ΚΕΦΑΛΑΙΟ ΤΡΙΤΟ

Εξέταση της σχέσης δυο μεταβλητών

Μία στατιστική ανάλυση δεν περιορίζεται ποτέ στη μελέτη μίας μεταβλητής, αλλά πάντοτε απαιτείται η μελέτη της σχέσης μεταξύ δύο ή και περισσότερων μεταβλητών. Στο κεφάλαιο αυτό θα δοθεί περιληπτικά ο τρόπο εξέτασης της σχέσης δύο μεταβλητών. Η τεχνική που ακολουθείται για την παραπάνω ανάλυση εξαρτάται αποκλειστικά από τη διάκριση των μεταβλητών σε ποιοτικές και ποσοτικές. Έτσι θα ασχοληθούμε με την εύρεση πιθανών σχέσεων μεταξύ α) δύο ποιοτικών μεταβλητών β) δύο ποσοτικών μεταβλητών και τέλος γ) ποσοτικής-ποιοτικής.

3.1 Δύο ποιοτικές μεταβλητές

Η εύρεση της πιθανής σχέσης μεταξύ δύο ποιοτικών μεταβλητών επιτυγχάνεται με το X^2 στατιστικό τεστ. Επιπρόσθετα, πλήθος στατιστικών μέτρων είναι διαθέσιμα ανάλογα με τη φύση των μεταβλητών για τον καθορισμό της έντασης της σχέσης μεταξύ των δύο ποιοτικών μεταβλητών (βλέπε σχετικά Παπαϊωάννου και Λουκάς, 2002, σελ. 289-292, Παπαϊωάννου και Φερεντίνος, 2000, σελ. 270-276). Η μεθοδολογία που χρησιμοποιείται για τη στατιστική ανάλυση ενός τέτοιου προβλήματος περιγράφεται στη συνέχεια.

1. Η εύρεση της πιθανής σχέσης μεταξύ δύο ποιοτικών μεταβλητών, επιτυγχάνεται μέσω της δημιουργίας του πίνακα συνάφειας (crosstabulation or contigency table), ο οποίος είναι διδιάστατος (στο επίπεδο) με r το πλήθος γραμμές, όσες οι κατηγορίες της μίας ποιοτικής μεταβλητής, και c στήλες όσες οι κατηγορίες της άλλης ποιοτικής μεταβλητής. Έτσι δημιουργούνται r x c κελιά (κυψελίδες), κάθε ένα από τα οποία παριστάνει ένα συνδυασμό των τιμών των δύο μεταβλητών και στα οποία καταγράφονται οι παρατηρούμενες συχνότητες εμφάνισής τους. Ο έλεγχος της ύπαρξης ή όχι ανεξαρτησίας μεταξύ δύο ποιοτικών μεταβλητών επιτυγχάνεται με το X^2 στατιστικό τεστ που δίνεται από τη σχέση:

$$X^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\left(O_{ij} - E_{ij}\right)^2}{E_{ij}},$$

όπου O_{ij} είναι η παρατηρούμενη συχνότητα του (i, j) κελιού (με άλλα λόγια ο αριθμός των περιπτώσεων που ανήκουν στην i και j κατηγορία της πρώτης και δεύτερης ποιοτικής μεταβλητής αντίστοιχα), E_{ij} η αναμενόμενη συχνότητα αυτού του κελιού (είναι ο αριθμός των περιπτώσεων κάθε κελιού αν οι προς μελέτη μεταβλητές ήταν στατιστικά ανεξάρτητες). Η αναμενόμενη συχνότητα E_{ij} δίνεται από τη σχέση: $E_{ij} = \frac{\sum_{i=1}^{r} O_{ij} \sum_{j=1}^{c} O_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}} = \frac{\sum_{i=1}^{r} O_{ij} \sum_{j=1}^{c} O_{ij}}{n}$, όπου n το μέγεθος του δείγματος. Είναι εύκολα κατανοητό ότι μεγάλες αποκλίσεις των αναμενόμενων τιμών από τις παρατηρούμενες τιμές υποδηλώνει πιθανή ύπαρξη σχέσης, εξάρτησης. Η υπόθεση της ανεξαρτησίας απορρίπτεται, σε επίπεδο σημαντικότητας α, όταν $X^2 \geq X^2_{(r-1)(c-1),a}$ (ή όταν p-τιμή <α). Σε περίπτωση που η υπόθεση της ανεξαρτησίας απορρίπτεται τότε προχωρούμε στο βήμα 2 και 3.

<u>Σχόλιο</u>: α) Το παραπάνω τεστ εφαρμόζεται υπό τις προϋποθέσεις ότι α) το μέγεθος του δείγματος είναι τετραπλάσιο του πλήθους των κελιών και β) οι αναμενόμενες συχνότητες δεν είναι μικρότερες του 1 και το 25% αυτών δεν είναι μικρότερες του 5. Αν δεν πληρούνται αυτές οι δύο προϋποθέσεις τότε στην περίπτωση των 2 X 2 κελιών χρησιμοποιείται το ακριβές στατιστικό του Fisher, ενώ σε κάθε άλλη περίπτωση πρέπει να γίνει συγχώνευση γειτονικών κελιών, κατά τέτοιο τρόπο ώστε να εξαλείφεται το παραπάνω πρόβλημα αλλά ταυτόχρονα να υπάρχει φυσική ερμηνεία των νέων κατηγοριών-κελιών. Η συγχώνευση των κελιών επιτυγχάνεται με επανακωδικοποίηση (recode) μίας εκ των δύο ποιοτικών μεταβλητών.

β) Στην περίπτωση 2 X 2 πινάκων χρησιμοποιείται αντί του κλασικού X^2 τεστ η διόρθωση συνεχείας του Yates (Continuity Correction).

2. Για να διαπιστωθεί ποια κελιά «δημιουργούν» το πρόβλημα της εξάρτησης των δύο μεταβλητών αρκεί να παρατηρήσουμε τις αναμενόμενες τιμές ή ακόμα καλύτερα τις τιμές των Adj. Standardized residuals: $d_{ij} = \frac{\left(O_{ij} - E_{ij}\right)/\sqrt{E_{ij}}}{\sqrt{\left(1-\frac{n_{i.}}{n}\right)\left(1-\frac{n_{.j}}{n}\right)}}$, τα οποία ακολουθούν κατά προσέγγιση κανονική κατανομή όταν οι μεταβλητές του πίνακα συνάφειας είναι ανεξάρτητες μεταξύ τους. Επομένως, μπορούν να θεωρηθούν ως z-τιμές και τιμές αυτών μεγαλύτερες κατά απόλυτη τιμή από το $1.96 = z_{0.025}$ υποδεικνύουν κελιά που διαφέρουν σαφώς από το μοντέλο της ανεξαρτησίας (για επίπεδο σημαντικότητας 5%).

3. Θέλοντας να διερευνηθεί η ένταση και η φύση της σχέσης των δύο μεταβλητών είναι διαθέσιμα πλήθος στατιστικών μέτρων. Κάποια από αυτά τα στατιστικά μέτρα είναι:

α) Ο συντελεστής συνάφειας ή σύμπτωσης (contigency coefficient),

$$C = \sqrt{\frac{X^2}{(X^2 + n)}}$$

που τιμές του κοντά στο 0 δηλώνουν ανεξάρτητες μεταβλητές, ενώ η μέγιστη τιμή του είναι μικρότερη του 1, αλλά εξαρτάται από τον αριθμό των κατηγοριών των δύο μεταβλητών,

β) ο συντελεστές Phi (αναφέρεται και ως συντελεστής του Pearson)

$$\Phi = \sqrt{\frac{X^2}{n}},$$

η μέγιστη τιμή του οποίου εξαρτάται από το μέγεθος του πίνακα, με την τιμή 0 να υποδηλώνει ανεξαρτησία των μεταβλητών.

γ) ο συντελεστής V του Cramer

$$V = \sqrt{\frac{X^2}{n \min(r-1, c-1)}}$$

που ταυτίζεται στη περίπτωση των 2 X 2 πινάκων με το συντελεστή Phi και παίρνει τιμές από 0 (ανεξαρτησία) έως 1 (απόλυτη συνάφεια),

δ) ο συντελεστής Lambda, επίσης γνωστός και ως *Goodman-Kruskal lambda* και οι συντελεστές αβεβαιότητας (uncertainty coefficient) γνωστοί και ως Theil's U.

Στην ειδική περίπτωση διατάξιμων (Ordinal) ποιοτικών μεταβλητών μπορούμε να χρησιμοποιήσουμε στατιστικά μέτρα (συντελεστές) που προσδιορίζουν και τη φύση της συνάφειας (θετική ή αρνητική). Τα μέτρα αυτά παίρνουν τιμές στο διάστημα [-1,1] με την τιμή –1 να αντιστοιχεί σε τέλεια αρνητική συνάφεια, η τιμή 0 σε μη ύπαρξη συνάφειας και η τιμή 1 σε τέλεια θετική συνάφεια. Μεταξύ άλλων τέτοιοι στατιστικοί συντελεστές είναι ο Gamma (ο zero-order για 2-way tables και ο conditional για 3-way έως 10-way tables), ο Kendall's tau-b (κατάλληλος για συμμετρικούς πίνακες), ο Kendall's tau-c (κατάλληλος για μη συμμετρικούς) και ο Somers' d (κατάλληλος για περιπτώσεις όπου η μία από τις δύο μεταβλητές μπορεί να θεωρηθεί εξαρτημένη, ενώ η άλλη ανεξάρτητη).

