ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΤΟΜΕΑΣ ΠΙΘΑΝΟΤΗΤΩΝ-ΣΤΑΤΙΣΤΙΚΗΣ & ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ

Στατιστική Ανάλυση Δεδομένων με το S.P.S.S.

Διδακτικές Σημειώσεις

Απόστολος Δ. Μπατσίδης

ΙΩΑΝΝΙΝΑ 2014

Στην Ελένη

ΠΕΡΙΕΧΟΜΕΝΑ

Σελ.

ΚΕΦΑΛΑΙΟ ΤΕΤΑΡΤΟ

Έλεγχος ότι η παράμετρος θέσης ενός πληθυσμού είναι ίση με δοθείσα γνωστή τιμή

ΚΕΦΑΛΑΙΟ ΠΕΜΠΤΟ

Έλεγχος για τις παραμέτρους θέσης δύο πληθυσμών με ανεξάρτητα δείγματα

ΚΕΦΑΛΑΙΟ ΕΚΤΟ

Έλεγχος για τις παραμέτρους θέσης δύο πληθυσμών με εξαρτημένα δείγματα

ΚΕΦΑΛΑΙΟ ΕΒΔΟΜΟ

Έλεγχος για τις παραμέτρους θέσης περισσοτέρων από δύο πληθυσμών με ανεξάρτητα δείγματα

ΚΕΦΑΛΑΙΟ ΟΓΔΟΟ

Γραμμική παλινδρόμηση

ΠΡΟΛΟΓΟΣ 1^{ης} έκδοσης

Η ευρεία χρήση της Στατιστικής και σε άλλα επιστημονικά πεδία π.χ. στις κοινωνικές επιστήμες, στην ιατρική, στις οικονομικές επιστήμες κ.α. οδήγησε στη δημιουργία λογισμικών για την εφαρμογή ποικίλων στατιστικών μεθόδων. Μέσω αυτών επιτυγχάνεται η καταγραφή και η περαιτέρω ανάλυση των δεδομένων που θα οδηγήσει στην εξαγωγή συμπερασμάτων για τον υπό μελέτη πληθυσμό. Ένα τέτοιο στατιστικό πρόγραμμα, ίσως το πιο ευρέως χρησιμοποιούμενο, είναι το S.P.S.S. (Statistical Package for Social Sciences), ή όπως αλλιώς πρόσφατα μετονομάστηκε PASW Statistics. Οι παρούσες σημειώσεις επιχειρούν να αποτελέσουν έναν απλό οδηγό μίας στοιχειώδους ανάλυσης δεδομένων και όχι έναν οδηγό για τη διεξαγωγή μίας πλήρους και εμπεριστατωμένης στατιστικής ανάλυσης. Μέσω εκπαιδευτικών συνόλων δεδομένων γίνεται προσπάθεια εφαρμογής όσων έχουν διδαχθεί οι φοιτητές του προπτυχιακού προγράμματος σπουδών του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων στα υπόλοιπα μαθήματα του γνωστικού αντικειμένου της Στατιστικής.

Στο παραπάνω πλαίσιο η διάρθρωση της ύλης έχει ως εξής.

Το πρώτο κεφάλαιο είναι εισαγωγικό. Μέσω ενός απλού παραδείγματος περιγράφεται επιγραμματικά ο τρόπος καταχώρησης, εισαγωγής και αποθήκευσης των δεδομένων, καθώς και οι δυνατότητες μετασχηματισμού και επανακωδικοποίησης αυτών.

Στο δεύτερο κεφάλαιο το ενδιαφέρον επικεντρώνεται στη συνοπτική παρουσίαση των δεδομένων μίας μεταβλητής μέσω γραφημάτων και περιγραφικών μέτρων. Αντικείμενο, δηλαδή, του δεύτερου κεφαλαίου είναι η Περιγραφική Στατιστική.

Στο τρίτο κεφάλαιο παρουσιάζονται οι τρόποι εξέτασης της σχέσης δύο μεταβλητών. Η εξέταση, σύμφωνα και με τις απαιτήσεις ενός προπτυχιακού μαθήματος, διακρίνεται σε τρεις περιπτώσεις. Στην πρώτη υποθέτουμε ότι οι μεταβλητές είναι ποιοτικές, στη δεύτερη ότι είναι ποσοτικές, ενώ στην τρίτη περίπτωση η μία μεταβλητή είναι ποσοτική και η άλλη ποιοτική.

Στο τέταρτο, πέμπτο, έκτο και έβδομο κεφάλαιο εφαρμόζονται, στο πλαίσιο της κλασικής Στατιστικής Συμπερασματολογίας, οι κλασικότεροι έλεγχοι υποθέσεων για έναν, δύο ή περισσότερους πληθυσμούς με μεθόδους της παραμετρικής και μη

παραμετρικής στατιστικής. Στο τέλος κάθε κεφαλαίου υπάρχουν λυμένα ή άλυτα παραδείγματα με τη βοήθεια εκπαιδευτικών συνόλων δεδομένων.

Στο όγδοο κεφάλαιο προσαρμόζεται το μοντέλο της γραμμικής παλινδρόμησης και αναπτύσσεται η ανάλυση των υπολοίπων.

Στο τέλος των σημειώσεων δίνεται η σχετική βιβλιογραφία.

Στο σημείο αυτό θα ήταν σημαντική παράλειψή μου να μην ευχαριστήσω θερμά τον κ. Κωνσταντίνο Καρακώστα, αφυπηρετήσαντα Αν. Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων, για τον πολύτιμο χρόνο που αφιέρωσε για την ανάγνωση της πρώτης έκδοσης αυτών των σημειώσεων, τις εύστοχες προτάσεις και παρατηρήσεις του, καθώς και για την παραχώρηση του υλικού διδασκαλίας και συνόλων δεδομένων του μαθήματος Στατιστική Ανάλυση Δεδομένων. Οι ασκήσεις που επισημαίνονται με * καθώς και τα αντίστοιχα σύνολα δεδομένα προέρχονται από το υλικό διδασκαλίας του κ. Κ. Καρακώστα (βλέπε Καρακώστας (2004)). Βέβαια κάθε παράβλεψη αυτών των σημειώσεων «βαραίνει» μόνο το συγγραφέα. Επιπλέον, θα ήθελα να ευχαριστήσω τον κ. Κωνσταντίνο Ζωγράφο, Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων, γιατί στα πλαίσια των μεταπτυχιακών μου σπουδών μου έδωσε τη δυνατότητα για τη διδασκαλία μεθοδολογιών ανάλυσης δεδομένων σε προπτυχιακό επίπεδο. Η παρότρυνσή του υπήρξε καθοριστική πάντοτε.

Ιωάννινα, Ιούλιος 2011

Απόστολος Δ. Μπατσίδης

ΠΡΟΛΟΓΟΣ 2^{ης} έκδοσης

Στην δεύτερη έκδοση έχει διατηρηθεί ο χαρακτήρας της πρώτης έκδοσης. Όμως από την πρώτη χρονιά συγγραφής αυτών των διδακτικών σημειώσεων, πολλά έχουν αλλάξει στο περιβάλλον του στατιστικού πακέτου προγραμμάτων SPSS, το οποίο πλέον έχει μετονομαστεί IBM SPSS Statistics. Στην έκδοση αυτή έχουν διορθωθεί πολλές αβλεψίες, ενώ δεν θεωρήθηκε σκόπιμο να τροποποιηθούν οι σημειώσεις ως προς τις εντολές του στατιστικού πακέτου.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω τον καθηγητή Πληροφορικής της Δευτεροβάθμιας Εκπαίδευσης, προπτυχιακό φοιτητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων, κατά το Ακ. Έτος 2013-2014, κ. Ερωτόκριτο Αλαμάνο για τις εύστοχες παρατηρήσεις του που οδήγησαν στη βελτίωση της πρώτης έκδοσης. Κλείνοντας θα ήθελα να επισημάνω ότι παρότι οι σημειώσεις εκδίδονται εκ νέου πάντοτε υπάρχουν αβλεψίες, λάθη και κενά. Κάθε παράβλεψη «βαραίνει» μόνο το συγγραφέα και κάθε παρατήρηση είναι ευπρόσδεκτη.

Ιωάννινα, Οκτώβριος 2014

Απόστολος Δ. Μπατσίδης

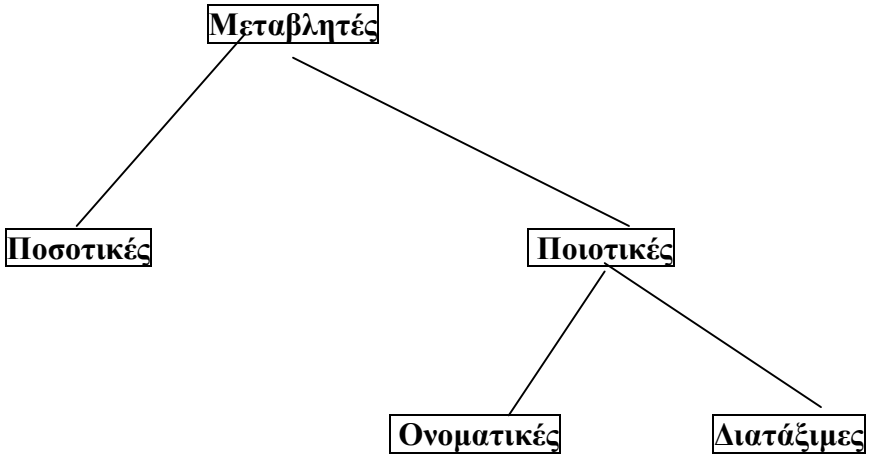
ΚΕΦΑΛΑΙΟ ΠΡΩΤΟ

Εισαγωγή και αποθήκευση δεδομένων-Τα βασικά του S.P.S.S.

1.1. Εισαγωγή δεδομένων στο S.P.S.S.

Για τη διεξαγωγή μίας στατιστικής μελέτης απαιτείται αρχικά η διατύπωση και ο σχεδιασμός του προς επίλυση προβλήματος. Έπειτα συλλέγονται, είτε από διαθέσιμες πηγές είτε με τη διεξαγωγή έρευνας με ερωτηματολόγιο, τα δεδομένα που αφορούν το προς επίλυση πρόβλημα. Το επόμενο καθοριστικό βήμα, για τη μετέπειτα ανάλυση των δεδομένων, την ερμηνεία των αποτελεσμάτων της ανάλυσης και την εξαγωγή συμπερασμάτων, είναι η καταχώρηση των δεδομένων που θα συλλέξουμε σε κάποιο λογισμικό πακέτο ανάλυσης δεδομένων.

Από τα παραπάνω γίνεται κατανοητό ότι η καταχώρηση των προς ανάλυση δεδομένων στο λογισμικό επιδρά στη μετέπειτα ανάλυση και στην εξαγωγή των συμπερασμάτων. Οι ερωτήσεις που διατυπώνονται σε ένα ερωτηματολόγιο αντιστοιχούν τις περισσότερες φορές σε μία μεταβλητή. Γνωρίζουμε από τη θεωρία της Στατιστικής ότι οι μεταβλητές διακρίνονται σε ποσοτικές και ποιοτικές, οι οποίες με τη σειρά τους διακρίνονται σε διατάξιμες και ονοματικές (βλέπε Ζωγράφος, 2003, σελ. 7-15).

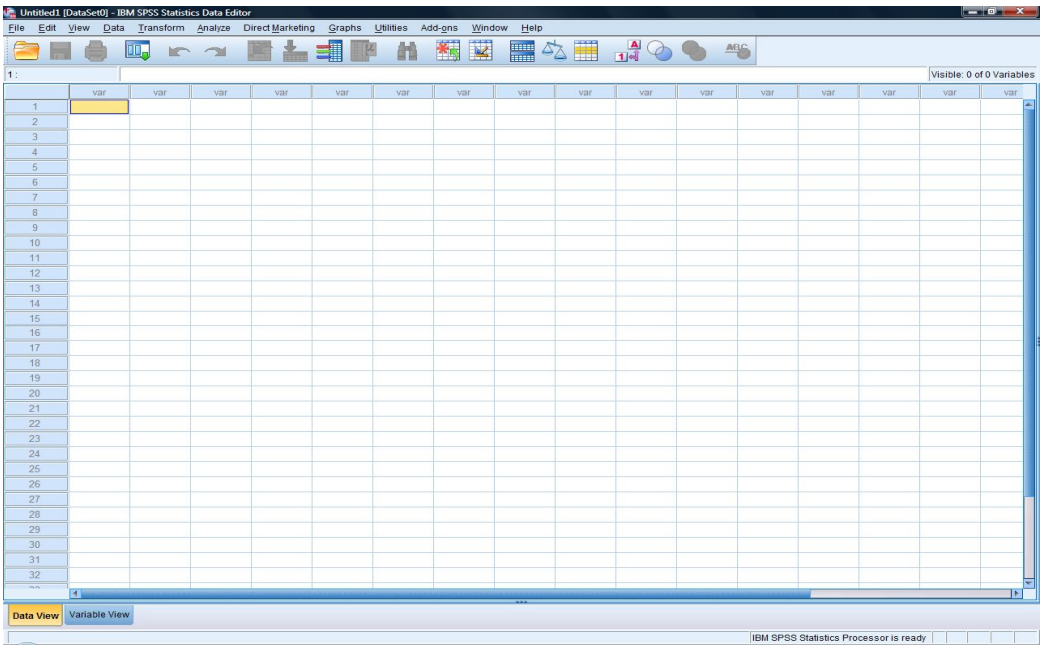


Ποσοτική μεταβλητή είναι εκείνη που μπορεί να μετρηθεί (έχει δηλ. αριθμητικές τιμές). Παραδείγματα ποσοτικών μεταβλητών είναι η ηλικία, το βάρος, το ύψος, η αξία μίας μετοχής, ο δείκτης νοημοσύνης κ.ά.

Ποιοτική μεταβλητή είναι εκείνη που περιγράφει χαρακτηριστικά του πληθυσμού που μεταβάλλονται κατά ποιότητα ή είδος, αλλά όχι κατά μέγεθος. Τέτοιες μεταβλητές είναι το φύλο, η διαγωγή ενός μαθητή, το χρώμα των ματιών-μαλλιών, η στάση υπέρ ή κατά ενός νομοσχεδίου κ.ο.κ. Από αυτές εκείνες που παρέχουν τη δυνατότητα διάταξης ονομάζονται διατάξιμες (π.χ. η διαγωγή ενός μαθητή).

1.1.1 Εισαγωγή δεδομένων από το πληκτρολόγιο

Το S.P.S.S. ή όπως αλλιώς έχει μετονομαστεί σε PASW Statistics, έχει ενσωματωμένο έναν ιδιαίτερα εύχρηστο τρόπο εισαγωγής δεδομένων. Όταν ενεργοποιούμε το λογισμικό εμφανίζεται το παρακάτω παράθυρο διαλόγου του Data Editor.



Το πρώτο βήμα θα γίνει με την ενεργοποίηση του Data Editor. Η ενεργοποίηση αυτή επιτυγχάνεται επιλέγοντας **File→New→Data**, ενώ στη συνέχεια θα πρέπει να ληφθεί υπόψη ότι οι οριζόντιες γραμμές αντιστοιχούν στις n περιπτώσεις-πειραματικές μονάδες, ενώ οι κατακόρυφες στήλες στις p υπό μελέτη μεταβλητές.

Στη συνέχεια, μέσω ενός παραδείγματος θα περιγραφεί πως εισάγουμε ποσοτικά και πως ποιοτικά δεδομένα στο S.P.S.S..

Παράδειγμα 1.1 (Ζωγράφος, 2003, σελ. 13-15)

Έστω ότι επιλέγεται ένα τυχαίο δείγμα 35 παιδιών προσχολικής ηλικίας. Για κάθε παιδί εξετάζεται ο δείκτης νοημοσύνης του, το ύψος του, ο χρόνος σε δευτερόλεπτα που διανύει τα 100 μέτρα, η συμπεριφορά του και η οικονομική κατάσταση της οικογένειας του. Τα δεδομένα παρουσιάζονται στον πίνακα που ακολουθεί όπου στη στήλη Φύλο Α= Αγόρι, Θ= Κορίτσι, στη στήλη Διαγωγή Α= Κοσμιωτάτη και Β= Κοσμία, στη στήλη Οικονομική Κατάσταση Α=0-450, Β= 450-600, Γ=600-900 και Δ= 900 ευρώ και άνω.

Φύλο	Διαγωγή	Οικον. Κατάσταση	IQ	Ύψος	Χρόνος
Α	Β	Β	111	95	22
Θ	Α	Γ	90	98	25
Θ	Α	Γ	90	92	18
Θ	Α	Γ	90	104	19
Α	Α	Α	104	85	21
Α	Α	Β	72	96	20
Θ	Β	Β	105	89	21
Α	Α	Δ	93	103	22
Α	Α	Γ	99	110	18
Α	Α	Β	93	85	27
Α	Α	Β	84	94	30
Θ	Α	Α	95	98	21
Θ	Α	Γ	93	96	24
Α	Α	Δ	78	99	26
Θ	Α	Β	108	83	19
Θ	Β	Β	100	87	27
Α	Α	Α	81	85	25
Α	Α	Γ	77	97	24
Α	Α	Γ	67	96	23
Α	Α	Α	100	107	28
Α	Α	Β	104	102	29
Θ	Α	Δ	111	106	19
Θ	Α	Δ	122	95	20
Α	Α	Β	99	82	28
Θ	Α	Β	108	94	31
Θ	Α	Α	126	90	19
Α	Β	Α	90	90	23
Α	Α	Γ	110	96	32
Α	Α	Γ	117	87	27
Θ	Α	Α	119	97	24
Θ	Α	Δ	105	90	22
Α	Β	Β	100	107	18
Θ	Α	Α	75	92	25
Α	Α	Γ	96	98	30
Θ	Α	Δ	81	95	23

Αρχικά προσδιορίζονται και διακρίνονται σε ποσοτικές και ποιοτικές οι μεταβλητές που χρησιμοποιούνται. Έπειτα καταχωρούνται τα δεδομένα του πίνακα στο S.P.S.S..

Οι μεταβλητές που χρησιμοποιούνται στο παράδειγμα αυτό αναφέρονται στο φύλο, τη διαγωγή, την οικονομική κατάσταση της οικογένειας (ποιοτικές μεταβλητές), το δείκτη νοημοσύνης και το ύψος ενός παιδιού, καθώς και το χρόνο σε δευτερόλεπτα που κάθε παιδί διανύει απόσταση 100 μέτρων (ποσοτικές μεταβλητές).

