ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

3. Στατιστική Συμπερασματολογία για ποιοτικές μεταβλητές

Η έννοια της Στατιστικής Συμπερασματολογίας (Statistical Inference)

- Συμπερασματολογία (Inference): εξαγωγή συμπεράσματος με βάση κάποια στοιχεία
- Στατιστική Συμπερασματολογία (Statistical inference): Ένα σύνολο από διαδικασίες με τις οποίες το μέγεθος του δείγματος και στατιστικά μέτρα που υπολογίζονται από το δείγμα χρησιμοποιούνται για την εκτίμηση παραμέτρων του πληθυσμού

Στατιστικά μέτρα και Παράμετροι

- Στατιστικά μέτρα (Statistics): τιμές που υπολογίζονται από το δείγμα
- Παράμετροι (Parameters): Τιμές που μπορούν να υπολογιστούν μόνο σε απογραφή πληθυσμού και αποτελούν ακριβείς μετρήσεις του πληθυσμού
- Οι παράμετροι αντιπροσωπεύουν «αυτό που θέλουμε να μάθουμε» για έναν πληθυσμό.
- Τα στατιστικά χρησιμοποιούνται στην εκτίμηση των παραμέτρων του πληθυσμού

Εκτίμηση παραμέτρων

- Εκτίμηση παραμέτρου: η διαδικασία χρήσης πληροφοριών από το δείγμα για τον υπολογισμό ενός διαστήματος που περιγράφει το εύρος των τιμών που μπορεί να πάρει μια παράμετρος του πληθυσμού με κάποια πιθανότητα
- Διάστημα εμπιστοσύνης – δ.ε. (Confidence interval): Ένα εύρος τιμών μέσα στο οποίο έχουμε εμπιστοσύνη ότι θα «πέσει» η άγνωστη παράμετρος. Η εμπιστοσύνη εκφράζεται με μια πιθανότητα (συνήθως 90%, 95%, 99%)

Κατηγορικές (Ποιοτικές) μεταβλητές

- Εύρεση δ.ε. για τα ποσοστά από τον πίνακα συχνοτήτων
- Σχέση ανάμεσα σε δύο ποιοτικές μεταβλητές – διαδικασία Crosstabs

Παράδειγμα (data: L_research.sav)

Type of Question	Variable	Label
Demographic	L_use	Use L_store
	N_use	Use N_store
	Dwell	Type of Dwelling
	sex	Respondent's Sex
	work	Work Status
	Commute	Pass by L & N stores on way to work?
Lifestyle	Bargain	Look for bargains
	cash	Always pay cash
	Quick	Like quick, easy shopping
	Knowme	Shop where they know my name
	Hurry	Always in a hurry

	Value	Label
L_use	0	Do Not Use Regularly
	1	Use Regularly
N_use	0	Do Not Use Regularly
	1	Use Regularly
dwell	1	Own Home
	2	Rent
sex	1	Male
	2	Female
work	1	Full-Time
	2	Part-Time
	3	Retired/Do Not Work
commute	0	No
	1	Yes
bargain	1	Disagree
	2	Neither Agree Nor Disagree
	3	Agree
cash	1	Disagree
	2	Neither Agree Nor Disagree
	3	Agree
quick	1	Disagree
	2	Neither Agree Nor Disagree
	3	Agree
knowme	1	Disagree
	2	Neither Agree Nor Disagree
	3	Agree
hurry	1	Disagree
	2	Neither Agree Nor Disagree
	3	Agree

Εκτίμηση ποσοστού

- Να βρεθεί δ.ε. για το ποσοστό των κατοίκων περιοχής που έχει δικό του σπίτι
- Από διαδικασία Frequencies:

Type of Dwelling

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Own Home	29	17,9	17,9	17,9
	Rent	133	82,1	82,1	100,0
	Total	162	100,0	100,0	

p=0.179

n=162

q=1-p=0.821

Υπολογισμός δ.ε. για ποσοστό

- Ορίζουμε επίπεδο εμπιστοσύνης (confidence level) – συνήθως 95%
- Από πίνακες κανονικής κατανομής:

  Για 95%δ.ε. $\Rightarrow$ z=1.96
- Τύποι υπολογισμού:

$$p - z\sqrt{\frac{pq}{n}} \qquad p + z\sqrt{\frac{pq}{n}}$$

Κάτω όριο ***Άνω όριο***

Υπολογισμός δ.ε.

- Υπολογισμός:

$$0.179 - 1.96\sqrt{\frac{0.179 \cdot 0.821}{162}} \qquad 0.179 + 1.96\sqrt{\frac{0.179 \cdot 0.821}{162}}$$