Στην περίπτωση που η μία ποιοτική μεταβλητή είναι ονοματική και η άλλη διαστηματική χρησιμοποιείται ο συντελεστής Eta που παίρνει τιμές στο [0,1], με την τιμή 0 να υποδεικνύει μη ύπαρξη σχέσης, ενώ η τιμή 1 υποδεικνύει υψηλού βαθμού σχέση. Ο συντελεστής αυτός είναι κατάλληλος όταν η εξαρτημένη μεταβλητή είναι διαστηματική (π.χ. το εισόδημα) και η ανεξάρτητη μεταβλητή έχει περιορισμένο αριθμό κατηγοριών (π.χ. το φύλο που έχει δύο κατηγορίες άνδρας γυναίκα). Δύο τιμές αυτού του συντελεστή υπολογίζονται από το λογισμικό, θεωρώντας εναλλάξ καθεμία από τις 2 υπό μελέτη μεταβλητές ως διαστημικές (άρα ο ερευνητής πρέπει να διαλέξει αυτή που αρμόζει στη φύση των δεδομένων του).

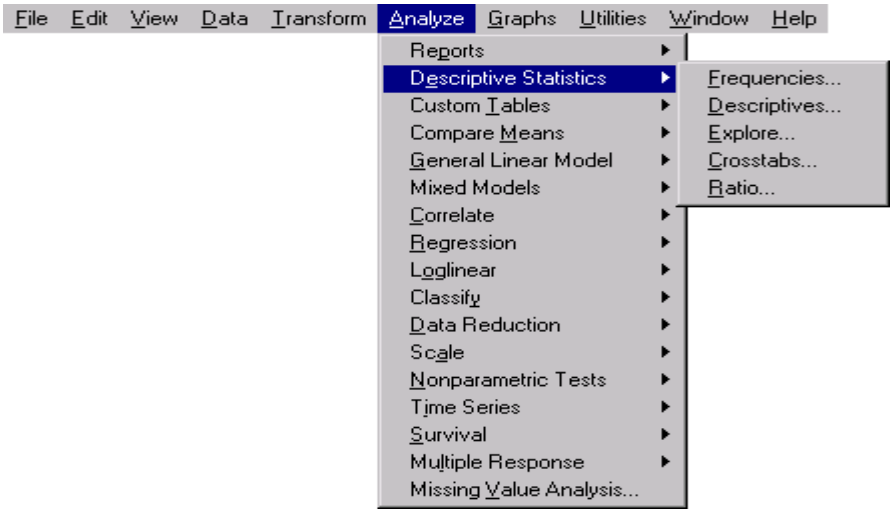
Ο συντελεστής Kappa του Kohen χρησιμοποιείται για πίνακες συνάφειας που έχουν τις ίδιες κατηγορίες στις στήλες και στις γραμμές. Παίρνει τιμές στο [-1,1]. Η τιμή 1 (-1 αντίστοιχα) υποδεικνύει πλήρη συμφωνία (πλήρη διαφωνία αντίστοιχα), ενώ η τιμή 0 υποδεικνύει ότι η συμφωνία είναι τυχαία.

Υλοποίηση στο S.P.S.S.

Σε συνέχεια του Παραδείγματος 1.1 να αποφανθείτε για την ύπαρξη ή όχι σχέσης μεταξύ των μεταβλητών Φύλο και Διαγωγή.

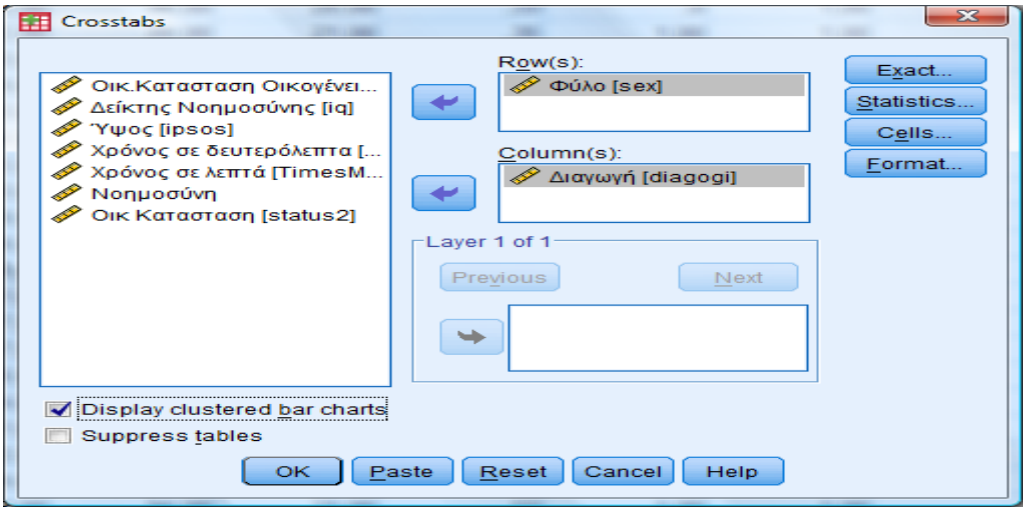
Η διαδικασία αυτή υλοποιείται ως εξής:

i. Analyze →Descriptive Statistics→ Crosstabs

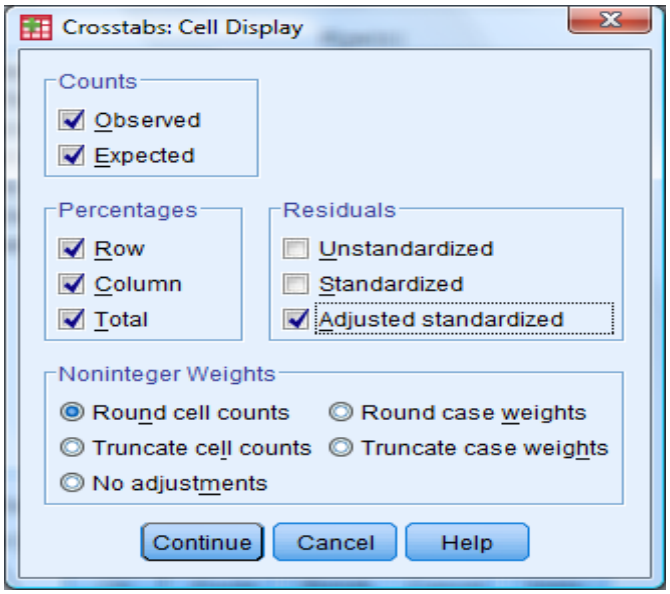


ii. Στο νέο παράθυρο διαλόγου που προκύπτει διαλέγουμε την ποιοτική μεταβλητή τις δυνατές τιμές της οποίας θέλουμε να έχουμε στις γραμμές (στήλες αντίστοιχα) του πίνακα συνάφειας και τη μετακινούμε στο πλαίσιο Rows (πλαίσιο Columns αντίστοιχα). Θέλοντας να κατασκευαστούν ομάδες ραβδογραμμάτων (bar charts) για κάθε τιμή της μεταβλητής που καθορίζεται στο πλαίσιο Rows, ενώ η μεταβλητή που καθορίζει το ύψος των ράβδων είναι αυτή που έχουμε καθορίσει στο πλαίσιο Columns επιλέγουμε στο αρχικό παράθυρο το πλαίσιο Display Cluster Bar Charts.

Σχόλιο: Καλό είναι να μην επιλέγουμε το πλαίσιο Suppress tables γιατί σε μία τέτοια περίπτωση δε θα εμφανίζεται ο πίνακας συνάφειας.



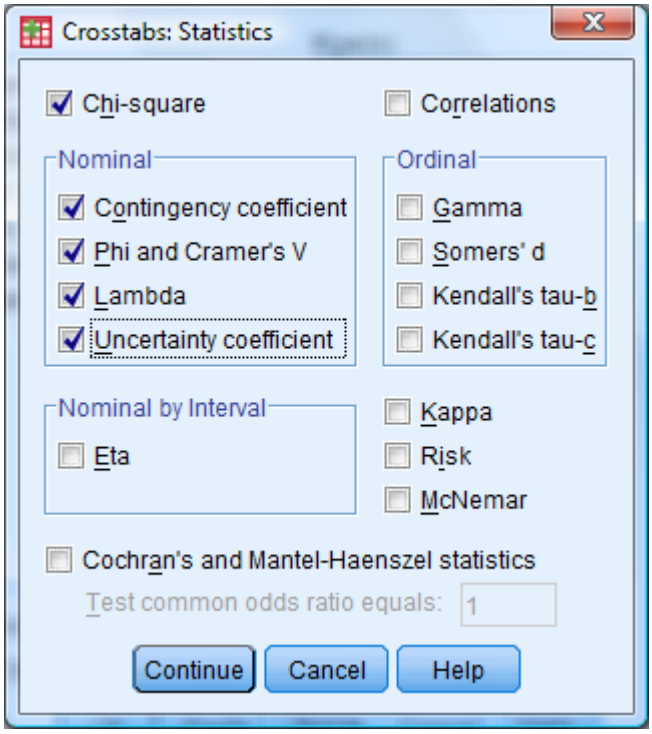
iii. Για να αποφανθούμε για την ύπαρξη, την ένταση και φύση της σχέσης των δύο μεταβλητών θα πρέπει να εμπλουτίσουμε τις πληροφορίες που μας δίνει το λογισμικό ως προεπιλογή. Αυτό μπορεί να επιτευχθεί αρχικά από την επιλογή Cells επιλέγοντας τα ακόλουθα:



Observed, Expected counts με τα οποία αποκτούμε τις παρατηρούμενες και αναμενόμενες αντίστοιχα συχνότητες σε κάθε κελί του πίνακα συνάφειας.

