Συσχέτιση μεταξύ ποσοτικών χαρακτηριστικών

Γιώτα Τουλούμη
Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
gtouloum@med.uoa.gr

Βάνα Σύψα
Καθηγήτρια Επιδημιολογίας και Ιατρικής Στατιστικής
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
vsipsa@med.uoa.gr

Στατιστική

Περιγραφική στατιστική:
Συνοπτική παρουσίαση
δεδομένων

Στατιστική συμπερασματολογία:
Συναγωγή συμπερασμάτων για τη
σχέση π.χ. μίας μεταβλητής με μία
άλλη

Με τη βοήθεια
στατιστικών δοκιμασιών
(t-test, X^2, κλπ)

Παράδειγμα

- Σε μία έρευνα συλλέγονται, μεταξύ άλλων, τα εξής δεδομένα για τους συμμετέχοντες:
    - Ηλικία (έτη)
    - Φύλο
    - Συστολική πίεση (mm Hg)
    - Κάπνισμα (ναι/όχι)

Πιθανά ερωτήματα

- **Διαφέρουν τα επίπεδα διαστολικής πίεσης μεταξύ ανδρών και γυναικών;**
  - Πίεση: ποσοτική μεταβλητή
  - Φύλο: ποιοτική μεταβλητή με 2 επίπεδα

  → **t-test (σύγκριση 2 μέσων τιμών)**

- **Διαφέρουν οι καπνιστικές συνήθειες μεταξύ ανδρών και γυναικών;**
  - Κάπνισμα: ποιοτική μεταβλητή
  - Φύλο: ποιοτική μεταβλητή

  → **X^2 test (σύγκριση ποσοστών)**

- **Υπάρχει σχέση μεταξύ ηλικίας και διαστολικής πίεσης;**
  - Ηλικία: ποσοτική μεταβλητή
  - Συστολική πίεση: ποσοτική μεταβλητή

  → **?**

Στατιστικές δοκιμασίες για τη διερεύνηση σχέσης μεταξύ 2 παραγόντων

Παράγοντας 1	Παράγοντας 2	
	Ποσοτική	**Ποιοτική**
Ποιοτική	t-test (για ποιοτική με 2 επίπεδα)	X^2-test
Ποσοτική	?	

“Υπάρχει σχέση μεταξύ ηλικίας και διαστολικής πίεσης;”
→ Τι εννοούμε «σχέση»;

- **Βαθμός συσχέτισης** μεταξύ των δύο μεταβλητών (παρουσία ή απουσία συσχέτισης και, αν υπάρχει, πόσο «δυνατή» είναι)

- **Αν υπάρχει συσχέτιση**, μας ενδιαφέρει:
  - Το είδος της συσχέτισης **(γραμμική ή μη γραμμική)**
  - Ποια είναι η **κατεύθυνσή της** (θετική ή αρνητική)

- Πως **εξαρτάται** η μία μεταβλητή από την άλλη,
  - Π.χ. αν αυξηθεί η ηλικία κατά 10 έτη, πόσο αναμένεται να αυξηθούν τα επίπεδα διαστολικής πίεσης;

Συσχέτιση και ερευνητική μονάδα

Οι δύο μεταβλητές των οποίων τη σχέση θέλουμε να εξετάσουμε **μετριούνται στην ίδια ερευνητική ομάδα,** η οποία μπορεί να είναι:

- Το άτομο

  Π.χ. συστολική πίεση και ηλικία, χρόνια εκπαίδευσης και εισόδημα

- Οικογένεια

  Π.χ. ύψος πατέρα και ύψος παιδιού, χρόνια εκπαίδευσης γονέων και αριθμός παιδιών

- Ευρύτερες ομάδες

  Π.χ. θνησιμότητα από καρκίνο και κατά κεφαλή κατανάλωση καπνού σε ένα πληθυσμό

Παράδειγμα

- Σε μελέτη για τη διερεύνηση της επίδρασης του μολύβδου στην σωματομετρική ανάπτυξη των παιδιών, μελετήθηκαν παιδιά σχολικής ηλικίας από τρεις περιοχές:
  - Λαύριο, Ελευσίνα και Λουτράκι

- Το συνολικό δείγμα αποτελείτο από 522 παιδιά, 274 αγόρια και 248 κορίτσια ηλικίας 6-9 χρονών. Μέρος των δεδομένων παρουσιάζεται στον πίνακα που ακολουθεί

(Kafourou et al, Archives of Environmental health, 1997; 52: 377- 383).

Κωδικός	Πόλη	Ηλικία (έτη)	Ανάστημα πατέρα (cm)	Μόλυβδος (μg/mL)	Ανάστημα παιδιού (cm)
353	2	8	172	23.42	116
419	2	.	165	51.17	107
19	1	8	152	.	114
26	1	7	177	5.94	122
506	2	7	155	20.21	119
683	3	8	170	4.16	117
612	3	7	164	9.78	112
97	1	8	164	.	121
504	2	7	172	17.29	113
469	2	9	170	26.98	124
498	2	7	160	13.24	110
565	2	8	168	22.94	123
140	1	8	162	2.86	115
374	2	6	155	26.59	112
673	3	6	172	5.69	119
644	3	8	167	11.87	123
507	2	8	177	10.19	125
711	3	7	165	4.15	124

Οπου υπάρχει . υποδεικνύει ελλείπουσα τιμή. Για την πόλη: 1 σημαίνει Λουτράκι, 2 Λαύριο και 3 Ελευσίνα.

leadold.sav [DataSet2] - IBM SPSS Statistics Data Editor

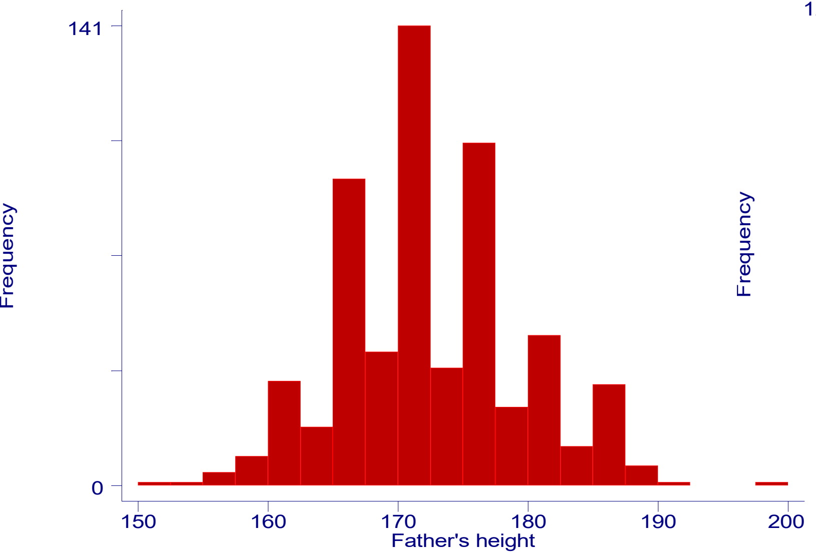
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	kode	sex	town	jobfa	jobmo	edufa	edumo	faheig	moheig	height	lead	age
1	1	0	1	2	10	12	12	189	172	129	,00	8,00
2	2	0	1	3	10	12	12	165	165	129	,00	7,00
3	4	0	1	2	10	15	8	174	159	120	,00	8,00
4	6	0	1	2	10	12	6	175	165	132	,00	8,00
5	8	0	1	1	10	9	9	182	165	131	,00	8,00
6	9	0	1	1	10	9	6	170	163	125	,00	8,00
7	11	0	1	2	10	12	7	180	160	131	3,55	8,00
8	15	1	1	1	10	6	6	166	160	114	5,87	8,00
9	17	0	1	1	10	6	6	171	156	123	9,29	7,00
10	18	0	1	2	10	8	6	173	163	127	7,07	8,00
11	19	1	1	2	10	6	6	152	145	114	,00	8,00
12	20	1	1	2	99	12	12	170	168	128	6,02	7,00
13	21	1	1	2	6	8	8	171	170	129	5,00	8,00
14	22	1	1	2	4	6	9	166	160	120	5,36	7,00
15	23	0	1	1	10	4	5	999	999	126	8,14	8,00
16	26	0	1	1	2	99	99	177	145	122	5,94	7,00
17	27	1	1	2	10	6	6	999	999	126	5,34	8,00