Κάτω όριο ***Άνω όριο***

- 95% δ.ε.:(0.120, 0.238) ή **(12%, 23.8%)**

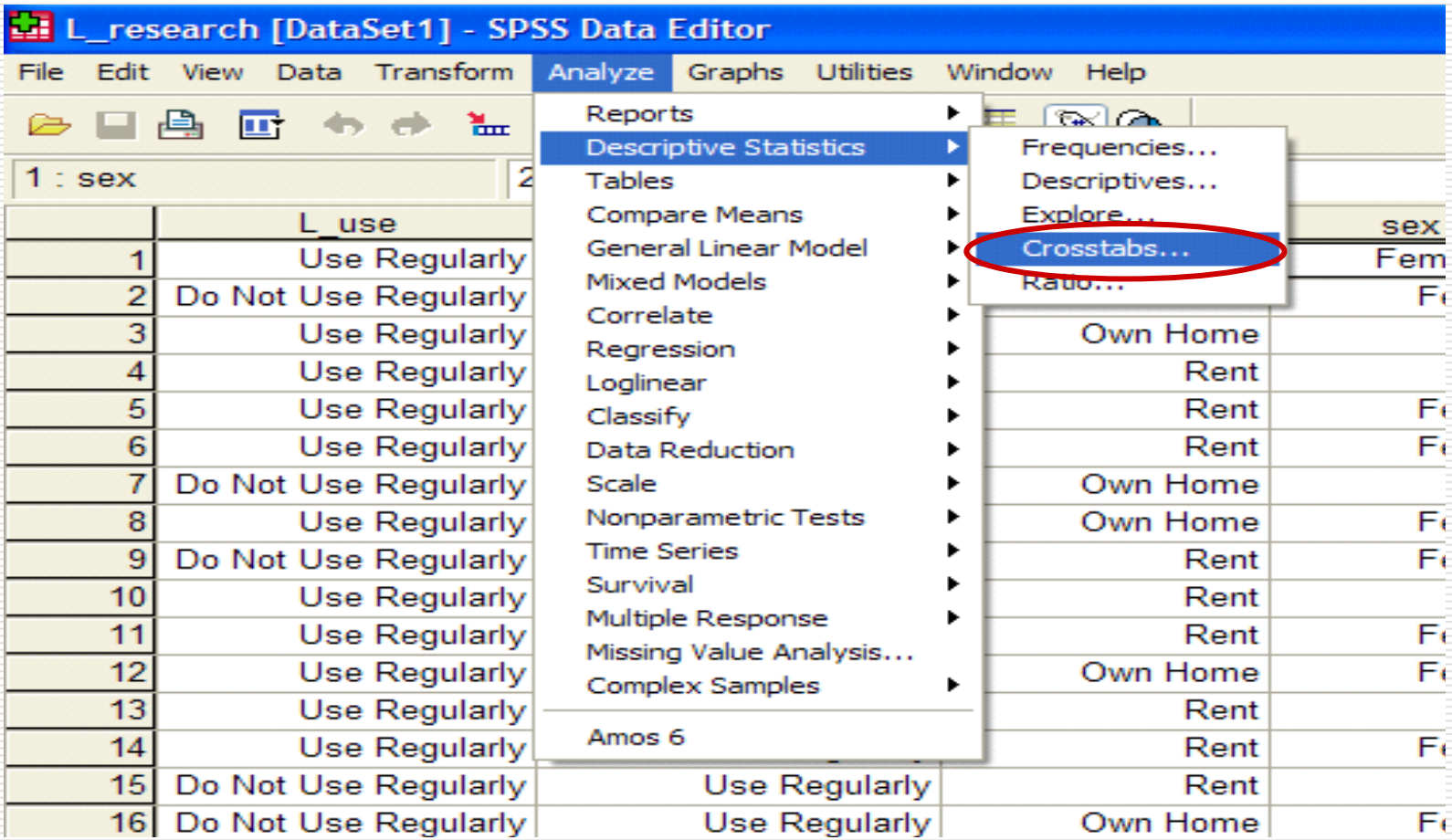
Σχέση δύο μεταβλητών

- Παραδείγματα:
  - Έχουν τα δύο καταστήματα (L και N) τους ίδιους πελάτες;
  - Ποιο είναι το δημογραφικό προφίλ των πελατών του κάθε καταστήματος (δηλ. ποια η σχέση των δημογραφικών μεταβλητών με την προτίμηση καταστήματος);
  - Ποιο είναι το προφίλ του τρόπου ζωής των πελατών του κάθε καταστήματος (όμοια για τις μεταβλητές τρόπου ζωής);

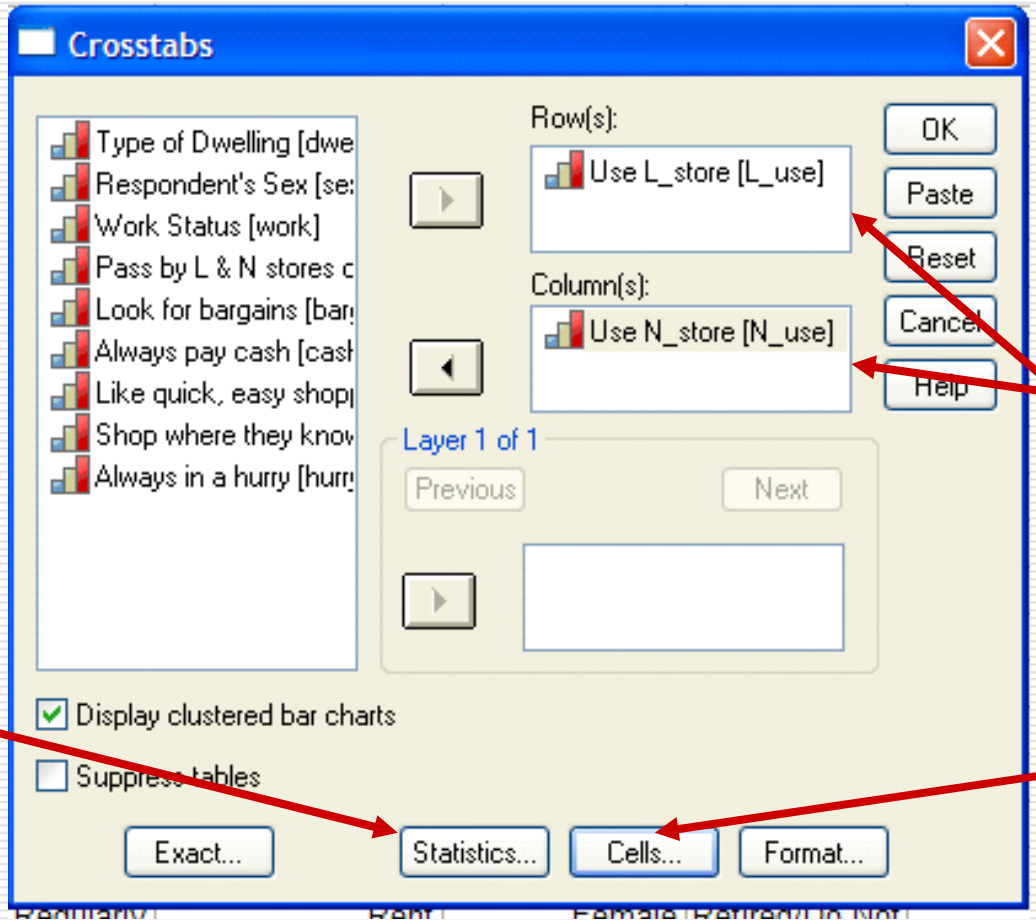
Παράδειγμα

- Οι μεταβλητές *L_use* και *N_use* έχουν τιμές:
  - 1=use regularly
  - 0=do not use regularly
- Μας ενδιαφέρει να πάρουμε την κοινή κατανομή τους
- Κατασκευή πίνακα συνάφειας **(contingency table)**