Percentages από όπου αποκτούμε τα ποσοστά εντός των γραμμών (Row), στηλών (Columns) καθώς και στο σύνολο των δεδομένων (Total). Τα ποσοστά εντός των γραμμών και στηλών αθροίζουν στο 100% κατά μήκος των αντίστοιχων γραμμών, στηλών αντίστοιχα, ενώ τα συνολικά ποσοστά αθροίζουν στο 100% μέσα σε όλα τα κελιά του πίνακα.

iv. Από την επιλογή Statistics έχουμε τη δυνατότητα όπως φαίνεται και στο πλαίσιο που ακολουθεί να πραγματοποιήσουμε τον έλεγχο ανεξαρτησίας, να αναζητήσουμε το βαθμό και τη φύση της συνάφειας καθώς και πλήθος στατιστικών μέτρων. Για το παράδειγμά μας είναι ορθό να επιλέξουμε τα ακόλουθα:



Σχόλιο: Για πίνακες με 2 γραμμές και 2 στήλες, δηλαδή για ποιοτικές μεταβλητές με δύο δυνατές τιμές η καθεμία, επιλέγοντας το Chi-square υπολογίζεται το X^2 του Pearson, το τεστ πηλίκου πιθανοφανειών (the likelihood-ratio chi-square), το Fisher's exact test (ένας έλεγχος ιδιαίτερα χρήσιμος για τις περιπτώσεις που δεν ικανοποιούνται οι προϋποθέσεις του X^2 τεστ ανεξαρτησίας), καθώς και το X^2 τεστ ανεξαρτησίας του Yates με διόρθωση συνεχείας (continuity correction). Για πίνακες συνάφειας μεγαλύτερης διάστασης υπολογίζονται μόνο το X^2 του Pearson και το τεστ πηλίκου πιθανοφανειών. Επιπλέον, το S.P.S.S μας πληροφορεί αν υπάρχουν κελιά με αναμενόμενη τιμή μικρότερη του 5. Υπενθυμίζεται ότι απαραίτητη προϋπόθεση για να χρησιμοποιηθεί το X^2 τεστ ανεξαρτησίας του Pearson είναι η μη ύπαρξη αναμενόμενων τιμών μικρότερων του 5. Σε αντίθετη περίπτωση συγχωνεύονται γειτονικά κελιά, εκτός από την περίπτωση των 2 X 2 πινάκων όπου καταφεύγουμε στο Fisher's exact test.

Ερμηνεία αποτελεσμάτων

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Φύλο * Διαγωγή	35	100,0%	0	,0%	35	100,0%

Ο παραπάνω πίνακας μας πληροφορεί ότι 35 παρατηρήσεις είναι διαθέσιμες ταυτόχρονα στις δύο μεταβλητές χωρίς την ύπαρξη ελλιπών τιμών, ενώ ο επόμενος πίνακας είναι ένας πίνακας διπλής εισόδου, γνωστός και ως πίνακας συνάφειας.

Φύλο * Διαγωγή Crosstabulation

			Διαγωγή		Total
			A	B	A
Φύλο	Αγόρι	Count	16	3	19
		Expected Count	16,3	2,7	19,0
		% within Φύλο	84,2%	15,8%	100,0%
		% within Διαγωγή	53,3%	60,0%	54,3%
		% of Total	45,7%	8,6%	54,3%
		Adjusted Residual	-,3	,3	
	Κορίτσι	Count	14	2	16
		Expected Count	13,7	2,3	16,0
		% within Φύλο	87,5%	12,5%	100,0%
		% within Διαγωγή	46,7%	40,0%	45,7%
		% of Total	40,0%	5,7%	45,7%
		Adjusted Residual	,3	-,3	
Total		Count	30	5	35
		Expected Count	30,0	5,0	35,0
		% within Φύλο	85,7%	14,3%	100,0%
		% within Διαγωγή	100,0%	100,0%	100,0%
		% of Total	85,7%	14,3%	100,0%

Ας ερμηνεύσουμε κάποια από τα αποτελέσματα του παραπάνω πίνακα συνάφειας. Παρατηρούμε ότι οι αναμενόμενες συχνότητες (Expected Count) είναι κοντά στις παρατηρούμενες συχνότητες (Count). Επιπλέον 84,2 % των αγοριών έχουν διαγωγή Κοσμιωτάτη (αφού το 84,2 βρίσκεται στο % within Φύλο και στη διασταύρωση αγοριού και διαγωγής Α), ενώ το 53,7% αυτών που έχουν διαγωγή Κοσμιωτάτη είναι αγόρια (αφού το 53,7% βρίσκεται στο % within Διαγωγή και στη διασταύρωση αγοριού και

διαγωγής Α). Ακόμη τα αγόρια με διαγωγή Κοσμιωτάτη αποτελούν το 45,7% των ερωτηθέντων (αφού το 45,7% βρίσκεται στο % of Total και στη διασταύρωση αγοριού και διαγωγής Α). Τέλος καμία από τις τιμές των Adj. Residuals δεν είναι μεγαλύτερη κατά απόλυτη τιμή από το 1.96.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,077(b)	1	,782		
Continuity Correction(a)	,000	1	1,000		
Likelihood Ratio	,077	1	,781		
Fisher's Exact Test				1,000	,585
Linear-by-Linear Association	,075	1	,785		
N of Valid Cases	35				

a Computed only for a 2x2 table
b 2 cells (50,0%) have expected count less than 5. The minimum expected count is 2,29.

Ο πίνακας Chi-Square Tests μας πληροφορεί για το αποτέλεσμα του ελέγχου της ανεξαρτησίας. Έτσι από την υποσημείωση b που μας δίνεται στον πίνακα αυτό πληροφορούμαστε ότι υπάρχουν δύο κελιά (50% των συνολικών) με αναμενόμενες συχνότητες μικρότερες του 5. Καθώς ο πίνακας συνάφειας είναι 2 Χ 2 θα χρησιμοποιηθεί το Fisher's exact test από όπου καταλήγουμε στο συμπέρασμα ότι η υπόθεση της ανεξαρτησίας φύλου και διαγωγής στο σχολείο δεν μπορεί να απορριφθεί καθώς η p-τιμή είναι μεγαλύτερη από 0,05.

Τέλος, στους παρακάτω πίνακες το λογισμικό μας παραθέτει τις τιμές των μέτρων συνάφειας. Οι τιμές για αυτούς τους δείκτες είναι αναμενόμενο να είναι κοντά στο μηδέν καθώς η υπόθεση της ανεξαρτησίας δεν έχει απορριφθεί.

Directional Measures

			Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,000	,000	.(c)	.(c)
		Φύλο Dependent	,000	,000	.(c)	.(c)
		Διαγωγή Dependent	,000	,000	.(c)	.(c)
	Goodman and Kruskal tau	Φύλο Dependent	,002	,016		,785(d)
		Διαγωγή Dependent	,002	,016		,785(d)
	Uncertainty Coefficient	Symmetric	,002	,014	,140	,781(e)
		Φύλο Dependent	,002	,011	,140	,781(e)
		Διαγωγή Dependent	,003	,019	,140	,781(e)

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Cannot be computed because the asymptotic standard error equals zero.
d Based on chi-square approximation
e Likelihood ratio chi-square probability.

Symmetric Measures

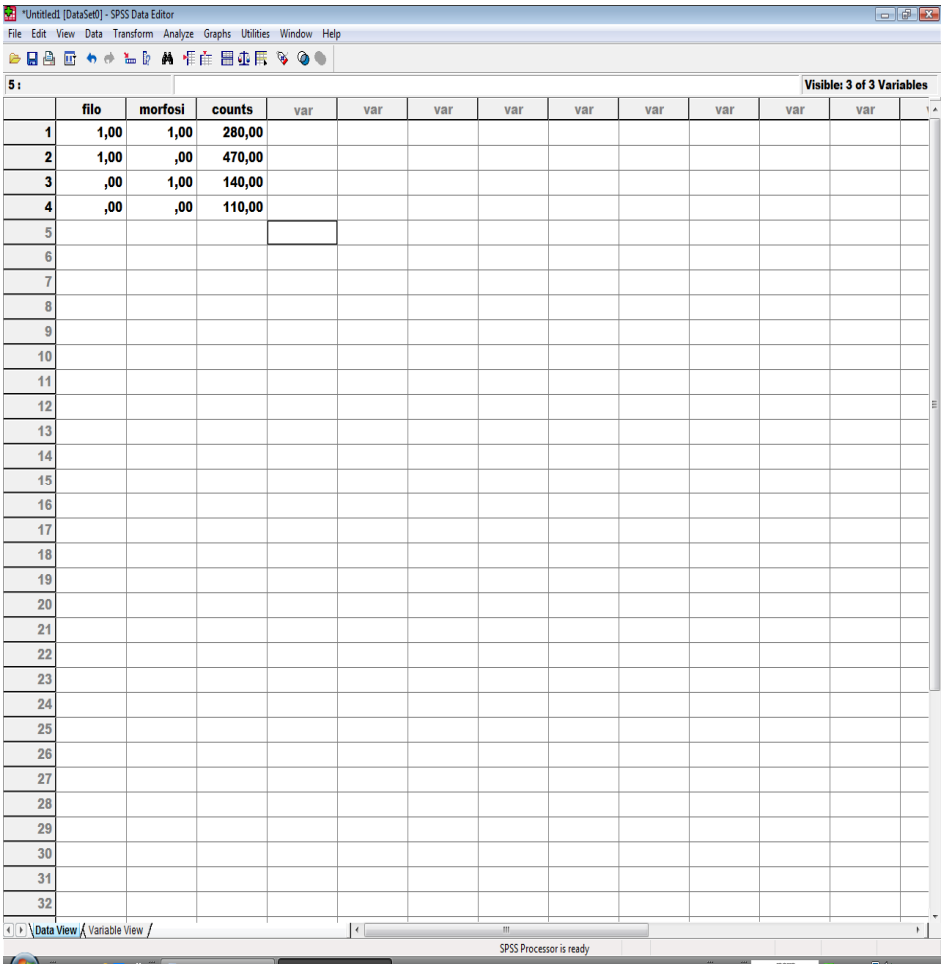
		Value	Approx. Sig.
Nominal by Nominal	Phi	-,047	,782
	Cramer's V	,047	,782
	Contingency Coefficient	,047	,782
N of Valid Cases		35	

