεστών συσχέτισης

```
#Ελεγχος υπόθεσης #
res1 <- cor.test(FinalData$score, FinalData$workload, method = "pearson")
res1
res2<-cor.test(FinalData$score, FinalData$workload, method = "spearman")
res2
res3<-cor.test(FinalData$score, FinalData$workload, method = "kendal")
res3
```

Γραφική αναπαράσταση του αποτελέσματος



Γραφική αναπαράσταση του αποτελέσματος

```
## Γραφική απεικόνιση ##
install.packages("ggpubr")
library("ggpubr")
ggscatter(FinalData, x = "score", y = "workload",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "GRAscore", ylab = "Hours of study")
```