Διαδικασία **Crosstabs** – επιλογή



Διαδικασία **Crosstabs**: **Ο**ρισμός μεταβλητών

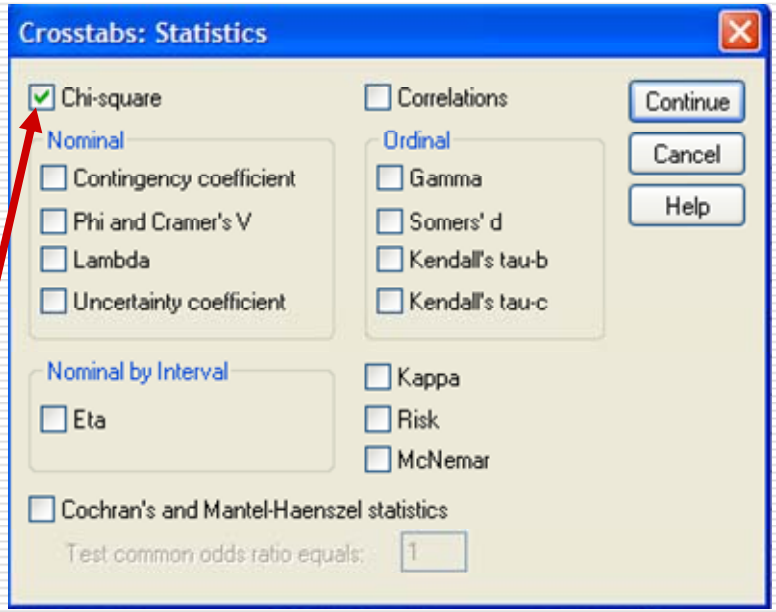


Στατιστικά μέτρα για έλεγχο της σχέσης

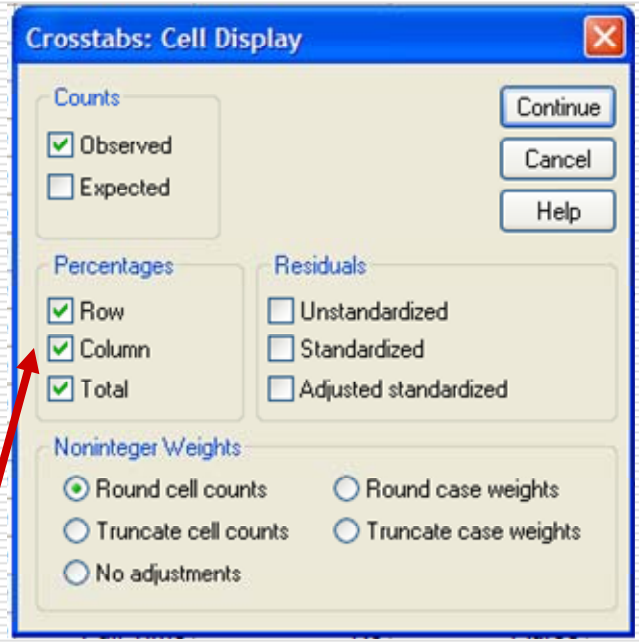
Ορισμός μεταβλητών στις γραμμές και στήλες του πίνακα

Μορφή κελιών πίνακα

Διαδικασία **Crosstabs**: ορισμός στατιστικών μέτρων



Στατιστικός έλεγχος ανεξαρτησίας χ^2



Εμφάνιση ποσοστών γραμμών, στηλών, συνολικό

Διαδικασία **Crosstabs**: Αποτελέσματα – πίνακας συνάφειας

Use L_store * Use N_store Crosstabulation

			Use N_store		
			Do Not Use Regularly	Use Regularly	Total
Use L_store	Do Not Use Regularly	Count	13	77	90
		% within Use L_store	14,4%	85,6%	100,0%
		% within Use N_store	33,3%	62,6%	55,6%
		% of Total	8,0%	47,5%	55,6%
	Use Regularly	Count	26	46	72
		% within Use L_store	36,1%	63,9%	100,0%
		% within Use N_store	66,7%	37,4%	44,4%
		% of Total	16,0%	28,4%	44,4%
Total		Count	39	123	162
		% within Use L_store	24,1%	75,9%	100,0%
		% within Use N_store	100,0%	100,0%	100,0%
		% of Total	24,1%	75,9%	100,0%

- Από αυτούς που χρησιμοποιούν τακτικά το L, το 63,9% χρησιμοποιεί τακτικά το N.
- Από αυτούς που χρησιμοποιούν τακτικά το N, το 37,4% χρησιμοποιεί τακτικά το L.
- Από το σύνολο των ερωτηθέντων, το 28.4% χρησιμοποιεί τακτικά και τα δύο μαγαζιά

Διαδικασία **Crosstabs**: Αποτελέσματα – Στατιστικός έλεγχος

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	10,273[b]	1	,001		
Continuity Correction[a]	9,122	1	,003		
Likelihood Ratio	10,311	1	,001		
Fisher's Exact Test				,002	,001
Linear-by-Linear Association	10,210	1	,001		
N of Valid Cases	162				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 17,33.