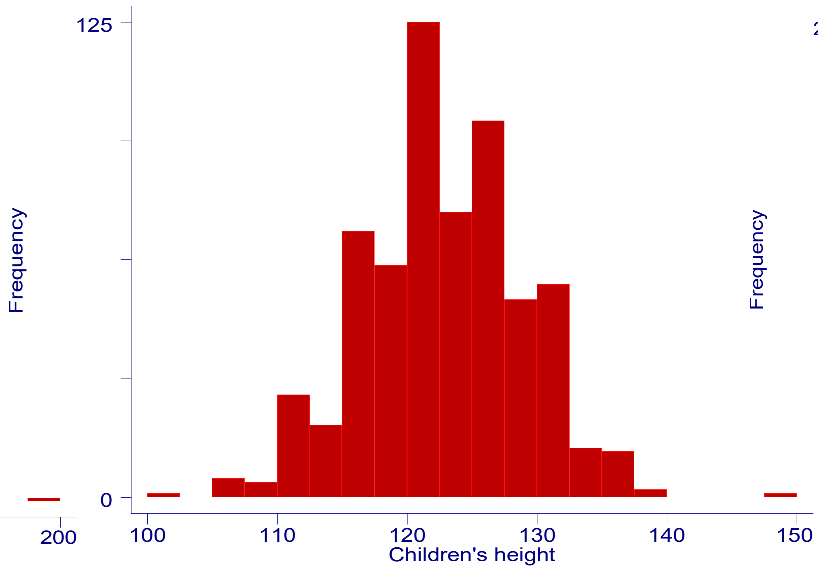
Παράδειγμα (συνέχεια)

- Πιθανά ερωτήματα:
  - Σχέση αναστήματος παιδιού με ανάστημα πατέρα
  - Σχέση αναστήματος παιδιού με ηλικία παιδιού
  - Σχέση αναστήματος παιδιού με επίπεδα μολύβδου

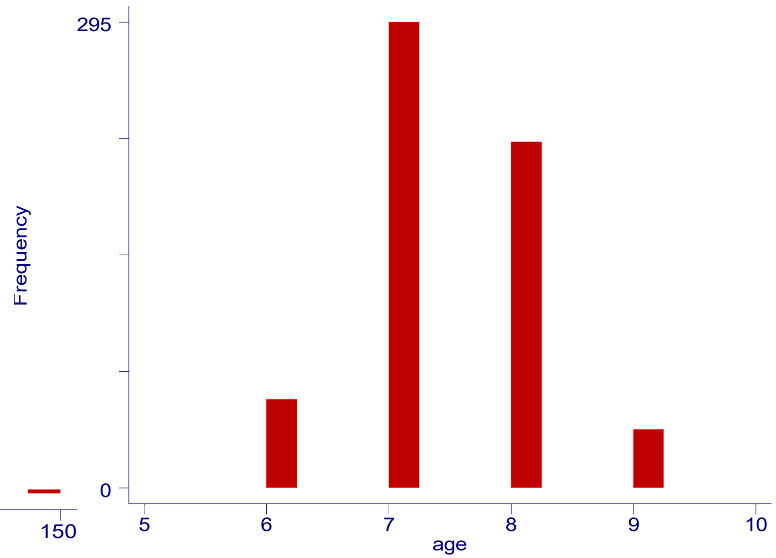
Κατανομή συχνοτήτων του αναστήματος των πατέρων



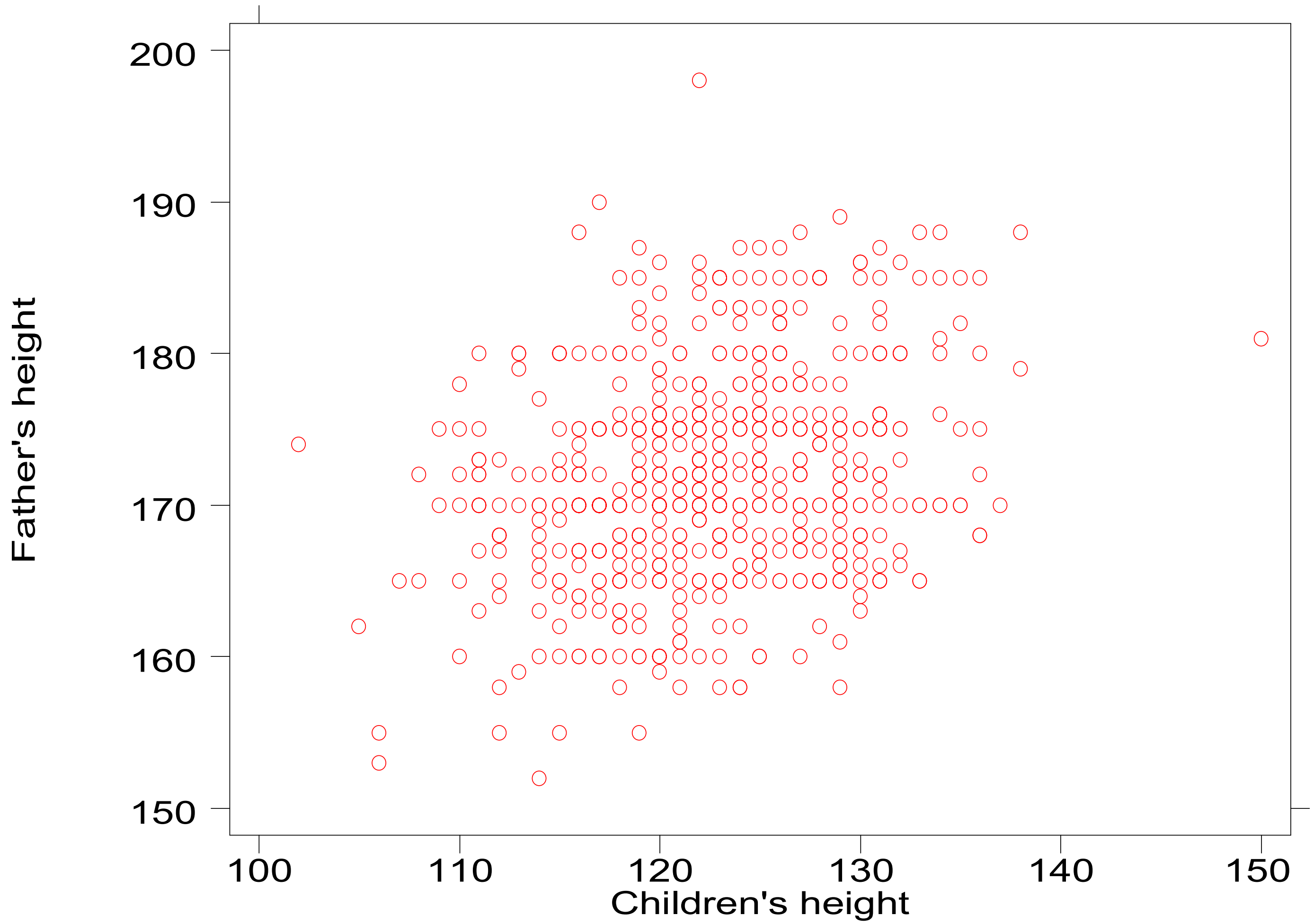
Κατανομή συχνοτήτων του αναστήματος των παιδιών



Κατανομή συχνοτήτων της ηλικίας των παιδιών

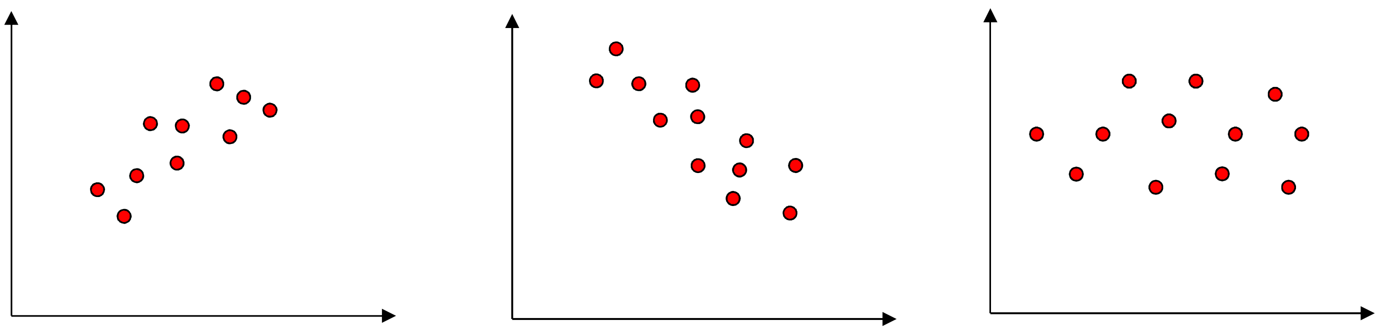


Διάγραμμα συσχέτισης (ή στικτόγραμμα – scatter plot) του αναστήματος του πατέρα με το ανάστημα του παιδιού