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

<u>Παρατήρηση</u>: Έστω ότι μας δινόταν ο ακόλουθος πίνακας διπλής εισόδου:

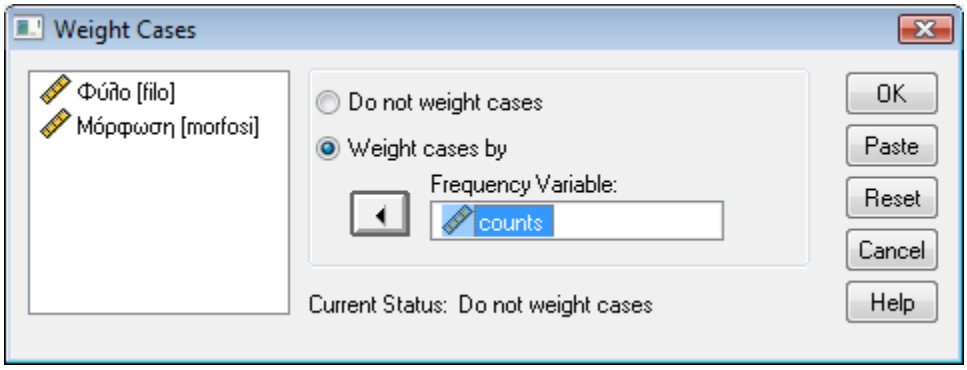
	Μη πτυχιούχοι	Πτυχιούχοι
Άνδρες	470	280
Γυναίκες	110	140

Το ερώτημα που τίθεται είναι αν το φύλο και η κατοχή πτυχίου είναι ανεξάρτητα. Πως θα χρησιμοποιηθεί το S.P.S.S. για τον υπολογισμό του X^2 τεστ ανεξαρτησίας. Σε μία τέτοια περίπτωση στο παράθυρο του Data View στις πρώτες δύο στήλες καταγράφουμε τους δυνατούς συνδυασμούς των δύο ποιοτικών μεταβλητών. Στο παράδειγμά μας καθώς αυτές είναι δίτιμες είναι αντιληπτό ότι οι δυνατοί συνδυασμοί είναι 4. Έτσι αν για τη μεταβλητή Φύλο: 1=άνδρας και 0=γυναίκα, και για τη μεταβλητή Μόρφωση: 1=πτυχιούχος και 0=μη πτυχιούχος, οι δυνατοί συνδυασμοί είναι (1,1), (1,0), (0,1) και (0,0). Στην τρίτη στήλη καταγράφουμε τις παρατηρούμενες συχνότητες για κάθε συνδυασμό. Είναι 280, 470, 140 και 110, αντίστοιχα.



Για να δηλωθεί στο λογισμικό ο ξεχωριστός ρόλος της τρίτης στήλης επιλέγουμε: Data→Weight Cases και στο νέο παράθυρο διαλόγου που προκύπτει αφού επιλέξουμε το

πλαίσιο Weight cases by τοποθετούμε στο πλαίσιο Frequency Variable τη μεταβλητή όπου καταγράφονται οι παρατηρούμενες συχνότητες και πατάμε OK.



Στη συνέχεια ακολουθούμε τα κλασικά βήματα για τον υπολογισμό του X^2 τεστ ανεξαρτησίας και προκύπτει ότι το φύλο και η κατοχή πτυχίου δεν είναι ανεξάρτητα ενδεχόμενα (p-τιμή του X^2 στατιστικού τεστ<0.05). Οι γυναίκες μη πτυχιούχοι είναι λιγότερες από το αναμενόμενο αποτέλεσμα υπό την ανεξαρτησία (Adj. Residual=-5.2).

Φύλο * Μόρφωση Crosstabulation

			Μόρφωση		Total
			Μη πτυχιούχος	Πτυχιούχος	Μη πτυχιούχος
Φύλο	Γυναίκα	Count	110	140	250
		% within Φύλο	44,0%	56,0%	100,0%
		% within Μόρφωση	19,0%	33,3%	25,0%
		% of Total	11,0%	14,0%	25,0%
		Adjusted Residual	-5,2	5,2	
	Άνδρας	Count	470	280	750
		% within Φύλο	62,7%	37,3%	100,0%
		% within Μόρφωση	81,0%	66,7%	75,0%
		% of Total	47,0%	28,0%	75,0%
		Adjusted Residual	5,2	-5,2	
Total		Count	580	420	1000
		% within Φύλο	58,0%	42,0%	100,0%
		% within Μόρφωση	100,0%	100,0%	100,0%
		% of Total	58,0%	42,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	26,820(b)	1	,000		
Continuity Correction(a)	26,059	1	,000		
Likelihood Ratio	26,560	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	26,793	1	,000		
N of Valid Cases	1000				

a Computed only for a 2x2 table
b 0 cells (,0%) have expected count less than 5. The minimum expected count is 105,00.

3.2 Δύο ποσοτικές μεταβλητές

Η μελέτη της σχέσης δύο ποσοτικών μεταβλητών μπορεί να γίνει:

α) με σκοπό την τεκμηρίωση της σχέσης που έχουν,

β) με σκοπό να καταλήξουμε σε μία μαθηματική σχέση που τις συνδέει και τέλος

γ) με σκοπό τη σύγκριση των πληθυσμιακών μέσων τιμών.

Στη δεύτερη περίπτωση έχουμε να κάνουμε με το μοντέλο της ανάλυσης παλινδρόμησης, με το οποίο θα ασχοληθούμε στο αντίστοιχο κεφάλαιο της Παλινδρόμησης. Για την τρίτη περίπτωση αναφερόμαστε αναλυτικά στο Όγδοο Κεφάλαιο. Στην παράγραφο αυτή θα ασχοληθούμε μόνο με το α), που επιτυγχάνεται:

- Με το γράφημα των τιμών των δύο ποσοτικών μεταβλητών (διάγραμμα διασποράς).
- Με το συντελεστή συσχέτισης.

Έστω ότι X και Y είναι δύο τυχαίες μεταβλητές και (x_i, y_i) είναι n το πλήθος ζεύγη αριθμητικών τιμών αυτών, και θέλουμε να εξετάσουμε την ύπαρξη ή μη γραμμικής εξάρτησης μεταξύ δύο ποσοτικών τυχαίων μεταβλητών και να υπολογίσουμε και το βαθμό αυτής της γραμμικής σχέσης. Η εξέταση της ύπαρξης ή μη γραμμικής εξάρτησης μπορεί να γίνει με γραφικούς, αλλά κυρίως στατιστικούς τρόπους. Οι στατιστικοί τρόποι ελέγχου στηρίζονται στους συντελεστές συσχέτισης που έχουν παρουσιαστεί στη βιβλιογραφία και είναι: ο συντελεστής συσχέτισης του Pearson, του Spearman και του

Kendall. Η επιλογή του συντελεστή συσχέτισης που θα χρησιμοποιηθεί εξαρτάται από το αν πληρούνται ή όχι κάποιες προϋποθέσεις, τις οποίες και πρέπει αρχικά να ελέγξει ο ερευνητής. Πιο συγκεκριμένα, ελέγχουμε αν:

α) το ποσοστό των ακραίων τιμών στις διαθέσιμες δειγματικές παρατηρήσεις ξεπερνά το 10% αυτών, και

β) αν ο πληθυσμός από τον οποίο λαμβάνεται το τυχαίο δείγμα (x_i, y_i), $i = 1,...,n$ μπορούμε να ισχυριστούμε ότι περιγράφεται ικανοποιητικά από τη διδιάστατη κανονική κατανομή.

Στη συνέχεια παρουσιάζονται όλα τα πιθανά αποτελέσματα των α) και β), τα διάφορα βήματα της ανάλυσης και οι αποφάσεις στις οποίες οδηγούμαστε.

