Βιοστατιστική
5ο Εργαστήριο SPSS

Διαστήματα εμπιστοσύνης
Έλεγχοι υποθέσεων
(δύο δείγματα)

χ^2 έλεγχος

Κατανομές Στατιστικών Συναρτήσεων

Δύο ανεξάρτητα δείγματα από κανονική κατανομή

Έστω $\underset{\sim}{X} = (X_1, X_2, ..., X_{n_1})$ **τ.δ. από** $N(\mu_1, \sigma_1^2)$ **μεγέθους** n_1 **και**

$\underset{\sim}{Y} = (Y_1, Y_2, ..., Y_{n_2})$ **τ.δ. από** $N(\mu_2, \sigma_2^2)$ **μεγέθους** n_2**, ανεξάρτητα, τότε**

$$T_1(\underset{\sim}{X}, \underset{\sim}{Y}) = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1), \quad T_2(\underset{\sim}{X}, \underset{\sim}{Y}) = \frac{S_X^2}{S_Y^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1, n_2-1},$$

Εάν $\sigma_1^2 = \sigma_2^2$ **τότε** $T_3(\underset{\sim}{X}, \underset{\sim}{Y}) = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_{pooled}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$

όπου $S_{pooled}^2 = \dfrac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{(n_1 + n_2 - 1)}$

Κατανομές Στατιστικών Συναρτήσεων

Δύο δείγματα από κανονική κατανομή

ζευγαρωτές παρατηρήσεις

(n_1= n_2=n, ίδιες δειγματικές μονάδες)

Ορίζουμε τις τ.μ. $\underset{\sim}{Z} = (Z_1, Z_2, ..., Z_n)$ **ως** $Z_i = X_i - Y_i$ **για i=1,...,n**

με $E(Z_i)=\mu_1-\mu_2$ **και** $Var(Z_i)=\sigma_1^2+\sigma_2^2$ **για i=1,2,...,n**

$$T_1(\underset{\sim}{X}, \underset{\sim}{Y}) = \frac{\overline{Z} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}} \sim N(0,1),$$

$$T_3(\underset{\sim}{X}, \underset{\sim}{Y}) = \frac{\overline{Z} - (\mu_1 - \mu_2)}{\frac{S_Z}{\sqrt{n}}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\frac{S_Z}{\sqrt{n}}} \sim t_{n-1}$$

Κατανομές Στατιστικών Συναρτήσεων

Δύο δείγματα (μεγάλα) από οποιαδήποτε κατανομή

Έστω $\underset{\sim}{X} = (X_1, X_2, ..., X_{n_1})$ **τ.δ.** $(n_1 \gg 25)$ **από κατανομή**

με $E(X_i)=\mu_1$ **και** $Var(X_i)=\sigma_1^2$ $i=1,2,...,n_1$

$\underset{\sim}{Y} = (Y_1, Y_2, ..., Y_{n_2})$ **τ.δ.** $(n_2 \gg 25)$**, ανεξάρτητο του** $\underset{\sim}{X}$**, από κατανομή**

με $E(Y_i)=\mu_2$ **και** $Var(Y_i)=\sigma_2^2$ $i=1,2,...,n_2$ **τότε (από το ΚΟΘ)**

$$T_1(\underset{\sim}{X}, \underset{\sim}{Y}) = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1),$$

$$T_3(\underset{\sim}{X}, \underset{\sim}{Y}) = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \sim N(0,1)$$

Διαστήματα εμπιστοσύνης

Δύο δείγματα από κανονική κατανομή

Διάστημα εμπιστοσύνης ίσων ουρών σ.ε. 1-α

για το μ_1-μ_2, όταν τα $\sigma_1^2=\sigma_2^2=\sigma^2$ είναι άγνωστα (βλέπε διαφάνεια 8)

$$\left(\bar{x}-\bar{y}-t_{n_1+n_2-2}(a/2)S_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \bar{x}-\bar{y}+t_{n_1+n_2-2}(a/2)S_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

όταν τα $\sigma_1^2 \neq \sigma_2^2$ και είναι άγνωστα

$$\left(\bar{x}-\bar{y}-t_v(a/2)\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}},\ \bar{x}-\bar{y}+t_v(a/2)\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}\right)$$

όπου $v=\dfrac{\left(\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}\right)^2}{\frac{\left(\frac{S_X^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_Y^2}{n_2}\right)^2}{n_2-1}}$ **(στρογγυλευμένο στον πλησιέστερο ακέραιο)**

όταν $n_1 \neq n_2$

και $v=2(n-1)$ **όταν** $n_1=n_2=n$

Διάστημα εμπιστοσύνης-Έλεγχοι υποθέσεων

Δύο δείγματα από κανονική κατανομή

ΧΡΟΝΟΣ ΠΟΥ ΑΠΑΙΤΗΘΗΚΕ (σε λεπτά)	
Συνήθης Μέθοδος Εκπαίδευσης X	Νέα Μέθοδος Εκπαίδευσης Y
32	35
37	31
35	29
28	25
41	34
44	40
35	27
31	32
34	31

Να κατασκευασθεί δ.ε. ίσων ουρών σ.ε. 90% για την διαφορά μεταξύ των μέσων χρόνων που απαιτούνται για την συναρμολόγηση του συγκεκριμένου προϊόντος μετά από περίοδο εκπαίδευσης δύο εβδομάδων με τη συνήθη και τη νέα μέθοδο εκπαίδευσης, από 18 διαφορετικούς εργάτες οι οποίοι εκπαιδεύτηκαν με τις δύο μεθόδους.