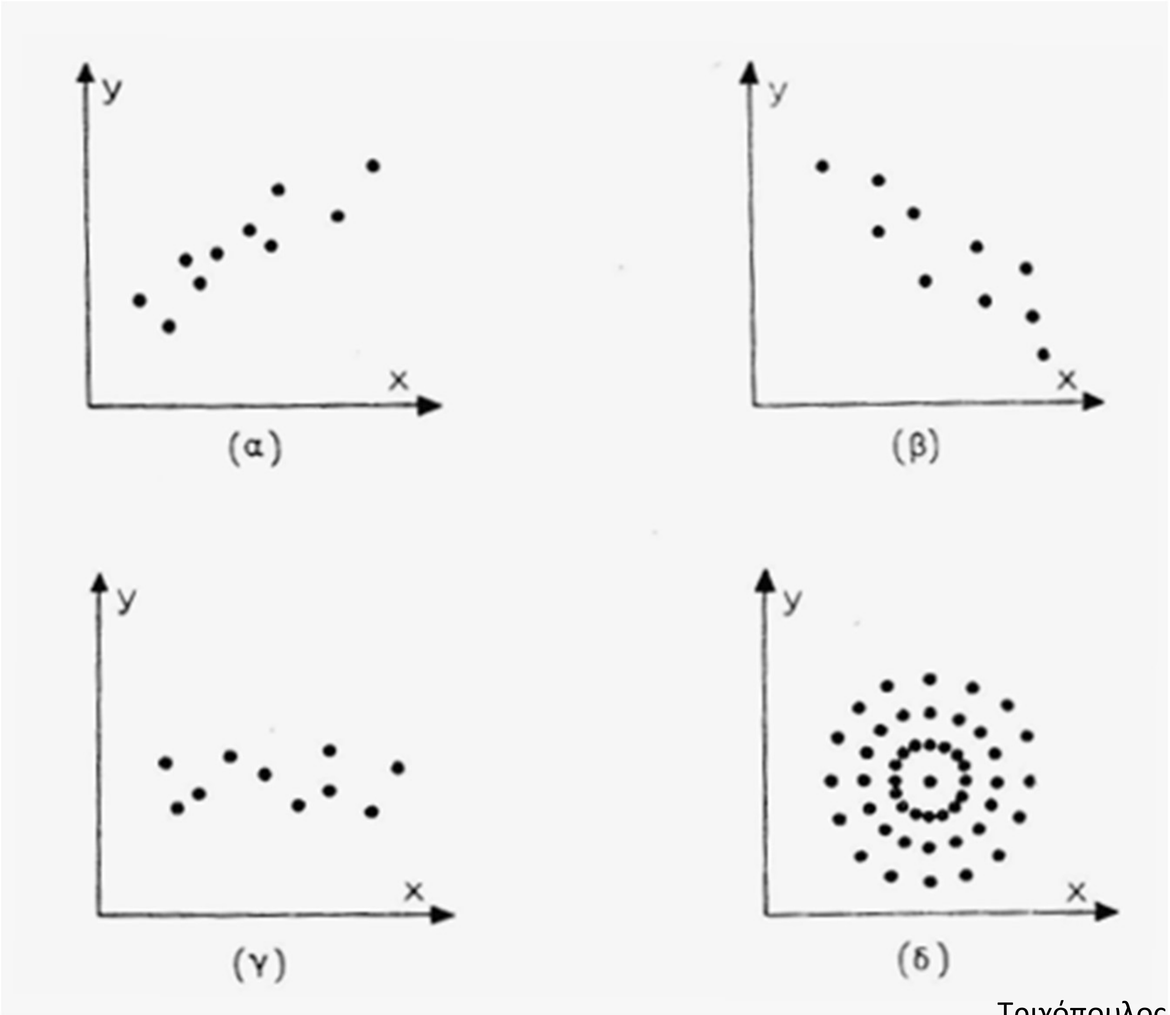


Αρχική διερεύνηση της σχέσης δύο ποσοτικών μεταβλητών

- Με διάγραμμα μπορούμε να έχουμε μία εικόνα για:
  - Την ύπαρξη γραμμικής ή άλλης μορφής συσχέτισης
  - Το βαθμό συμμεταβολής των 2 μεταβλητών
  - Αν η συσχέτιση είναι θετική/αρνητική



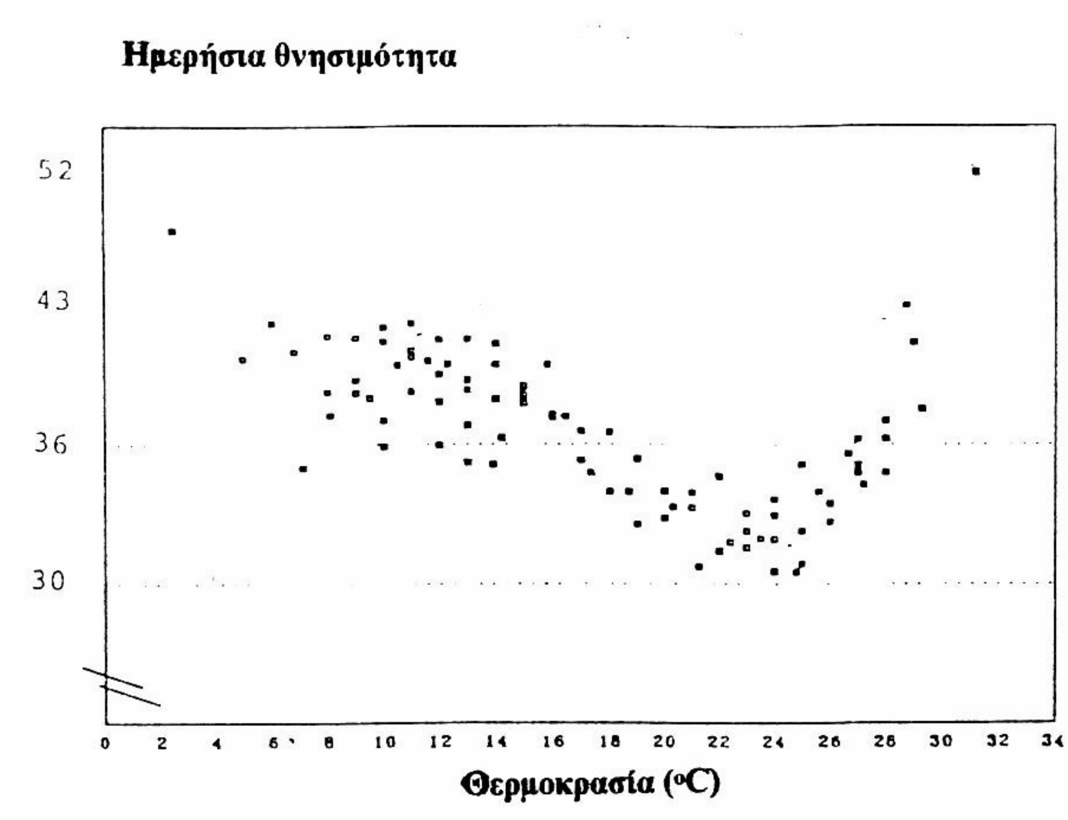
Διαγράμματα συσχέτισης: (α) θετική συσχέτιση, (β) αρνητική συσχέτιση, (γ) και (δ) απουσία συσχέτισης



Τριχόπουλος Δ, Τζώνου Α, Κατσουγιάννη Κ. Βιοστατιστική

Μη γραμμικές συσχετίσεις

Μπορεί να υπάρχει σχέση μεταξύ δύο μεταβλητών η οποία όμως δεν είναι γραμμική π.χ. θνησιμότητα και θερμοκρασία

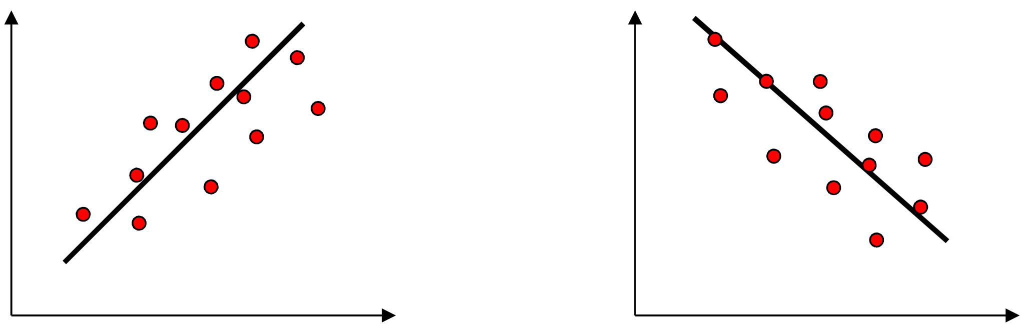


Touloumi et al, Int J of Epidemiolo,1994

Γραμμική (ευθύγραμμη) συσχέτιση

Γραμμική (ευθύγραμμη) συσχέτιση

- Θα ασχοληθούμε με τη διερεύνηση ύπαρξης **γραμμικών (ευθύγραμμων)** συσχετίσεων



- Ένας δείκτης που αποτελεί μέτρο του βαθμού της <u>γραμμικής</u> σχέσης μεταξύ δύο ποσοτικών μεταβλητών είναι ο **συντελεστής συσχέτισης Pearson.**