Έλεγχος χ^2

- Επειδή sig.=0.001<0.05, προκύπτει ότι υπάρχει σημαντική σχέση ανάμεσα στις μεταβλητές
- Απορρίπτεται η υπόθεση της ανεξαρτησίας

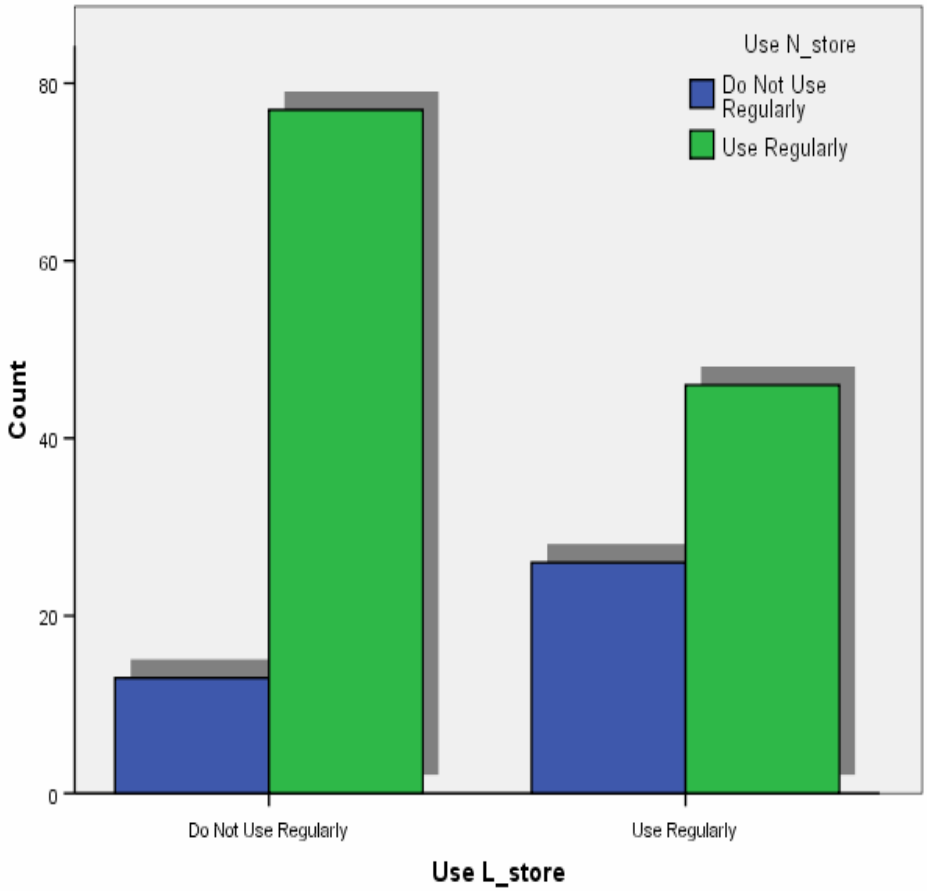
Διαδικασία **Crosstabs**: Αποτελέσματα – Ερμηνεία του ελέγχου

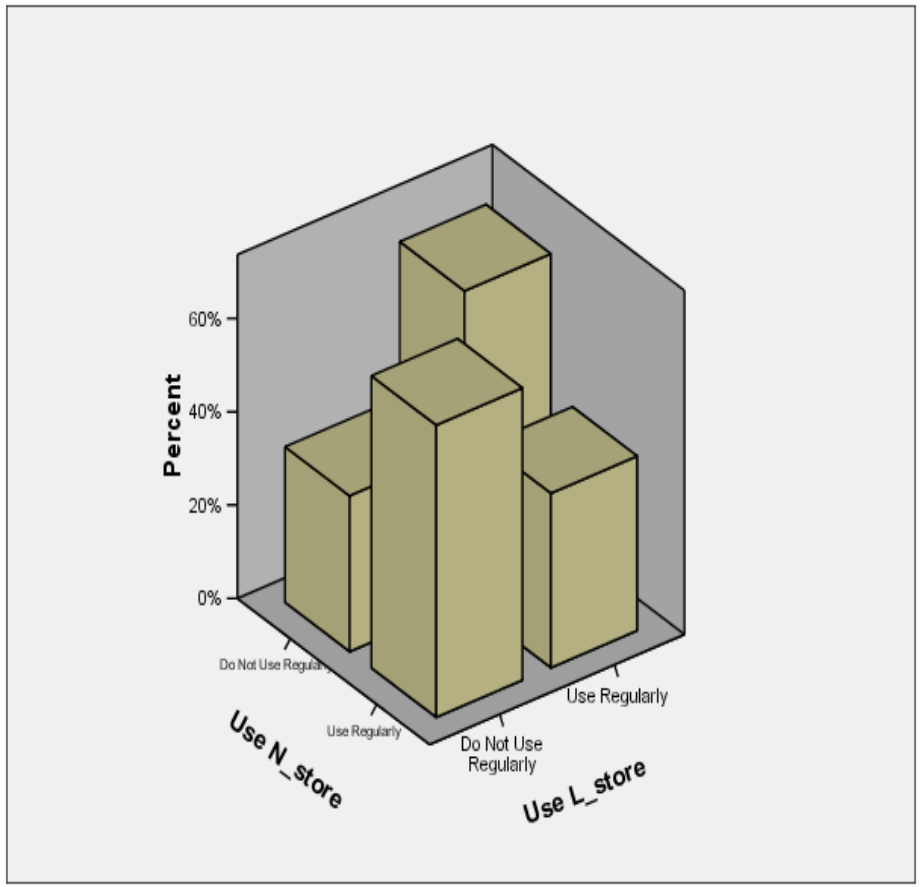
- Από την πλευρά του L:
  - Από αυτούς που δεν το επισκέπτονται συχνά (90 άτομα) το 85,6% χρησιμοποιεί συχνά το N
  - Από αυτούς που το επισκέπτονται συχνά (72 άτομα) το ποσοστό που χρησιμοποιεί συχνά το N είναι 63,9%
- Οι τακτικοί πελάτες του L συμπεριφέρονται ως προς το N διαφορετικά από τους περιστασιακούς (οι κατανομές διαφέρουν)

Διαδικασία **Crosstabs**: Αποτελέσματα – Ερμηνεία του ελέγχου

- Από την πλευρά του Ν:
  - Από αυτούς που δεν το επισκέπτονται συχνά (39 άτομα) το 66.7% χρησιμοποιεί συχνά το L
  - Από αυτούς που το επισκέπτονται συχνά (123 άτομα) το ποσοστό που χρησιμοποιεί συχνά το L είναι 37,4%.
- Οι τακτικοί πελάτες του Ν συμπεριφέρονται ως προς το L διαφορετικά από τους περιστασιακούς (οι κατανομές διαφέρουν)