Διάστημα εμπιστοσύνης-Έλεγχοι υποθέσεων

Δύο δείγματα από κανονική κατανομή

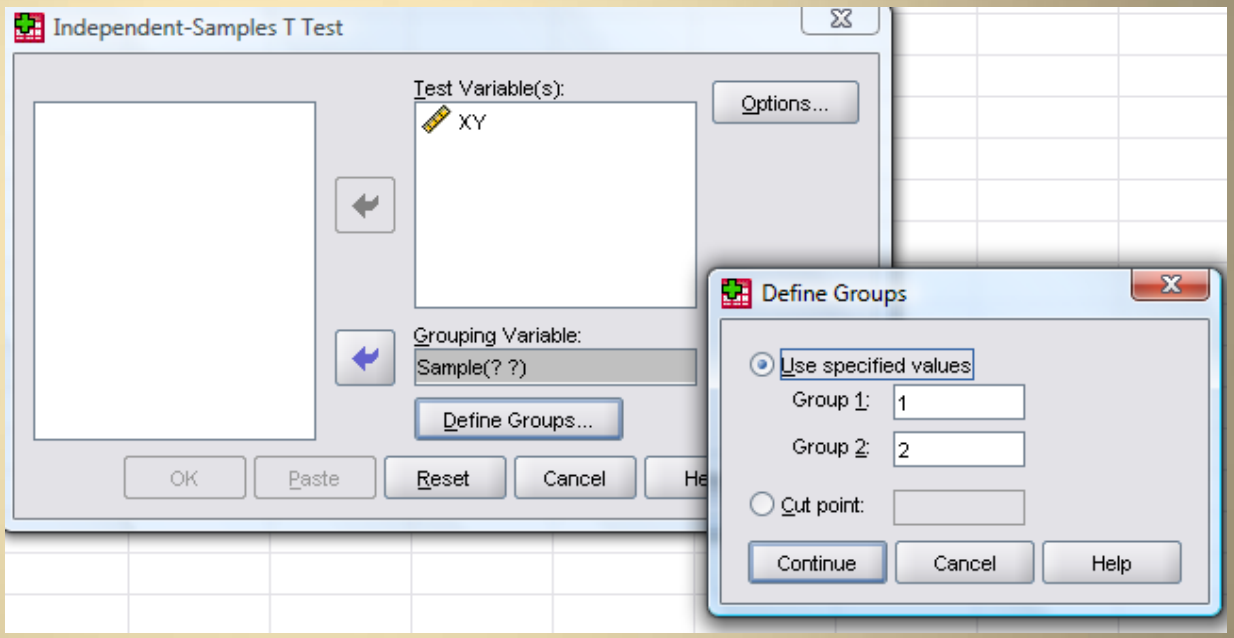
*Untitled1 [DataSet0] - SPSS Statistics Data E

File Edit View Data Transform Analyze

19 : XY

	XY	Sample
1	32,00	1
2	37,00	1
3	35,00	1
4	28,00	1
5	41,00	1
6	44,00	1
7	35,00	1
8	31,00	1
9	34,00	1
10	35,00	2
11	31,00	2
12	29,00	2
13	25,00	2
14	34,00	2
15	40,00	2
16	27,00	2
17	32,00	2
18	31,00	2
19		
20		

Click Analyze >Compare Means
>Independent-Samples T Test



Διάστημα εμπιστοσύνης-Έλεγχοι υποθέσεων

Δύο δείγματα από κανονική κατανομή

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
XY	Equal variances assumed	,061	,807	1,649	16
	Equal variances not assumed			1,649	15,844

$H_0 : \sigma_1^2 = \sigma_2^2$ αμφίπλευρος

$H_1 : \sigma_1^2 \neq \sigma_2^2$

ο Levene έλεγχος εφαρμόζεται για οποιαδήποτε δείγματα. Ενώ ο έλεγχος διασπορών που βασίζεται στην F κατανομή (διαφ. 2) προϋποθέτει τα δείγματα να προέρχονται από κανονική κατανομή

Independent Samples Test

		t-test for Equality of Means		
		Sig. (2-tailed)	Mean Difference	Std. Error Difference
XY	Equal variances assumed	,119	3,66667	2,22292
	Equal variances not assumed	,119	3,66667	2,22292

t=Mean Differ/Std. Error Diff

$H_0 : \mu_1 = \mu_2$ (σε ε.σ.≤ 11.9%)

$H_1 : \mu_1 \neq \mu_2$ αμφίπλευρος

Independent Samples Test

		t-test for Equality of Means	
		90% Confidence Interval of the Difference	
		Lower	Upper
XY	Equal variances assumed	-,21429	7,54762
	Equal variances not assumed	-,21664	7,54997

άρα το 90% δ.ε. περιέχει το 0

Group Statistics

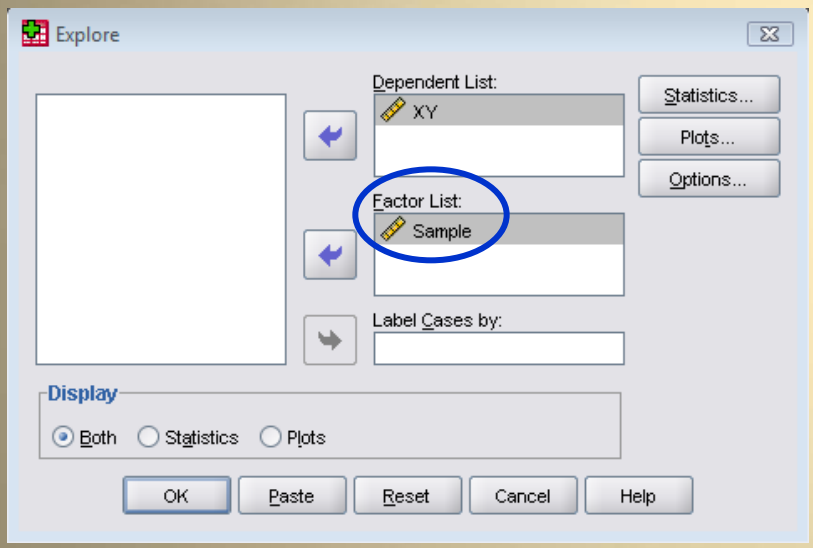
	Sample	N	Mean	Std. Deviation	Std. Error Mean
XY	1	9	35,2222	4,94413	1,64804
	2	9	31,5556	4,47524	1,49175

Διάστημα εμπιστοσύνης-Έλεγχοι υποθέσεων

Δύο δείγματα από κανονική κατανομή

Για τα δείγματα που μελέτησα, επειδή $n_1 = n_2 = 9 << 25$, πρέπει να κάνω ελέγχους κανονικότητας.