Προϋποθέσεις για τη χρήση του

- Οι δύο μεταβλητές είναι ποσοτικές
- Η κατανομή των τιμών των μεταβλητών είναι κατά προσέγγιση κανονική
- Η συλλογή των παρατηρήσεων και για τις δύο μεταβλητές έγινε τυχαία

Συντελεστής συσχέτισης Pearson (Pearson correlation coefficient)

- Αν X και Y δύο μεταβλητές σε δείγμα μεγέθους n, τότε ο **συντελεστής συσχέτισης r** εκφράζεται ως εξής:

Βαθμός στον οποίο οι X και Y συμμεταβάλλονται

$$r = \frac{\boxed{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}}{\sqrt{\boxed{\sum_{i=1}^{n}(X_i - \bar{X})^2}\boxed{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}}$$

Βαθμός μεταβλητότητας X * Βαθμός μεταβλητότητας Y

Συντελεστής συσχέτισης Pearson

$$\boxed{\sum_{i=1}^{n}(X_i-\bar{X})^2 =}$$

$$\sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X}^2)=$$

$$\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\bar{X}^2 =$$

$$\sum_{i=1}^{n}X_i^2-2\sum_{i=1}^{n}\frac{X_i}{n}\sum_{i=1}^{n}X_i+n\bar{X}^2 =$$

$$\sum_{i=1}^{n}X_i^2-2\frac{(\sum_{i=1}^{n}X_i)^2}{n}+n\frac{(\sum_{i=1}^{n}X_i)^2}{n^2}=$$

$$\boxed{\sum_{i=1}^{n}X_i^2-\frac{(\sum_{i=1}^{n}X_i)^2}{n}}$$

Όμοια:

- $\boxed{\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}Y_i^2-\frac{(\sum_{i=1}^{n}Y_i)^2}{n}}$

- $\boxed{\sum_{i=1}^{n}(X_i-\bar{X})^2(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}X_iY_i-\frac{(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{n}}$

(δεν χρειάζεται να τα μάθετε)

Συντελεστής συσχέτισης Pearson (Pearson correlation coefficient)

- Τελικά

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

$$= \frac{\text{Συμμεταβλητότητα (X, Y)}}{\sqrt{\text{Διακύμανση(X)} * \text{Διακύμανση(Y)}}} = \frac{\text{Cov(X, Y)}}{S_x S_y}$$

$$\boxed{= \frac{\sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}) * (\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n})}}}$$

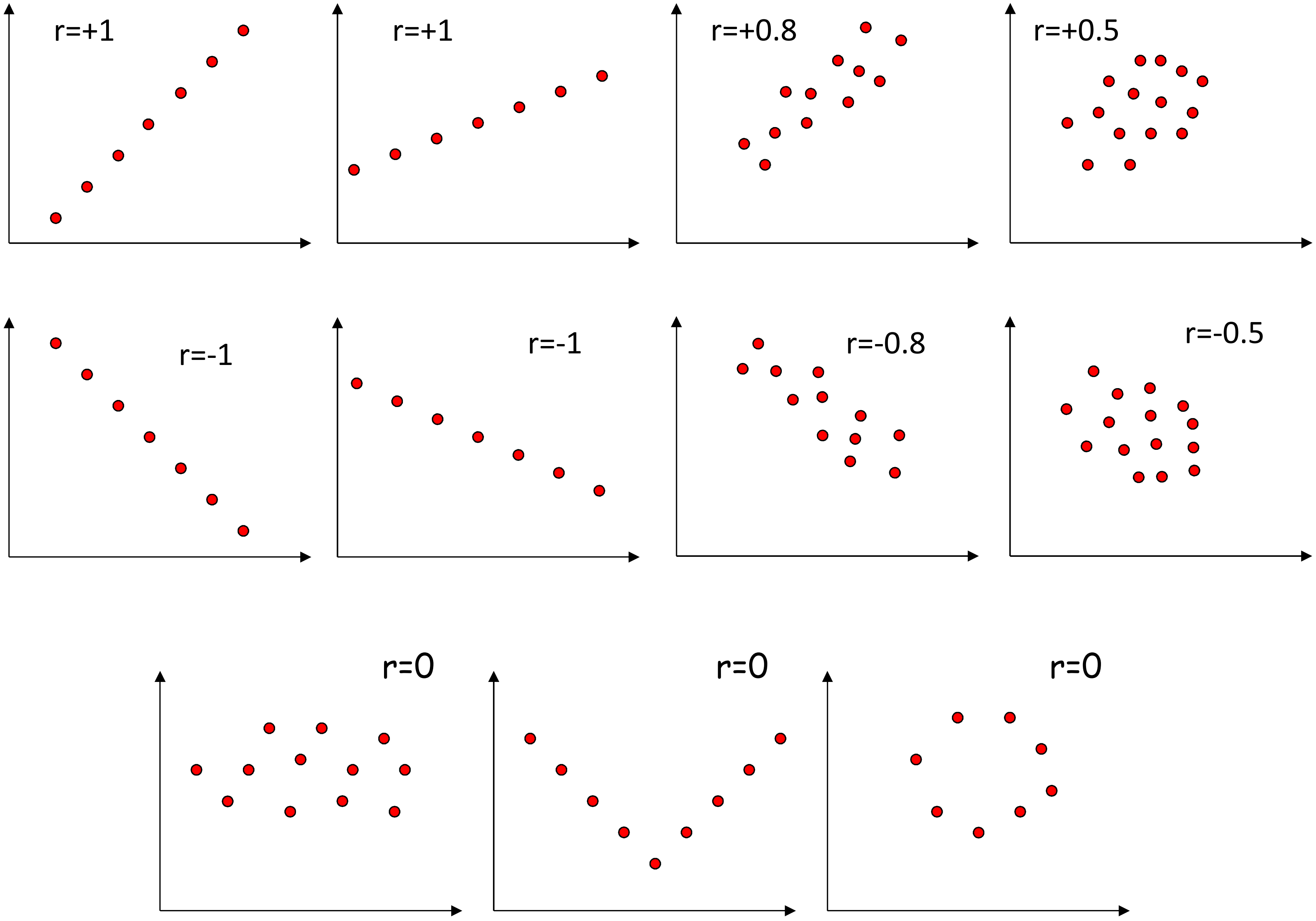
(δεν χρειάζεται να θυμάστε τον τύπο απέξω)

Ιδιότητες του συντελεστή συσχέτιση

- Ο συντελεστής συσχέτισης είναι καθαρός αριθμός (δεν έχει μονάδες)
- Παίρνει τιμές από -1 έως 1
- Μετρά μόνο την ευθύγραμμη συσχέτιση

Ερμηνεία του συντελεστή συσχέτιση

- Ερμηνεύουμε την **τιμή του και το πρόσημό του**
  - Τιμές κοντά στο 0 δείχνουν **απουσία** συσχέτισης
  - Τιμές κοντά στο 1 δείχνουν **ύπαρξη** **εὐθύγραμμης** συσχέτισης
    - Όσο πιο κοντά στο +1 → Τόσο πιο δυνατή **θετική συσχέτιση**
    - Όσο πιο κοντά στο -1 → Τόσο πιο δυνατή **αρνητική συσχέτιση**


r=+1
r=+1
r=+0.8
r=+0.5
r=-1
r=-1
r=-0.8
r=-0.5
r=0
r=0
r=0

r και r^2

- Ο συντελεστής συσχέτισης r αποτελεί μέτρο του βαθμού ευθύγραμμης συσχέτισης μεταξύ 2 ποσοτικών μεταβλητών, αλλά η αντιστοιχία δεν είναι αναλογική
  - π.χ. αν X και Y: r=0.4 / X και Z: r=0.8

  → **Δε σημαίνει ότι η συσχέτιση του X και Z είναι δύο φορές πιο ισχυρή από του X και Y**

- Καλύτερο μέτρο το τετράγωνό του r, δηλ. το r^2

X και Y: r=0.4 → $\mathbf{r^2=0.16}$

X και Z: r=0.8 → $\mathbf{r^2=0.64}$

Άρα η σχέση X και Z είναι 4 **φορές ισχυρότερη** από των X και Y (0.64/0.16)

Υπολογισμός 95% ορίων αξιοπιστίας του r

- Γενικά, σε κανονική κατανομή: 95% όρια αξιοπιστίας μίας εκτίμησης →

  εκτίμηση ±1.96*SE(εκτίμησης)

- Η δειγματοληπτική κατανομή του r δεν είναι κανονική → μετασχηματισμός κατά Fisher του r σε z

ΕΚΤΟΣ ΥΛΗΣ!

$$z = \frac{1}{2} ln(\frac{1+r}{1-r})$$

- Η ποσότητα z ακολουθεί ασυμπτωματικά την κανονική κατανομή με $SE(z) = \frac{1}{\sqrt{n-3}}$

  άρα 95% CI για z: $z \pm 1.96*SE(z)$

- 95% CI για το r είναι:

$$\boxed{r_L = \frac{e^{2z_L} - 1}{e^{2z_l} + 1}} \quad \boxed{r_U = \frac{e^{2z_U} - 1}{e^{2z_U} + 1}}$$