Διαδικασία **Crosstabs**: Αποτελέσματα – Γραφική παράσταση





Άλλα ερωτήματα

- Υπάρχει σχέση ανάμεσα στις άλλες δημογραφικές μεταβλητές και στις *L_use* και *N_use*;
- Υπάρχει σχέση ανάμεσα στις μεταβλητές τρόπου ζωής και στις *L_use* και *N_use*;
- Να γίνουν οι απαραίτητοι έλεγχοι

Γενικά για σχέση ανάμεσα σε μεταβλητές

Ερωτήματα:

1. Υπάρχει; Έχουμε ενδείξεις σχέσης ανάμεσα σε δύο μεταβλητές που μελετάμε;
2. Ποια η φύση της; Είναι θετική ή αρνητική;
3. Ποια η ισχύς της; Πόσο ισχυρή είναι η σχέση;

Πίνακες συνάφειας

- Πίνακας συνάφειας: αποτελείται από γραμμές και στήλες που ορίζονται από τις κατηγορίες των δύο μεταβλητών
- Σε κάθε κελί υπάρχουν
  - Συχνότητα
  - Ποσοστό γραμμής
  - Ποσοστό στήλης
  - Ποσοστό στο σύνολο

Χρησιμότητα

- Οι πίνακες συνάφειας είναι ιδιαίτερα χρήσιμοι όταν έχουμε ονομαστικές μεταβλητές και θέλουμε να ελέγξουμε αν είναι συσχετισμένες
- Η ύπαρξη συστηματικής σχέσης ανιχνεύεται με τον έλεγχο χ^2 (Chi-Square test)

Ο έλεγχος χ^2

- Βασίζεται στον υπολογισμό ενός μέτρου από τις συχνότητες του πίνακα συνάφειας
- Η αρχική (μηδενική) υπόθεση είναι ότι οι δύο μεταβλητές δεν είναι συσχετισμένες
- Από τις παρατηρούμενες (Observed) συχνότητες υπολογίζονται οι αναμενόμενες (expected frequencies):

$$\text{Expected cell frequency} = \frac{\text{Cell column total} \times \text{Cell row total}}{\text{Grand total}}$$

Παράδειγμα

- Υπάρχει σχέση ανάμεσα στη μελέτη και στην επίδοση σε διαγώνισμα;
- Θα χρησιμοποιήσουμε τον έλεγχο χ^2 για να ελέγξουμε αν υπάρχει σημαντική συσχέτιση
- Θα προσπαθήσουμε να ορίσουμε τη φύση της σχέσης από τα ποσοστά γραμμών και στηλών

Μεταβλητές

- Did you study for the midterm test? __yes __no
- How did you perform on the midterm test? __pass __fail

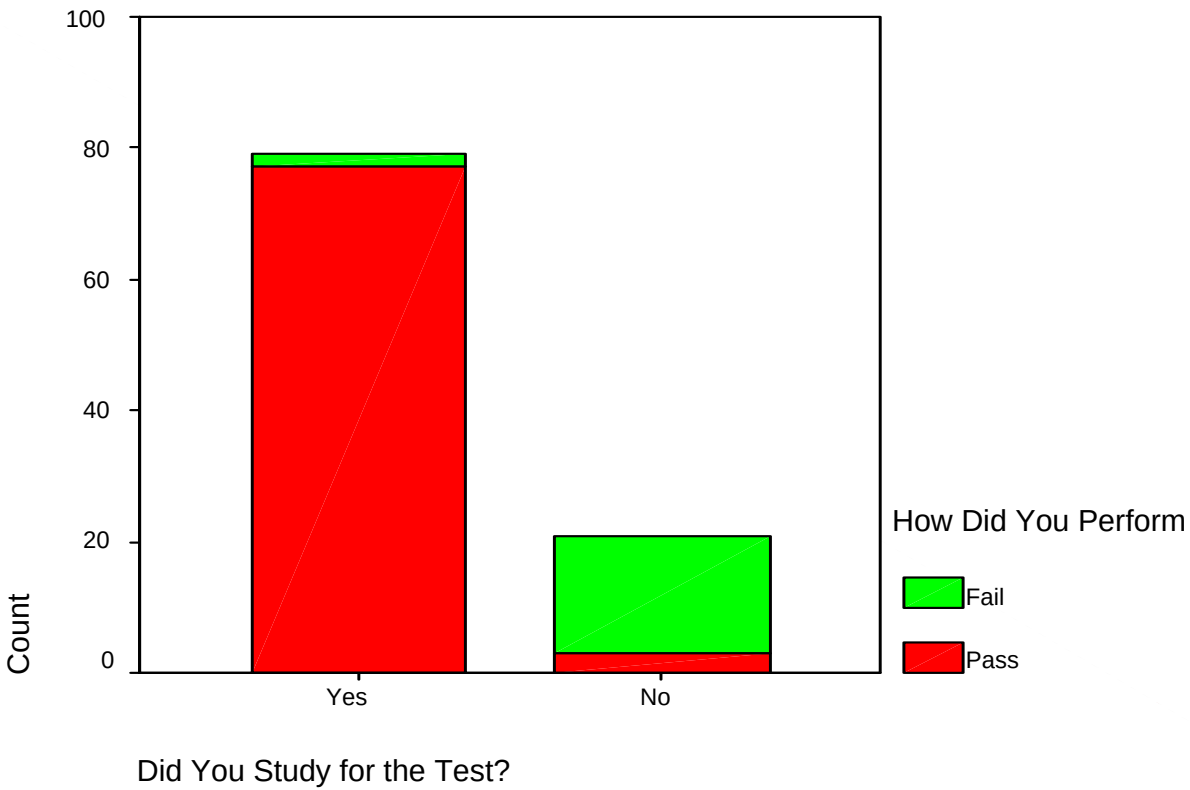
Did You Study for the Test? * How Did You Perform on the Test? Crosstabulation

Count

		How Did You Perform on the Test?		Total
		Pass	Fail	
Did You Study for the Test?	Yes	71	6	77
	No	7	16	23
Total		78	22	100