Εάν δεν προέρχονται από κανονικές κατανομές τα προηγούμενα συμπεράσματα για τη διαφορά των μέσων δεν είναι έγκυρα.



Tests of Normality

	Sample	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
XY	1	,185	9	,200*	,966	9	,863
	2	,127	9	,200*	,977	9	,946

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

άρα και τα δύο δείγματα προέρχονται από κανονική κατανομή

Διάστημα εμπιστοσύνης-Έλεγχοι υποθέσεων

Δύο δείγματα από κανονική κατανομή

ζευγαρωτές παρατηρήσεις

(n_1= n_2=n, ίδιες δειγματικές μονάδες)

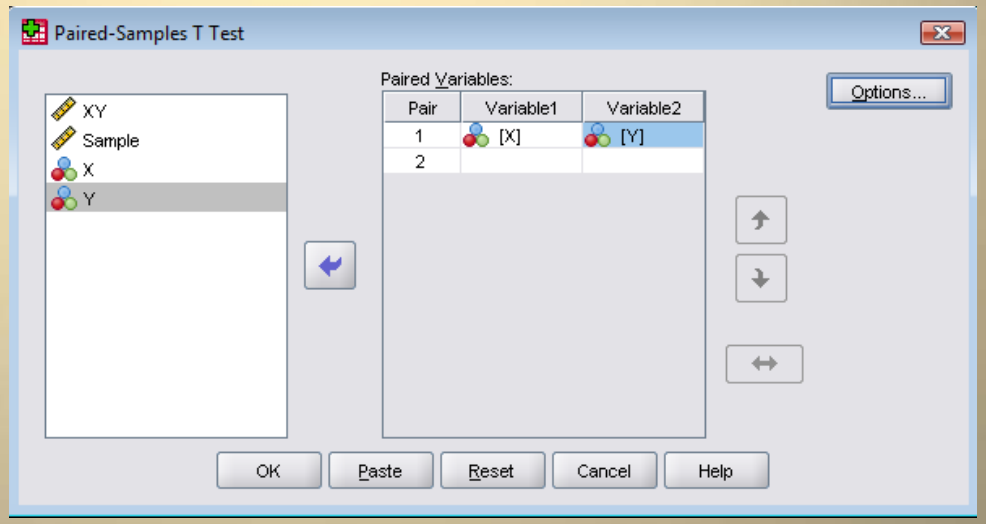
(εδώ καταχρηστικά, με την υπόθεση ότι έχω μόνο 9 εργάτες οι οποίοι αρχικά εκπαιδεύτηκαν και χρονομετρήθηκαν με τη συνήθη μέθοδο και στη συνέχεια εκπαιδεύτηκαν και χρονομετρήθηκαν με τη νέα μέθοδο)

Data Editor

Analyze Graphs Utilities

X	Y
32,00	35,00
37,00	31,00
35,00	29,00
28,00	25,00
41,00	34,00
44,00	40,00
35,00	27,00
31,00	32,00
34,00	31,00
.	.

Click Analyze >Compare Means
>Paired-Samples T Test



Διάστημα εμπιστοσύνης-Έλεγχοι υποθέσεων

Δύο δείγματα από κανονική κατανομή

ζευγαρωτές παρατηρήσεις

(n_1= n_2=n, ίδιες δειγματικές μονάδες)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	X	35,2222	9	4,94413	1,64804
	Y	31,5556	9	4,47524	1,49175

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	X & Y	9	,700	,036

Paired Samples Test

		Paired Differences				
					90% Confidence Interval of the Difference	
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper
Pair 1	X - Y	3,66667	3,67423	1,22474	1,38919	5,94414

το 90% δ.ε. δεν περιέχει το 0

Paired Samples Test

		t	df	Sig. (2-tailed)
Pair 1	X - Y	2,994	8	,017

έχω διαφορετικό συμπέρασμα

$H_0 : \mu_1 = \mu_2$ αμφίπλευρος

$H_1 : \mu_1 \neq \mu_2$ (αποδ. H_1 σε ε.σ.≥ 1.7%)

Έλεγχος για ισότητα μέσων

Δύο ανεξάρτητα δείγματα μικρού μεγέθους $n_1, n_2 <<$

1. Ελέγχουμε εάν τα δείγματα προέρχονται από κανονική κατανομή, χρησιμοποιώντας τους ελέγχους Kolmogorov-Smirnov και Shapiro-Wilk
2. Εάν αποδεχθούμε αυτήν την υπόθεση (αλλιώς έλεγχος δεν μπορεί να γίνει), με τον έλεγχο Levene εξετάζουμε εάν οι διασπορές είναι ίσες

3α. Εάν οι διασπορές είναι ίσες, χρησιμοποιούμε στο t-test το δ.ε. και το p-value (Sign.) της $1^{ης}$ γραμμής

3β. Εάν οι διασπορές είναι άνισες, χρησιμοποιούμε στο t-test το δ.ε. και το p-value (Sign.) της $2^{ης}$ γραμμής

Έλεγχος για ισότητα μέσων

Ζευγαρωτές παρατηρήσεις δείγμα μικρού μεγέθους $n <<$

1. Ελέγχουμε εάν τα δείγματα προέρχονται από κανονική κατανομή
2. Εάν αποδεχθούμε αυτήν την υπόθεση (αλλιώς έλεγχος δεν μπορεί να γίνει), κάνουμε στο paired t-test.