<u>Μεθοδολογία</u>

1. Αρχικά ελέγχουμε αν υπάρχουν ακραίες τιμές στις διαθέσιμες δειγματικές τιμές. Αν το ποσοστό των ακραίων τιμών, οι οποίες αφαιρούνται μία-μία, δε ξεπερνά το 10%, τότε προχωρούμε στο βήμα 2. Αν το ποσοστό των ακραίων τιμών ξεπερνά το 10%, τότε δοκιμάζουμε μήπως ο μετασχηματισμός του λογαρίθμου διορθώνει το πρόβλημα. Αν το πρόβλημα αυτό διορθώνεται τότε μεταβαίνουμε στο βήμα 2, σε διαφορετική περίπτωση συμπεραίνουμε ότι θα χρησιμοποιηθεί ο μη παραμετρικός συντελεστής συσχέτισης (βλέπε βήμα 4).
2. Στο βήμα 2, καθώς τα διάφορα στατιστικά προγράμματα δεν μας δίνουν τη δυνατότητα για ελέγχους της διδιάστατης κανονικότητας, προχωρούμε σε ελέγχους της μονοδιάστατης κανονικότητας για καθένα από τα δείγματα $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_n$. Επομένως, χρησιμοποιώντας το τεστ των Shapiro-Wilk καθώς και γραφικούς τρόπους, ελέγχουμε αν οι διαθέσιμες δειγματικές παρατηρήσεις (είτε οι αρχικές είτε οι μετασχηματισμένες του βήματος 1) προέρχονται από πληθυσμούς που περιγράφονται ικανοποιητικά από την κανονική κατανομή. Αν ο έλεγχος της κανονικότητας μας υποδεικνύει ότι η υπόθεσης της κανονικότητας δεν απορρίπτεται (p-τιμή >α), τότε η ανάλυση θα συνεχιστεί με επιφύλαξη με τον παραμετρικό συντελεστή συσχέτισης (βλέπε βήμα 3). Αν η υπόθεση της κανονικότητας απορρίπτεται για έναν ή και τους δύο πληθυσμούς (τεστ Shapiro-Wilk, p-τιμή <α), τότε ελέγχουμε αν το πρόβλημα της μη

κανονικότητας διορθώνεται μετασχηματίζοντας τα δεδομένα (Box-Cox μετασχηματισμός) και επανελέγχοντας την ύπαρξη ακραίων τιμών, δηλαδή ξεκινώντας την ανάλυση από το βήμα 1. Αν με κάποιο μετασχηματισμό των δεδομένων επιτυγχάνεται η κανονικότητα συνεχίζουμε την ανάλυση παραμετρικά, με την επιφύλαξη αν η από κοινού κατανομή ακολουθεί διδιάστατη κανονική (βήμα 3). Σε αντίθετη περίπτωση, αν το πλήθος των δειγματικών παρατηρήσεων (μη λαμβάνοντας υπόψη αυτές που έχουν αφαιρεθεί στο βήμα 1) του πληθυσμού ή των πληθυσμών για τους οποίους απορρίπτεται η υπόθεση ότι περιγράφονται ικανοποιητικά από την κανονική κατανομή είναι μεγάλο (συνήθως μεγαλύτερο του 30) κάνοντας χρήση του Κεντρικού Οριακού Θεωρήματος, προβαίνουμε στον παραμετρικό έλεγχο της υπό έλεγχο υπόθεσης (βλέπε βήμα 3). Σε αυτήν την περίπτωση η p-τιμή του ελέγχου θα είναι προσεγγιστική. Αν η υπόθεση της κανονικότητας απορρίπτεται τόσο για τις αρχικές όσο και για τις μετασχηματισμένες δειγματικές τιμές (τεστ Shapiro-Wilk, p-τιμή <α), και ταυτόχρονα το πλήθος των δειγματικών παρατηρήσεων (μη λαμβάνοντας υπόψη αυτές που έχουν αφαιρεθεί στο βήμα 1) είναι μικρό (συνήθως μικρότερο του 30), συνεχίζεται η περαιτέρω ανάλυση μη παραμετρικά (βήμα 4).

3. **Συντελεστής συσχέτισης του Pearson**: Ο συντελεστής συσχέτισης του Pearson μας δίνει το βαθμό γραμμικής (και μόνο) εξάρτησης δύο ποσοτικών τυχαίων μεταβλητών και δίνεται από τη σχέση:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Πρόκειται για έναν καθαρό αριθμό μεταξύ του -1 και 1. Όταν r=0 δεν υπάρχει γραμμική σχέση μεταξύ των X και Y, χωρίς αυτό βέβαια να αποκλείει την ύπαρξη κάποιας σχέσης άλλης μορφής π.χ. εκθετικής. Όταν r=+1 υπάρχει θετική γραμμική εξάρτηση (αύξηση των τιμών της μιας επιφέρει αύξηση στις τιμές της άλλης), ενώ όταν r=-1 υπάρχει αρνητική γραμμική εξάρτηση (αύξηση των τιμών της μιας επιφέρει μείωση στις τιμές της άλλης). Τιμές κοντά στο –1 ή στο 1 υποδηλώνουν αρνητική/ θετική συσχέτιση, αντίστοιχα, ενώ τιμές κοντά στο 0 μη ύπαρξη γραμμικής σχέσης.

Απόλυτες τιμές του συντελεστή αυτού στο [0,0.3] υποδηλώνουν ασθενή γραμμική εξάρτηση, στο (0.3,0.6] μεσαία, ενώ στο (0.6,1] ισχυρή.

Αποδεικνύεται ότι (βλέπε Παπαΐωάννου και Λουκάς, 2002, σελ. 179) η υπόθεση της μη ύπαρξης γραμμικής εξάρτησης ελέγχεται με το στατιστικό τεστ

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

όπου n το μέγεθος του δείγματος. Η υπόθεση της μη ύπαρξης γραμμικής σχέσης απορρίπτεται, με επίπεδο σημαντικότητας α, όταν $|t| \geq t_{n-2,a/2}$.

4. <u>Μη παραμετρικοί συντελεστές συσχέτισης:</u> Στην περίπτωση αυτή μεταξύ άλλων έχουν προταθεί ο συντελεστής συσχέτισης του Spearman και του Kendall. Κάποιες πληροφορίες για αυτούς παρατίθενται στη συνέχεια.

<u>Συντελεστής συσχέτισης του Spearman</u>: Ο συντελεστής συσχέτισης του Spearman, ο οποίος συμβολίζεται με r_s, δεν είναι τίποτε άλλο παρά ο συντελεστής συσχέτισης του Pearson όταν αυτός εφαρμόζεται στις τάξεις $R_1,...,R_n$ και $S_1,...,S_n$, δηλαδή

$$r_s = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}},$$

όπου $\overline{R} = \frac{\sum_{i=1}^{n} R_i}{n}$ και $\overline{S} = \frac{\sum_{i=1}^{n} S_i}{n}$.

Στην περίπτωση ύπαρξης δεσμών (ties) μεταξύ των X_i ή των Y_i η κλασική αντιμετώπιση όπως έχει ήδη αναφερθεί είναι ο υπολογισμός, σε κάθε μία από τις ίσες αυτές τιμές, του μέσου όρου των τάξεων που θα είχαν αν δεν ταυτίζονταν. Αν $u_1, u_2, ...$ και $v_1, v_2, ...$ είναι οι τάξεις των δειγματικών τιμών X_i και Y_i, αντίστοιχα, όπου έχουμε δεσμούς, τότε ο συντελεστής συσχέτισης εναλλακτικά υπολογίζεται από τη σχέση:

$$r = \frac{n\left(n^2-1\right) - 6\sum\left(R_i - S_i\right)^2 - 6\left(U+V\right)}{\left[\left\{n\left(n^2-1\right)-U\right\}\left\{n\left(n^2-1\right)-V\right\}\right]^{1/2}},$$

όπου $U = \sum\left(u_i^3 - u_i\right)$ και $V = \sum\left(v_i^3 - v_i\right)$.

Αποδεικνύεται ότι $Z = \left(r_s - \rho\right)\sqrt{n-1} \underset{H_0}{\overset{\pi\rho o\sigma\varepsilon\gamma.}{\sim}} N\left(0,1\right)$, το οποίο και χρησιμοποιείται για τον έλεγχο της υπόθεσης $H_0 : \rho = 0$.

Συντελεστής συσχέτισης του Kendall: Ο Kendall πρότεινε το στατιστικό:

$$\tau = \frac{n_c - n_d}{n(n-1)/2},$$

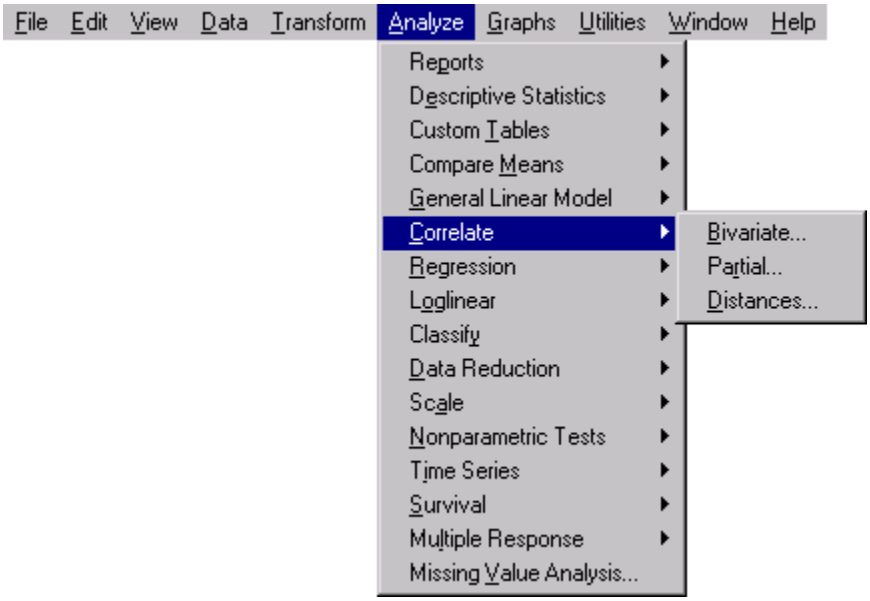
όπου n_c και n_d είναι το πλήθος των σύμφωνων και ασύμφωνων ζευγαριών, αντίστοιχα, δηλαδή των τιμών $\left(X_i, Y_i\right)$ και $\left(X_j, Y_j\right)$, $i \neq j$, $i, j = 1,...,n$, που το πρόσημο της διαφοράς $X_i - X_j$ είναι σύμφωνο, δεν είναι σύμφωνο, αντίστοιχα, με το πρόσημο της διαφοράς $Y_i - Y_j$. Για μεγάλο μέγεθος δείγματος αποδεικνύεται ότι:

$$Z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \underset{H_0}{\overset{\alpha\sigma\upsilon\mu\pi.}{\sim}} N\left(0,1\right).$$