Cross-Tabulations

- Ύπαρξη συσχέτισης φανερή:
  - Επιτυχία με μελέτη



Σημαντικότητα της σχέσης

- Είναι η σχέση στατιστικά σημαντική;
- Είναι δηλ. συστηματική ή απλά έτυχε;
- Χρησιμοποιούμε τον έλεγχο χ^2
- χ^2: μέτρο απόστασης ανάμεσα στις παρατηρούμενες και τις αναμενόμενες συχνότητες
  - Παρατηρούμενες (Observed): οι συχνότητες των κελιών
  - Αναμενόμενες (Expected): Υπολογίζονται κάτω από την υπόθεση ότι δεν υπάρχει σχέση ανάμεσα στις δύο μεταβλητές

Υπολογισμός του χ^2

- Τύπος υπολογισμού:

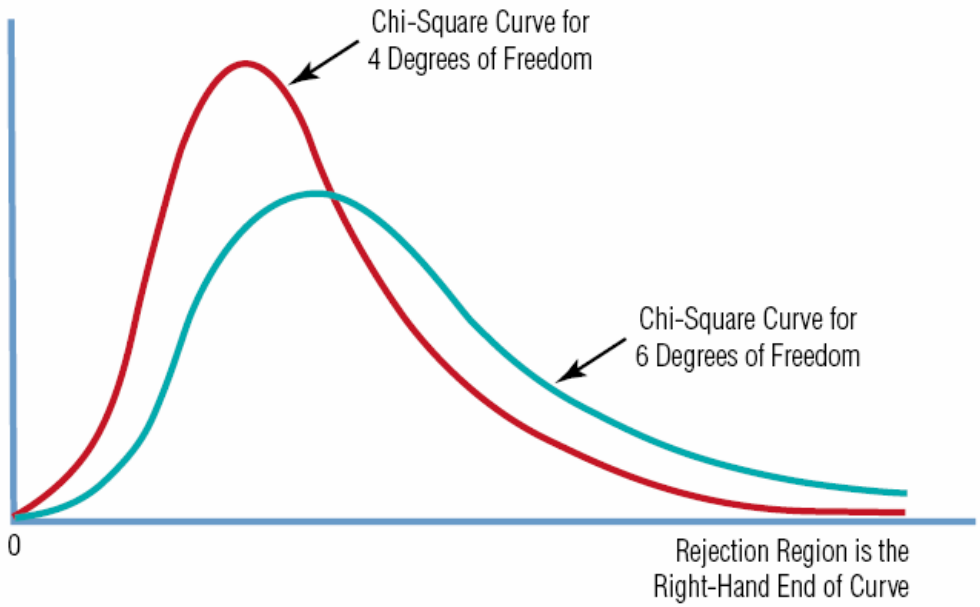
$$\chi^2 = \sum_{i-1}^{n} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

where

Observed_i = observed frequency in cell i
Expected_i = expected frequency in cell i
n = number of cells

Κατανομή του χ^2

- Ακολουθεί την χ^2 κατανομή της οποίας το σχήμα εξαρτάται από τους βαθμούς ελευθερίας
- Η τιμή του χ^2 που υπολογίζεται συγκρίνεται με τιμή από πίνακα για να φανεί η στατιστική σημαντικότητα



Ερμηνεία του χ^2 ελέγχου

- Πώς ερμηνεύεται η σημαντικότητα (significance) p;
  - Είναι η πιθανότητα να βρούμε στοιχεία που να στηρίζουν την αρχική υπόθεση (μη ύπαρξη σχέσης) αν η διαδικασία επαναληφθεί πολλές φορές με ανεξάρτητα δείγματα
  - Αν η τιμή p είναι ≤ 0.05, η πιθανότητα στήριξης της αρχικής υπόθεσης είναι μικρή
  - Επομένως υπάρχει σημαντική σχέση ανάμεσα στις δύο μεταβλητές

Ανάγνωση των αποτελεσμάτων του SPSS

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	39.382[b]	1	.000		
Continuity Correction[a]	35.865	1	.000		
Likelihood Ratio	34.970	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	100				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.06.

P=0.000... <0.05 $\Rightarrow$ σημαντική σχέση

Φύση της σχέσης από τα ποσοστά

Did You Study for the Test? * How Did You Perform on the Test? Crosstabulation

			How Did You Perform on the Test?		
			Pass	Fail	Total
Did You Study for the Test?	Yes	Count	71	6	77
		% within Did You Study for the Test?	92.2%	7.8%	100.0%
	No	Count	7	16	23
		% within Did You Study for the Test?	30.4%	69.6%	100.0%
Total		Count	78	22	100
		% within Did You Study for the Test?	78.0%	22.0%	100.0%

- 92% από αυτούς που μελέτησαν πέρασαν
- 70% από αυτούς που δε μελέτησαν απέτυχαν

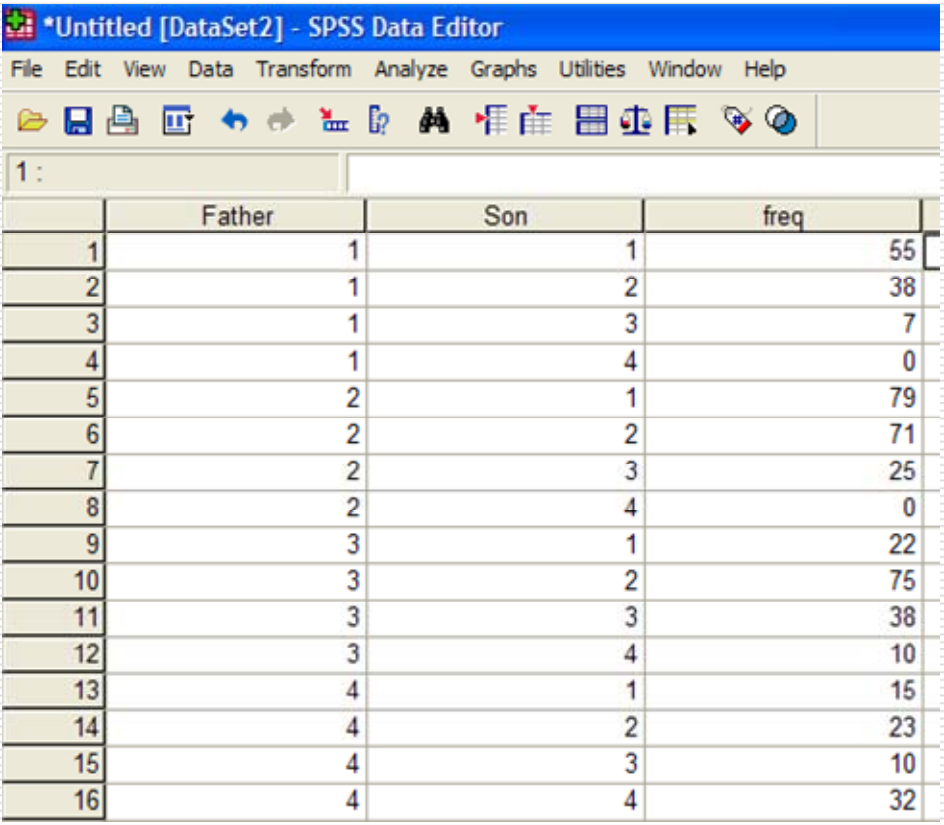
Εισαγωγή πίνακα συνάφειας στον επεξεργαστή δεδομένων