Έλεγχος για ισότητα μέσων

Δύο ανεξάρτητα δείγματα μεγάλου μεγέθους $n_1, n_2 >> 25, 30$

1. Δεν χρειάζεται έλεγχος για το εάν προέρχονται από κανονική κατανομή
2. Δεν εξετάζουμε εάν οι διασπορές είναι ίσες
3. Χρησιμοποιούμε στο t-test το δ.ε. και το p-value (Sign.) της 2ης γραμμής

Έλεγχος για ισότητα μέσων

Ζευγαρωτές παρατηρήσεις
δείγμα μεγάλου μεγέθους $n >> 25, 30$

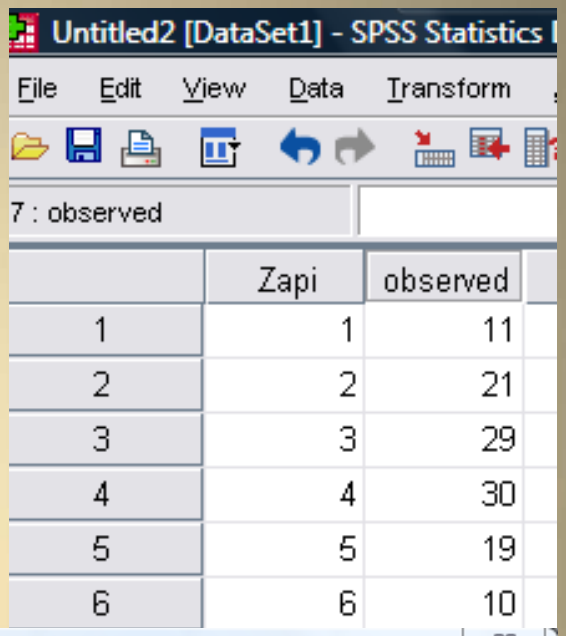
1. Δεν χρειάζεται έλεγχος για το εάν προέρχονται από κανονική κατανομή
2. Κάνουμε στο paired t-test.

Pearson's X^2 έλεγχος καλής προσαρμογής

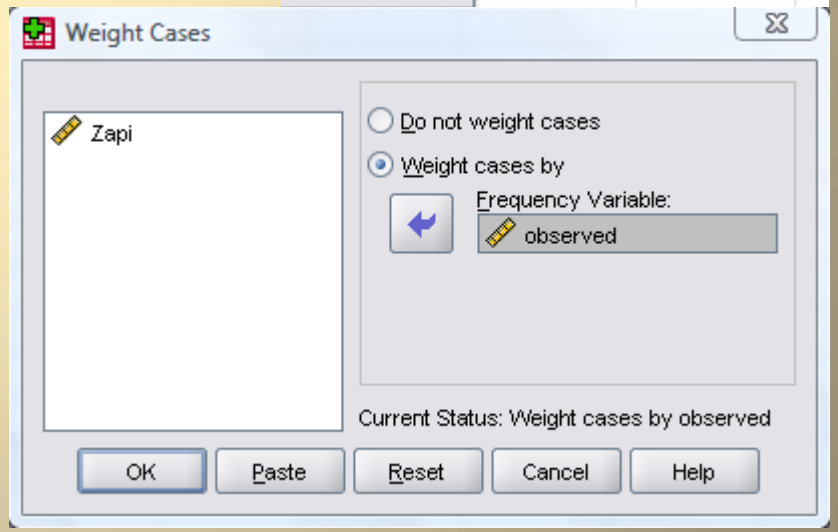
Ρίχνουμε ένα ζάρι 120 φορές και παίρνουμε

Αποτέλεσμα (Ai)	1	2	3	4	5	6
Αρ. Εμφανίσ. (πi)	11	21	29	30	19	10

Θέλουμε να ελέγξουμε εάν το ζάρι είναι δίκαιο

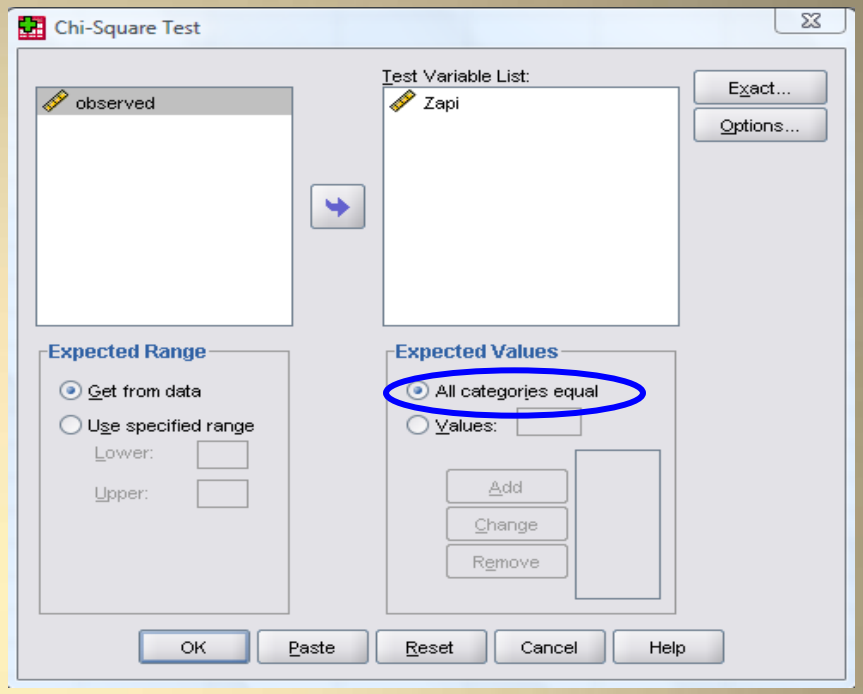


a. Εκχωρώ τιμές στις μεταβλητές Zapi και observed
b. Click Data
c. Weight Cases
d. Επιλέγω Weight cases by observed
e. Ok



Pearson's X^2 έλεγχος καλής προσαρμογής

a. Click Analyze
b. Nonparametric Tests
c. Chi-square Test
d. Επιλέγω Test Variable List Zapi
e. Επιλέγω All categories equal
f. Ok



Zapi

	Observed N	Expected N	Residual
1	11	20,0	-9,0
2	21	20,0	1,0
3	29	20,0	9,0
4	30	20,0	10,0
5	19	20,0	-1,0
6	10	20,0	-10,0
Total	120		

Test Statistics

	Zapi
Chi-Square	18,200[a]
df	5
Asymp. Sig.	,003

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 20,0.