1. Εισαγωγή δεδομένων ποσοτικής μεταβλητής

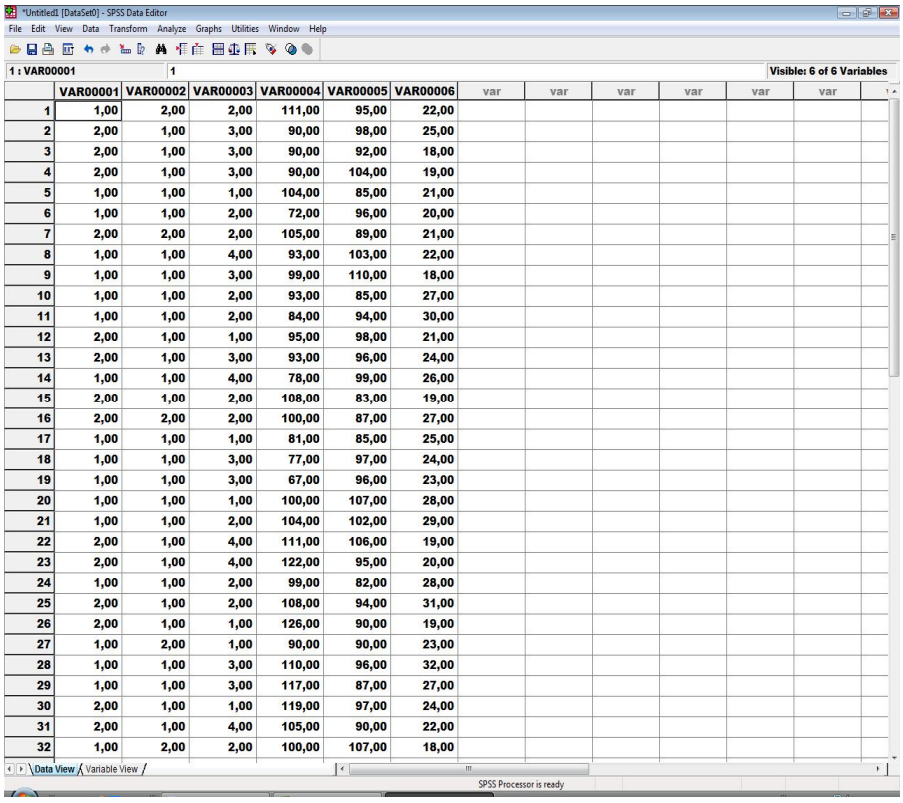
Όταν τα δεδομένα είναι ποσοτικά, η εισαγωγή τους είναι πολύ απλή και εύκολη. Για να ξεκινήσουμε την εισαγωγή των δεδομένων της ποσοτικής μεταβλητής επιλέγουμε ένα από τα υπάρχοντα κελιά (εμφανίζεται ένα μαύρο πλαίσιο στο επιλεχθέν κελί), πληκτρολογούμε την αντίστοιχη τιμή και μετά πατάμε Enter. Κατά αυτόν τον τρόπο η αριθμητική τιμή εισέρχεται στην προκαθορισμένη θέση. Επαναλαμβάνουμε τη διαδικασία αυτήν και για τα n κελιά που δημιουργούνται από τις n γραμμές και την στήλη. Με τον τρόπο αυτό θα έχουμε εισάγει τα δεδομένα μιας ποσοτικής μεταβλητής. Η ίδια διαδικασία υλοποιείται και για τις υπόλοιπες στήλες-ποσοτικές μεταβλητές. Επομένως, κάθε γραμμή των δεδομένων αντιστοιχεί σε ένα εκ των 35 υποκειμένων και η κάθε στήλη σε μία εκ των μεταβλητών (π.χ. σε μία ερώτηση ή σε ένα υποερώτημα).

2. Εισαγωγή δεδομένων ποιοτικής μεταβλητής

Η εισαγωγή δεδομένων ποιοτικής μεταβλητής διαχωρίζεται στην καταγραφή τους σε μορφή είτε αριθμητικών δεδομένων (που είναι επικρατέστερο να γίνεται) είτε χαρακτήρων. Στη συνέχεια θα αναφέρουμε τον πρώτο τρόπο και ως παρατήρηση θα δοθεί ο δεύτερος τρόπος στο βήμα 4.

Η εισαγωγή δεδομένων ποιοτικής μεταβλητής σε μορφή αριθμητικών δεδομένων προϋποθέτει μία προεργασία πάνω στο ερωτηματολόγιο. Η εν λόγω προεργασία περιλαμβάνει την αντιστοίχιση κωδικών (αριθμητικών τιμών) σε όλες τις πιθανές κατηγορίες-απαντήσεις κάθε ποιοτικής μεταβλητής. Μέσω της κωδικοποίησης αυτής κάθε απάντηση-τιμή της ποιοτικής μεταβλητής αντιστοιχεί σε έναν κωδικό. Στη συνέχεια εισάγουμε στα κελιά τους κωδικούς αυτούς ακολουθώντας την ίδια διαδικασία που

περιγράφηκε και κατά την καταχώρηση των ποσοτικών δεδομένων. Για το λόγο αυτό κάνουμε τη σύμβαση ότι για τη μεταβλητή Φύλο θα κατοχυρώνουμε την τιμή 1 όταν έχουμε Α=Αγόρι και την τιμή 2 όταν είναι Θ=Κορίτσι. Με το ίδιο σκεπτικό, για τη Διαγωγή θα κατοχυρώνουμε την τιμή 1 όταν έχουμε Α και την τιμή 2 όταν είναι Β, ενώ για την Οικονομική Κατάσταση θα εισάγουμε την τιμή 1=Α, 2=Β, 3=Γ και 4=Δ. Προκύπτει το ακόλουθο σύνολο δεδομένων στο S.P.S.S. (δίνεται τμήμα του).


*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : VAR00001 | 1 | Visible: 6 of 6 Variables

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	var	var	var	var	var	var
1	1,00	2,00	2,00	111,00	95,00	22,00						
2	2,00	1,00	3,00	90,00	98,00	25,00						
3	2,00	1,00	3,00	90,00	92,00	18,00						
4	2,00	1,00	3,00	90,00	104,00	19,00						
5	1,00	1,00	1,00	104,00	85,00	21,00						
6	1,00	1,00	2,00	72,00	96,00	20,00						
7	2,00	2,00	2,00	105,00	89,00	21,00						
8	1,00	1,00	4,00	93,00	103,00	22,00						
9	1,00	1,00	3,00	99,00	110,00	18,00						
10	1,00	1,00	2,00	93,00	85,00	27,00						
11	1,00	1,00	2,00	84,00	94,00	30,00						
12	2,00	1,00	1,00	95,00	98,00	21,00						
13	2,00	1,00	3,00	93,00	96,00	24,00						
14	1,00	1,00	4,00	78,00	99,00	26,00						
15	2,00	1,00	2,00	108,00	83,00	19,00						
16	2,00	2,00	2,00	100,00	87,00	27,00						
17	1,00	1,00	1,00	81,00	85,00	25,00						
18	1,00	1,00	3,00	77,00	97,00	24,00						
19	1,00	1,00	3,00	67,00	96,00	23,00						
20	1,00	1,00	1,00	100,00	107,00	28,00						
21	1,00	1,00	2,00	104,00	102,00	29,00						
22	2,00	1,00	4,00	111,00	106,00	19,00						
23	2,00	1,00	4,00	122,00	95,00	20,00						
24	1,00	1,00	2,00	99,00	82,00	28,00						
25	2,00	1,00	2,00	108,00	94,00	31,00						
26	2,00	1,00	1,00	126,00	90,00	19,00						
27	1,00	2,00	1,00	90,00	90,00	23,00						
28	1,00	1,00	3,00	110,00	96,00	32,00						
29	1,00	1,00	3,00	117,00	87,00	27,00						
30	2,00	1,00	1,00	119,00	97,00	24,00						
31	2,00	1,00	4,00	105,00	90,00	22,00						
32	1,00	2,00	2,00	100,00	107,00	18,00						

Data View | Variable View

SPSS Processor is ready

3. Ονομασία μεταβλητών

Αφού εισάγουμε τα δεδομένα, διαπιστώνουμε ότι το S.P.S.S. εξ ορισμού ονομάζει τις μεταβλητές →VAR00001, VAR00002 κ.ο.κ. Κάτι τέτοιο φυσικά δεν είναι καθόλου εύχρηστο. Θα ορίσουμε σε κάθε μεταβλητή το όνομα που εμείς επιθυμούμε έχοντας ως γνώμονα τους ακόλουθους κανόνες. Το όνομα κάθε μεταβλητής:

α) μπορεί να "καταλαμβάνει" 64 bytes, που σημαίνει 64 χαρακτήρες στις συνήθεις γλώσσες,

β) είναι μοναδικό,

γ) πρέπει να ξεκινά με γράμμα ή με έναν από τους χαρακτήρες @, #, ή $,

δ) δεν πρέπει να περιέχει σημεία στίξης πλην της τελείας, αστεράκια καθώς και κενά,

ε) δεν μπορεί να περικλείονται οι λέξεις All,Ne, Eq, To, Le, Lt, By, Or, Gt, And, Not, Ge, With,

στ) μπορεί να γραφεί τόσο με μικρά όσο και με κεφαλαία γράμματα,

ζ) καλό θα ήταν να μην έχει ως τελευταίο χαρακτήρα την τελεία και την κάτω παύλα, τέλος

η) δεν επιτρέπεται να ξεκινά με το σύμβολο $ ονομασία μεταβλητής που ορίζεται από το χρήστη, ενώ είναι επιτρεπτή για παράδειγμα η ακόλουθη ονομασία A._$@#1.

Η αλλαγή ενός ονόματος (μετονομασία) επιτυγχάνεται επιλέγοντας τη μεταβλητή που θέλουμε να επεξεργαστούμε και κάνοντας διπλό κλικ στο όνομα της.

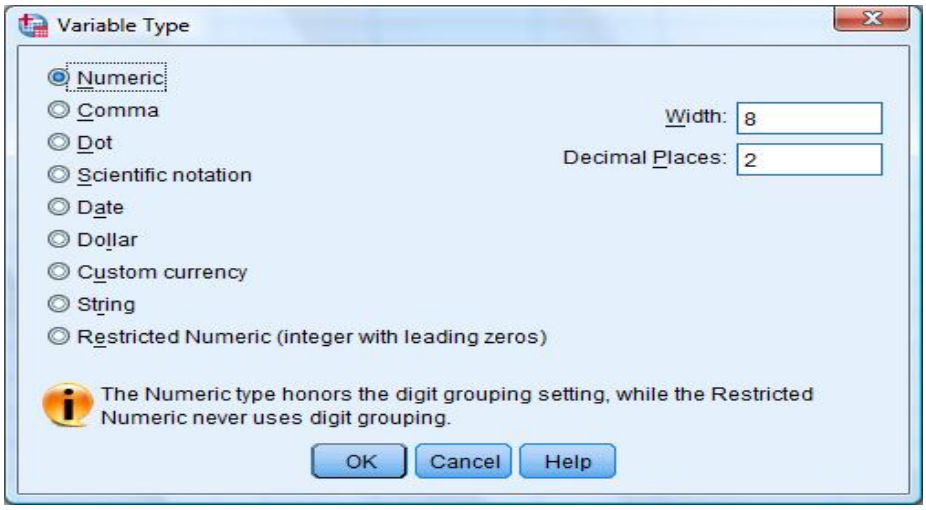
Σχόλιο: Εναλλακτικά μπορεί να επιλεγεί το πλαίσιο Variable View (εμφανίζεται στο κάτω αριστερό άκρο της προηγούμενης εικόνας) .

Στο παράθυρο **Variable View** και στο πλαίσιο **Name** εισάγουμε τα κωδικοποιημένα ονόματα των μεταβλητών μας, έστω Sex, Diagogi, Status, Iq, Ipsos, Time. Επιπλέον, στο πεδίο Label δηλώνεται η πλήρης περιγραφή του ονόματος της μεταβλητής που βοηθά στην καλύτερη παρουσίαση των αποτελεσμάτων μας. Κατά αυτό τον τρόπο δηλώνουμε την ονομασία που θα εμφανίζεται στους πίνακες των αποτελεσμάτων των αναλύσεων που θα ακολουθήσουν π.χ. Φύλο, Διαγωγή, Οικον. Κατάσταση, Δείκτης Νοημοσύνης, Ύψος, Χρόνος σε δευτερόλεπτα.

4. Καθορισμός του τύπου της μεταβλητής

Στο πλαίσιο **Variable Type** το λογισμικό με βάση τις τιμές που πληκτρολογούμε καθορίζει αυτόματα τον τύπο της μεταβλητής, έχοντας ως προεπιλογή να τις εμφανίζει αριθμητικές (numeric) με 2 δεκαδικά ψηφία (Decimals Places) και συνολικό μήκος (δηλώνεται στο πλαίσιο Width) 8 θέσεων. Για τον υπολογισμό του μήκους μίας μεταβλητής λαμβάνονται υπόψη το πρόσημο, το ακέραιο μέρος, η δεκαδική τελεία

καθώς και το δεκαδικό μέρος της. Σε περίπτωση που τα δεδομένα μας είναι τέτοια που παραβιάζονται αυτές οι προεπιλογές πρέπει να τις τροποποιήσουμε κατάλληλα.



Άλλοι δυνατοί τύποι δεδομένων είναι οι ακόλουθοι:

Comma. Αριθμητικές τιμές που έχουν το κόμμα «,» ανά τρεις θέσεις και την τελεία «.» ως υποδιαστολή, π.χ. 6,900.38.

Dot. Αριθμητικές τιμές που έχουν τελεία «.» ανά τρεις θέσεις και το κόμμα «,» ως υποδιαστολή, π.χ. 6.900,38.

Scientific notation. Ποσοτική μεταβλητή της οποίας οι αριθμητικές τιμές γράφονται σε επιστημονική μορφή. Για παράδειγμα στη μορφή 123, 1.23E2, 1.23D2, 1.23E+2, και 1.23+2.

Date. Ημερομηνίες.

Dollar. Τιμές δολαρίου, με ή χωρίς το σύμβολο ($), με την τελεία ως υποδιαστολή και το κόμμα ανά τρεις θέσεις, π.χ. 7,355.38 δολλάρια.

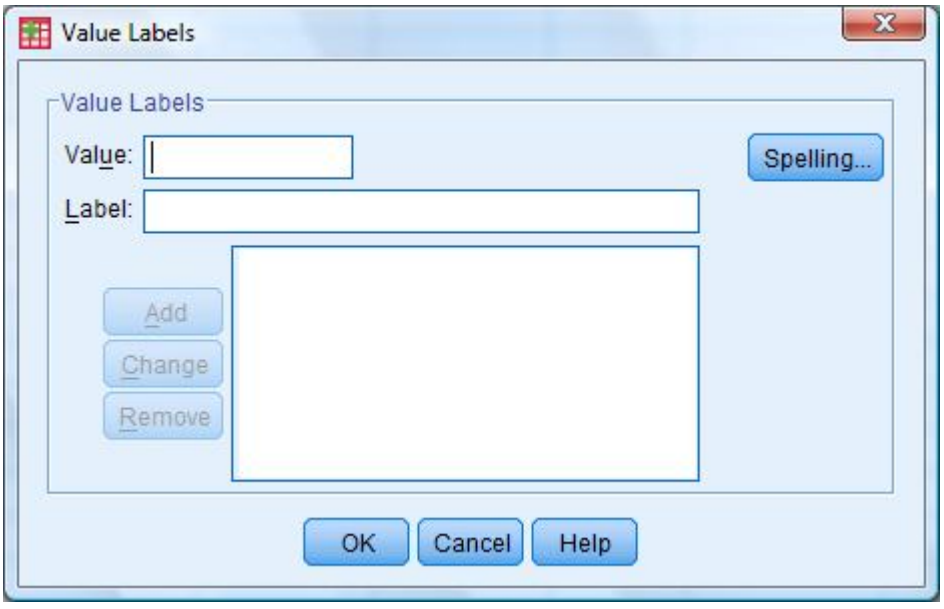
Custom currency. Αριθμητικές τιμές των οποίων ο τρόπος εμφάνισης καθορίζεται από το χρήστη.

String. Μεταβλητή που δεν είναι αριθμητική και επομένως δε χρησιμοποιείται στους υπολογισμούς.
Restricted numeric. Μεταβλητή της οποίας οι τιμές είναι μη αρνητικοί ακέραιοι.

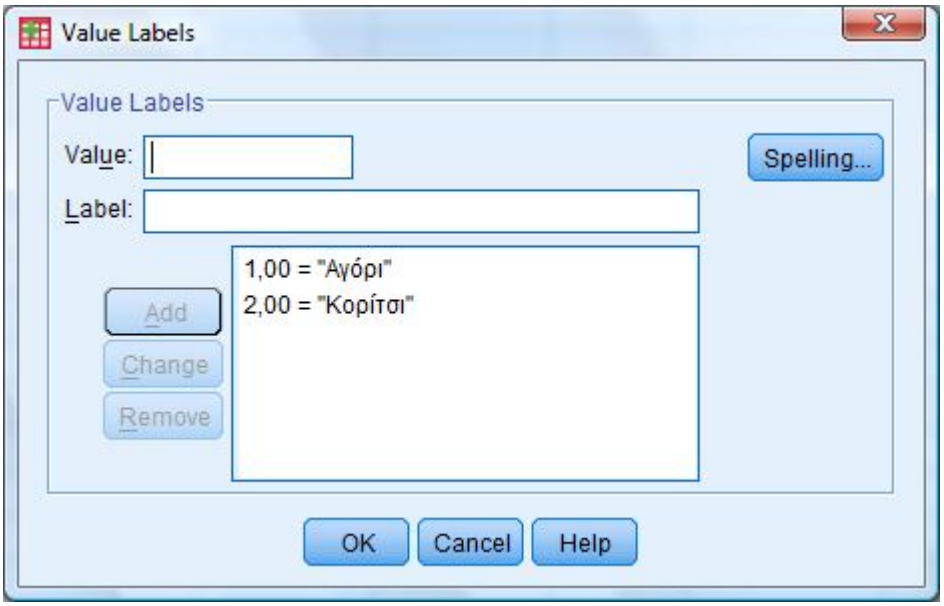
Παρατήρηση: Είναι δυνατή και η εισαγωγή δεδομένων με μη αριθμητικά δεδομένα αρκεί να δημιουργηθεί μία μεταβλητή, η οποία μέσω της καρτέλας του Variable Type του Variable View να οριστεί ως αλφαριθμητική, δηλαδή ως String.