Στατιστική αξιολόγηση

- Το ερώτημα είναι αν υπάρχει συσχέτιση μεταξύ 2 μεταβλητών στον πληθυσμό
  - **ρ** → συντελεστής συσχέτισης των 2 μεταβλητών στον **πληθυσμό**
  - **r** → συντελεστής συσχέτισης των 2 μεταβλητών στο **δείγμα**

(κατ'αναλογία με το μ και το $\bar{X}$)

- Αξιολογούμε το r βασιζόμενοι σε ένα δείγμα π.χ. ηλικία και συστολική πίεση σε δείγμα 10 ατόμων

→ r=0.70

- Φαίνεται να υπάρχει θετική συσχέτιση. Είναι το εύρημα αυτό στατιστικά σημαντικό; Πώς θα διατυπώσουμε τη μηδενική υπόθεση;

H_0: ρ=0
H_1: ρ≠0

Στατιστική αξιολόγηση

- Ελέγχουμε την τιμή |r| στον αντίστοιχο πίνακα με τις οριακές τιμές (συνήθως με την τιμή που αντιστοιχεί στο 5% επίπεδο σημαντικότητας) σε n-2 βαθμούς ελευθερίας (degrees of freedom: df ή ΒΕ) όπου n ο αριθμός των παρατηρήσεων (ζευγών)

- Αν |r|≥οριακή τιμή → απορρίπτω H_0 και συμπεραίνω ότι η σχέση είναι στατιστικά σημαντική

- Αν |r|<οριακή τιμή → η σχέση δεν είναι στατιστικά σημαντική

- Π.χ. σε δείγμα 10 ατόμων, δύο μεταβλητές έχουν r=0.6 → θετική συσχέτιση
- Είναι στατιστικά σημαντική η συσχέτιση;

- B.E= n-2 = 10-2 = 8
- r=0,6 <0,632

→Δεν είναι στατιστικά σημαντικό στο 5% επίπεδο σημαντικότητας (p>0,05)

- Επειδή 0,549 < r <0,632

→ p>5% αλλά p<10%

Σχέση Ασαφής – Μη στατιστικά σημαντική

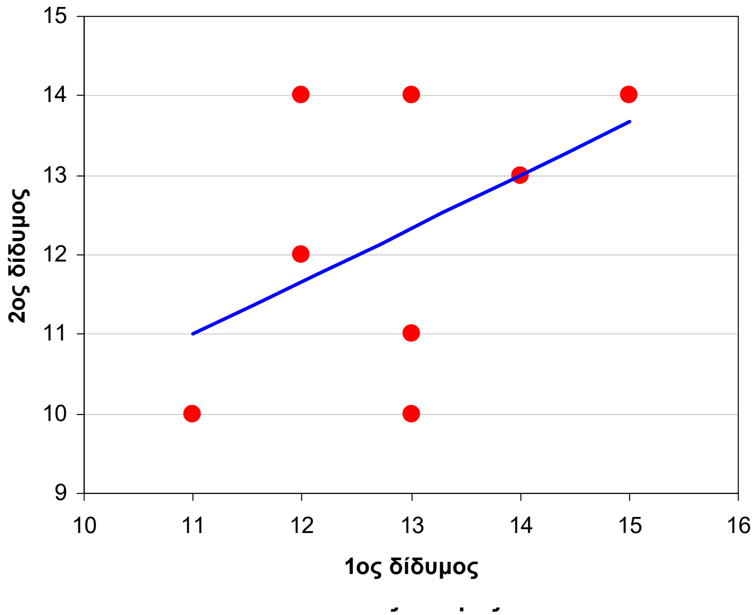
Επίπεδο σημαντικότητας

DF	.10	.05	.02	.01
1	.9877	.9969	.9995	.9999
2	.9000	.9500	.9800	.9900
3	.8054	.8783	.9343	.9587
4	.7293	.8114	.8822	.9172
5	.6694	.7545	.8329	.8745
6	.6215	.7067	.7887	.8343
7	.5822	.6664	.7498	.7977
8	$\boxed{.5494}$	$\boxed{.6319}$	.7155	.7646
9	.5214	.6021	.6851	.7348
10	.4973	.5760	.6581	.7079
11	.4762	.5529	.6339	.6835
12	.4575	.5324	.6120	.6614
13	.4409	.5140	.5923	.6411
14	.4259	.4973	.5742	.6226
15	.4124	.4821	.5577	.6055
16	.4000	.4683	.5425	.5897
17	.3887	.4555	.5285	.5751
18	.3783	.4438	.5155	.5614
19	.3687	.4329	.5034	.5487
20	.3598	.4227	.4921	.5368
21	.3515	.4132	.4815	.5256
22	.3438	.4044	.4716	.5151
23	.3365	.3961	.4622	.5052
24	.3297	.3882	.4534	.4958
25	.3233	.3809	.4451	.4869
26	.3172	.3739	.4372	.4785
27	.3115	.3673	.4297	.4705
28	.3061	.3610	.4226	.4629
29	.3009	.3550	.4158	.4556
30	.2960	.3494	.4093	.4487
31	.2913	.3440	.4032	.4421

Παράδειγμα υπολογισμού και στατιστικής αξιολόγησης του συντελεστή συσχέτισης Pearson

- Συστολική αρτηριακή πίεση σε ζεύγη μονο-ωογενών διδύμων.

(χ) 1ος δίδυμος	(ψ) 2ος δίδυμος
12	14
11	10
13	11
15	14
13	14
12	12
13	10
14	13



Παράδειγμα (συνέχεια)

- Συστολική αρτηριακή πίεση σε ζεύγη μονο-ωογενών διδύμων.

(χ) 1ος δίδυμος	(ψ) 2ος δίδυμος	χ*ψ	χ^2	ψ^2
12	14	168	144	196
11	10	110	121	100
13	11	143	169	121
15	14	210	225	196
13	14	182	169	196
12	12	144	144	144
13	10	130	169	100
14	13	182	196	169
Σχ=103	Σψ=98	Σχψ=1269	$\Sigma\chi^2$=1337	$\Sigma\psi^2$=1222

n=8

Παράδειγμα (συνέχεια)

$$r = \frac{\sum \chi\psi - \frac{\sum \chi * \sum \psi}{n}}{\sqrt{\left\{ \Sigma\chi^2 - \frac{(\Sigma\chi)^2}{n} \right\} * \left\{ (\Sigma\psi^2 - \frac{(\Sigma\psi)^2}{n} \right\}}}$$

$$= \frac{1269 - \frac{103*98}{8}}{\sqrt{\left\{ 1337 - \frac{103^2}{8} \right\} * \left\{ 1222 - \frac{98^2}{8} \right\}}}$$

$$= \frac{1269 - 1261{,}75}{\sqrt{10{,}88 * 21{,}5}} = \frac{7{,}25}{15{,}29} = 0{,}474$$

Παράδειγμα (συνέχεια)

- r=0.474
- Αξιολόγηση στο 5% επίπεδο σημαντικότητας στους 6 ΒΕ (n-2=8-2=6)

Επίπεδο σημαντικότητας

DF	.10	.05	.02	.01
1	.9877	.9969	.9995	.9999
2	.9000	.9500	.9800	.9900
3	.8054	.8783	.9343	.9587
4	.7293	.8114	.8822	.9172
5	.6694	.7545	.8329	.8745
6	.6215	.7067	.7887	.8343
7	.5822	.6664	.7498	.7977
8	.5494	.6319	.7155	.7646
9	.5214	.6021	.6851	.7348
10	.4973	.5760	.6581	.7079

B.E.	10%	5%	1%	1‰
6	**0,622**	**0,707**	**0,834**	**0,925**

0.474

|r| = 0.474 <0.707

→Όχι στατιστικά σημαντική συσχέτιση (p>5% & p>10%)