$H_0 : p_1 = p_2 = \ldots = p_6 = 1/6$

H_1 : κάποιο p_i διαφέρει

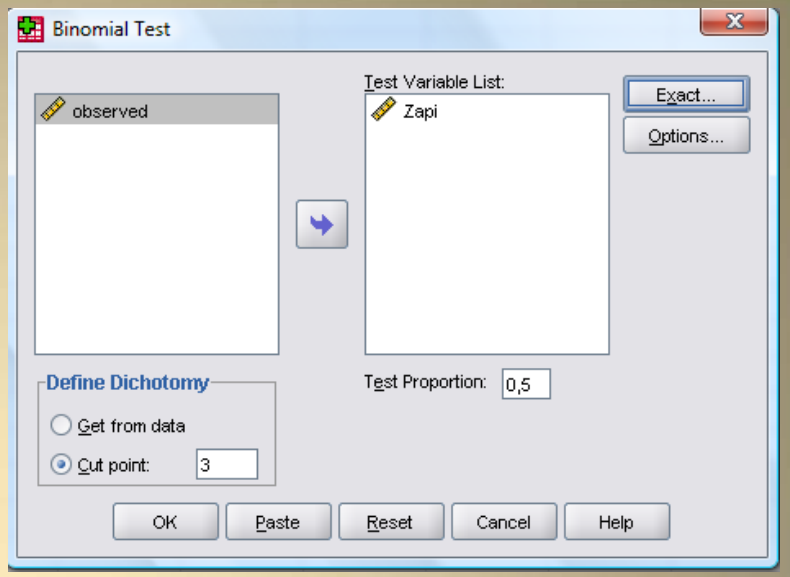
(αποδ. H_1 σε ε.σ.≥ 0.3%

Αμφίπλευρος

όπου $p_i = P(Z=i)$ $i=1,\ldots,6$

Έλεγχος για ποσοστό

a. Click Analyze
b. Nonparametric Tests
c. Binomial
d. Επιλέγω Test Variable List Zapi
e. Επιλέγω Test Proportion 0,5 (0,333)
f. Επιλέγω Cut point 3 (2)
g. Ok



Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
Zapi	Group 1	<= 3	61	,51	,50	,927[a]
	Group 2	> 3	59	,49		
	Total		120	1,00		

a. Based on Z Approximation.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Zapi	Group 1	<= 2	32	,267	,333	,072[a,b]
	Group 2	> 2	88	,733		
	Total		120	1,000		

a. Alternative hypothesis states that the proportion of cases in the first group < ,333.

b. Based on Z Approximation.

$H_0 : P(Z=1,2,3)=0.5$

(αποδ. H_0 σε ε.σ.≤ 92.7%)

$H_1 : P(Z=1,2,3)\neq 0.5$ αμφίπλευρος

$H_0 : P(Z=1,2)\geq 1/3$ μονόπλευρος

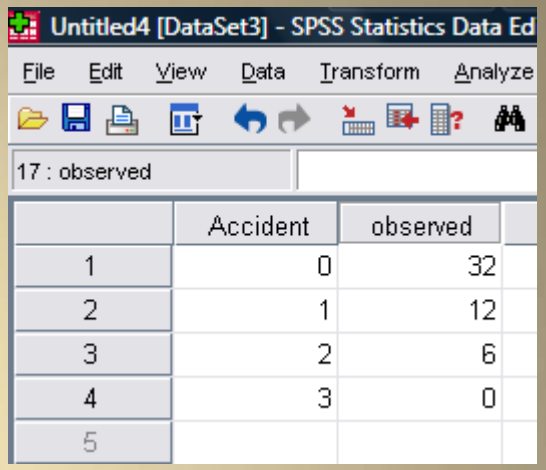
$H_1 : P(Z=1,2)<1/3$ (αποδ. H_1 σε ε.σ.≥ 7.2%)

Pearson's X^2 έλεγχος καλής προσαρμογής

Ο αριθμός των ατυχημάτων κατά εβδομάδα σε μια διασταύρωση μετρήθηκε για 50 εβδομάδες και έδωσε

αριθ. ατυχ	0	1	2	≥ 3
αριθ. εβδομ	32	12	6	0

Θέλουμε να ελέγξουμε εάν το πλήθος των ατυχημάτων ακολουθούν Poisson κατανομή



	Accident	observed
1	0	32
2	1	12
3	2	6
4	3	0
5		

Όμως η παράμετρος της Poisson κατανομής δεν μας δίδεται και άρα την εκτιμούμε από τον δειγματικό μέσο

$$\hat{\lambda} = \bar{x} = \frac{0\times 32 + 1\times 12 + 2\times 6}{50} = 0.48$$

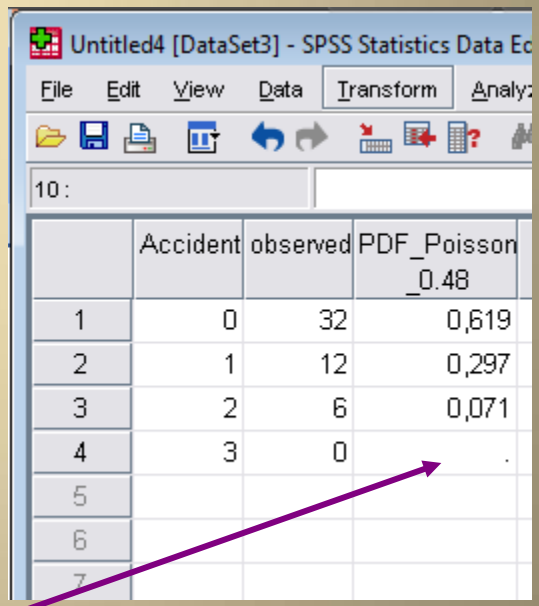
και υπολογίζουμε τις πιθανότητες που μας ενδιαφέρουν