5. Ετικέτες τιμών μιας μεταβλητής

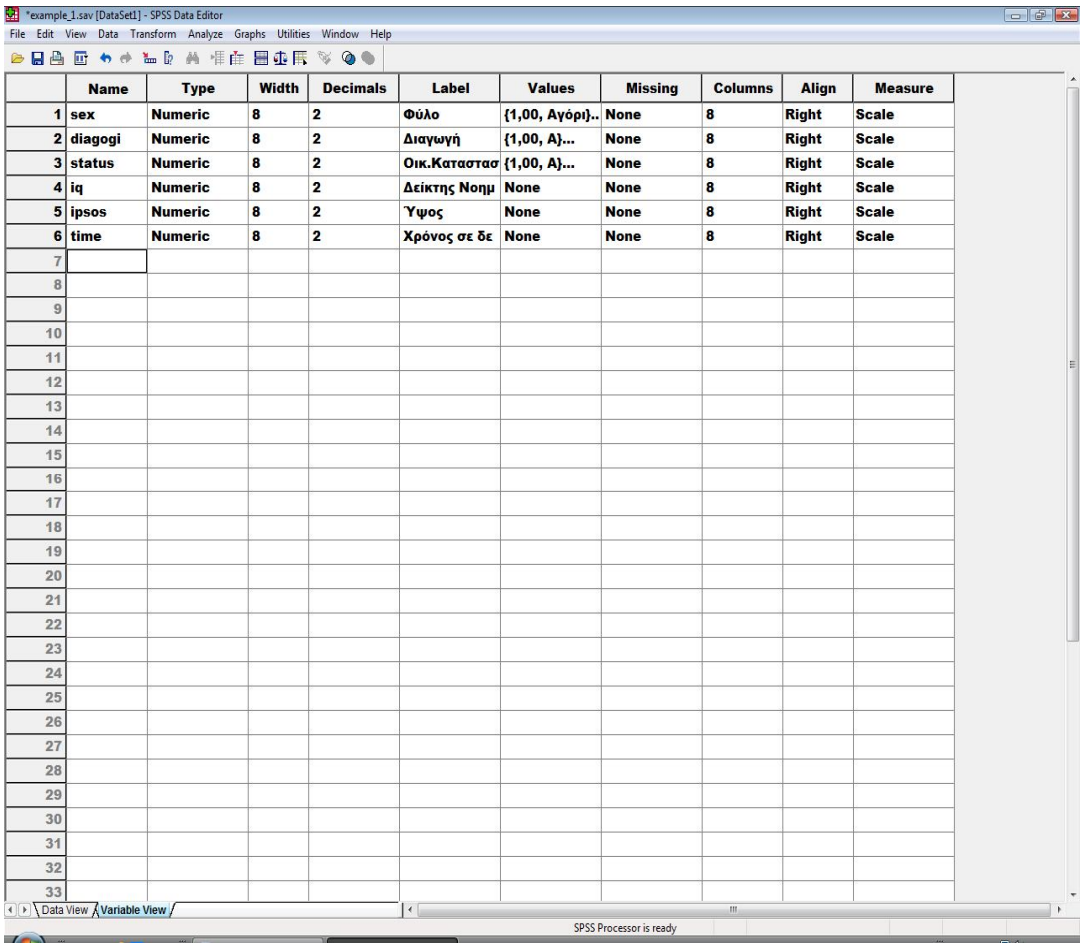
Για να μην ανατρέχουμε συνεχώς στο ερωτηματολόγιο προκειμένου να θυμηθούμε τι σημαίνει ο κάθε κωδικός είναι χρήσιμο όλοι οι κωδικοί των μεταβλητών της μελέτης να καταγράφονται στο πλαίσιο **Values** του παραθύρου Variable View. Επομένως, στο πεδίο αυτό ουσιαστικά εισάγουμε στο λογισμικό τις συμβάσεις τις οποίες κάναμε κατά την καταχώρηση των δεδομένων. Αυτό επιτυγχάνεται κλικάροντας το κάτω δεξί άκρο του κελιού, το οποίο σχηματίζεται από την μεταβλητή και τη στήλη Values. Προκύπτει τότε το παράθυρο διαλόγου:



Εισάγουμε μία-μία τις τιμές της, συνηθέστερα, κατηγορικής μεταβλητής στο πλαίσιο **Value** πληκτρολογώντας την τιμή που αντιπροσωπεύει την κάθε κατηγορία της μεταβλητής, και έπειτα στο πλαίσιο **Value Label** δίνεται η περιγραφή της π.χ. Αγόρι. Επιλέγουμε το Add και επαναλαμβάνουμε την παραπάνω διαδικασία έως ότου προστεθεί κάθε άλλη δυνατή τιμή και ονομασία της κατηγορικής μεταβλητής. Όταν ολοκληρωθεί η παραπάνω διαδικασία πατάμε το πλήκτρο OK.



Η παραπάνω διαδικασία επαναλαμβάνεται για όλες τις ποιοτικές μεταβλητές του παραδείγματός μας. Αν ακολουθηθεί η παραπάνω διαδικασία θα προκύψει το ακόλουθο παράθυρο.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	sex	Numeric	8	2	Φύλο	{1,00, Αγόρι}..	None	8	Right	Scale
2	diagogi	Numeric	8	2	Διαγωγή	{1,00, Α}...	None	8	Right	Scale
3	status	Numeric	8	2	Οικ.Καταστασ	{1,00, Α}...	None	8	Right	Scale
4	iq	Numeric	8	2	Δείκτης Νοημ	None	None	8	Right	Scale
5	ipsos	Numeric	8	2	Ύψος	None	None	8	Right	Scale
6	time	Numeric	8	2	Χρόνος σε δε	None	None	8	Right	Scale

6. Επιπλέον δυνατότητες από το παράθυρο Variable View

Στο πλαίσιο **Missing Values** καθορίζουμε τις τιμές των ελλιπών τιμών μίας μεταβλητής. Για το σκοπό αυτό το λογισμικό μας δίνει τις ακόλουθες επιλογές:

α) No missing values (προεπιλογή). Δεν θεωρείται ελλιπής τιμή καμία παρατήρηση εκτός αυτών με τα κενά κελιά.

β) Discrete missing values. Δηλώνονται στα τρία πλαίσια οι 3 διαφορετικές τιμές που όταν καταγράφονται κατά την εισαγωγή των δεδομένων θα σημαίνουν ελλιπή τιμή π.χ. -1, -99, 0. Η επιλογή αυτή είναι ιδιαίτερα χρήσιμη αν για παράδειγμα θέλουμε να γίνει διαχωρισμός μεταξύ των ελλιπών τιμών που εμφανίζονται λόγω του ότι: (i) η

συγκεκριμένη ερώτηση δεν υποβλήθηκε στον ερωτώμενο, (ii) ο ερωτώμενος δεν απάντησε στο συγκεκριμένο ερώτημα και (iii) ο ερωτώμενος είχε αποχωρήσει από την έρευνα ήδη.

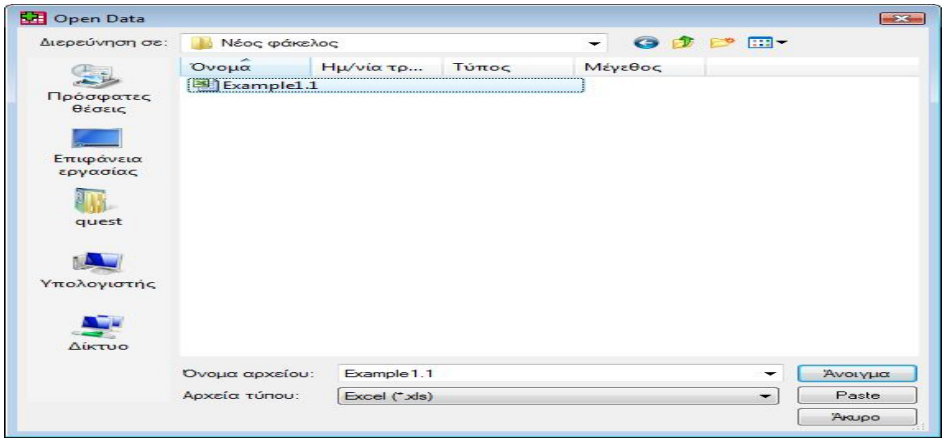
γ) Range plus one optional discrete missing. Δηλώνεται ένα διάστημα, με κάτω και άνω άκρο τις δηλωθείσες τιμές στα πλαίσια Low και High αντίστοιχα, καθώς και μία διακριτή τιμή. Κάθε τιμή που καταγράφεται εντός του διαστήματος καθώς και η διακριτή τιμή λαμβάνεται ως ελλιπής.

Στο πλαίσιο Columns καθορίζουμε το μήκος κάθε στήλης, ενώ στο πλαίσιο Align επιλέγουμε την επιθυμητή στοίχιση των δεδομένων εντός των κελιών (αριστερά, δεξιά, στο κέντρο). Τέλος, από το πλαίσιο Measure καθορίζουμε το είδος της μεταβλητής (Scale=Ποσοτική, Ordinal=Διατάξιμη, Nominal=Ονοματική).

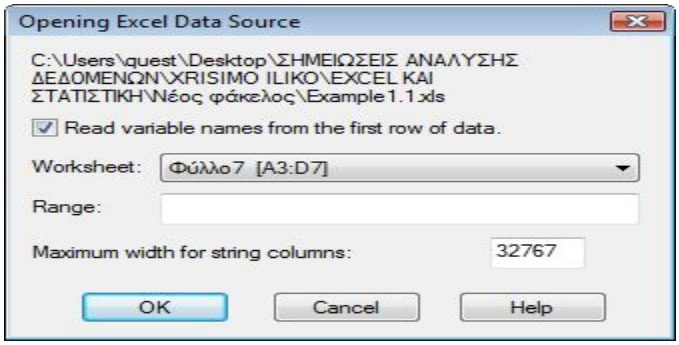
1.1.2 Εισαγωγή δεδομένων από αρχείο του Microsoft Excel

Έστω ότι θέλουμε να εισάγουμε στο S.P.S.S τα δεδομένα ενός αρχείου Excel π.χ. του Example1.1.xlsx (η κατάληξη xlsx ή xls είναι κοινή για αρχεία του Excel, ανάλογα με την έκδοση του). Αυτό επιτυγχάνεται ως εξής:

1. Από το αρχικό παράθυρο διαλόγου του Data View του S.P.S.S. επιλέγουμε File→Open→Data και επιλέγουμε το αρχείο Excel που αναζητούμε. Ζητούμε να ανοίξουμε (Open) αυτό το αρχείο, σύμφωνα και με όσα φαίνονται στην παρακάτω εικόνα



2. Τότε προκύπτει το ακόλουθο παράθυρο διαλόγου που μας επιτρέπει να καθορίζουμε πότε τα ονόματα των μεταβλητών περιέχονται ή όχι στην πρώτη γραμμή του φύλλου εργασίας του Excel (Read variable names from the first row of data). Επιπλέον, καθορίζουμε τα δεδομένα που θέλουμε να μεταφερθούν στο S.P.S.S (δηλώνονται στο Range).



Επιπλέον, είναι επιτρεπτό να καθορίσουμε επιπρόσθετα και ποια φύλλα του αρχείου μας θέλουμε να μετατρέψουμε σε αρχείο δεδομένων του S.P.S.S (δηλώνονται στο Worksheet).

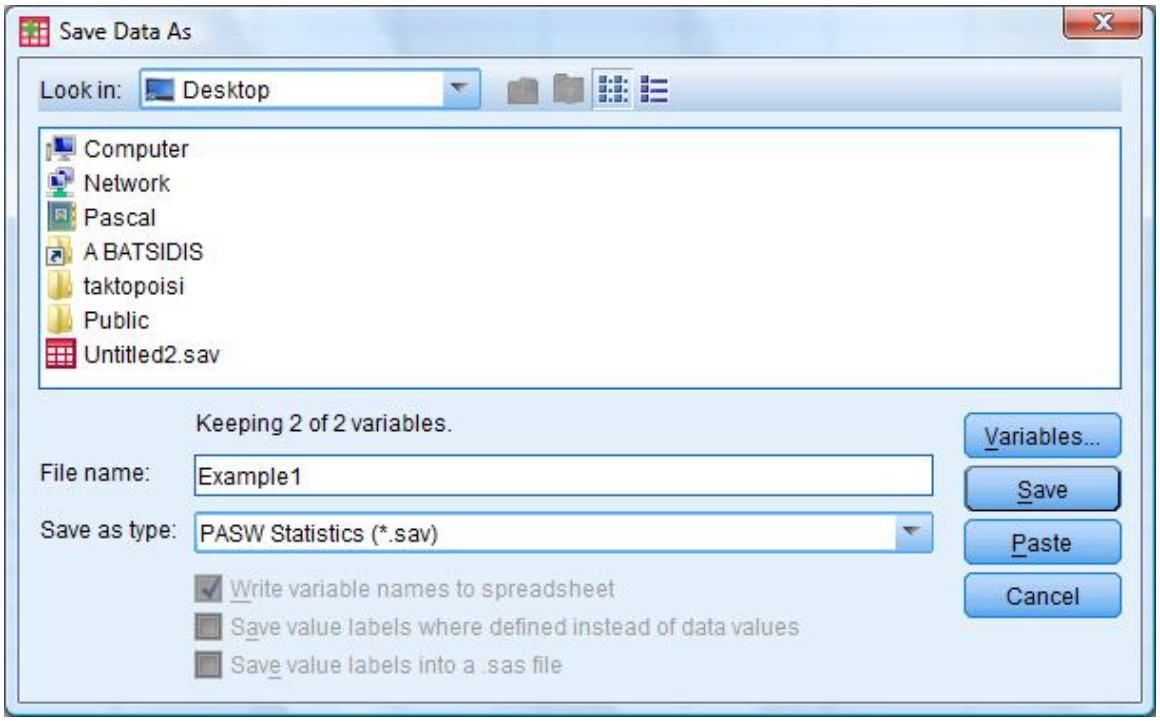
Παρατήρηση

Αν στο αρχείο Excel τα ονόματα των μεταβλητών δηλώνονται στην πρώτη γραμμή των δεδομένων τότε επιλέγοντας το πλαίσιο Read variable names from first row of data αν οι ονομασίες αυτές δεν συμφωνούν με τους κανόνες ονομασίας των αρχείων του S.P.S.S μετατρέπονται σε τέτοιες έτσι ώστε να ικανοποιούνται και οι αυθεντικές ονομασίες αποθηκεύονται στο πλαίσιο Variables labels.

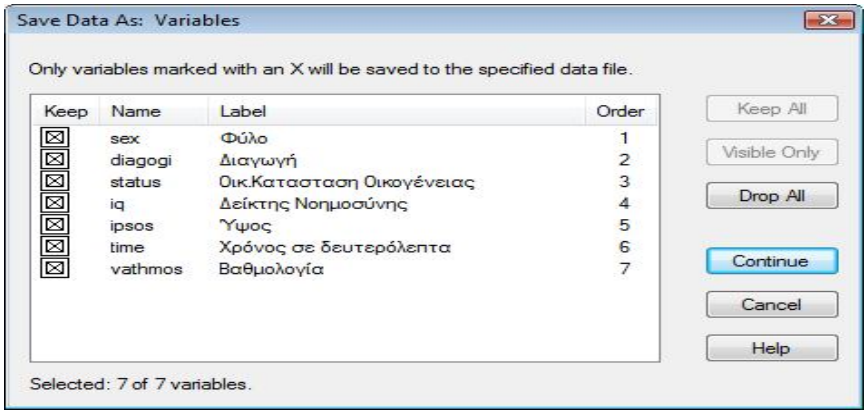
1.2 Αποθήκευση Δεδομένων

Η αποθήκευση των δεδομένων επιτυγχάνεται, όπως σε όλα τα προγράμματα, με τις επιλογές **File→Save** για υπάρχον αρχείο δεδομένων και **File→Save as,** όταν πρόκειται για ένα νέο αρχείο δεδομένων. Στη δεύτερη περίπτωση δηλώνεται και το όνομα του νέου αρχείου καθώς και σε ποιο φάκελο δεδομένων του υπολογιστή θα

αποθηκευτεί. Έτσι, το αρχείο του παραδείγματος αποθηκεύτηκε με το όνομα example_1.sav (sav είναι η κατάληξη των αρχείων δεδομένων του S.P.S.S.) στο φάκελο Data.



Το λογισμικό του S.P.S.S. δίνει τη δυνατότητα τόσο εξ ολοκλήρου αποθήκευσης ενός συνόλου δεδομένων (μέσω της επιλογής Αποθήκευση ή Save) όσο και ενός τμήματος αυτού (μέσω της επιλογής Variables). Ειδικότερα, κλικάροντας στο πλαίσιο Variables προκύπτει το παρακάτω παράθυρο διαλόγου όπου από τη στήλη Keep επιλέγονται οι μεταβλητές τα δεδομένα των οποίων επιθυμούμε να αποθηκευτούν για τη συνέχιση της στατιστικής ανάλυσης



1.3 Μετασχηματισμός και επανακωδικοποίηση δεδομένων

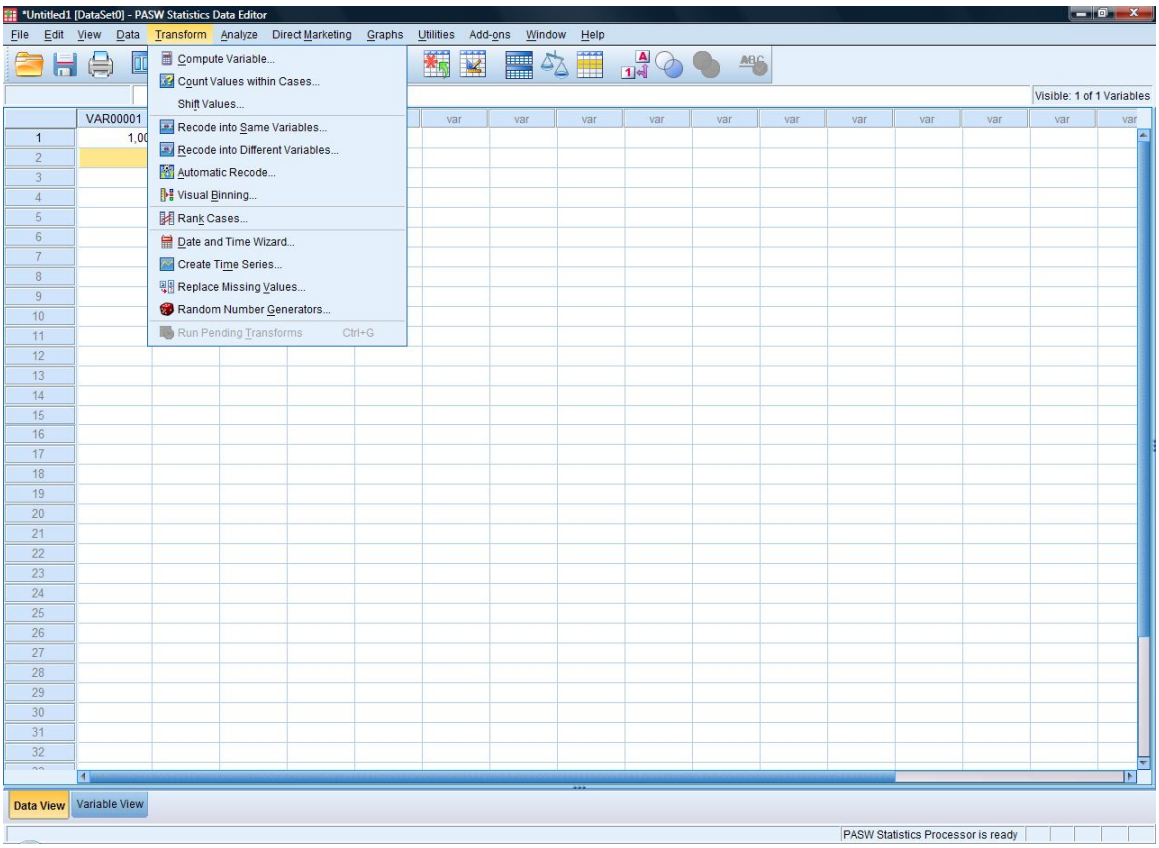
Μερικές φορές σε μία ανάλυση, κρίνεται απαραίτητος ο μετασχηματισμός των δεδομένων π.χ. για την επίτευξη της κανονικότητας όπως θα δούμε σε επόμενο εδάφιο, ή η επανακωδικοποίηση των δεδομένων π.χ. για τη συγχώνευση γειτονικών κελιών στους πίνακες συνάφειας. Ας δούμε περιληπτικά πως υλοποιούνται στο S.P.S.S. μέσω απλών παραδειγμάτων

Μετασχηματισμός δεδομένων

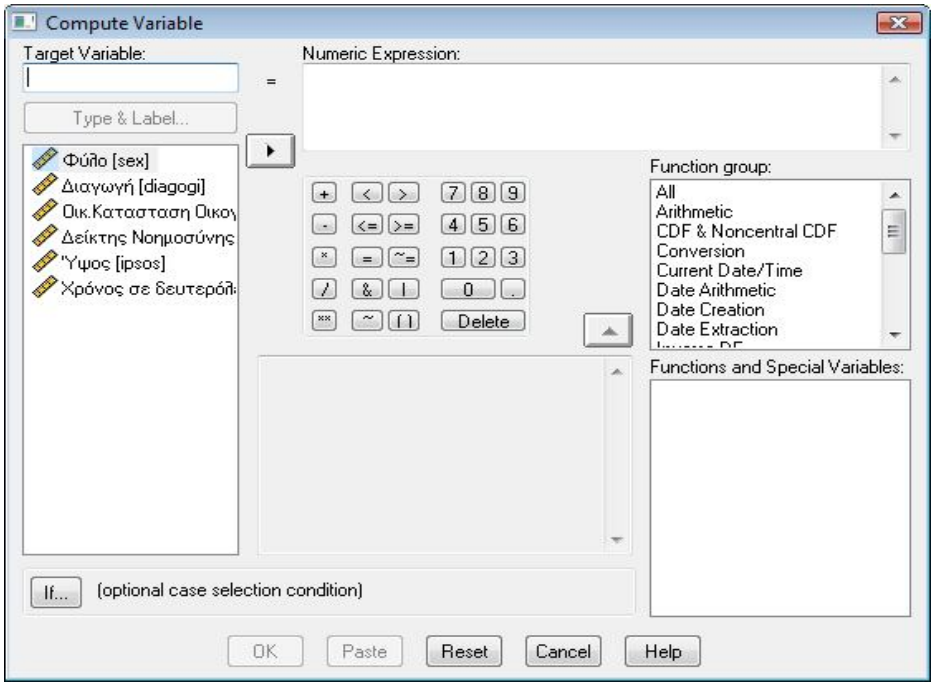
Για κάθε παιδί του Παραδείγματος 1.1. καταγράφεται ο χρόνος σε δευτερόλεπτα που διανύει τα 100 μέτρα καθώς και το ύψος του σε εκατοστά. Ζητείται να δημιουργηθεί μία νέα μεταβλητή που θα μετρά τον προαναφερθέντα χρόνο με μονάδα μέτρησης το λεπτό.