Επιφυλάξεις ως προς τη χρήση του συντελεστή συσχέτισης

1. Συσχέτιση και συμφωνία τιμών

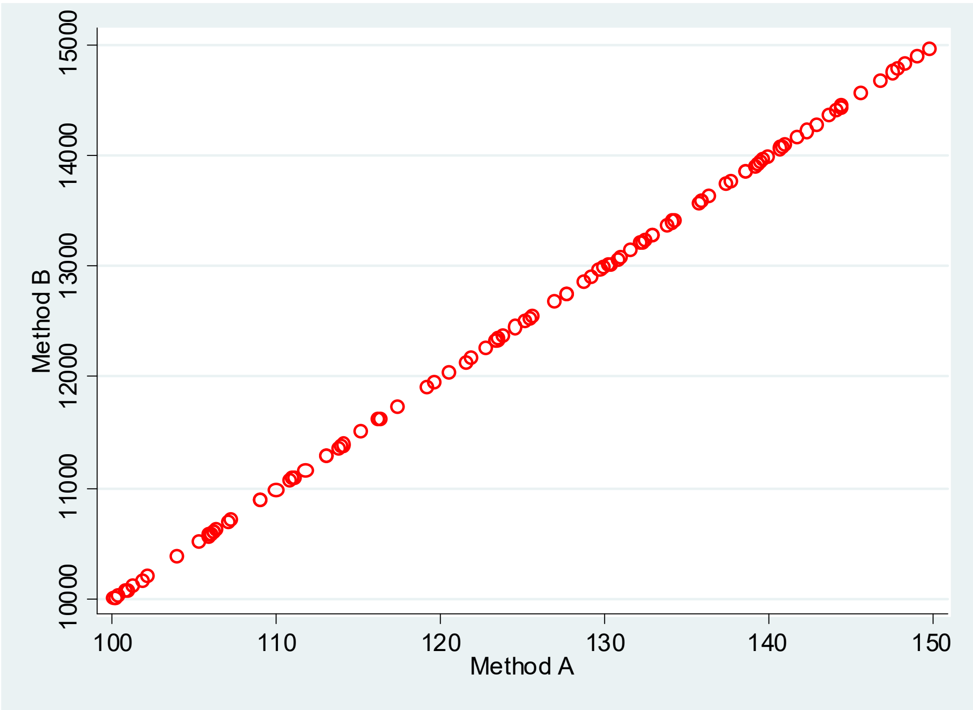
2. Συσχέτιση και αιτιότητα (αυτό ισχύει γενικά σε όλες τις σχέσεις που διερευνούμε)

3. Επίδραση μεμονωμένων παρατηρήσεων στο συντελεστή συσχέτισης

1. Συσχέτιση και συμφωνία τιμών

Αν έχουμε δύο μεταβλητές που μετρούν το **ίδιο** μέγεθος π.χ.

X=Μέθοδος Α και Y=Μέθοδος Β για τη μέτρηση των επιπέδων ενός ποσοτικού μεγέθους

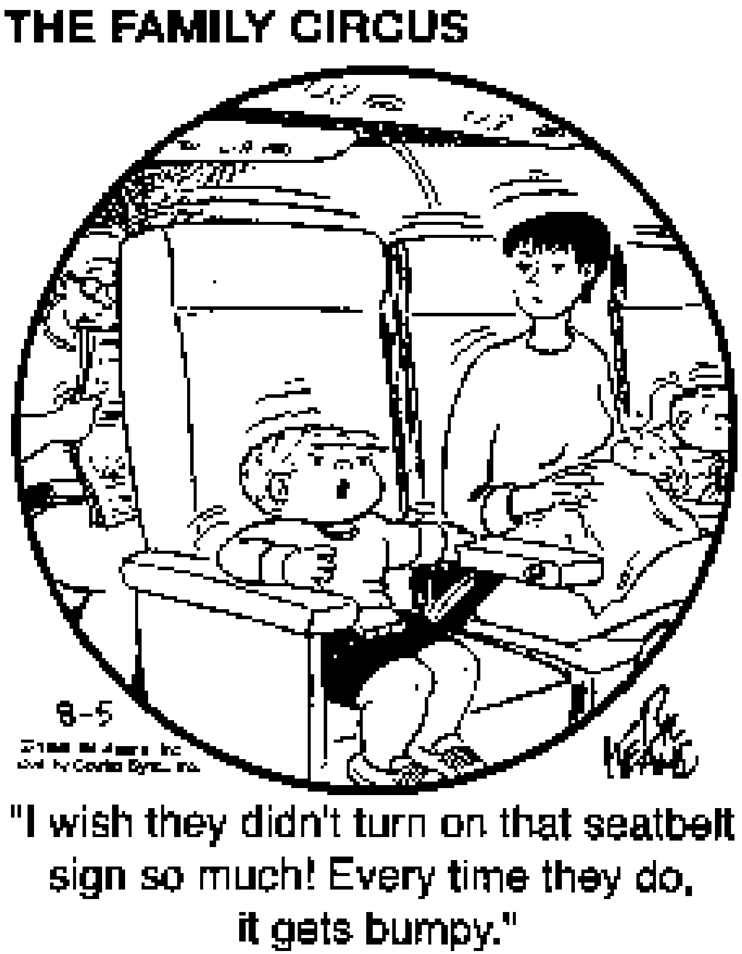


r=1
αλλά προφανώς δεν υπάρχει ταύτιση τιμών (η μέθοδος Β έχει 100πλάσιες τιμές από μέθοδο Α)

2. Συσχέτιση και αιτιολογική σχέση

- Η ύπαρξη συσχέτισης μεταξύ δύο μεταβλητών δε σημαίνει απαραίτητα ότι υπάρχει αιτιολογική σχέση μεταξύ τους,

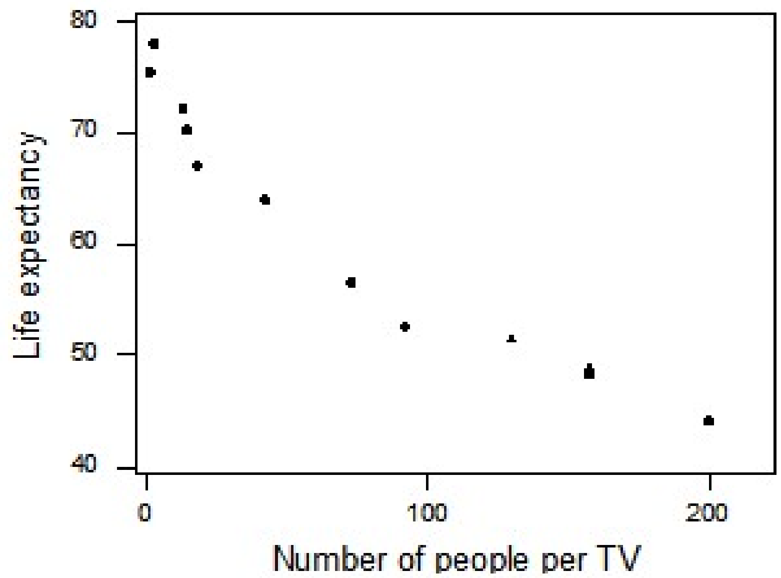
  π.χ. αύξηση των επιπέδων της μίας μεταβλητής οδηγούν σε αύξηση των επιπέδων της άλλης;



Παράδειγμα

- Έστω ότι σε 12 χώρες συλλέγουμε δεδομένα για
    - X: προσδόκιμο επιβίωσης
    - Y: αριθμός ατόμων που αντιστοιχεί σε μία τηλεόραση

  Υπάρχει συσχέτιση μεταξύ των 2 μεταβλητών;



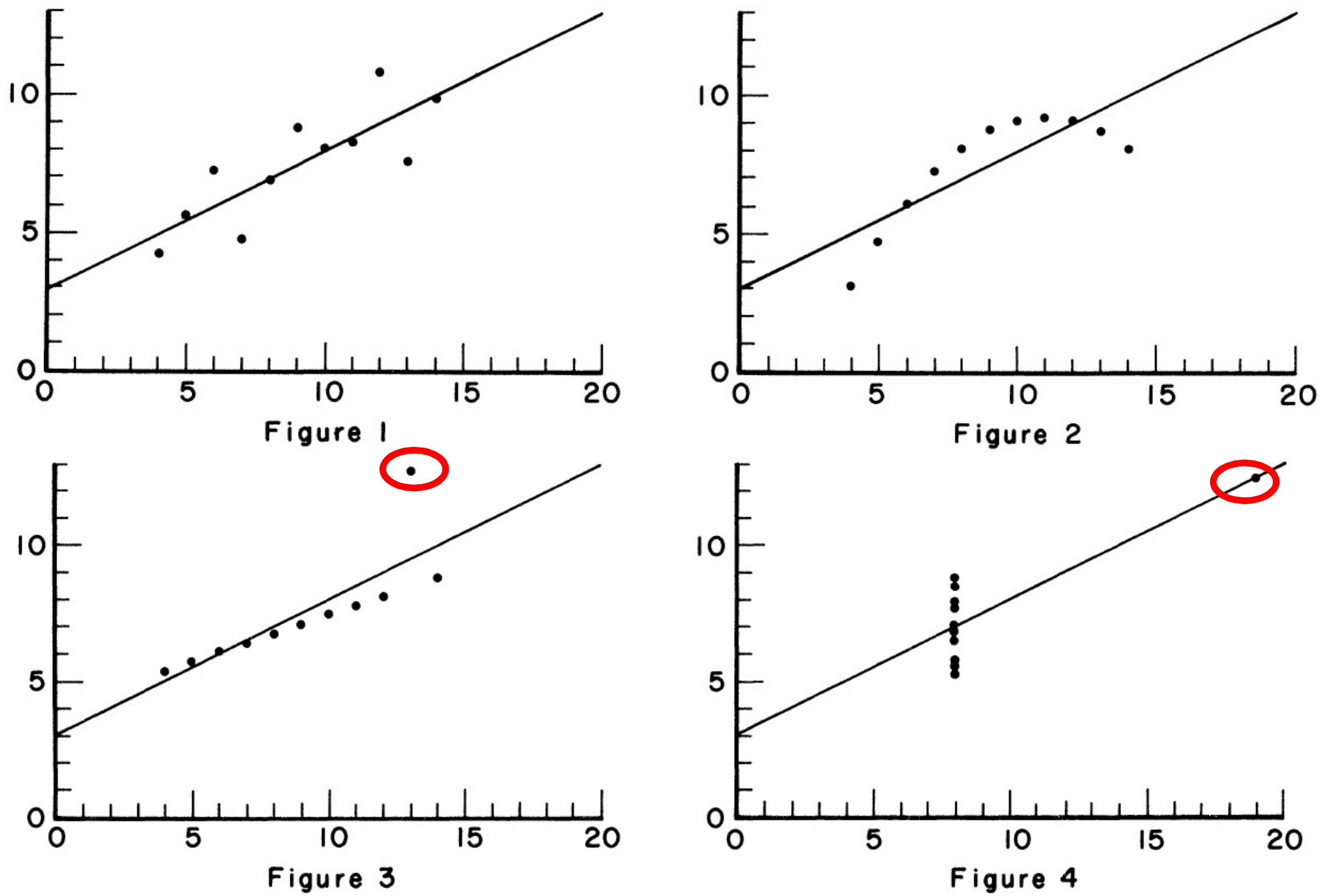
Παράδειγμα (συνέχεια)