$\hat{p}_1 = P(X=0) = e^{-\lambda} = 0.619$

$\hat{p}_2 = P(X=1) = e^{-\lambda}\lambda = 0.297$

$\hat{p}_3 = P(X=2) = e^{-\lambda}\frac{\lambda^2}{2!} = 0.071$

$\hat{p}_4 = P(X\geq 3) = 1-\hat{p}_1-\hat{p}_2-\hat{p}_3 = 0.013$

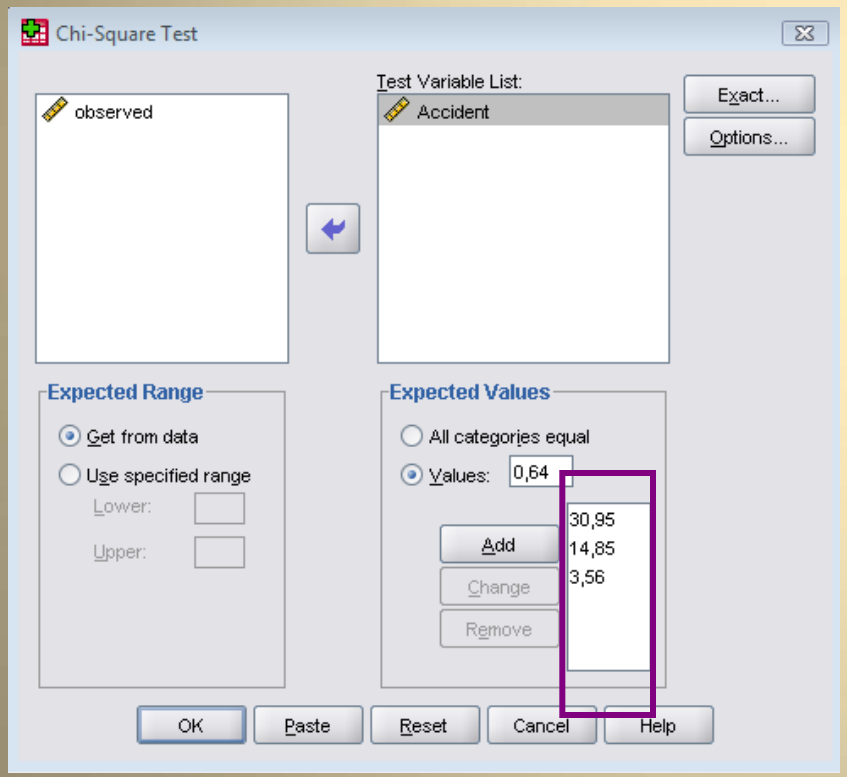


	Accident	observed	PDF_Poisson_0.48
1	0	32	0,619
2	1	12	0,297
3	2	6	0,071
4	3	0	.
5			
6			

Pearson's X^2 έλεγχος καλής προσαρμογής

κ = 4	A_1	A_2	A_3	A_4	Άθ.
X_i	0	1	2	≥ 3	
O_i	32	12	6	0	50
p_i	0,619	0,297	0,063	0,021	1
θ_i	30,95	14,85	3.56	0.64	50

Οι θεωρητικές τιμές προκύπτουν ως $\theta_i = n * p_i$



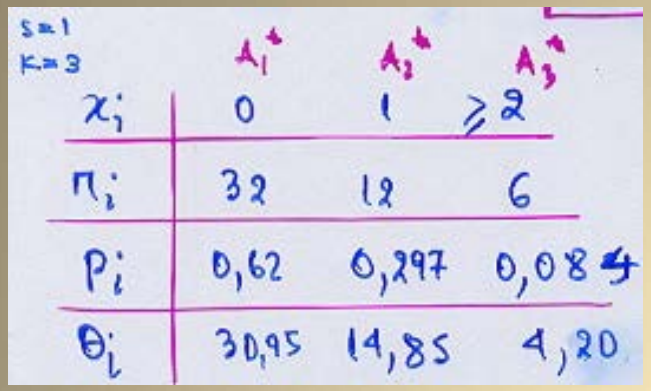
Accident

	Observed N
0	32
1	12
2	6
Total	50[a]

a. 3 cells were generated, but 4 expected frequencies were specified. Chi-Square Test cannot be performed.

οπότε συνενώνουμε τις δύο τελευταίες κατηγορίες

Pearson's X^2 έλεγχος καλής προσαρμογής



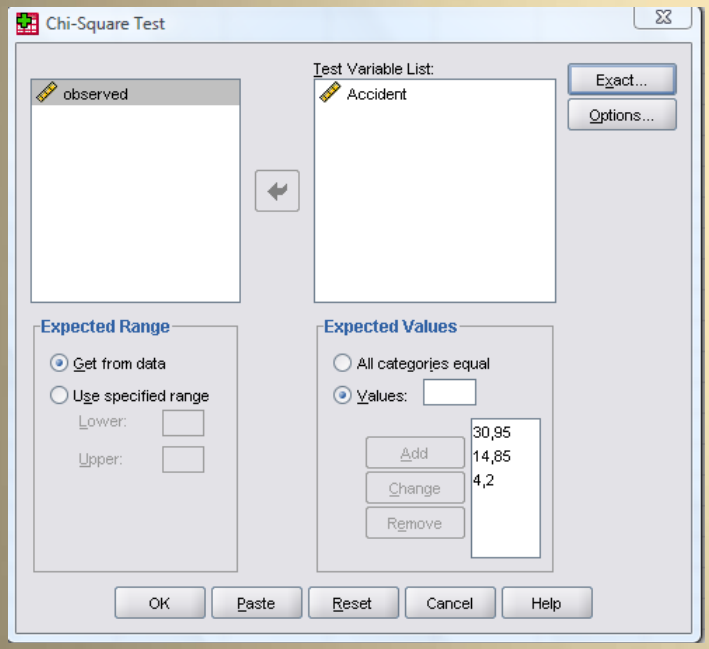
Accident

	Observed N	Expected N	Residual
0	32	31,0	1,1
1	12	14,9	-2,8
2	6	4,2	1,8
Total	50		

Test Statistics

	Accident
Chi-Square	1,354[a]
df	2
Asymp. Sig.	,508

a. 1 cells (33,3%) have expected frequencies less than 5. The minimum expected cell frequency is 4,2.