Έχοντας αποφασίσει τον επιθυμητό μετασχηματισμό ακολουθούμε τα ακόλουθα βήματα:

1. Επιλέγουμε από το βασικό μενού **Transform→Compute Variable.**



2. Στο παράθυρο **Compute Variable** στο πλαίσιο **Target Variable** δηλώνουμε το όνομα της νέας μεταβλητής, έστω TimeMinutes, ενώ έπειτα από το πλαίσιο **Type & Label** δίνεται η δυνατότητα να ορίσουμε πιο λεπτομερή περιγραφή της.



Στο πλαίσιο Numeric Expression σχηματίζουμε τον κατάλληλο μετασχηματισμό κάνοντας χρήση του Calculator Pad (αν χρειαστεί να χρησιμοποιήσουμε δεκαδικούς τότε κάνουμε χρήση της τελείας) και των συναρτήσεων που δίνονται στο πλαίσιο **Function Group**. Στο πλαίσιο αυτό μεταξύ άλλων έχουμε την ακόλουθη ομαδοποίηση των συναρτήσεων:

α) All: δίνονται όλες οι συναρτήσεις σε αλφαβητική σειρά διάταξης.

β) Arithmetic: δίνονται αριθμητικές συναρτήσεις όπως η απόλυτη τιμή (Abs), το συνημίτονο (Cos), το ημίτονο (Sin), ο δεκαδικός λογάριθμος (Lg10), ο φυσικός λογάριθμος (Ln), η τετραγωνική ρίζα (Sqrt) κ.ά.

γ) CDF and Noncentral CDF: δίνονται οι τιμές των αθροιστικών συναρτήσεων κατανομών (cdf=cumulative distribution function) και των μη κεντρικών αθροιστικών συναρτήσεων κατανομών ειδικών, γνωστών κατανομών όπως η διωνυμική, η εκθετική, η κανονική, η μη κεντρική t κατανομή κ.ά.

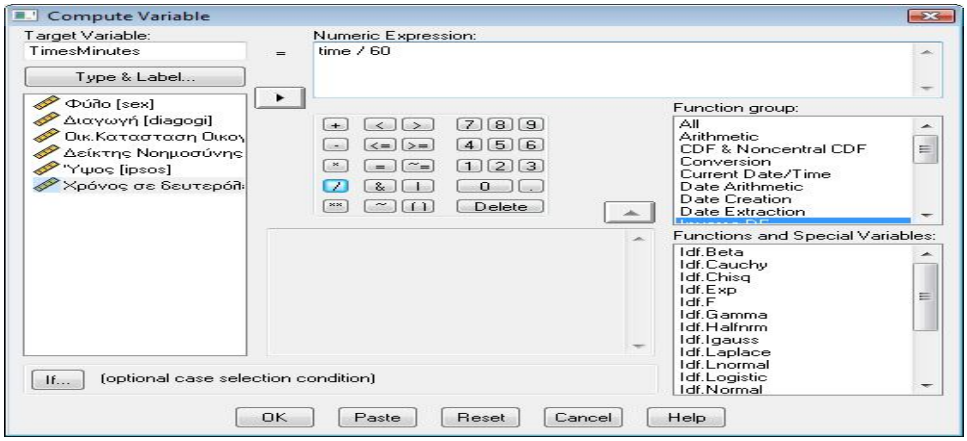
δ) Inverse DF: δίνει την τιμή της κατανομής για την οποία η αθροιστική συνάρτηση κατανομής είναι ίση με προκαθορισμένη πιθανότητα.

ε) PDF and NonCentral PDF: μας δίνει την τιμή της συνάρτησης πυκνότητας πιθανότητας ή της συνάρτησης πιθανότητας για γνωστές τιμές των παραμέτρων της κατανομής σε προκαθορισμένη τιμή.

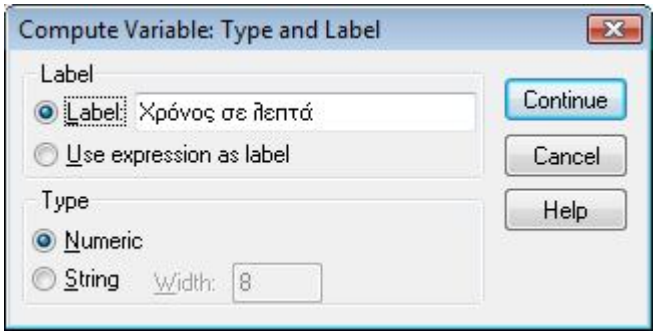
στ) Random Numbers: δημιουργεί μία στήλη δεδομένων που αποτελούν ένα τυχαίο δείγμα από διάφορους πληθυσμούς π.χ. από έναν εκθετικό ή κανονικό πληθυσμό.

ζ) Statistical: δίνονται στατιστικές συναρτήσεις όπως είναι η μέση τιμή, η τυπική απόκλιση, η διακύμανση, ο συντελεστής μεταβλητότητας κ.ά.

Για το παράδειγμά μας έχουμε:

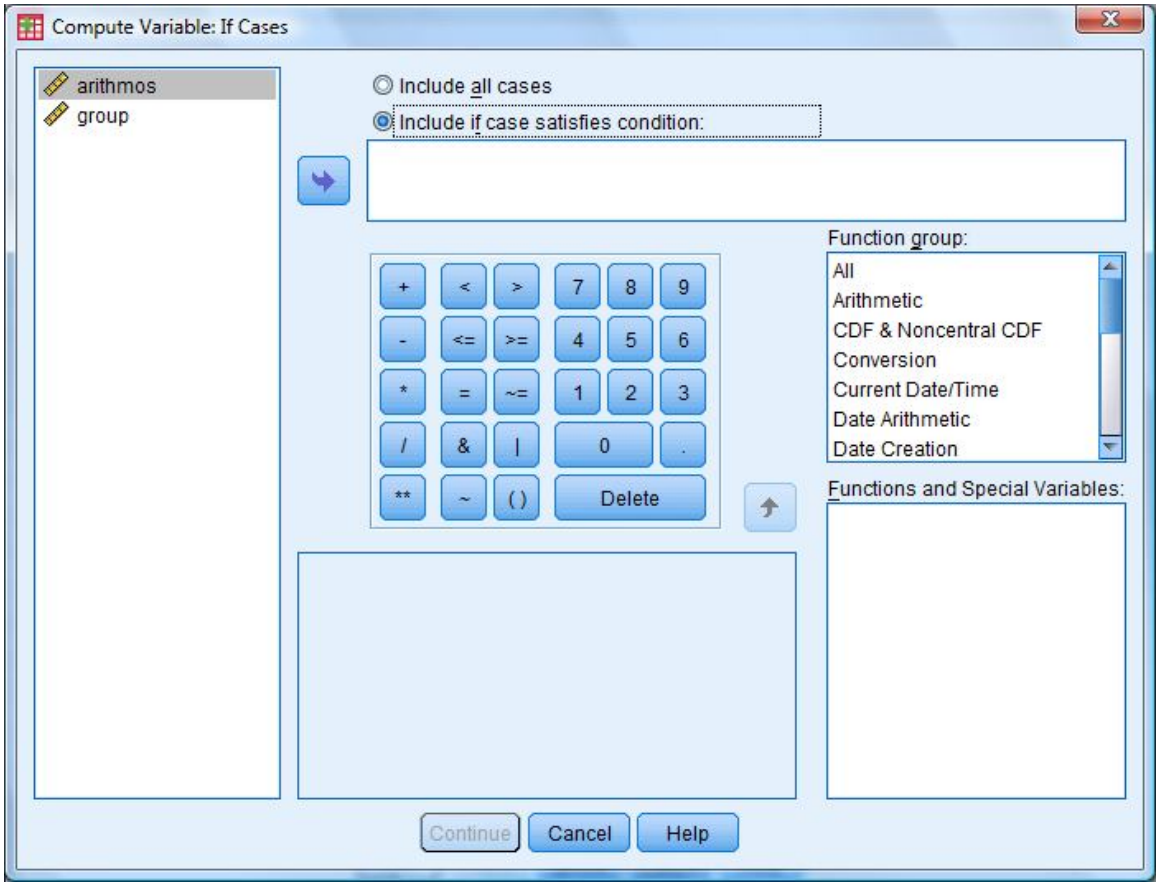


Από το πλαίσιο Type & Label οδηγούμαστε σε ένα παράθυρο διαλόγου όπου μπορούμε να δηλώσουμε το όνομα και τον τύπο της μεταβλητής που δημιουργήσαμε. Ειδικότερα, από το πεδίο Label είτε δίνουμε τη νέα ονομασία είτε χρησιμοποιούμε τη μαθηματική έκφραση ως ονομασία (Use expression as label), ενώ στο πεδίο Type δηλώνουμε αν η νέα μεταβλητή είναι αριθμητική ή αλφαριθμητική (string).



Τέλος, προχωρούμε στην ενεργοποίηση της επιλογής **If…(optional case selection condition)** αν η ύπαρξη τιμών της νέας μεταβλητής εξαρτάται από την ικανοποίηση ή όχι μίας συνθήκης ή έκφρασης μίας άλλης μεταβλητής.

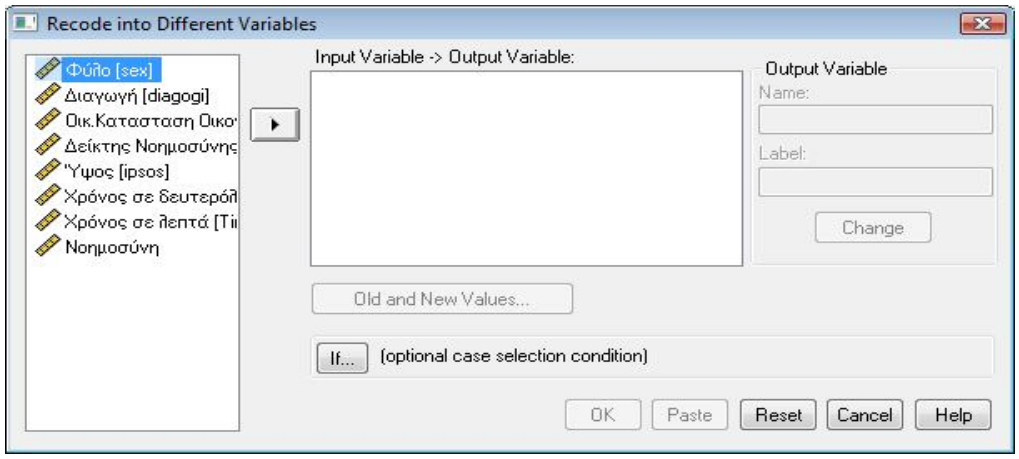
Σχόλιο: Εφόσον θέλουμε να χρησιμοποιηθούν οι πειραματικές μονάδες που ικανοποιούν κάποια συνθήκη επιλέγουμε το πλαίσιο Include if case satisfies condition και στο πλαίσιο που ακολουθεί δηλώνεται, σχηματίζεται η επιθυμητή συνθήκη όπως φαίνεται στην παρακάτω εικόνα. Αφού δηλωθεί η αναγκαία συνθήκη πατάμε **Continue**.



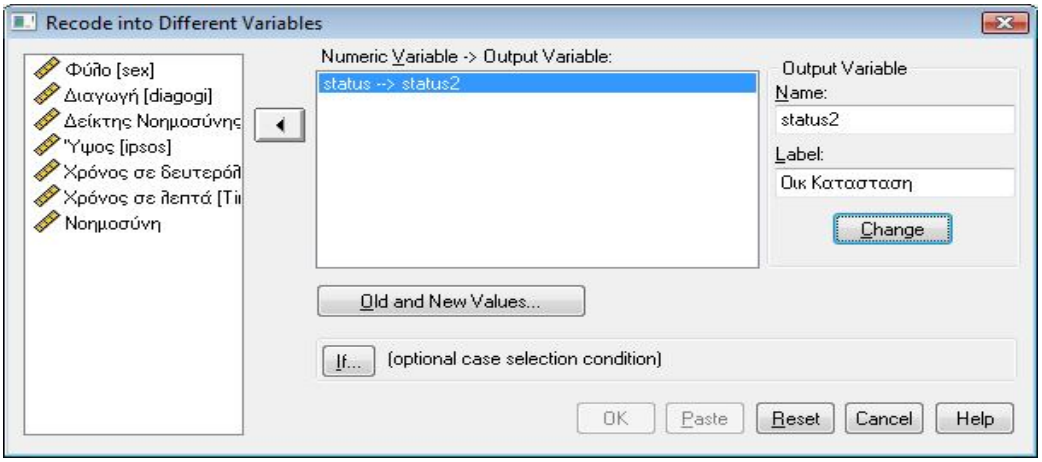
Πατώντας το Continue επανερχόμαστε στο προηγούμενο παράθυρο διαλόγου και εφόσον η νέα μεταβλητή έχει δημιουργηθεί πατάμε το πλαίσιο OK.

Επανακωδικοποίηση Μεταβλητών

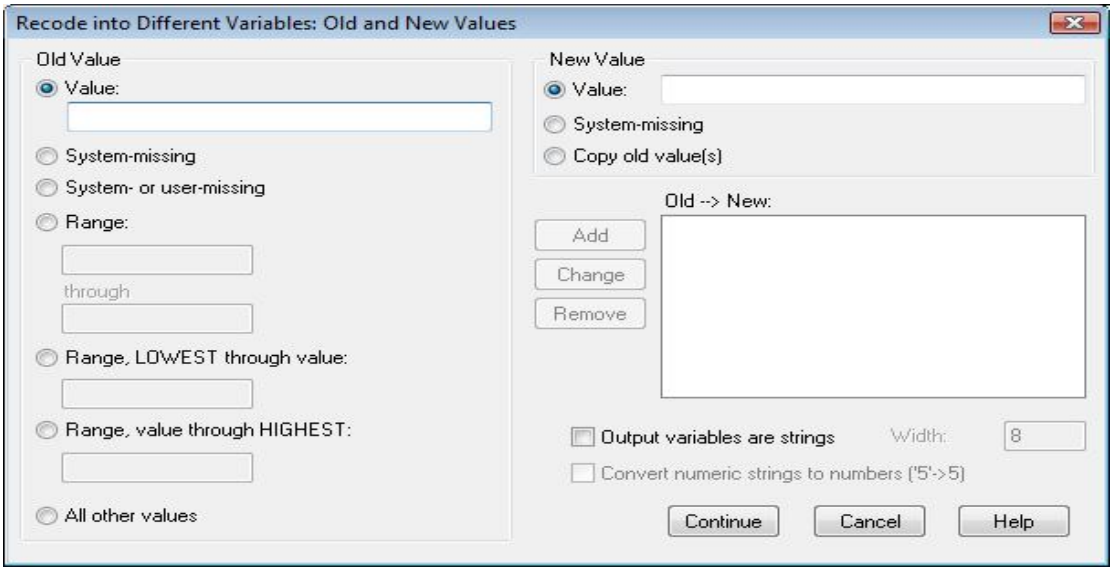
Πολλές φορές εκτός από το μετασχηματισμό των δεδομένων μίας μεταβλητής κρίνεται σκόπιμη η επανακωδικοποίηση της. Αυτό είναι προτιμότερο να γίνεται με τη διαδικασία Transform→Recode Into Different Variables. Τότε ενεργοποιείται ένα νέο παράθυρο διαλόγου, στο οποίο τοποθετούμε στο πλαίσιο Input Variable→Output Variable τη μεταβλητή προς επανακωδικοποίηση καθώς και το όνομα της νέας το οποίο δηλώνεται στο πλαίσιο Output Variable. Μία πλήρης περιγραφή της νέας μεταβλητής δηλώνεται στο πλαίσιο Output Variable Label. Η αλλαγή επιτυγχάνεται πατώντας το πλαίσιο Change και δηλώνοντας την επιθυμητή επανακωδικοποίηση στο παράθυρο διαλόγου που προκύπτει.



Ας δούμε την εν λόγω διαδικασία με χρήση του Παραδείγματος 1.1. Έστω ότι θέλουμε μία νέα κωδικοποίηση για τη μεταβλητή Οικονομική Κατάσταση, σύμφωνα με την οποία όσοι ανήκουν στις κατηγορίες 1 και 2 θα ανήκουν σε μία νέα κατηγορία και όσοι στις 3 και 4 σε μία άλλη. Δηλαδή, όσων παιδιών οι οικογένειες έχουν εισόδημα από 0-600 Ευρώ αποτελούν μία κατηγορία, ενώ τα υπόλοιπα μία άλλη. Ονομάζουμε τη νέα μεταβλητή status2 με ετικέτα Οικ. Κατάσταση.