- Παράδειγμα:
- Τα αγόρια έχουν την τάση να ακολουθούν το επάγγελμα του πατέρα τους;
- Έρευνα σε 500 άνδρες με ερωτηματολόγια
- Μεταβλητές:
  - Επάγγελμα πατέρα (father)
  - Επάγγελμα γιου (son)

Δεδομένα με μορφή πίνακα

		son			
		Professional or Business	Skilled	Unskilled	Farmer
father	Professional or Business	55	38	7	0
	Skilled	79	71	25	0
	Unskilled	22	75	38	10
	Farmer	15	23	10	32

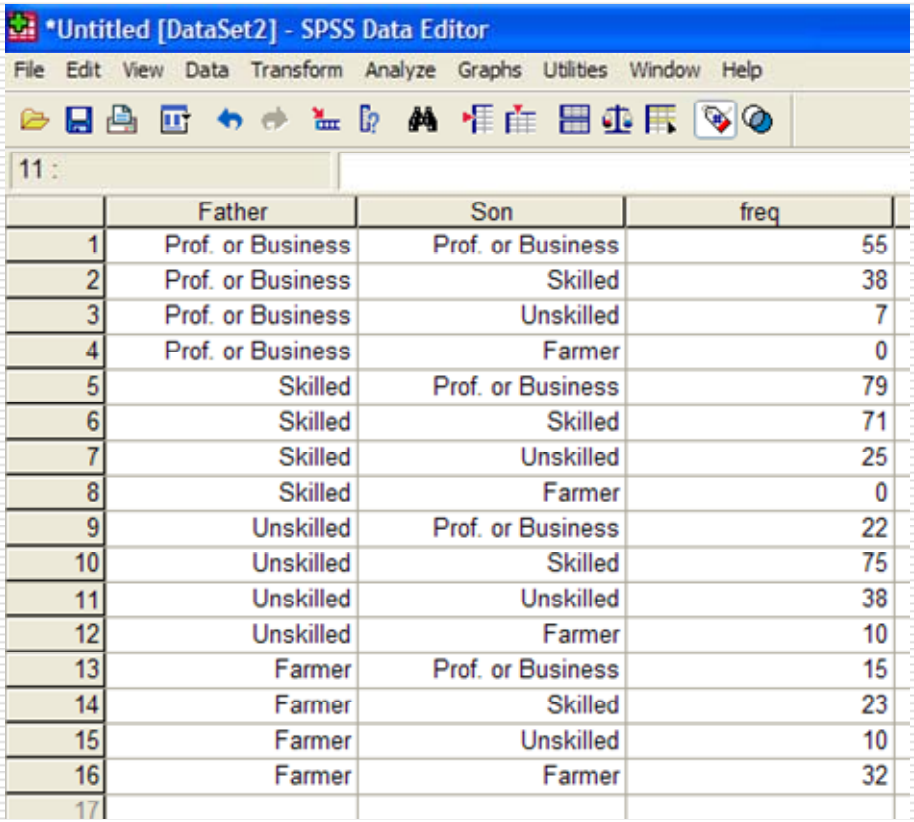
Εισαγωγή του πίνακα


*Untitled [DataSet2] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 :

	Father	Son	freq
1	1	1	55
2	1	2	38
3	1	3	7
4	1	4	0
5	2	1	79
6	2	2	71
7	2	3	25
8	2	4	0
9	3	1	22
10	3	2	75
11	3	3	38
12	3	4	10
13	4	1	15
14	4	2	23
15	4	3	10
16	4	4	32