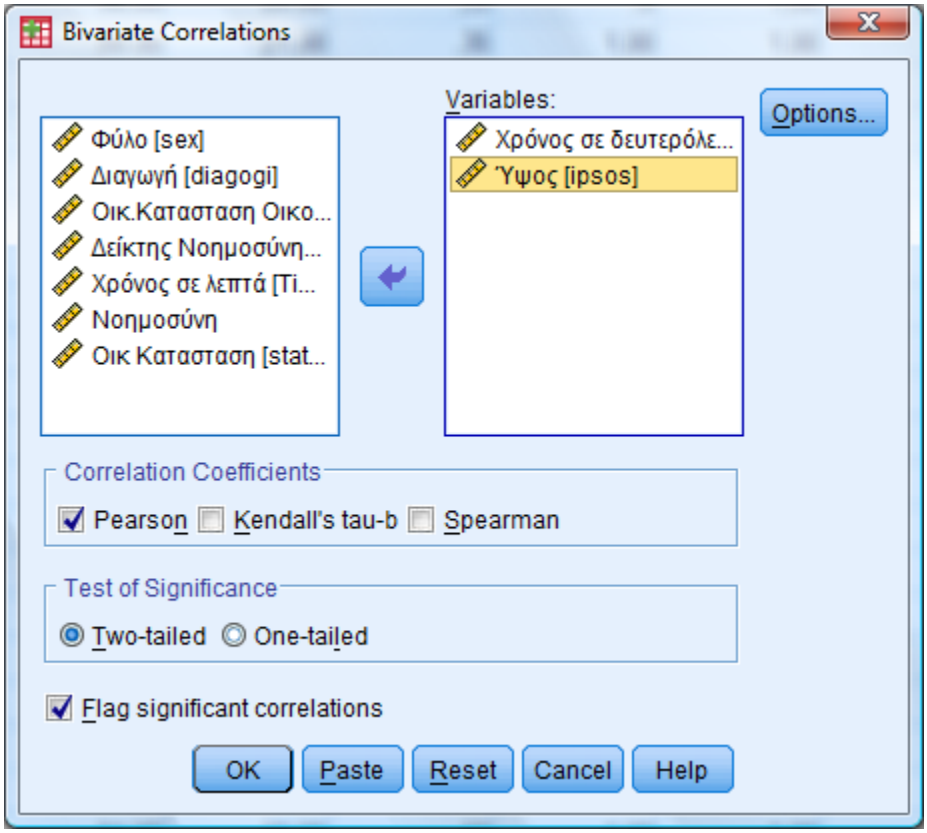
Σχόλιο: Βασικό πλεονέκτημα των συντελεστών συσχέτισης του Spearman και του Kendall είναι ότι μπορούν να χρησιμοποιηθούν και για διατάξιμες μεταβλητές, είναι ανθεκτικοί στην ύπαρξη ακραίων τιμών, και ως μη παραμετρικοί συντελεστές δεν απαιτούν καμία υπόθεση για τους πληθυσμούς. Από την άλλη μεριά βασικό μειονέκτημα είναι ότι δεν υπολογίζονται από τις πραγματικές τιμές, αλλά από τις τάξεις.

Υλοποίηση των 3-4 στο S.P.S.S.

i. Analyze→Correlate→Bivariate



ii. Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε τις δύο ποσοτικές μεταβλητές που μελετούμε και τις μετακινούμε στο πλαίσιο Variables. Αν μετακινήσουμε περισσότερες από δύο τότε οι υπολογισμοί θα γίνουν για κάθε συνδυασμό ανά δύο και επιλέγουμε το συντελεστή συσχέτισης που πρέπει και επιθυμούμε να υπολογιστεί (έστω εδώ του Pearson). Στο πλαίσιο Test of Significance προχωρούμε σε μονόπλευρο (One-Tailed) ή δίπλευρο έλεγχο (Two-tailed) για τον πληθυσμιακό συντελεστή συσχέτισης (θα πρέπει να μην υπάρχει πρόβλημα ύπαρξης ακραίων τιμών και να ισχύει η υπόθεση της διδιάστατης κανονικότητας, επομένως θα πρέπει τουλάχιστον να έχουμε την κανονικότητα για καθεμία εκ των περιθωρίων). Με ενεργοποιημένο το πλαίσιο Flag Significant Correlations το λογισμικό μας υποδεικνύει τις στατιστικά σημαντικές συσχετίσεις. Τέλος από την επιλογή Options έχουμε τη δυνατότητα να ζητήσουμε από το λογισμικό να υπολογίσει τη μέση τιμή και την τυπική απόκλιση κάθε μεταβλητής (Means and standard deviations) καθώς επίσης και τις μεταξύ τους διακυμάνσεις (Cross product deviations and covariances). Τέλος μπορούμε να καθορίσουμε τον τρόπο χειρισμού των ελλιπών τιμών.



Ερμηνεία αποτελεσμάτων

Από τον πίνακα των αποτελεσμάτων συμπεραίνουμε ότι δεν υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ του ύψους των παιδιών και του χρόνου που διανύουν τα 100 μέτρα. Αυτό διότι ο συντελεστής συσχέτισης του Pearson είναι -0.166, δηλαδή κοντά στο μηδέν, και επιπλέον η p-τιμή για το δίπλευρο έλεγχο είναι ίση με 0.342>0.05. Άρα η υπόθεση της μη ύπαρξης γραμμικής συσχέτισης δεν μπορεί να απορριφθεί (υπό την προϋπόθεση της κανονικότητας και της μη ύπαρξης ακραίων τιμών, έλεγχοι που πρέπει να προηγούνται της ανάλυσης, όπως έχουμε ήδη αναφέρει στα βήματα 1-2).

Correlations

		Ύψος	Χρόνος σε δευτερόλεπτα
Ύψος	Pearson Correlation	1	-,166
	Sig. (2-tailed)		,342
	N	35	35
Χρόνος σε δευτερόλεπτα	Pearson Correlation	-,166	1
	Sig. (2-tailed)	,342	
	N	35	35

Σχόλιο: Προσοχή η μη ύπαρξη γραμμικής σχέσης μεταξύ του ύψους των παιδιών και του χρόνου σε δευτερόλεπτα που διανύουν τα 100 μέτρα δεν αποκλείει την ύπαρξη κάποιας σχέσης άλλης μορφής.

5. Τέλος, η εξέταση της ύπαρξης ή μη γραμμικής εξάρτησης μπορεί να γίνει με γραφικό τρόπο μέσω του διαγράμματος διασποράς. Το διάγραμμα διασποράς δεν είναι τίποτε άλλο παρά το γράφημα των τιμών των δύο ποσοτικών μεταβλητών. Στον οριζόντιο άξονα τοποθετούνται οι τιμές εκείνης της μεταβλητής που ενδέχεται να έχει το ρόλο της ανεξάρτητης, ενώ στον κατακόρυφο οι τιμές της εξαρτημένης. Η απεικόνιση αυτή μας βοηθά να έχουμε μία πρώτη υπόνοια για την ύπαρξη ή όχι κάποιας μαθηματικής σχέσης. Επιπλέον, στην περίπτωση ύπαρξης σχέσης δύναται να αναγνωριστεί η μορφή αυτής (αν είναι γραμμική, τετραγωνική, εκθετική κ.ο.κ.).

Υλοποίηση στο S.P.S.S.