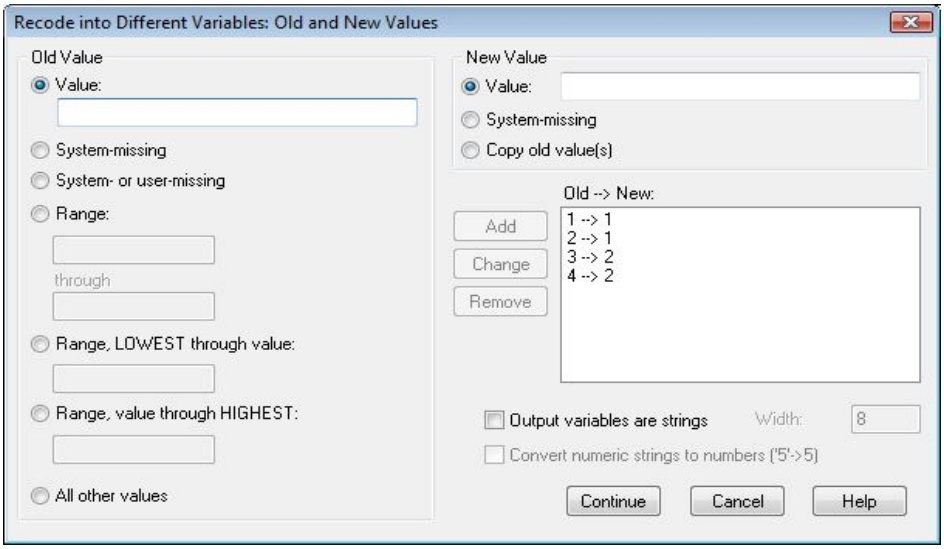
Από την επιλογή Old and New Values οδηγούμαστε σε ένα παράθυρο διαλόγου που μας επιτρέπει την επανακωδικοποίηση μίας ποιοτικής μεταβλητής αλλά και την κωδικοποίηση μίας ποσοτικής σε ποιοτική.



Θα πρέπει να έχουμε όμως υπόψη δύο πολύ βασικούς κανόνες:

- Δεν θα πρέπει να υπάρχουν τομές στις κατηγορίες π.χ. τα διαστήματα [25,50], [50,75] δεν μπορούν να αποτελούν δύο νέες κατηγορίες μίας υπάρχουσας ποσοτικής μεταβλητής.
- Θα πρέπει να υπάρχουν κατάλληλες κατηγορίες για κάθε τιμή των αρχικών δεδομένων π.χ. οι κατηγορίες 25-49, 50-75 δε πρέπει να χρησιμοποιηθούν αν στα αρχικά δεδομένα υπάρχουν τιμές μεταξύ 49 και 50.

Έτσι για το παράδειγμά μας, θα πρέπει να δηλώσουμε ότι οι παλιές τιμές 1 και 2 αντιστοιχούν σε μία νέα κατηγορία με τιμή έστω 1, ενώ οι παλιές 3 και 4 σε μία νέα, με τιμή έστω 2. Έτσι θα πρέπει να προκύψει το ακόλουθο παράθυρο διαλόγου



Πληκτρολογούμε Continue και OK.

<u>Σχόλιο:</u> Στο αριστερό μέρος της παραπάνω εικόνας (πλαίσιο Old Value) παρατηρούμε ότι δίνονται οι ακόλουθες επιλογές:

α) Value: πληκτρολογείτε μία παλιά τιμή και αντιστοιχίζεται είτε σε μία νέα τιμή (δηλώνεται στο πλαίσιο Value του New Value) είτε σε ελλιπή τιμή της νέας μεταβλητής (δηλώνεται στο πλαίσιο System missing του New Value).

β) System missing: δηλώνουμε πως θα επανακωδικοποιηθούν οι ελλιπείς τιμές-κενά κελιά.

γ) System-or user missing: δηλώνουμε πως θα επανακωδικοποιηθούν οι ελλιπείς τιμές τόσο του συστήματος όσο και αυτές που έτσι κατοχυρώθηκαν από το χρήστη.

δ) Range: δηλώνεται πως θα επανακωδικοποιηθεί ένα εύρος, διάστημα τιμών το κάτω και άνω άκρο του οποίου δίνεται στα πλαίσια που ακολουθούν.

ε) Range, lowest through value: δηλώνεται πως θα επανακωδικοποιηθούν οι τιμές από την μικρότερη ως αυτή που δηλώνεται στο πλαίσιο που ακολουθεί.

στ) Range, value through highest: δηλώνεται πως θα επανακωδικοποιηθούν οι τιμές από αυτή που δηλώνεται στο πλαίσιο που ακολουθεί ως τη μεγαλύτερη.

ζ) All other values: δηλώνεται πως θα επανακωδικοποιηθούν όλες οι υπόλοιπες τιμές.

ΚΕΦΑΛΑΙΟ ΔΕΥΤΕΡΟ

Εξερευνώντας τα δεδομένα μας-Περιγραφική Στατιστική

Το πρώτο βήμα στην ανάλυση ενός συνόλου δεδομένων, που αποτελούν μετρήσεις ενός δείγματος είναι η παρουσίαση και σύνοψη των πληροφοριών του δείγματος για τις μεταβλητές που περιλαμβάνονται σε αυτό, χρησιμοποιώντας μεθόδους της Περιγραφικής Στατιστικής. Το S.P.S.S. έχει ενσωματωμένες διαδικασίες για το σκοπό αυτό τόσο για ποιοτικές όσο και για ποσοτικές μεταβλητές. Στις ενότητες που ακολουθούν παραθέτουμε τη διαδικασία για τη συνοπτική παρουσίαση ποιοτικών δεδομένων και έπειτα ποσοτικών δεδομένων.

2.1 Ποιοτικές μεταβλητές

Η συνοπτική παρουσίαση των δεδομένων μίας ποιοτικής μεταβλητής (βλέπε σχετικά Ζωγράφος, 2003, σελ. 18-31, 41-43) επιτυγχάνεται α) με τον πίνακα συχνοτήτων των δεδομένων και β) με τις γραφικές τους παραστάσεις (ραβδόγραμμα, κυκλικό διάγραμμα).

Ο πίνακας συχνοτήτων μιας ποιοτικής μεταβλητής προκύπτει από την απαρίθμηση και καταγραφή των δειγματικών τιμών στην αντίστοιχη κατηγορία. Ένας ολοκληρωμένος πίνακας συχνοτήτων μίας ποιοτικής μεταβλητής περιλαμβάνει τη στήλη των Συχνοτήτων (η συχνότητα παριστάνει τον αριθμό των φορών που μία κατηγορία της ποιοτικής μεταβλητής εμφανίζεται στο δείγμα) και τη στήλη των Σχετικών συχνοτήτων (η σχετική συχνότητα παριστάνει το ποσοστό επί τοις εκατό των φορών εμφάνισης μίας τιμής στο δείγμα). Επιπλέον, μπορούν να συμπεριληφθούν στον πίνακα συχνοτήτων διατάξιμων μόνο ποιοτικών μεταβλητών, η στήλη των Αθροιστικών συχνοτήτων (παριστάνει το πλήθος των τιμών του δείγματος που είναι μικρότερες ή το πολύ ίσες από μία τιμή) και η στήλη των Αθροιστικών σχετικών συχνοτήτων (παριστάνει το ποσοστό επί τοις εκατό των τιμών του δείγματος που είναι μικρότερες ή ίσες από μία τιμή).

Ένας τρόπος άμεσης κατανόησης των χαρακτηριστικών της κατανομής των συχνοτήτων επιτυγχάνεται με μία ειδική γραφική παράσταση που ονομάζεται ραβδόγραμμα. Στον οριζόντιο άξονα ενός ραβδογράμματος συχνοτήτων (εναλλακτικά ενός ραβδογράμματος σχετικών συχνοτήτων) σημειώνονται οι κατηγορίες στις οποίες τα μέλη του πληθυσμού κατατάσσονται, ενώ στον κατακόρυφο άξονα οι αντίστοιχες συχνότητες (εναλλακτικά οι αντίστοιχες σχετικές συχνότητες).

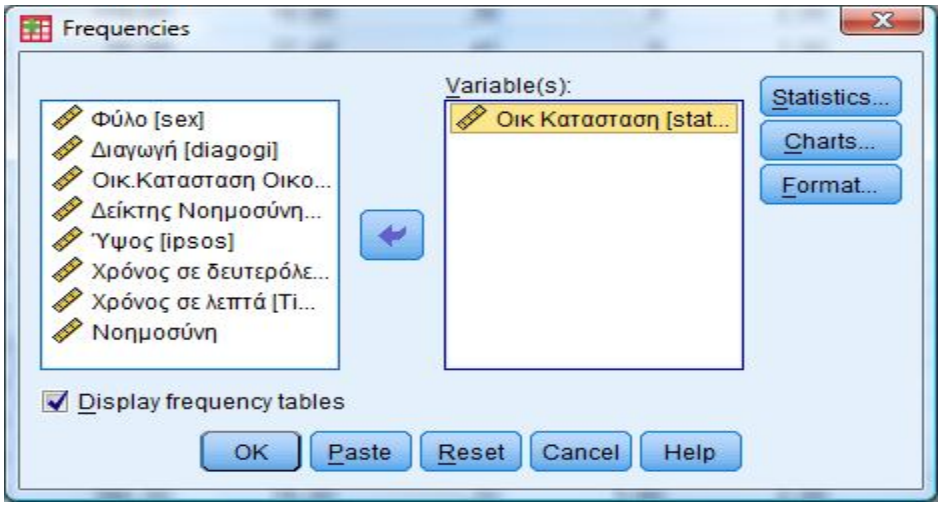
Το κυκλικό διάγραμμα είναι ένας κυκλικός δίσκος χωρισμένος σε τομείς, όσες και οι κατηγορίες στις οποίες τα μέλη του πληθυσμού κατατάσσονται. Το εμβαδό κάθε τομέα απεικονίζει το ποσοστό των ατόμων που ανήκουν στην αντίστοιχη κατηγορία.

Υλοποίηση στο S.P.S.S.

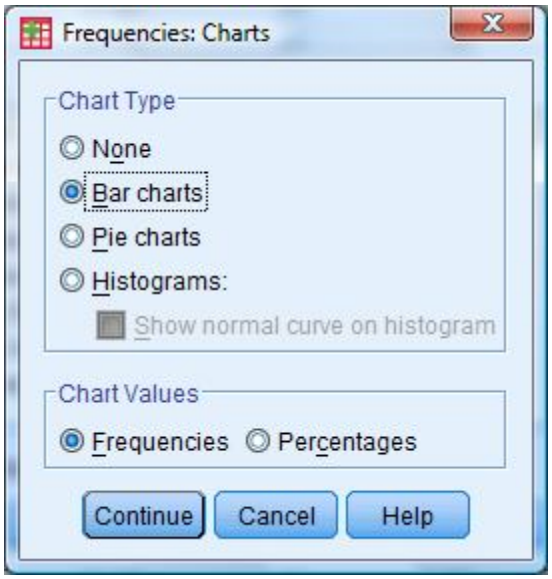
Σε συνέχεια του Παραδείγματος 1.1 θα γίνει ο πίνακας συχνοτήτων, το ραβδόγραμμα και το κυκλικό διάγραμμα της μεταβλητής που περιγράφει την οικονομική κατάσταση της οικογένειας.

Η συνοπτική παρουσίαση των δεδομένων ποιοτικών μεταβλητών γίνεται με την ακόλουθη διαδικασία:

1. Analyze→Descriptive Statistics→Frequencies.
2. Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε τις προς ανάλυση ποιοτικές μεταβλητές και τις μεταφέρουμε στο κουτί Variable(s). Έχοντας επιλέξει μόνο το πλαίσιο Display frequency tables θα παραχθούν μόνο οι πίνακες συχνοτήτων.



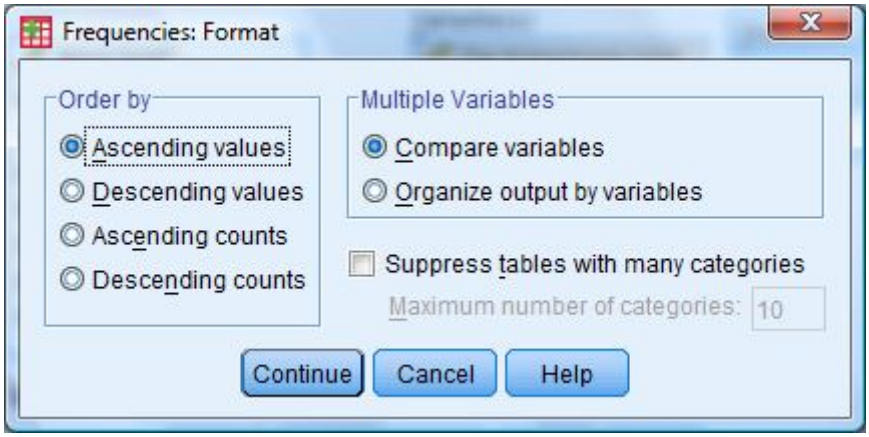
3. Από την επιλογή Charts μπορούμε να κατασκευάσουμε: Ραβδογράμματα (Bar charts), Κυκλικά Διαγράμματα (Pie charts). Τα ιστογράμματα (Histograms), όπως θα δούμε και στην επόμενη ενότητα, αφορούν την περίπτωση ποσοτικών μεταβλητών. Δυστυχώς κάθε φορά έχουμε τη δυνατότητα μίας επιλογής μεταξύ του Bar Charts και Pie Charts. Επιλέγοντας π.χ. την κατασκευή ραβδογράμματος (ή κυκλικού διαγράμματος), ενεργοποιείται η επιλογή Chart Values από όπου επιλέγοντας Frequencies ή Percentages καθορίζουμε αν στον κατακόρυφο άξονα των υπό κατασκευή ραβδογραμμάτων ή κυκλικών διαγραμμάτων θα εμφανίζονται οι απόλυτες συχνότητες (Frequencies) ή οι σχετικές συχνότητες (Percentages), αντίστοιχα.



4. Τέλος, από την επιλογή Format του κεντρικού παραθύρου διαλόγου Frequencies καθορίζουμε αν ο πίνακας συχνοτήτων θα εμφανιστεί είτε σε αύξουσα ή φθίνουσα σειρά εμφάνισης των διαφορετικών κατηγοριών της ποιοτικής μεταβλητής (Order by Ascending or Descending values) είτε σύμφωνα με τη συχνότητα εμφάνισης των διαφορετικών κατηγοριών (Order by Ascending or Descending Counts).

Επιπλέον, αν στο πλαίσιο Variable(s) του κεντρικού παραθύρου διαλόγου Frequencies έχουν δηλωθεί περισσότερες από μία μεταβλητές μπορούμε είτε να αποκτούμε τα αποτελέσματα σε ένα πίνακα για όλες (Compare Variables) είτε να γίνεται η ανάλυση ξεχωριστά για καθεμία (Organize output by variables).

Τέλος, η επιλογή Suppress tables with more than n categories εμποδίζει την εμφάνιση πινάκων με περισσότερες από n κατηγορίες (που είναι ο μέγιστος αριθμός κατηγοριών που δηλώνεται στο πλαίσιο Maximum number of categories).



Ερμηνεία αποτελεσμάτων

Statistics

Οικ.Κατασταση Οικογένειας

N	Valid	35
	Missing	0

Οικ.Κατασταση Οικογένειας

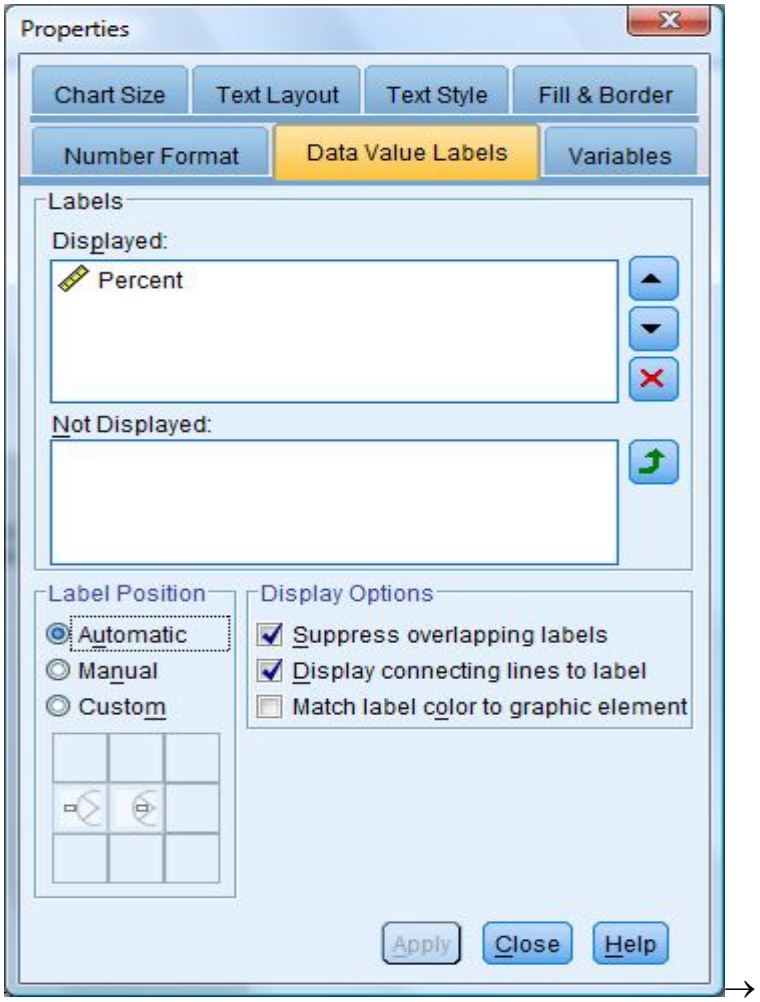
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Α	8	22,9	22,9	22,9
	Β	11	31,4	31,4	54,3
	Γ	10	28,6	28,6	82,9
	Δ	6	17,1	17,1	100,0
	Total	35	100,0	100,0	

Από τον πρώτο πίνακα πληροφορούμαστε ότι οι διαθέσιμες δειγματικές τιμές είναι 35 (Valid=35) και δεν υπάρχουν ελλιπείς τιμές (Missing=0). Ο δεύτερος πίνακας είναι ουσιαστικά ο πίνακας συχνοτήτων για την Οικονομική Κατάσταση της οικογένειας (από εδώ και στο εξής Οικ. Κατάσταση Οικογένειας). Έτσι, από τη στήλη Frequency (Στήλη Συχνοτήτων) προκύπτει ότι 8, 11, 10 και 6 από τις 35 συνολικά οικογένειες είναι οικονομικής κατάστασης Α, Β, Γ, Δ αντίστοιχα. Επιπλέον, από τη στήλη Percent (Στήλη Σχετικών Συχνοτήτων) έχουμε π.χ. ότι 22,9% των ερωτηθέντων ανήκει στην κατηγορία Α. Επισημαίνεται ότι το ποσοστό στη στήλη Percent υπολογίζεται στο σύνολο των ερωτηθέντων συμπεριλαμβανομένου και των πιθανών ελλιπών τιμών. Από την άλλη μεριά το ποσοστό στη στήλη Valid Percent υπολογίζεται στο σύνολο αυτών που έχουν

απαντήσει. Εδώ προφανώς προκύπτει ισότητα καθώς δεν έχουμε ελλιπείς παρατηρήσεις. Τέλος, από τη στήλη Cumulative Percent (Στήλη Αθροιστικών Σχετικών Συχνοτήτων) προκύπτει για παράδειγμα ότι 82,9% των ερωτηθέντων έχουν εισόδημα μικρότερο ή ίσο των 900 Ευρώ (Γ=600-900 Ευρώ). Η στήλη αυτή όπως γνωρίζουμε από τη θεωρία έχει νόημα μόνο για διατάξιμες ποιοτικές μεταβλητές, όπως αυτή του παραδείγματος. Τέλος, έχουμε το παρακάτω ραβδόγραμμα.