Προσοχή: εδώ οι βαθμοί ελευθερίας εμφανίζονται ως 2 γεγονός που δε λαμβάνει υπόψη τη μια εκτιμώμενη παράμετρο της Poisson.

Έτσι το p-value (0.508) είναι λανθασμένο

p-value=1-CDF.CHISQ(1.354,2-1)= 0.2446

H_0 : το δείγμα προέρχεται από Poisson κατανομή (σε ε.σ.≤24.46%)

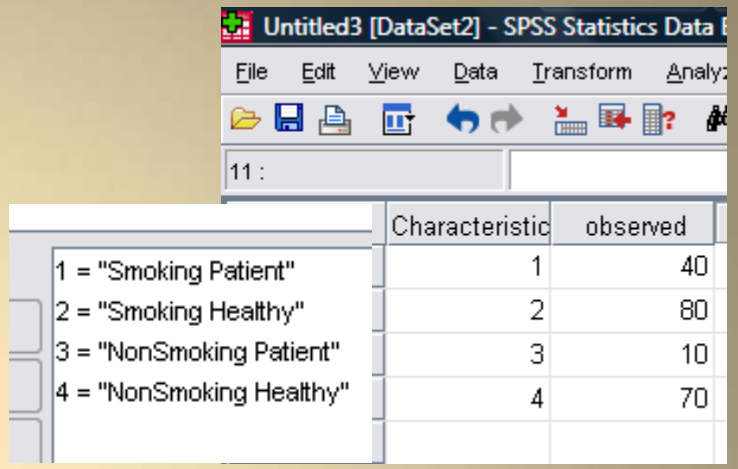
H_1 : δεν προέρχεται από Poisson

Pearson's X^2 έλεγχος ανεξαρτησίας

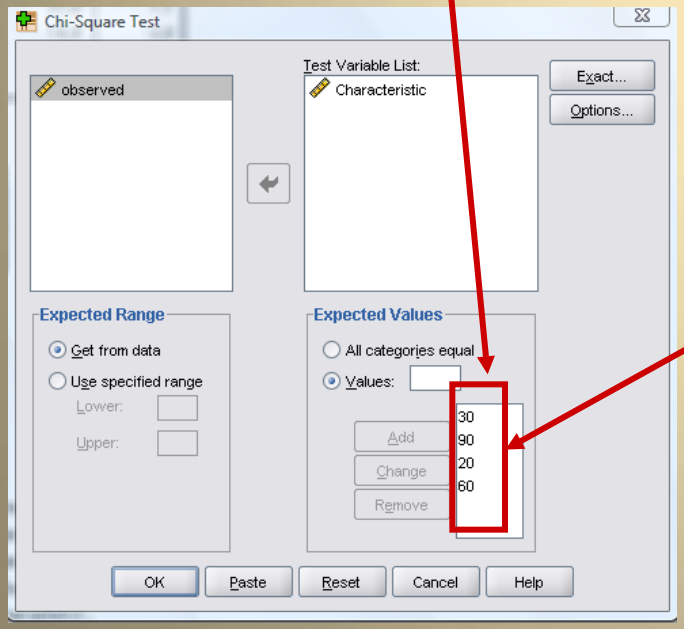
Οι καπνοβιομηχανίες ισχυρίζονται ότι δεν υπάρχει σχέση μεταξύ του καπνίσματος και ασθένειας των πνευμόνων Σε ένα δείγμα 200 ατόμων ηλικίας 40-50 ετών είχαμε τα εξής αποτελέσματα

	Ασθενείς	Υγιείς		
Καπνιστές	40	80	120	$\pi_{1.}$
Μη Καπν.	10	70	80	$\pi_{2.}$
	50	150	200	
	$\pi_{.1}$	$\pi_{.2}$		

Μπορούμε να ισχυριστούμε ότι είναι ανεξάρτητα ;



Οι θεωρητικές τιμές προκύπτουν ως $\theta_{ij}=\pi_{i.}*\pi_{.j}/n$



Characteristic

	Observed N	Expected N	Residual
Smoking Patient	40	30,0	10,0
Smoking Healthy	80	90,0	-10,0
NonSmoking Patient	10	20,0	-10,0
NonSmoking Healthy	70	60,0	10,0
Total	200		

βλέπε επόμενες διαφάνειες

Test Statistics

	Characteristic
Chi-Square	11,111[a]
df	3
Asymp. Sig.	,011

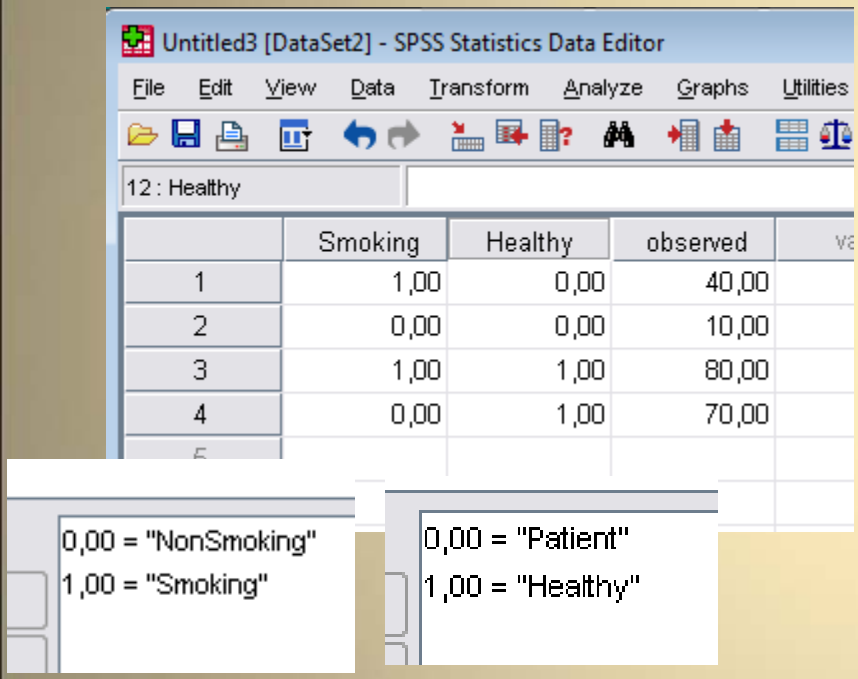
a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 20,0.