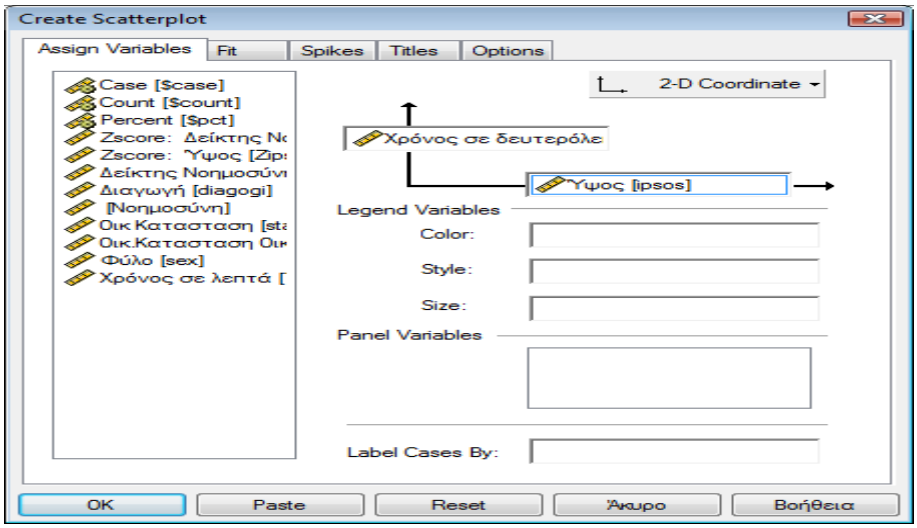
Για να αποκτήσουμε το διάγραμμα διασποράς μπορούμε να ακολουθήσουμε μία από τις παρακάτω δύο διαδικασίες:

Διαδικασία Scatterplot

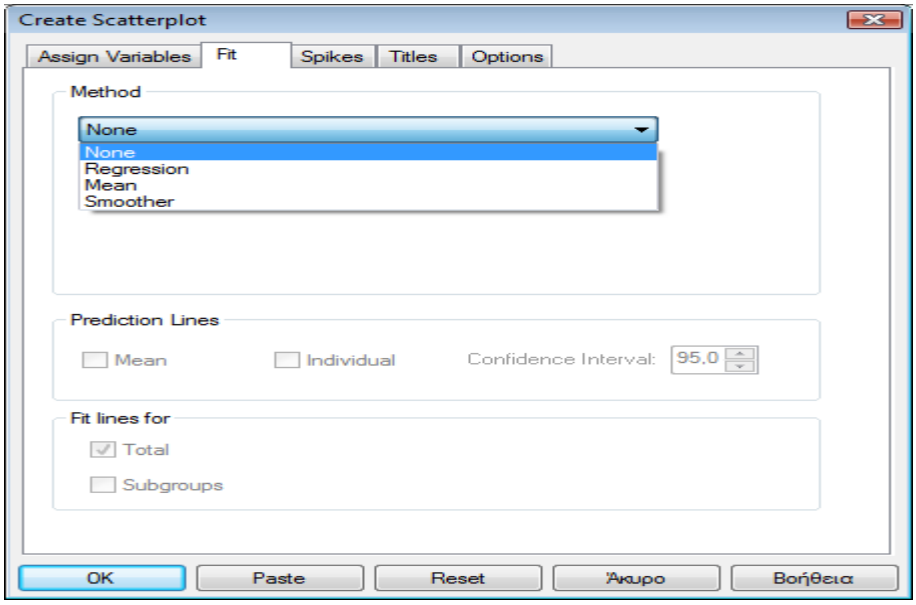
i. Από τη βασική ράβδο του λογισμικού επιλέγουμε Graphs→Interactive→Scatterplot.

ii. Στο νέο παράθυρο διαλόγου που προκύπτει μπορούμε να επιλέξουμε τον τύπο του γραφήματος που θέλουμε να κατασκευάσουμε διδιάστατο ή τρισδιάστατο. Όταν

σκοπός μας είναι να κατασκευάσουμε ένα σύνηθες διάγραμμα διασποράς επιλέγουμε 2-D Coordinate, δηλαδή διδιάστατο γράφημα. Μετακινούμε τη μεταβλητή (συνήθως την εξαρτημένη) στον κατακόρυφο άξονα του ορθογωνίου συστήματος αξόνων, ενώ στον οριζόντιο άξονα μετακινούμε την άλλη ποσοτική μεταβλητή (συνήθως την ανεξάρτητη). Επιπρόσθετα στο πλαίσιο Color μπορούμε να δηλώσουμε μία ποιοτική μεταβλητή ελέγχου έτσι ώστε στο διάγραμμα διασποράς να υπάρχει διάκριση των σημείων αντίστοιχη με τις κατηγορίες της ποιοτικής μεταβλητής. Έτσι για παράδειγμα μπορούμε να τοποθετήσουμε τη μεταβλητή Φύλο.

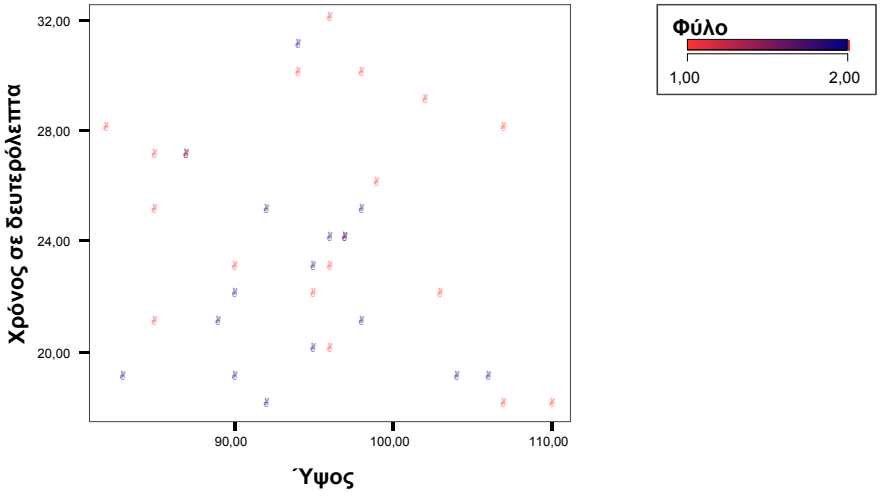


Από το πλαίσιο Fit έχουμε τη δυνατότητα μεταξύ των άλλων να προσαρμόσουμε την ευθεία της εξίσωσης παλινδρόμησης με ή χωρίς σταθερό όρο. Αυτό επιτυγχάνεται με την επιλογή Regression από το πλαίσιο Method. Τέλος από την επιλογή Fit lines for μπορούμε να ζητήσουμε την προσαρμογή της ευθείας είτε συνολικά (Total) είτε σε υποομάδες (Subgroups).



Από τα πλαίσια Spikes και Options έχουμε δυνατότητα περαιτέρω επεξεργασίας του τρόπου εμφάνισης των αποτελεσμάτων στο παράθυρο Output, ενώ από το πλαίσιο Titles καθορίζουμε τους τίτλους, τους υπότιτλους και τις επικεφαλίδες.

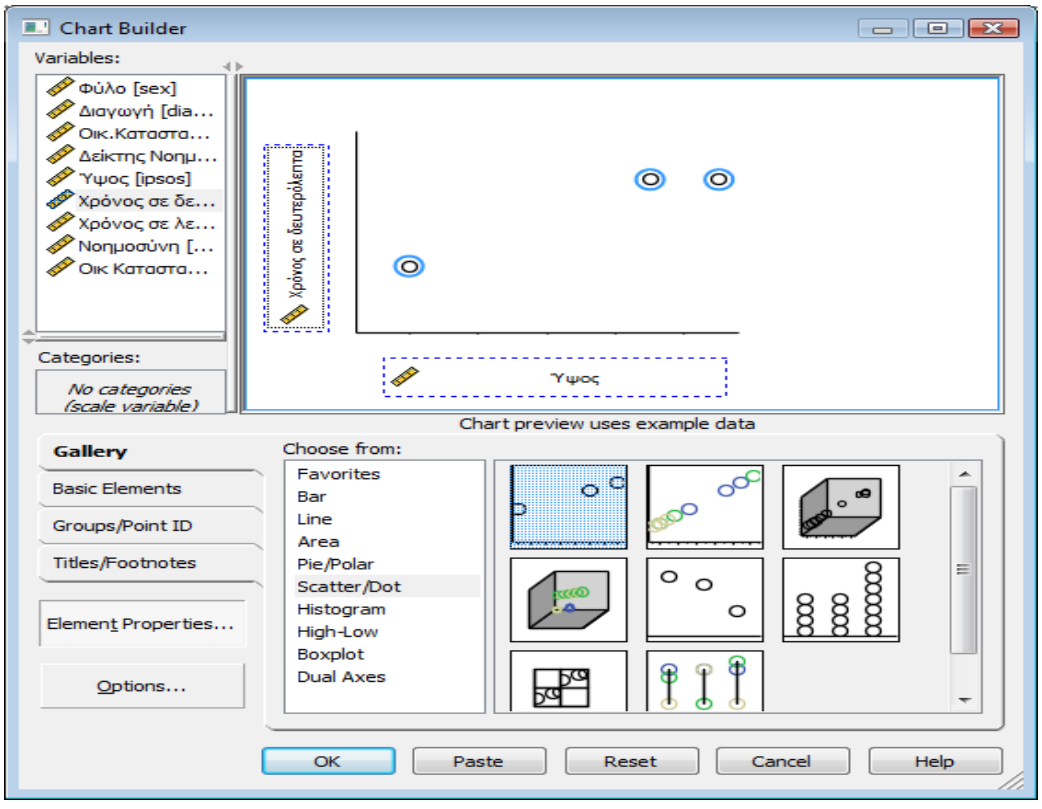
Έτσι για παράδειγμα αποκτούμε το ακόλουθο διάγραμμα διασποράς από το οποίο δε φαίνεται μία ξεκάθαρη γραμμική σχέση.