Κάνοντας διπλό κλικ στο ραβδόγραμμα ή στο κυκλικό διάγραμμα που προκύπτει στο Output (δηλαδή στο παράθυρο των αποτελεσμάτων) έχουμε τη δυνατότητα περαιτέρω επεξεργασίας του (ως προς το χρώμα, τον τρόπο εμφάνισης, τους τίτλους, τους υπότιτλους κ.ά.).

Ειδικότερα, θέλοντας να εμφανίζονται τα ποσοστά της κάθε κατηγορίας στο κυκλικό διάγραμμα κάνουμε διπλό κλικ σε αυτό και στο νέο παράθυρο επιλέγουμε Elements→Show Data Labels και στο επόμενο παράθυρο διαλόγου κάτω από το πλαίσιο Displayed ζητούμε να εμφανίζεται το Percent.



2.2 Ποσοτικές μεταβλητές

Η συνοπτική παρουσίαση των δεδομένων ποσοτικών μεταβλητών περιλαμβάνει τον υπολογισμό των τιμών διάφορων στατιστικών μέτρων, όπως η μέση τιμή (Mean), η τυπική απόκλιση (Std Deviation), οι συντελεστές κύρτωσης και λοξότητας (Kurtosis, Skewness, αντίστοιχα), η διάμεσος (median), η επικρατούσα τιμή (mode), το εύρος (range), τα ποσοστιαία σημεία (Percentile values) κ.ά. Το δεύτερο στάδιο περιλαμβάνει την πιθανή κατασκευή του ιστογράμματος (histogram) και θηκογράμματος (boxplot) της υπό εξέτασης ποσοτικής μεταβλητής, τον έλεγχο ύπαρξης ακραίων τιμών στις δειγματικές τιμές της υπό εξέταση μεταβλητής, καθώς και τον έλεγχο αν οι διαθέσιμες δειγματικές τιμές μπορούν να θεωρηθούν ότι προέρχονται από έναν πληθυσμό που

περιγράφεται ικανοποιητικά από την κανονική κατανομή (βλέπε σχετικά Ζωγράφος, 2003, σελ. 45-55).

Περιγραφικά μέτρα

Μια συνοπτική παρουσίαση των δεδομένων ποσοτικών μεταβλητών επιτυγχάνεται με τα περιγραφικά μέτρα, που διακρίνονται σε μέτρα θέσης και μέτρα διασποράς.

Ένα μέτρο θέσης είναι μία αριθμητική τιμή ενδεικτική της θέσης, του σημείου γύρω από το οποίο ένα σύνολο δεδομένων συγκεντρώνεται. Τέτοια είναι η μέση τιμή $\bar{X}$ (μέσος όρος των μετρήσεων), η διάμεσος (η τιμή εκείνη που χωρίζει τα δεδομένα σε δύο ίσα μέρη έτσι ώστε το πλήθος των μετρήσεων που βρίσκονται αριστερά της να είναι ίσο με το πλήθος των μετρήσεων που βρίσκεται δεξιά της) και η επικρατούσα τιμή ή κορυφή (η τιμή με τη μεγαλύτερη συχνότητα).

Ένα μέτρο διασποράς είναι μία αριθμητική τιμή ενδεικτική του τρόπου με τον οποίο τα δεδομένα κατανέμονται γύρω από τη μέση τιμή. Τέτοια μέτρα είναι το εύρος (παριστάνει τη διαφορά της ελάχιστης από τη μέγιστη τιμή), η διακύμανση S^2 (εκφράζει τη μεταβλητότητα ενός συνόλου αριθμητικών δεδομένων από τη μέση τους τιμή), η τυπική απόκλιση S (η θετική τετραγωνική ρίζα της διακύμανσης).

Άλλα περιγραφικά μέτρα, μεταξύ άλλων, είναι ο συντελεστής λοξότητας και κύρτωσης αντίστοιχα, που μετρούν την ασυμμετρία της κατανομής και την «αιχμηρότητα» της, αντίστοιχα. Ορίζονται από τις σχέσεις $b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{nS^3}$ και $b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{nS^4}$, αντίστοιχα, όπου $X_1, \ldots, X_n$ είναι οι διαθέσιμες δειγματικές τιμές (n

το μέγεθος του δείγματος), $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ η δειγματική μέση τιμή και $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ η δειγματική διακύμανση.

Με βάση την λοξότητα οι κατανομές διακρίνονται σε συμμετρικές όταν $b_1 = 0$ (σε αυτές ανήκει η κανονική κατανομή), σε θετικά ασύμμετρες (ή λοξές δεξιά) όταν $b_1 > 0$ και σε αρνητικά ασύμμετρες (ή λοξές αριστερά) όταν $b_1 < 0$.

Με βάση την κύρτωση οι κατανομές διακρίνονται σε λεπτόκυρτες όταν $b_2 > 3$, σε μεσόκυρτες όταν $b_2 = 3$ (σε αυτές ανήκει η κανονική κατανομή) και σε πλατύκυρτες όταν $b_2 < 3$. Το λογισμικό του S.P.S.S. υπολογίζει την τιμή $b_2 - 3$, έτσι ώστε η σύγκριση και η εξέταση για ενδείξεις αποκλίσεων από την κανονικότητα να γίνεται με το μηδέν.

Άλλα περιγραφικά μέτρα είναι:

α) τα εκατοστιαία ποσοστιαία σημεία. Το p-οστό εκατοστιαίο σημείο έχει την ιδιότητα p% των μετρήσεων να είναι μικρότερες ή ίσες από αυτό, και τέλος,

β) τα τεταρτημόρια που έχουν την ιδιότητα να χωρίζουν το σύνολο των μετρήσεων σε τέσσερα ίσα μέρη και δεν είναι τίποτε άλλο από το 25^o, 50^o, 75^o ποσοστιαίο σημείο.

Ιστόγραμμα συχνοτήτων

Πολλές φορές οι τιμές μιας ποσοτικής μεταβλητής είναι πολυάριθμες και για τη συνοπτική παρουσίασή τους κρίνεται σκόπιμη η ομαδοποίησή τους. Οι ομάδες έχουν τη μορφή κλειστών συνεχόμενων διαστημάτων. Το ιστόγραμμα συχνοτήτων συνίσταται από ένα σύνολο συγγενών ορθογώνιων παραλληλόγραμμων, των οποίων το ύψος είναι ανάλογο με τη συχνότητα κάθε ομάδας και το μήκος τους ανάλογο με το μήκος της ομάδας. Οι τιμές της μεταβλητής (ουσιαστικά τα άκρα των ομάδων) τοποθετούνται στον οριζόντιο άξονα, ενώ οι συχνότητες στον κατακόρυφο άξονα.

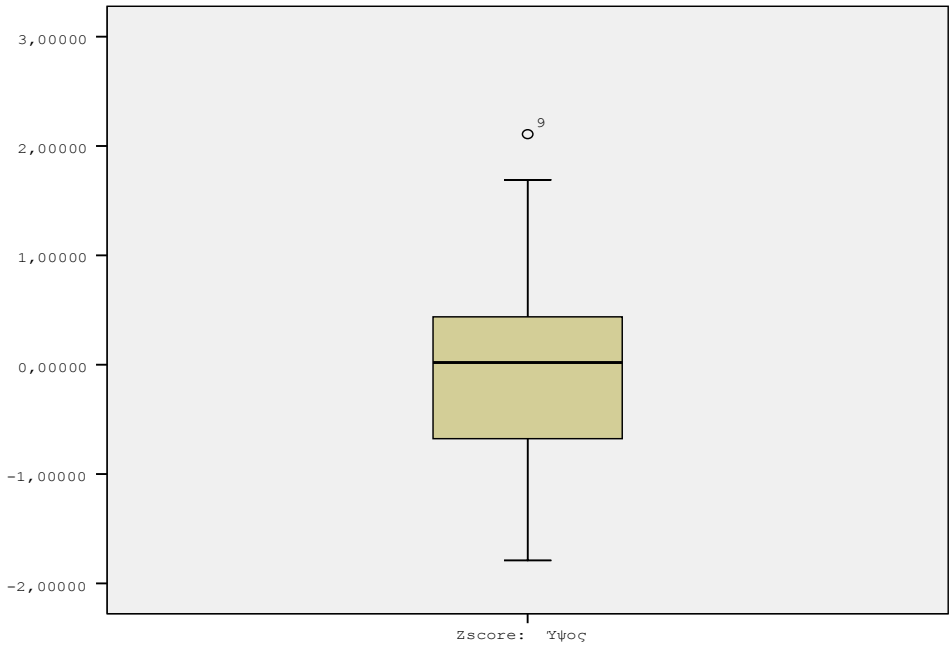
Φυλλογράφημα

Μία παραλλαγή του ιστογράμματος είναι το φυλλογράφημα (steam and leaf plot). Το φυλλογράφημα δεν είναι τίποτε άλλο παρά το αποτέλεσμα της περιστροφής κατά 90^o

του ιστογράμματος συχνοτήτων. Επομένως γίνεται αντιληπτό ότι το φυλλογράφημα επιδεικνύεται προς μία μεριά (δεξιά). Το μήκος κάθε γραμμής αντιστοιχεί στον αριθμό των παρατηρήσεων που ανήκουν στο διάστημα. Η κύρια διαφοροποίηση του φυλλογραφήματος σε σχέση με το ιστόγραμμα συχνοτήτων είναι ότι αναπαρίσταται κάθε τιμή με μία αληθινή τιμή.

Θηκόγραμμα

Στο θηκόγραμμα παριστάνονται περιγραφικά μέτρα όπως η διάμεσος, το 25^{ο} και 75^{ο} ποσοστιαίο σημείο και οι ακραίες τιμές («αντιφατικές» τιμές σε σχέση με τις υπόλοιπες παρατηρούμενες τιμές του συνόλου δεδομένων).



Το κάτω άκρο του κουτιού είναι το 25^{ο} ποσοστιαίο σημείο και το πάνω άκρο το 75^{ο}. Η διάμεσος παριστάνεται από μία οριζόντια γραμμή μέσα στο κουτί. Στην αρχή και στην κορυφή του σχήματος σημειώνονται δύο οριζόντιες γραμμές, που αναφέρονται ως φράχτες (whiskers). Το θηκόγραμμα μας βοηθά στο να δούμε αν υπάρχουν ακραίες τιμές (τιμές πέρα από τους whiskers, επισημαίνονται με «ο» και είναι ακραίες, ενώ με * επισημαίνονται οι extreme) καθώς και πιθανές αποκλίσεις από την κανονική κατανομή (αν η διάμεσος είναι πιο κοντά στην κορυφή ή στην αρχή του κουτιού και όχι στο

κέντρο). Ο κάτω και άνω φράκτης καθορίζονται, υπό κάποιες προϋποθέσεις, από τις σχέσεις:

$1^{ο}$ τεταρτημόριο – 1.5 * ενδοτεταρτημοριακό εύρος

και

$3^{ο}$ τεταρτημόριο +1.5 * ενδοτεταρτημοριακό εύρος

αντίστοιχα, όπου το ενδοτεταρτημοριακό εύρος είναι η διαφορά του $3^{ου}$ από το $1^{ο}$ τεταρτημόριο.

Παρατήρηση: Ο όρος ακραία τιμή αναφέρεται σε μία παρατήρηση η οποία κατά μία έννοια είναι «αντιφατική» σε σχέση με τις υπόλοιπες παρατηρούμενες τιμές του συνόλου δεδομένων. Οι ακραίες τιμές αρχικά θα πρέπει να επισημαίνονται και αφού διαπιστωθεί ότι δεν πρόκειται για λάθη κατά την πληκτρολόγηση των δεδομένων να μελετώνται. Δε συνιστάται ο αυτόματος αποκλεισμός τους από την έρευνα χωρίς καμία διάκριση, καθώς πολλές φορές και οι ακραίες τιμές περικλείουν εξίσου σημαντικές πληροφορίες. Επισημαίνεται ότι κάθε φορά αποφασίζουμε για την ύπαρξη μίας ακραίας τιμής και αφού την αποκλείσουμε προβαίνουμε σε έλεγχο ύπαρξης επιπρόσθετης ακραίας τιμής. Η μεθοδολογία αυτή θα αναπτυχθεί διεξοδικά σε επόμενη ενότητα καθώς και στο Κεφάλαιο 4.

Έλεγχος κανονικότητας

Η υπόθεση της κανονικότητας είναι μία από τις υποθέσεις πάνω στις οποίες έχει θεμελιωθεί η στατιστική συμπερασματολογία. Οι περισσότερες από τις μεθοδολογίες της Παραμετρικής Στατιστικής υποθέτουν, προϋποθέτουν ότι τα δεδομένα προέρχονται από έναν πληθυσμό, ο οποίος περιγράφεται ικανοποιητικά από την κανονική κατανομή. Για το λόγο αυτό πολλοί τρόποι ελέγχου έχουν εμφανιστεί στη βιβλιογραφία για την υπόθεση της κανονικότητας, τόσο στατιστικοί όσο και γραφικοί. Από τους στατιστικούς τρόπους ελέγχου ξεχωρίζει το στατιστικό τεστ που προτάθηκε από τους Shapiro-Wilk και οι επεκτάσεις αυτού. Η βασική γραφική μέθοδος για τον έλεγχο της κανονικότητας είναι το Q-Q (quantile-quantile) γράφημα, το οποίο συγκρίνει τα ποσοστιαία σημεία (quantile) του δείγματος έναντι των πληθυσμιακών ποσοστιαίων σημείων της κανονικής

κατανομής. Αν τα σημεία είναι κοντά σε ευθεία γραμμή δεν υπάρχει ένδειξη για απόκλιση από την κανονικότητα. Παρεκκλίσεις από την ευθεία γραμμή δηλώνουν μη κανονικότητα. Ο τύπος της μη γραμμικότητας μπορεί να υποδηλώνει το τρόπο απόκλισης από την κανονικότητα. Το S.P.S.S. για κάθε Q-Q γράφημα που κατασκευάζει μας δίνει και μία γραφική παράσταση που ονομάζεται Detrended Q-Q Plot. Η γραφική αυτή μέθοδος δείχνει τις ατομικές αποκλίσεις μεταξύ παρατηρούμενων και εκτιμώμενων αθροιστικών τιμών (ή εκατοστημορίων). Τα σημεία αυτά κατανέμονται γύρω από μία οριζόντια γραμμή που αντιστοιχεί στο 0.

<u>Παρατήρηση 1</u>: Στα στατιστικά πακέτα η απόφαση για την αποδοχή ή απόρριψη μιας στατιστικής υπόθεσης δεν γίνεται εξετάζοντας αν η τιμή του στατιστικού ανήκει στην **περιοχή απόρριψης** (γνωστή και ως **κρίσιμη περιοχή**), αλλά στη βάση των p-τιμών (p-value ή Sig.) Η **p-τιμή** ενός στατιστικού τεστ είναι η μικρότερη τιμή του επιπέδου σημαντικότητας για την οποία απορρίπτεται η μηδενική υπόθεση. Εύκολα προκύπτει τότε ότι **απορρίπτουμε την προς έλεγχο μηδενική υπόθεση αν η p-τιμή είναι μικρότερη από το προκαθορισμένο επίπεδο σημαντικότητας (συνήθως το 0.05).**

<u>Παρατήρηση 2</u>: Έστω $Y_1,\ldots,Y_n$ είναι οι n το πλήθος διαθέσιμες δειγματικές τιμές της υπό μελέτη μεταβλητής, οι οποίες αποκλίνουν από την κανονικότητα. Ο μετασχηματισμός Box-Cox (βλέπε Box and Cox (1964)) δίνεται από τη σχέση

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda}-1}{\lambda\left(\dot{Y}\right)^{\lambda-1}}, & \lambda \neq 0 \\ \left(\dot{Y}\right)\ln\left(Y_i\right),.....\lambda = 0 \end{cases}$$

όπου $\dot{Y} = \left(Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n\right)^{1/n}$ και υποθέτει ότι για κάποια τιμή της παραμέτρου λ τα μετασχηματισμένα δεδομένα ικανοποιούν την υπόθεση της κανονικότητας.

Υλοποίηση στο S.P.S.S.

Για την υλοποίηση των παραπάνω μπορούν να χρησιμοποιηθούν δύο διαδικασίες του λογισμικού, οι διαδικασίες Descriptives και Frequencies, η καθεμία εκ των οποίων μας δίνει διαφορετικές δυνατότητες και επιλογές. Αν έχουμε όμως ως στόχο την πιο αναλυτική παρουσίαση των δεδομένων μας χρησιμοποιούμε μία πιο σύνθετη διαδικασία, τη διαδικασία Explore. Στη συνέχεια θα παραθέσουμε τον τρόπο υλοποίησης των παραπάνω με τη διαδικασία Explore και απλώς θα αναφέρουμε τις επιπλέον δυνατότητες που δίνουν οι άλλες δύο διαδικασίες.

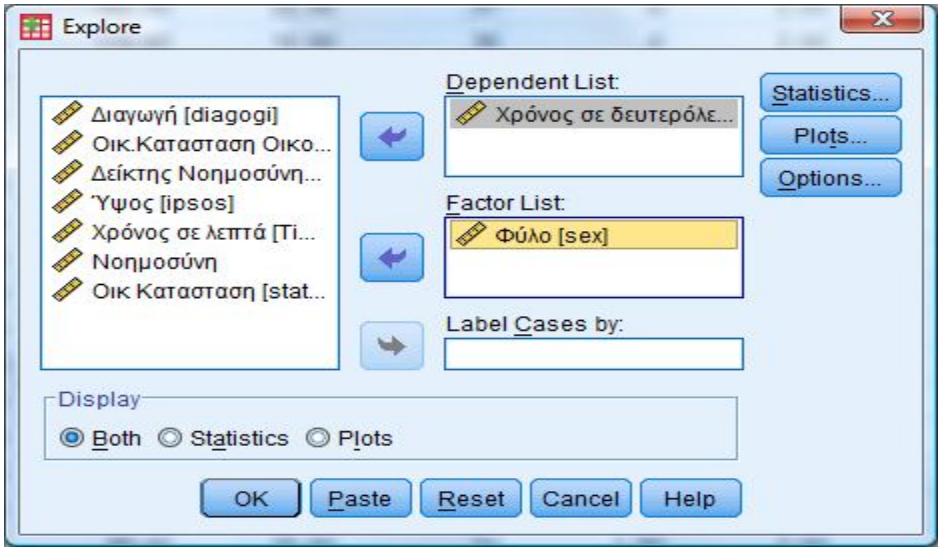
Διαδικασία Explore

Η διαδικασία Explore μπορεί να χρησιμοποιηθεί για την απόκτηση πλήθους στατιστικών μέτρων καθώς και γραφικών παραστάσεων τόσο για το σύνολο των δεδομένων όσο και ξεχωριστά για κατηγορίες αυτών.