p-value= 0.0009=

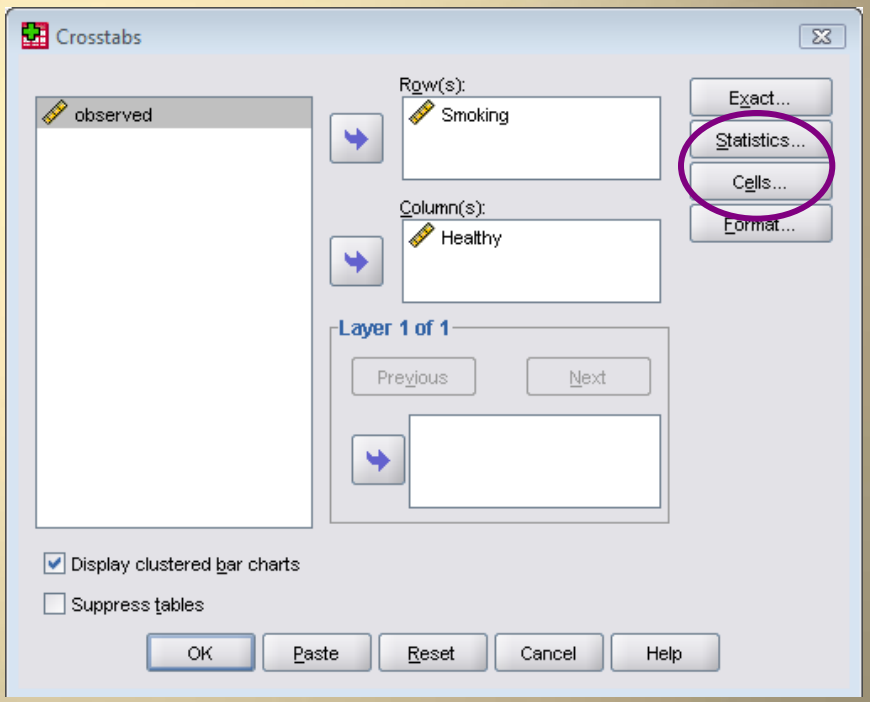
1-CDF.CHISQ(11.111,3-1-1)

Pearson's X^2 έλεγχος ανεξαρτησίας

Για να αποφύγουμε το πρόβλημα της εισαγωγής των θεωρητικών τιμών καθώς και τους λανθασμένους βαθμούς ελευθερίας στον X^2 έλεγχο κάνουμε τα εξής:

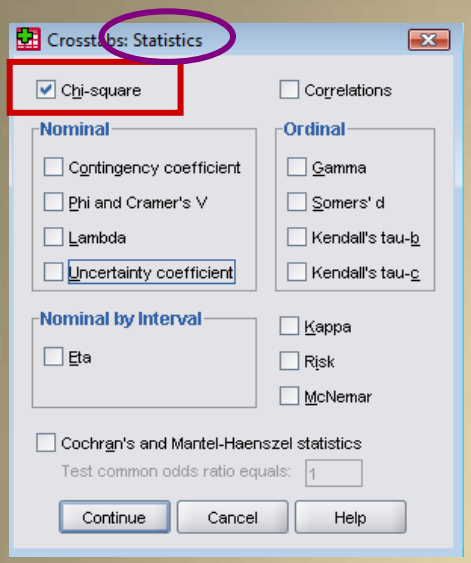


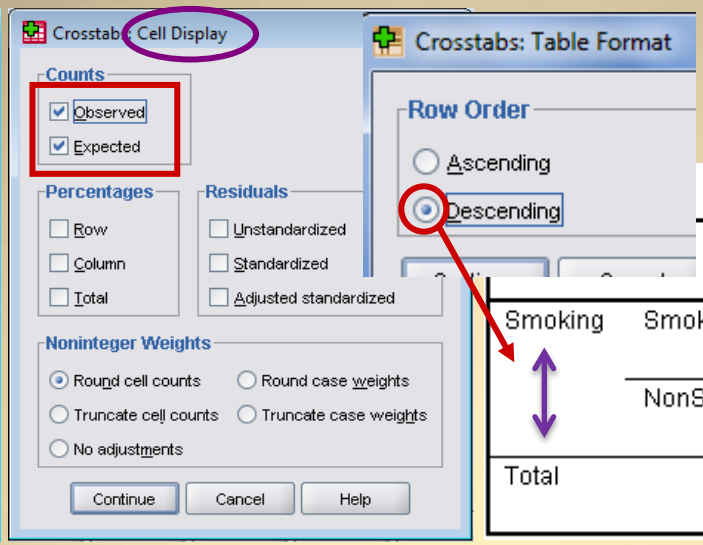
Click Analyze >Descriptive Statistics >Crosstabs



Weight cases by observed

Pearson's X^2 έλεγχος ανεξαρτησίας





Smoking * Healthy Crosstabulation

			Healthy		
			Patient	Healthy	Total
Smoking	Smoking	Count	40	80	120
		Expected Count	30,0	90,0	120,0
	NonSmoking	Count	10	70	80
		Expected Count	20,0	60,0	80,0
Total		Count	50	150	200
		Expected Count	50,0	150,0	200,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	11,111[a]	1	,001		
Continuity Correction[b]	10,028	1	,002		
Likelihood Ratio	11,887	1	,001		
Fisher's Exact Test				,001	,001
Linear-by-Linear Association	11,056	1	,001		
N of Valid Cases	200				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 20,00.

b. Computed only for a 2x2 table

παίρνουμε τους σωστούς βαθμούς ελευθερίας και p-value, οπότε:

H_0 : τα χαρακτηριστικά κάπνισμα και ασθένεια των πνευμόνων είναι ανεξάρτητα

H_1 : είναι εξαρτημένα

(αποδ. H_1 σε ε.σ. ≥ 0.1%)