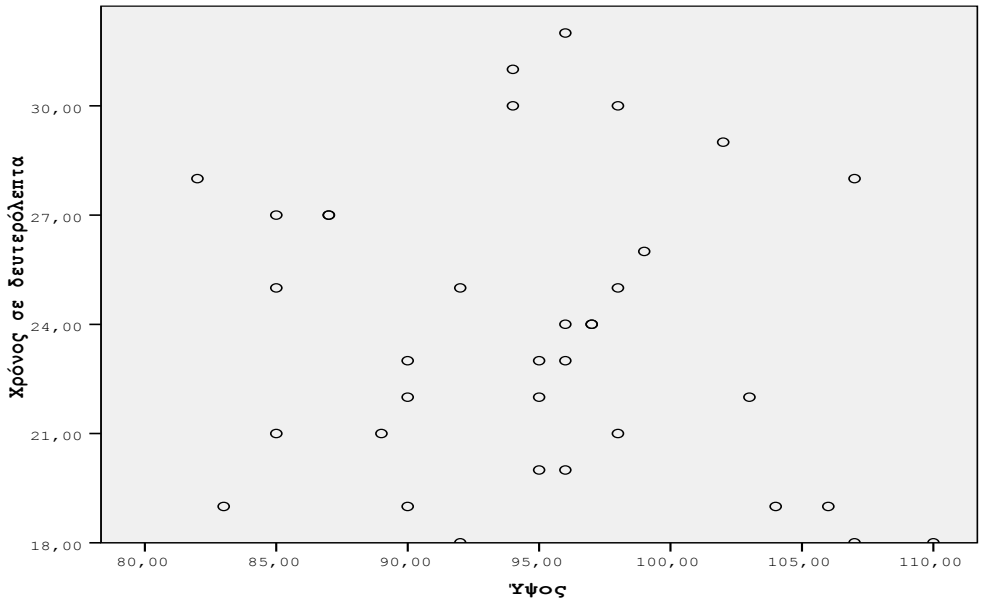
Διαδικασία Chart Builder

i. Από τη βασική ράβδο του λογισμικού επιλέγουμε Graphs→Chart Builder.

ii. Στο νέο παράθυρο διαλόγου που προκύπτει μπορούμε να επιλέξουμε τον τύπο του γραφήματος που θέλουμε να κατασκευάσουμε. Καθώς επιθυμούμε να κατασκευαστεί ένα διάγραμμα διασποράς επιλέγουμε Scatter/Dot. Έπειτα, όταν σκοπός μας είναι να κατασκευάσουμε ένα σύνηθες διάγραμμα διασποράς επιλέγουμε Simple Scatter. Μετακινούμε στο πλαίσιο Y Axis τη μεταβλητή (συνήθως την εξαρτημένη) που θα έχουμε στον κατακόρυφο άξονα του ορθογωνίου συστήματος αξόνων, ενώ στο πλαίσιο X Axis μετακινούμε την άλλη ποσοτική μεταβλητή (συνήθως την ανεξάρτητη). Τέλος από τη επιλογή Options το λογισμικό μας δίνει τη δυνατότητα χειρισμού των ελλιπών τιμών, ενώ από την επιλογή Titles/Footnotes έχουμε την δυνατότητα να ορίσουμε τίτλο, υπότιτλο καθώς και υποσημείωση για το διάγραμμα διασποράς που θα κατασκευαστεί.



Παρατηρούμε από το διάγραμμα διασποράς ότι δεν είναι ξεκάθαρη η ύπαρξη γραμμικής σχέσης.

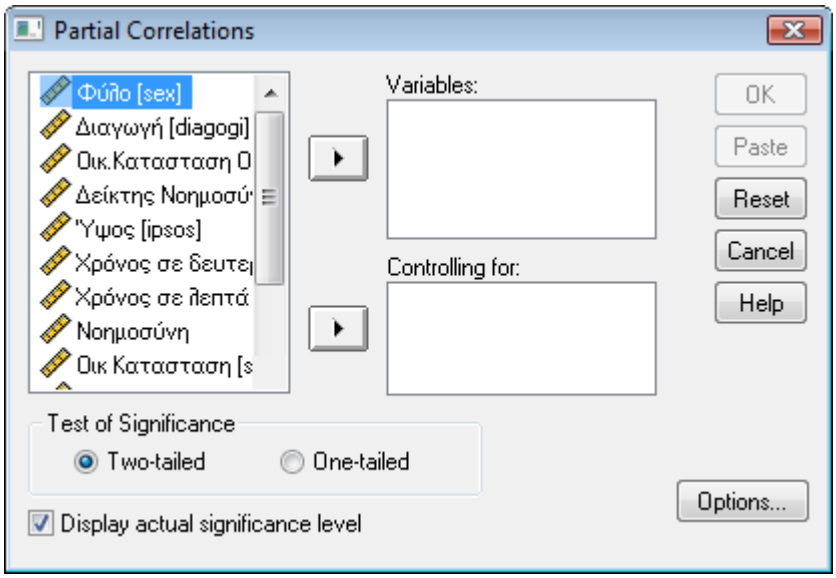


Μερικός συντελεστής συσχέτισης

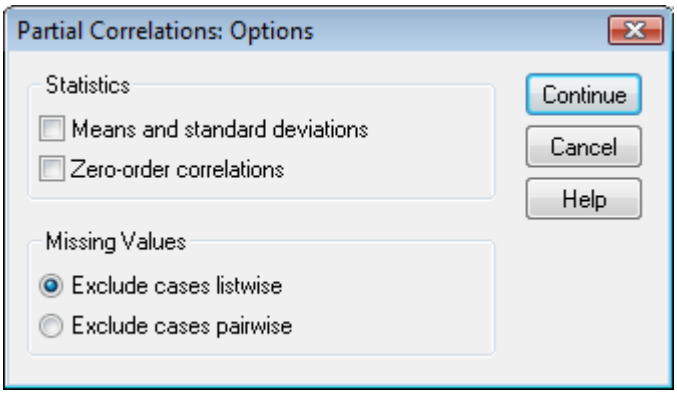
Κλείνοντας τούτη την παράγραφο θα ήταν ίσως παράλειψη να μην αναφέρουμε ότι όταν θέλουμε να εξετάσουμε την ένταση της γραμμικής σχέσης μεταξύ δύο μεταβλητών υπό την επίδραση μίας ή περισσοτέρων μεταβλητών ελέγχου χρησιμοποιούμε το μερικό συντελεστή συσχέτισης. Με τον μερικό συντελεστή συσχέτισης αυτό που προσπαθούμε να κάνουμε είναι να διαπιστώσουμε αν η μεταβλητή ελέγχου είναι εκείνη που προκαλεί τη γραμμική σχέση μεταξύ των μεταβλητών μας. Για την υλοποίηση αυτών (που έπεται όσων αναφέρθηκαν πρωτύτερα) ακολουθούμε τη διαδικασία:

i. Analyze→Correlate→Partial

ii. Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε τις μεταβλητές των οποίων τη σχέση θέλουμε να μελετήσουμε και τις τοποθετούμε στο πλαίσιο Variables, ενώ στο πλαίσιο Controlling for τοποθετούμε την ποσοτική μεταβλητή που υποψιαζόμαστε ότι δημιουργεί τη γραμμική εξάρτηση των υπό μελέτη μεταβλητών. Στο πλαίσιο Test of Significance προχωρούμε σε μονόπλευρο (One-Tailed) ή δίπλευρο έλεγχο (Two-Tailed) για το μερικό πληθυσμιακό συντελεστή συσχέτισης. Έχοντας επιλέξει το πλαίσιο Display actual significance level για κάθε συντελεστή συσχέτισης εμφανίζονται οι βαθμοί ελευθερίας και οι p-τιμές του ελέγχου ότι ο αντίστοιχος πληθυσμιακός συντελεστής συσχέτισης είναι ίσος με μηδέν. Αν δεν το επιλέξουμε οι συντελεστές που είναι στατιστικά σημαντικοί σε επίπεδο σημαντικότητας 0.05 υποδηλώνονται με ένα αστεράκι, ενώ αυτοί που είναι στατιστικά σημαντικοί σε επίπεδο 0.01 υποδηλώνονται με διπλό αστεράκι.



Από την επιλογή Options μας δίνεται η δυνατότητα να υπολογίσουμε τις μέσες τιμές και τυπικές αποκλίσεις (Means and standard deviations). Επιπρόσθετα επιλέγοντας το πλαίσιο Zero-order correlations υπολογίζονται οι απλοί συντελεστές συσχέτισης μεταξύ όλων των μεταβλητών (συμπεριλαμβανομένου και της μεταβλητής που έχουμε μετακινήσει στο πλαίσιο Controlling for). Τέλος, μπορούμε να καθορίσουμε τον τρόπο χειρισμού των ελλιπών τιμών.



Σχόλιο: Η διαδικασία του μερικού συντελεστή συσχέτισης υποθέτει ότι κάθε ζεύγος μεταβλητών ακολουθεί διδιάστατη κανονική κατανομή και είναι ευαίσθητη στην ύπαρξη ακραίων τιμών.

3.3 Ποσοτική-ποιοτική μεταβλητή

Έστω ότι μας ενδιαφέρει να αναζητήσουμε τη σχέση μεταξύ μίας ποσοτικής μεταβλητής και μίας ποιοτικής μεταβλητής, με δύο ή περισσότερες κατηγορίες. Ουσιαστικά αυτό που συνήθως θέλουμε να ελέγξουμε είναι αν οι πληθυσμιακές μέσες τιμές (της ποσοτικής μεταβλητής) δύο ή περισσότερων ομάδων (που καθορίζονται από την ποιοτική μεταβλητή) δε διαφέρουν στατιστικά σημαντικά. Οι έλεγχοι αυτοί αποτελούν αντικείμενο μελέτης των Κεφαλαίων 4-7.