- Αρνητική συσχέτιση: όσο περισσότερα άτομα/τηλεόραση (δηλ. λιγότερες τηλεοράσεις στον πληθυσμό) τόσο χαμηλότερο προσδόκιμο επιβίωσης

- Αιτιολογική συσχέτιση;;;
  - Π.χ. αν αυξήσω τον αριθμό τηλεοράσεων θα βελτιώσω το προσδόκιμο επιβίωσης;

- Προφανώς όχι
  - Και οι 2 μεταβλητές σχετίζονται με κάποιο τρίτο χαρακτηριστικό
    - Π.χ. το προσδόκιμο επιβίωσης και ο αριθμός τηλεοράσεων πιθανόν να σχετίζονται με το οικονομικό επίπεδο της χώρας

3. Επίδραση μεμονωμένων παρατηρήσεων στο συντελεστή συσχέτισης

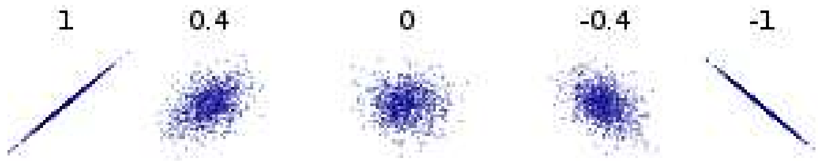


**r=0.816
και στα 4
σχήματα**

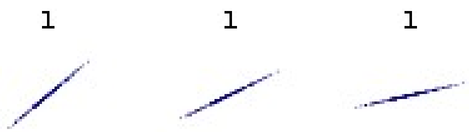
Anscombe’s quartet (The American Statistician, 1973)

Συμπερασματικά:

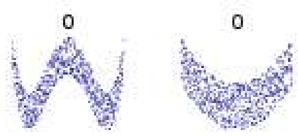
1. Ο συντελεστής συσχέτισης εκφράζει τη διεύθυνση και τη δύναμη μίας γραμμικής συσχέτισης



- **Όχι** όμως την κλίση της γραμμικής συσχέτισης



- **Όχι** άλλου είδους μη γραμμικές συσχετίσεις



Συμπερασματικά:

2. Σημαντική συσχέτιση δεν υπονοεί και αιτιότητα (όπως και στα προηγούμενα τεστ που έχουμε συζητήσει)

3. Υψηλή συσχέτιση μεταξύ των τιμών δύο μεταβλητών που μετρούν το ίδιο μέγεθος δεν σημαίνει απαραίτητα και καλή συμφωνία των τιμών τους

4. Γενικά ο συντελεστής συσχέτισης θα πρέπει να ερμηνεύεται σε συνδυασμό με γραφική απεικόνιση της σχέσης των δύο μεταβλητών
   - Διερεύνηση μη γραμμικών συσχετίσεων
   - Μεμονωμένες παρατηρήσεις με μεγάλη επίδραση στο τελικό αποτέλεσμα

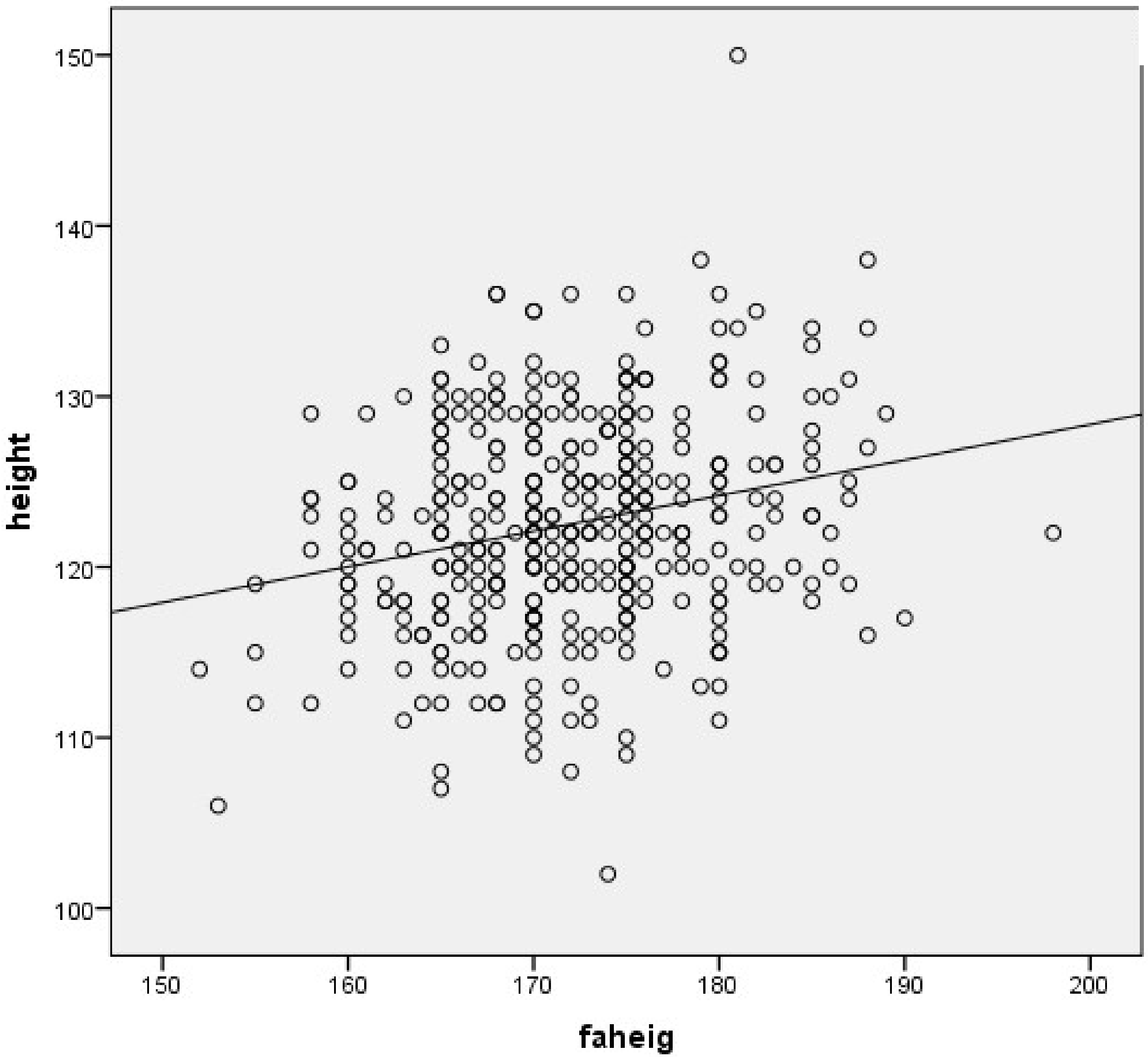
Στατιστικές δοκιμασίες για τη διερεύνηση σχέσης μεταξύ 2 παραγόντων