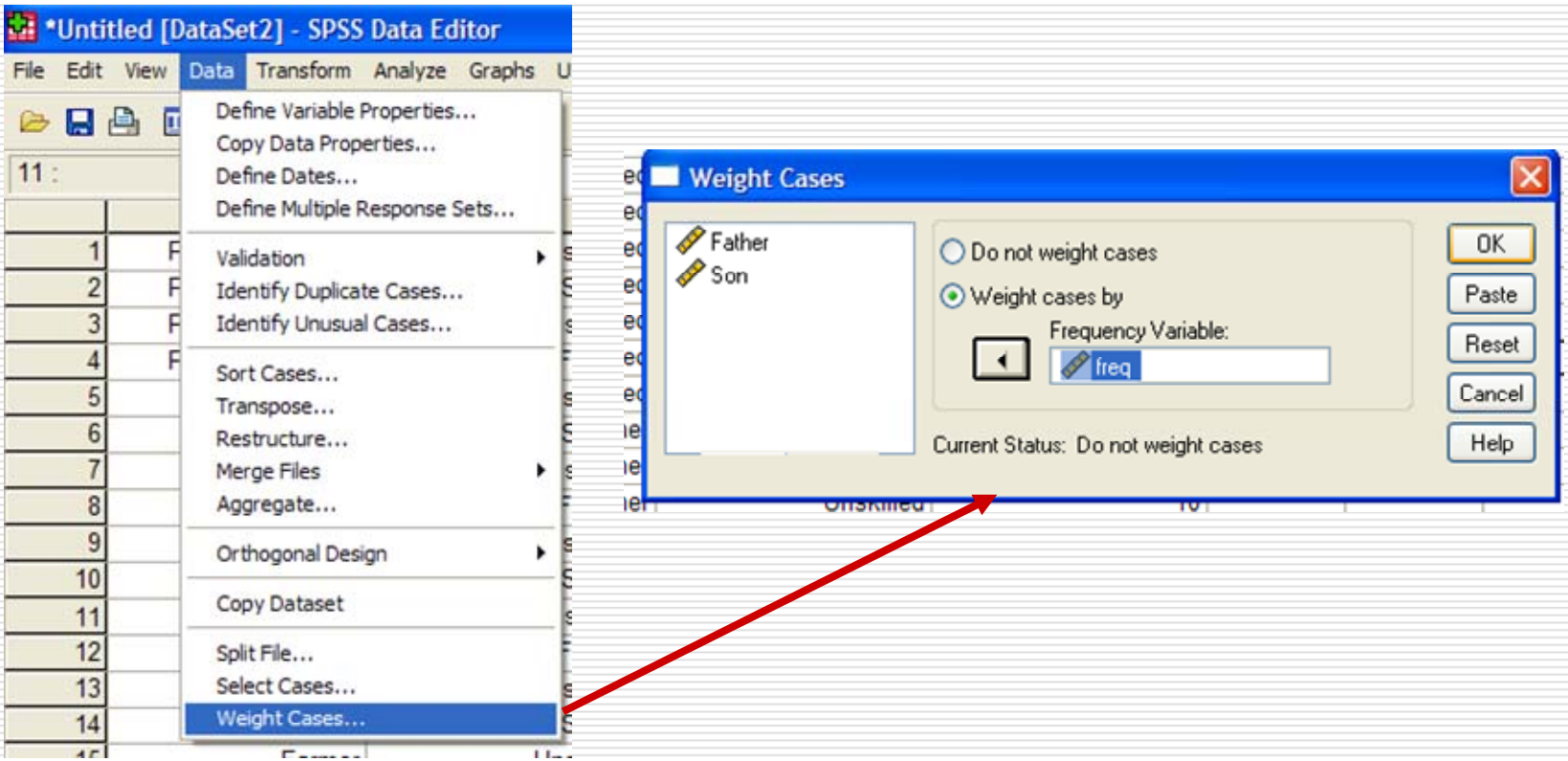

*Untitled [DataSet2] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

11 :

	Father	Son	freq
1	Prof. or Business	Prof. or Business	55
2	Prof. or Business	Skilled	38
3	Prof. or Business	Unskilled	7
4	Prof. or Business	Farmer	0
5	Skilled	Prof. or Business	79
6	Skilled	Skilled	71
7	Skilled	Unskilled	25
8	Skilled	Farmer	0
9	Unskilled	Prof. or Business	22
10	Unskilled	Skilled	75
11	Unskilled	Unskilled	38
12	Unskilled	Farmer	10
13	Farmer	Prof. or Business	15
14	Farmer	Skilled	23
15	Farmer	Unskilled	10
16	Farmer	Farmer	32
17			

Στάθμιση των cases



Αποτελέσματα crosstabs

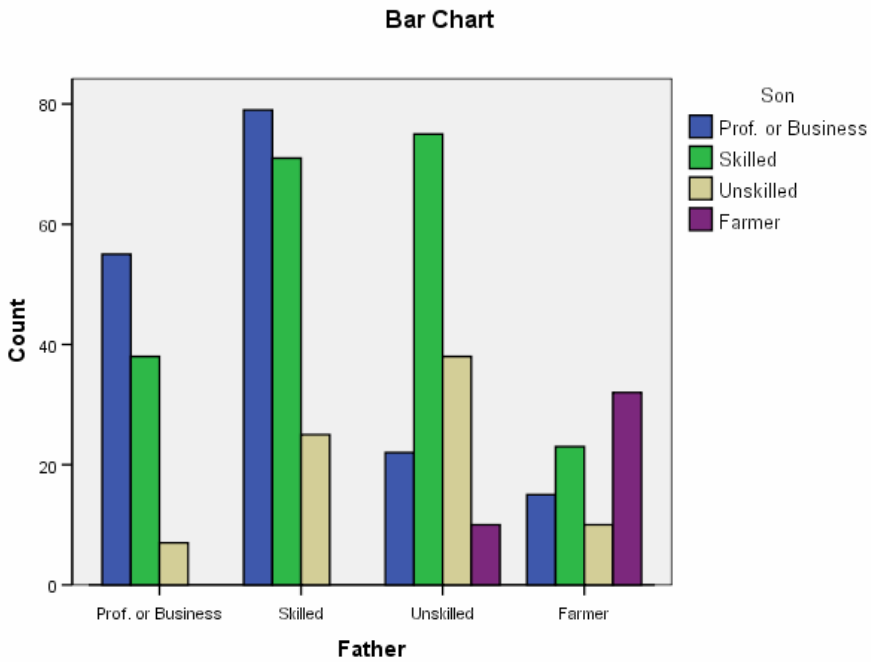
Father * Son Crosstabulation

			Son				Total
			Prof. or Business	Skilled	Unskilled	Farmer	
Father	Prof. or Business	Count	55	38	7	0	100
		% within Father	55,0%	38,0%	7,0%	,0%	100,0%
	Skilled	Count	79	71	25	0	175
		% within Father	45,1%	40,6%	14,3%	,0%	100,0%
	Unskilled	Count	22	75	38	10	145
		% within Father	15,2%	51,7%	26,2%	6,9%	100,0%
	Farmer	Count	15	23	10	32	80
		% within Father	18,8%	28,8%	12,5%	40,0%	100,0%
Total		Count	171	207	80	42	500
		% within Father	34,2%	41,4%	16,0%	8,4%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	180,874[a]	9	,000
Likelihood Ratio	160,832	9	,000
Linear-by-Linear Association	103,955	1	,000
N of Valid Cases	500		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,72.



Ερμηνεία αποτελεσμάτων

- Υπάρχει στατιστικά σημαντική σχέση ανάμεσα στα επαγγέλματα πατέρα – γιου
- Ερμηνεία για κάθε επάγγελμα πατέρα
- Π.χ. Το 40% των παιδιών από πατέρα αγρότη είναι επίσης αγρότες
- Η σχέση είναι πολύ ισχυρή