Χωρίς βλάβη της γενικότητας στη συνέχεια περιγράφεται η μεθοδολογία που ακολουθείται αν το ενδιαφέρον επικεντρώνεται στη συνοπτική παρουσίαση και αρχική μελέτη του χρόνου σε δευτερόλεπτα που διανύει ένα παιδί τα 100 μέτρα ως προς το φύλο του.

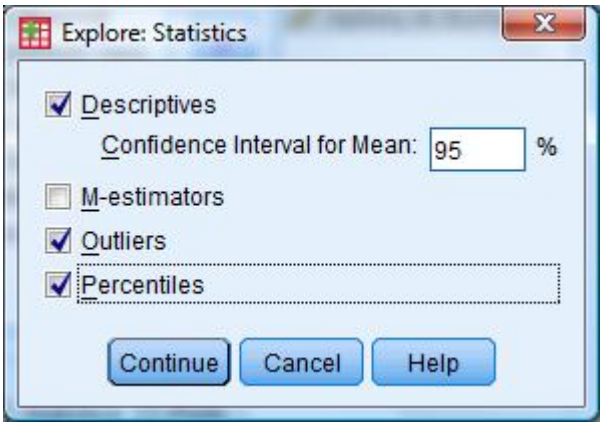
Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

1. Analyze →Descriptive Statistics→ Explore.
2. Στο παράθυρο διαλόγου που προκύπτει τοποθετούμε στο πλαίσιο Dependent τις ποσοτικές (υποχρεωτικά και μόνο) μεταβλητές που θέλουμε να αναλύσουμε. Στο πλαίσιο Factor list τοποθετούμε τις πιθανές ποιοτικές-κατηγορικές μεταβλητές (και μόνο) ως προς τις κατηγορίες των οποίων θέλουμε να προχωρήσουμε την ανάλυση μας, π.χ. έτος σπουδών, φύλο κ.ο.κ.



Επιπρόσθετα διατηρώντας την προεπιλογή Display Both (στο κάτω αριστερό άκρο του παραθύρου) έχουμε τη δυνατότητα απόκτησης τόσο στατιστικών μέτρων όσο και γραφημάτων. Το πλαίσιο Label Cases By το αφήνουμε ως έχει κενό, έτσι ώστε το S.P.S.S να χρησιμοποιήσει την προεπιλογή του αύξοντα αριθμού παρατήρησης.

3. Από την επιλογή Statistics επιλέγουμε τα ακόλουθα



Descriptives (προεπιλογή): απόκτηση των κυριότερων περιγραφικών μέτρων, όπως η διάμεσος, η μέση τιμή, η τυπική απόκλιση κ.ά. καθώς και ενός π.χ. 95% διαστήματος εμπιστοσύνης για την πληθυσμιακή μέση τιμή του υπό μελέτη χαρακτηριστικού (που έχει δηλωθεί στο πλαίσιο Dependent List). Το διάστημα αυτό υπολογίζεται υπό την υπόθεση της κανονικότητας. Επομένως χρειάζεται προσοχή στην περίπτωση αποκλίσεων από την κανονικότητα.

Outliers: το λογισμικό θα μας δώσει τις πέντε μικρότερες και πέντε μεγαλύτερες τιμές κάθε μεταβλητής που έχει δηλωθεί στο πλαίσιο Dependent List, ως προς τις κατηγορίες της μεταβλητής που έχει δηλωθεί στο πλαίσιο Factor List.

Percentiles: υπολογίζει το λογισμικό το 5^{o} –95^{o} ποσοστιαίο σημείο.

4. Από την επιλογή Plots έχουμε τη δυνατότητα για τα ακόλουθα:

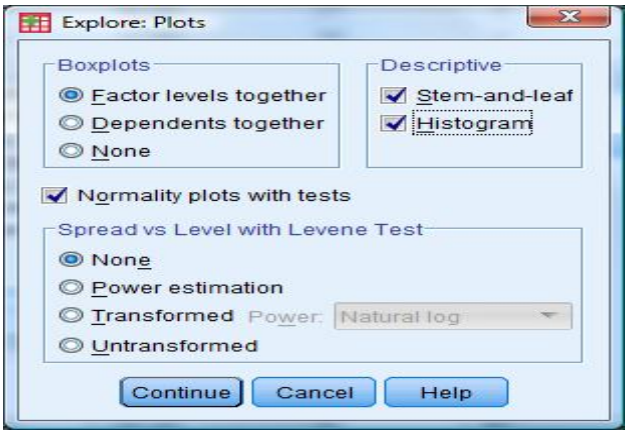
Boxplots: αποκτούμε τα θηκογράμματα. Σημειώνουμε επιπρόσθετα ότι η επιλογή *Factor levels together* δημιουργεί ένα ξεχωριστό πλαίσιο θηκογραμμάτων για καθεμία μεταβλητή που έχουμε δηλώσει στο πλαίσιο Dependent Variables, ως προς κάθε κατηγορία της ποιοτικής μεταβλητής που έχουμε δηλώσει στο πλαίσιο Factor List, ενώ η επιλογή *Dependents together* δημιουργεί ένα ξεχωριστό πλαίσιο θηκογραμμάτων για καθεμία κατηγορία της ποιοτικής μεταβλητής που έχουμε δηλώσει στο πλαίσιο Factor list ως προς καθεμία από τις ποσοτικές μεταβλητές που έχουν δηλωθεί στο πλαίσιο Dependent Variable. Είναι προτιμότερο να επιλέγουμε το Factor levels together.

Descriptive: έχουμε διαθέσιμες τις επιλογές *Steam-and-Leaf* και *Histogram*, από όπου δηλαδή μπορούμε να αποκτήσουμε το φυλλογράφημα και το ιστόγραμμα για τις ποσοτικές μεταβλητές.

Με την επιλογή Normality plots with tests αποκτούμε τόσο γραφικούς τρόπους ελέγχου της κανονικότητας (normal probability και detrended normal probability plots) όσο και στατιστικά τεστ ελέγχου (το Kolmogorov-Smirnov στατιστικό, με τη διόρθωση του Lilliefors καθώς και το Shapiro-Wilk στατιστικό τεστ, το οποίο και είναι προτιμότερο να εμπιστευόμαστε).

Spread vs. Level with Levene Test: μας δίνει τρόπο να ελέγξουμε την υπόθεση ότι η εξαρτημένη μεταβλητή (έχει δηλωθεί στο πλαίσιο Dependent List) έχει την ίδια διακύμανση μέσα σε δύο ή περισσότερους πληθυσμούς (που προκύπτουν από τις κατηγορίες της μεταβλητής που υπεισέρχονται στο πεδίο Factor List). Η παραπάνω υπόθεση της ισότητας των διακυμάνσεων ή ομοσκεδαστικότητας, όπως θα αναφερθεί σε επόμενα κεφάλαια, είναι αρκετά σημαντική για την εφαρμογή κάποιων μεθοδολογιών. Ο έλεγχος αυτής της υπόθεσης επιτυγχάνεται με το στατιστικό τεστ του *Levene* και επιλέγοντας το *Untransformed data*. Αν η ισότητα απορριφθεί, επαναλαμβάνοντας τα παραπάνω βήματα, επιλέγοντας το πλαίσιο *Power Estimation* το λογισμικό μας προσδιορίζει τον καλύτερο μετασχηματισμό. Έπειτα χρησιμοποιώντας από το πλαίσιο

Transformed τον μετασχηματισμό που μας έχει υποδειχθεί πρωτύτερα θα πάρουμε το καινούριο γράφημα και θα πραγματοποιηθεί ο έλεγχος της ομοσκεδαστικότητας για τα μετασχηματισμένα δεδομένα.



5. Από την επιλογή Options καθορίζουμε τον τρόπο χειρισμού των ελλιπών τιμών. Για τους τρόπους χειρισμού των ελλιπών δεδομένων έχει γραφεί πληθώρα ερευνητικών εργασιών και συγγραμμάτων. Στα πλαίσια του μαθήματός μας θα αναφέρουμε ότι θα διατηρούμε την (προ)επιλογή *Exclude cases listwise.* Η τεχνική αυτή χειρισμού των ελλιπών δεδομένων περιορίζει την ανάλυση σε εκείνες τις πειραματικές μονάδες (γραμμές) όπου είναι διαθέσιμες οι παρατηρούμενες τιμές σε όλες τις υπό μελέτη μεταβλητές (στήλες).

Ερμηνεία αποτελεσμάτων

Case Processing Summary

Φύλο		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Χρόνος σε δευτερόλεπτα	Αγόρι	19	100,0%	0	,0%	19	100,0%
	Κορίτσι	16	100,0%	0	,0%	16	100,0%

Ο πίνακας αυτός μας πληροφορεί ότι από τους 35 συμμετέχοντες 19 ήταν αγόρια και 16 κορίτσια χωρίς να υπάρχουν ελλιπείς τιμές.

Στον πίνακα Descriptives μας δίνονται διάφορα περιγραφικά μέτρα (και όχι μόνο) για τη μεταβλητή που περιγράφει το χρόνο σε δευτερόλεπτα που διένυσαν τα 100 μέτρα. Χρήζουν ιδιαίτερης προσοχής και σχολιασμού τα ακόλουθα.

Descriptives

	Φύλο			Statistic	Std. Error
Χρόνος σε δευτερόλεπτα	Αγόρι	Mean		24,8947	,94867
		95% Confidence Interval for Mean	Lower Bound	22,9017	
			Upper Bound	26,8878	
		5% Trimmed Mean		24,8830	
		Median		25,0000	
		Variance		17,099	
		Std. Deviation		4,13514	
		Minimum		18,00	
		Maximum		32,00	
		Range		14,00	
		Interquartile Range		6,00	
		Skewness		-,106	,524
		Kurtosis		-,943	1,014
	Κορίτσι	Mean		22,3125	,88844
		95% Confidence Interval for Mean	Lower Bound	20,4188	
			Upper Bound	24,2062	
		5% Trimmed Mean		22,0694	
		Median		21,5000	
		Variance		12,629	
		Std. Deviation		3,55375	
		Minimum		18,00	
		Maximum		31,00	
		Range		13,00	
		Interquartile Range		5,75	
		Skewness		,959	,564
		Kurtosis		,765	1,091

Η μέση τιμή (Mean) του χρόνου στα αγόρια είναι μεγαλύτερη από ότι στα κορίτσια (24,8947 έναντι 22,315). Το λογισμικό μας δίνει το 95% διάστημα εμπιστοσύνης (95% Confidence Interval for Mean, Lower and Upper Bound) το οποίο είναι αξιόπιστο με την προϋπόθεση ότι δεν υπάρχουν ακραίες τιμές και τα δεδομένα του χρόνου σε δευτερόλεπτα για κάθε ένα από τους δύο πληθυσμούς (αγόρια και κορίτσια) προέρχονται από πληθυσμούς που περιγράφονται από την κανονική κατανομή.

Από τους συντελεστές λοξότητας και κύρτωσης δεν μπορούμε να αποφανθούμε για το αν τα δεδομένα προέρχονται από κανονική κατανομή, καθώς οι τιμές αυτές δεν αποκλίνουν πολύ από το μηδέν. Επομένως απαιτούνται περισσότεροι γραφικοί και κυρίως στατιστικοί τρόποι ελέγχου της υπόθεσης της κανονικότητας.

Παρατηρούμε ότι ο μέσος χρόνος σε δευτερόλεπτα τόσο των αγοριών όσο και των κοριτσιών είναι περίπου ίσος με τη διάμεσο (median) του χρόνου, επομένως τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από συμμετρικό πληθυσμό.

Επιπλέον, στον πίνακα Percentiles εμφανίζονται τα ποσοστιαία σημεία, ενώ στη στήλη Extreme Values οι χρόνοι των 5 πιο αργών και πιο γρήγορων στα 100 μέτρα αγοριών και κοριτσιών.

Percentiles

		Φύλο	Percentiles						
			5	10	25	50	75	90	95
Weighted Average (Definition 1)	Χρόνος σε δευτερόλεπτα	Αγόρι	18	18	22	25	28	30,	.
		Κορίτσι	18	18,7	19	21,5	24,75	28,2	.
Tukey's Hinges	Χρόνος σε δευτερόλεπτα	Αγόρι			22	25,	28		
		Κορίτσι			19	21,5	24,5		

Extreme Values

	Φύλο			Case Number	Value
Χρόνος σε δευτερόλεπτα	Αγόρι	Highest	1	28	32,00
			2	11	30,00
			3	34	30,00
			4	21	29,00
			5	20	28,00(a)
		Lowest	1	32	18,00
			2	9	18,00
			3	6	20,00
			4	5	21,00
			5	8	22,00(b)
	Κορίτσι	Highest	1	25	31,00
			2	16	27,00
			3	2	25,00
			4	33	25,00
			5	13	24,00(c)
		Lowest	1	3	18,00
			2	26	19,00
			3	22	19,00
			4	15	19,00
			5	4	19,00

a Only a partial list of cases with the value 28,00 are shown in the table of upper extremes.
b Only a partial list of cases with the value 22,00 are shown in the table of lower extremes.
c Only a partial list of cases with the value 24,00 are shown in the table of upper extremes.

Στον πίνακα Tests of Normality αποφασίζουμε για την αποδοχή ή όχι της υπόθεσης της κανονικότητας με βάση τις p-τιμές του ελέγχου που δίνονται στη στήλη Sig. Έτσι έχοντας ως επίπεδο σημαντικότητας α =5% προκύπτει ότι δεν απορρίπτουμε την υπόθεση ότι τα δεδομένα του χρόνου για κάθε έναν από τους δύο πληθυσμούς (αγόρια και κορίτσια) προέρχονται από πληθυσμούς που περιγράφονται ικανοποιητικά από την κανονική κατανομή (p-τιμή του Shapiro-Wilk=0.694 και 0.126 μεγαλύτερες του 0.05, αντίστοιχα για αγόρια και κορίτσια). Στο ίδιο συμπέρασμα καταλήγουμε χρησιμοποιώντας και τους γραφικούς τρόπους ελέγχους (Normal Q-Q Plot και Detrended Normal Q-Q Plot).

Tests of Normality

	Φύλο	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Χρόνος σε δευτερόλεπτα	Αγόρι	,116	19	,200(*)	,966	19	,694
	Κορίτσι	,144	16	,200(*)	,912	16	,126

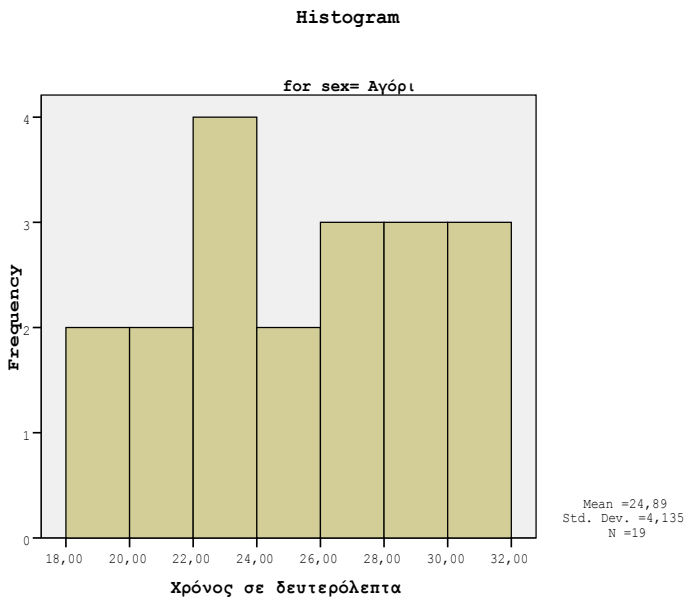
* This is a lower bound of the true significance. a Lilliefors Significance Correction

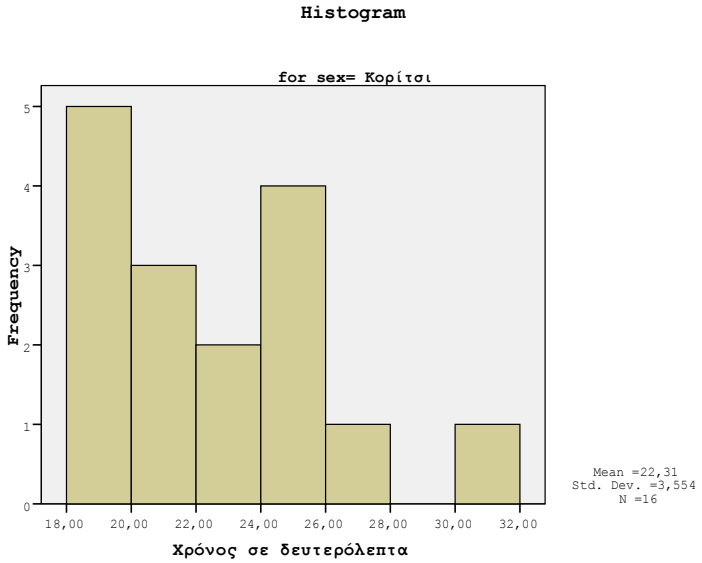
Στο πλαίσιο Test of Homogeneity of Variance δίνεται το στατιστικό τεστ του Levene για τον έλεγχο της υπόθεσης των ίσων διακυμάνσεων. Προκύπτει ότι δεν απορρίπτεται η υπόθεση των ίσων πληθυσμιακών διακυμάνσεων καθώς η p-τιμή του ελέγχου είναι ίση με 0.371>0.05.

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Χρόνος σε δευτερόλεπτα	Based on Mean	,823	1	33	,371
	Based on Median	,823	1	33	,371
	Based on Median and with adjusted df	,823	1	32,896	,371
	Based on trimmed mean	,883	1	33	,354

Επιπλέον έχουμε το ιστόγραμμα και το φυλλογράφημα της μεταβλητής Χρόνος σε δευτερόλεπτα ως προς το φύλο.