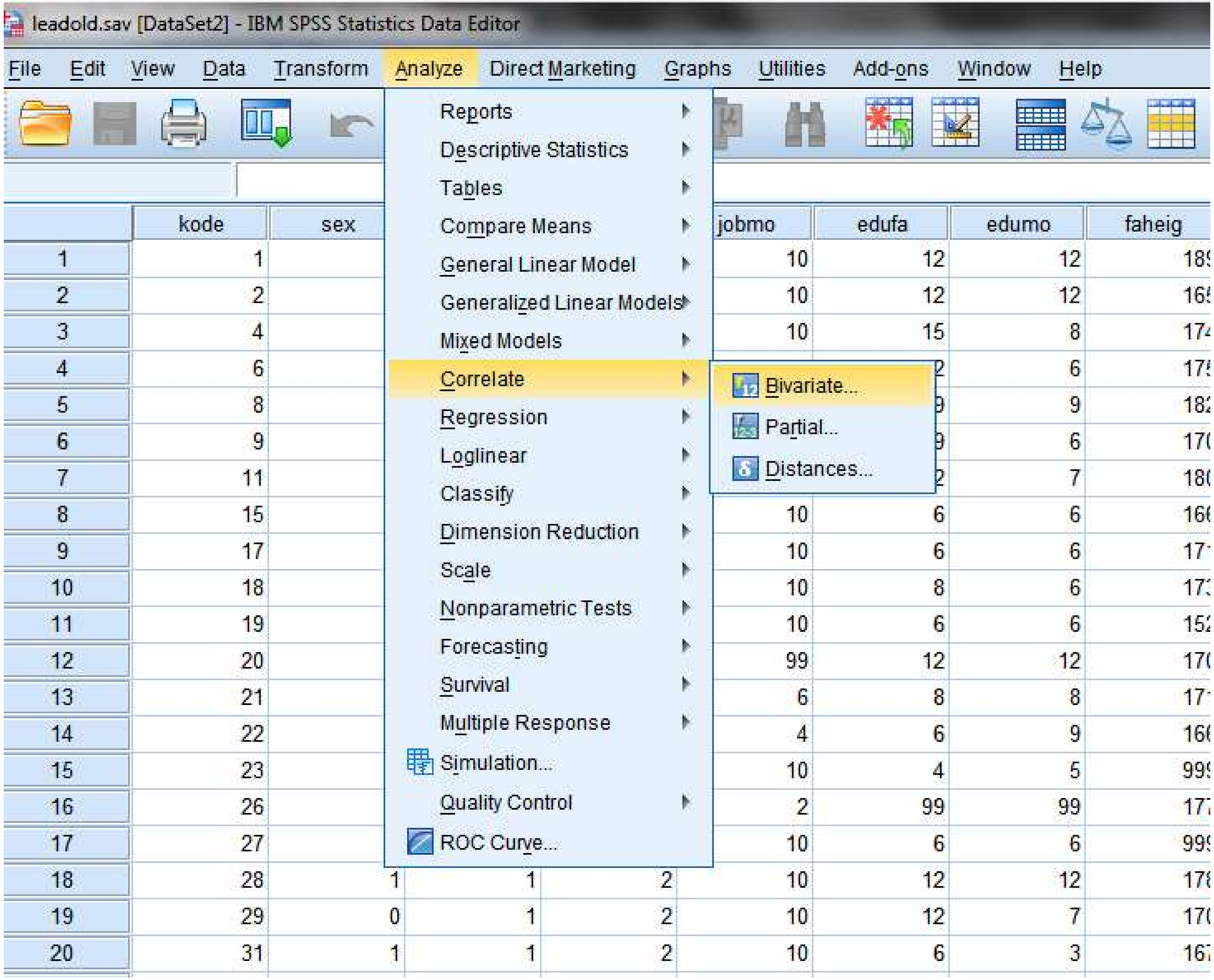
Παράγοντας 1	**Παράγοντας 2**	
	Ποσοτική	**Ποιοτική**
Ποιοτική	t-test (για ποιοτική με 2 επίπεδα)	X^2-test
Ποσοτική	Correlation coefficient με αξιολόγησή του	

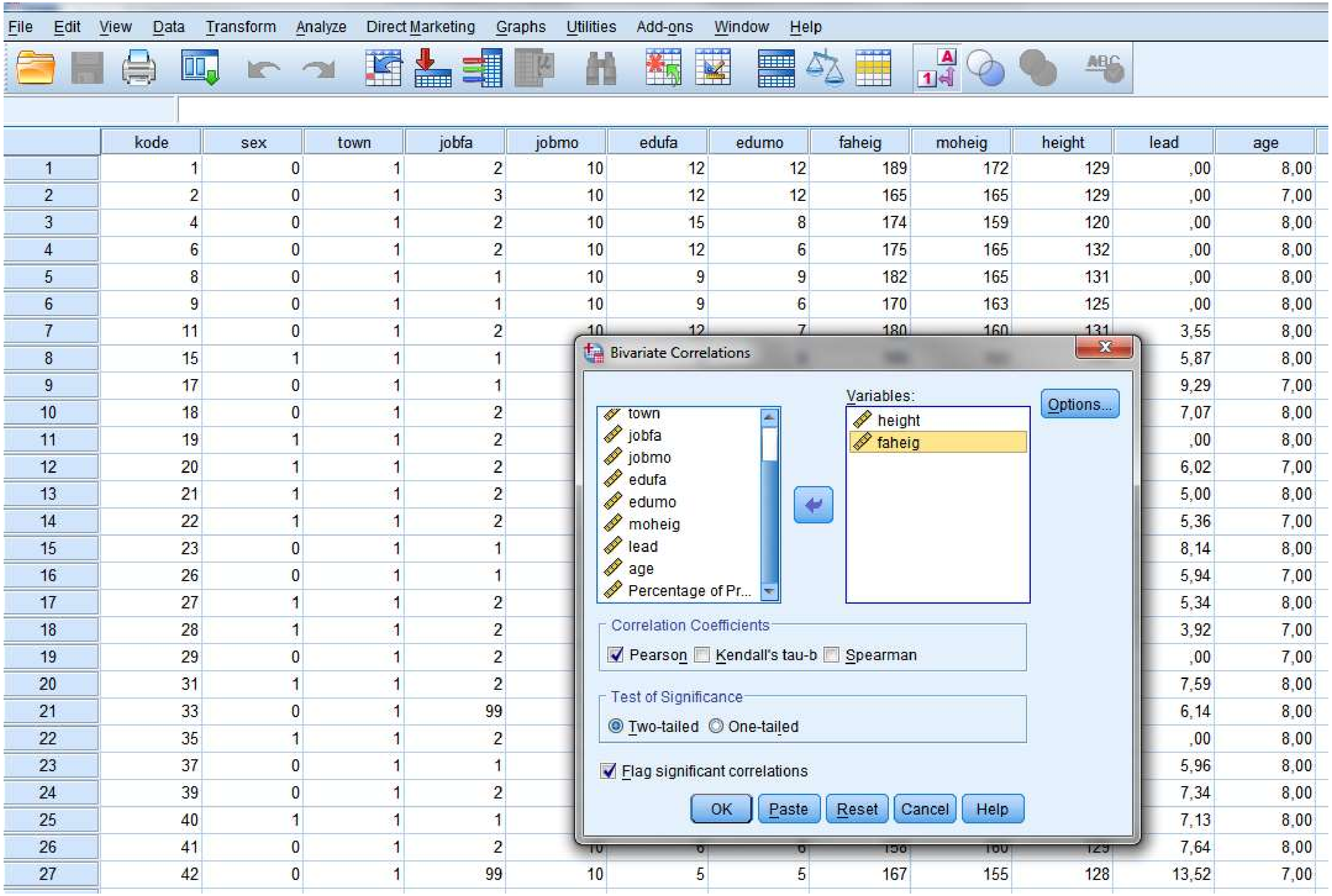
Σημείωση: Οι περισσότερες δοκιμασίες για **ποσοτικά** χαρακτηριστικά υποθέτουν την **κανονική** κατανομή των χαρακτηριστικών

Παράδειγμα εφαρμογής σε στατιστικό πακέτο

- SPSS
- Σχετίζεται το ανάστημα του παιδιού με το ανάστημα του πατέρα;


150
140
130
120
110
100
height
150
160
170
180
190
200
faheig


leadold.sav [DataSet2] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Simulation...
Quality Control
ROC Curve...
Bivariate...
Partial...
Distances...
kode
sex
jobmo
edufa
edumo
faheig


File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
kode sex town jobfa jobmo edufa edumo faheig moheig height lead age
Bivariate Correlations
Variables:
height
faheig
Options...
town
jobfa
jobmo
edufa
edumo
moheig
lead
age
Percentage of Pr...
Correlation Coefficients
Pearson
Kendall's tau-b
Spearman
Test of Significance
Two-tailed
One-tailed
Flag significant correlations
OK
Paste
Reset
Cancel
Help

Correlations

[DataSet2] C:\Τα έγγραφά μου\ΜΑΘΗΜΑΤΑ ΠΑΝΕΠΙΣΤΗΙ

Correlations

		height	faheig
height	Pearson Correlation	1	,238**
	Sig. (2-tailed)		,000
	N	398	381
faheig	Pearson Correlation	,238**	1
	Sig. (2-tailed)	,000	
	N	381	381

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