Χρόνος σε δευτερόλεπτα Stem-and-Leaf Plot for
sex= Αγόρι

Frequency	Stem &	Leaf
2,00	1 .	88
7,00	2 .	0122334
7,00	2 .	5677889
3,00	3 .	002

Stem width: 10,00
Each leaf: 1 case(s)

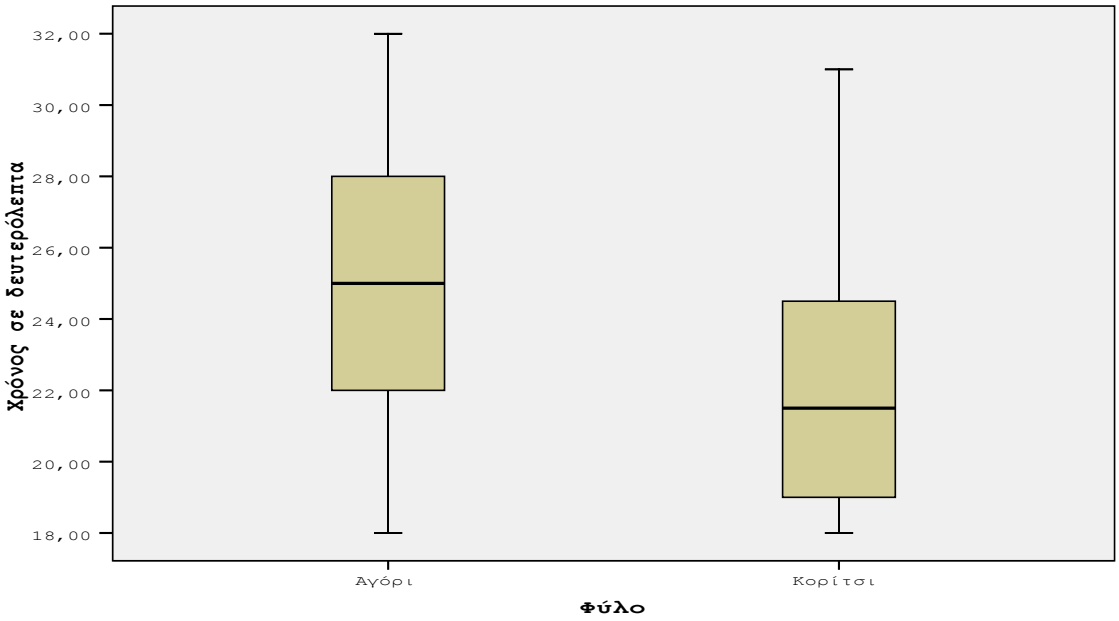
Χρόνος σε δευτερόλεπτα Stem-and-Leaf Plot for
sex= Κορίτσι

Frequency	Stem &	Leaf
5,00	1 .	89999
7,00	2 .	0112344
3,00	2 .	557
1,00	3 .	1

Stem width: 10,00
Each leaf: 1 case(s)

Στη συνέχεια παραθέτουμε το θηκόγραμμα (ως προς το φύλο) της μεταβλητής που περιγράφει το χρόνο που διανύουν τα παιδιά τα 100 μέτρα. Το θηκόγραμμα όπως παρατηρούμε μας δίνει τη δυνατότητα να συγκρίνουμε άμεσα τη διάμεσο, το 25^{o} και 75^{o} ποσοστιαίο σημείο, την μέγιστη και ελάχιστη παρατηρούμενη τιμή. Επιπλέον πιθανές

ακραίες τιμές δηλώνονται με ένα ο, ενώ οι extreme (πολύ ακραίες) με ένα *. Στο συγκεκριμένο παράδειγμα προκύπτει ότι δεν έχουμε ακραίες τιμές, η διάμεσος, η μέγιστη και η ελάχιστη τιμή του χρόνου σε δευτερόλεπτα των αγοριών είναι μεγαλύτερη από τις αντίστοιχες τιμές για τα κορίτσια.



<u>Επιπλέον δυνατότητες της διαδικασίας Descriptives</u>

Ακολουθούμε την πορεία: Analyze →Descriptive Statistics→ Descriptives, στο νέο παράθυρο διαλόγου που προκύπτει επιλέγοντας το πλαίσιο Save standardized values as variables καθίσταται δυνατή η αποθήκευση των τυποποιημένων τιμών (standardized values) για τις μεταβλητές του καταλόγου Variable(s). Αν $X_1,\ldots,X_n$ είναι οι διαθέσιμες δειγματικές τιμές της μεταβλητής που δηλώθηκε στο πλαίσιο Variable τότε δημιουργείται μία νέα στήλη με τιμές $Z_1,\ldots,Z_n$ που υπολογίζονται από τη σχέση:

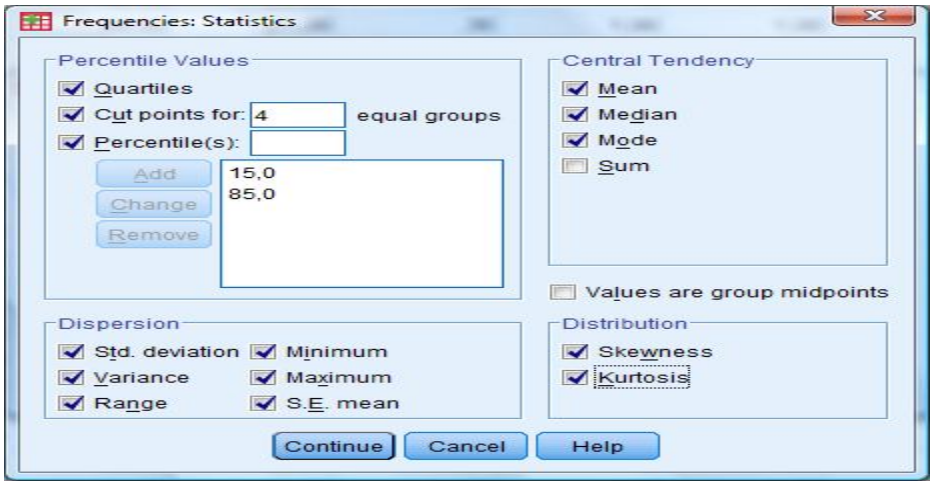
$$Z_i = \frac{X_i - \bar{X}}{S}.$$

Οι τυποποιημένες τιμές ή Z-scores είναι μερικές φορές χρήσιμες για περαιτέρω ανάλυση. Με αυτές μπορούμε για παράδειγμα να συγκρίνουμε δείγματα από διαφορετικούς πληθυσμούς ή μετρήσεις μεταβλητών σε διαφορετικές μονάδες μέτρησης.

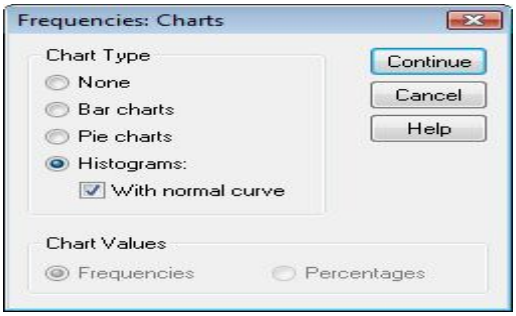
Επιπλέον δυνατότητες της διαδικασίας Frequencies

Αρχικά επιλέγουμε Analyze →Descriptive Statistics→ Frequencies.

Από την επιλογή Statistics επιλέγοντας το πλαίσιο Cut points for: ζητούμε την εμφάνιση των σημείων εκείνων για το διαχωρισμό των δεδομένων σε τόσες ομάδες όσες ο αριθμός που θα δηλωθεί (π.χ. cut points for 4 equal groups), καθώς και την εμφάνιση π.χ. του 15^{ου} και 85^{ου} ποσοστιαίου σημείου. Τέλος επιλέγουμε το πλαίσιο Values are group midpoints αν οι τιμές των δεδομένων μας είναι το μέσο ενός διαστήματος. Για παράδειγμα αν πρόκειται για ηλικίες και για όσους είναι από 30-40 είχαμε καταχωρήσει στην αντίστοιχη μεταβλητή το μέσο του διαστήματος δηλαδή την τιμή 35, ενώ για εκείνους με ηλικία από 40-50 αντίστοιχα την τιμή 45 κ.ο.κ.



Επιπλέον, από την επιλογή Charts έχει νόημα μόνο η κατασκευή ιστογράμματος (histogram) ζητώντας παράλληλα να σχεδιαστεί και η κανονική καμπύλη (normal curve). Η επιλογή αυτή μας δίνει τη δυνατότητα να διαπιστώσουμε αν έχουμε ενδείξεις για αποκλίσεις από την κανονική κατανομή



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.

Σχόλιο: α) Το παραπάνω τεστ εφαρμόζεται υπό τις προϋποθέσεις ότι ι) το μέγεθος του δείγματος είναι τετραπλάσιο του πλήθους των κελιών και ιι) 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).

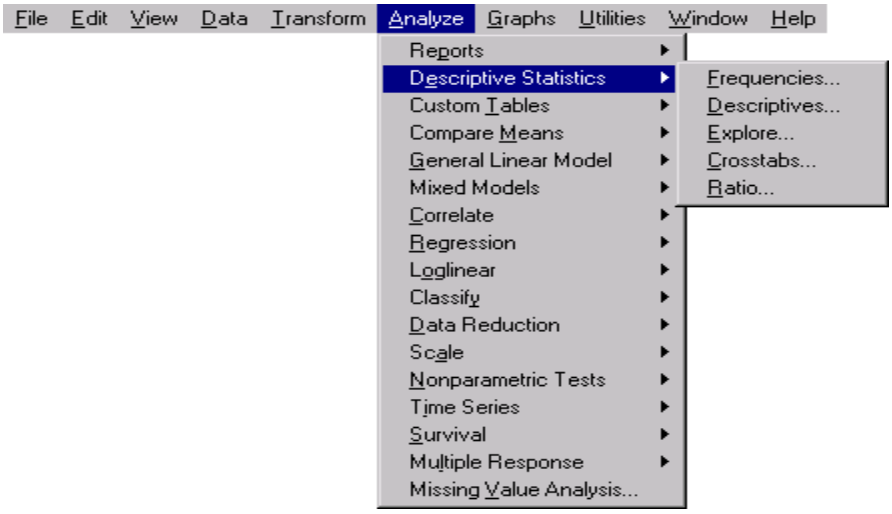
Ο συντελεστής Kappa του Kohen χρησιμοποιείται για πίνακες συνάφειας που έχουν τις ίδιες κατηγορίες στις στήλες και στις γραμμές. Παίρνει τιμές στο [-1,1]. Η τιμή 1 (-1 αντίστοιχα) υποδεικνύει πλήρη συμφωνία (πλήρη διαφωνία αντίστοιχα), ενώ η τιμή 0 υποδεικνύει ότι η συμφωνία είναι τυχαία.

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.

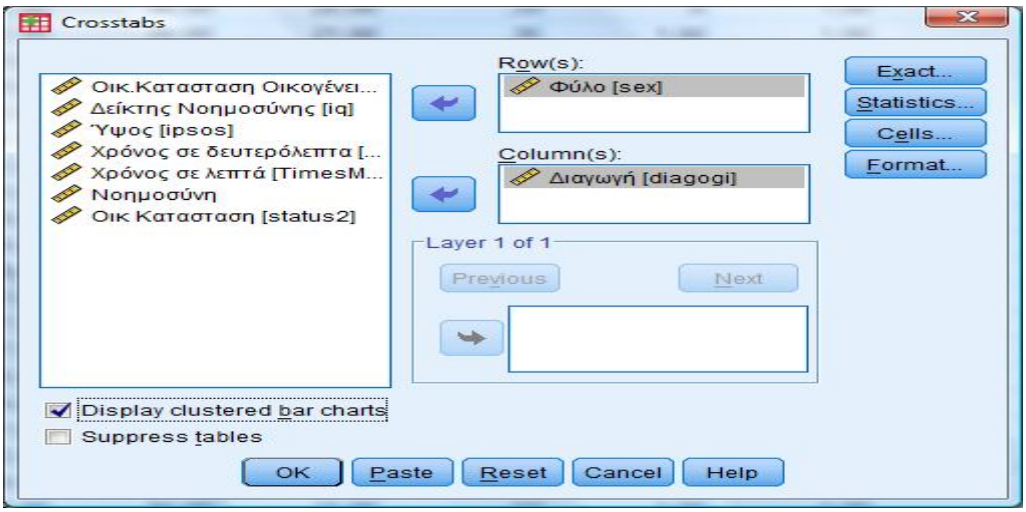
Η διαδικασία αυτή υλοποιείται ως εξής:

i. Analyze →Descriptive Statistics→ Crosstabs

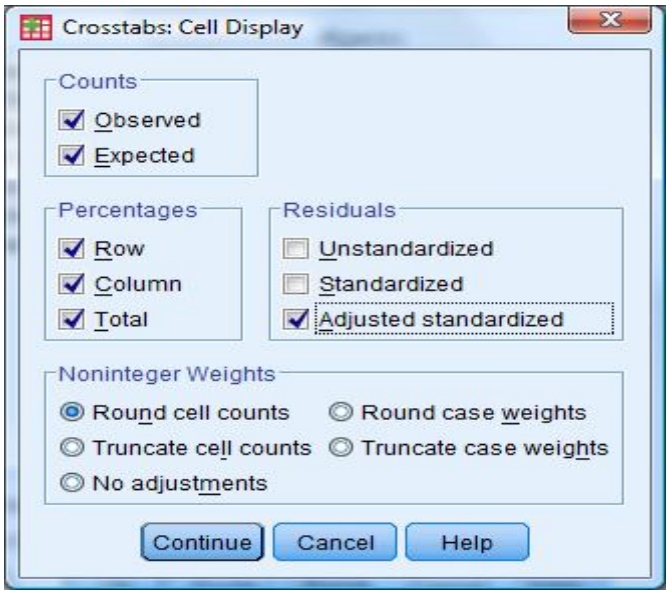


ii. Στο νέο παράθυρο διαλόγου που προκύπτει διαλέγουμε την ποιοτική μεταβλητή τις δυνατές τιμές της οποίας θέλουμε να έχουμε στις γραμμές (στήλες αντίστοιχα) του πίνακα συνάφειας και τη μετακινούμε στο πλαίσιο Rows (πλαίσιο Columns αντίστοιχα). Θέλοντας να κατασκευαστούν ομάδες ραβδογραμμάτων (bar charts) 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.



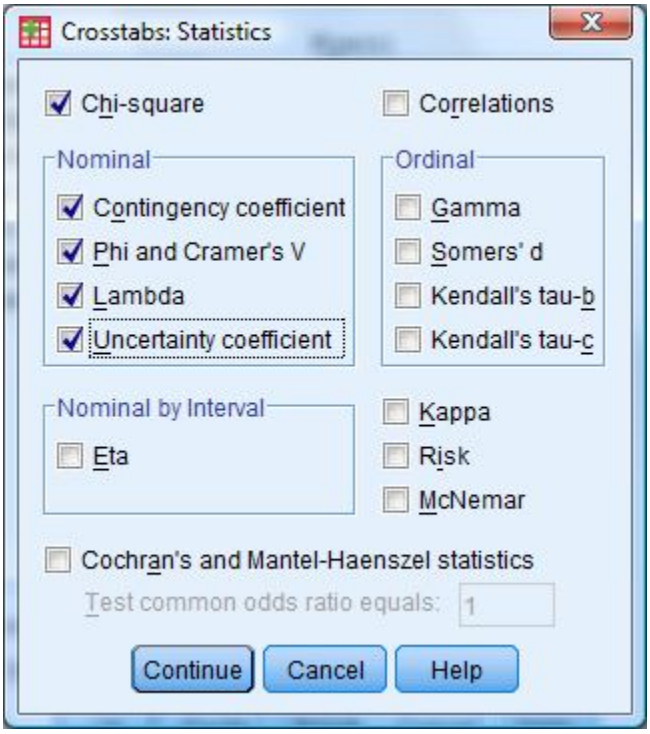
iii. Για να αποφανθούμε για την ύπαρξη, την ένταση και φύση της σχέσης των δύο μεταβλητών θα πρέπει να εμπλουτίσουμε τις πληροφορίες που μας δίνει το λογισμικό ως προεπιλογή. Αυτό μπορεί να επιτευχθεί αρχικά από την επιλογή Cells επιλέγοντας τα ακόλουθα:



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		
			A	B	Total
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 53,3% αυτών που έχουν διαγωγή Κοσμιωτάτη είναι αγόρια (αφού το 53,3% 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	

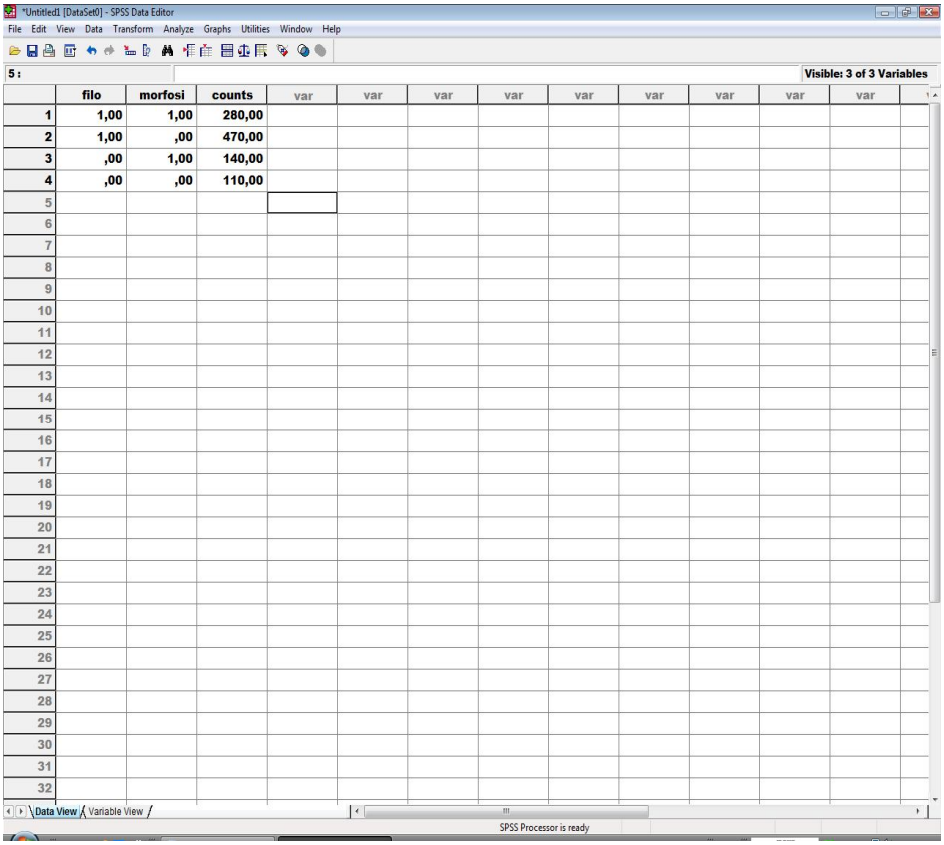
a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

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

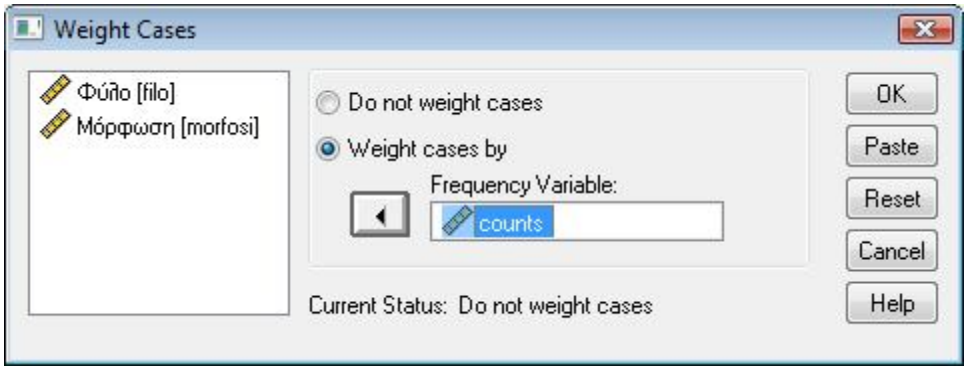
	Μη πτυχιούχοι	Πτυχιούχοι
Άνδρες	470	280
Γυναίκες	110	140

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 σε επόμενα κεφάλαια. Στην παράγραφο αυτή θα ασχοληθούμε μόνο με την πρώτη περίπτωση, δηλαδή με την τεκμηρίωση της σχέσης που έχουν δύο ποσοτικές μεταβλητές. Για αυτό το σκοπό θα εξετάσουμε:

- το γράφημα των τιμών των δύο ποσοτικών μεταβλητών (διάγραμμα διασποράς).
- 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 ή ίσο 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. **<u>Συντελεστής συσχέτισης του Pearson</u>**: 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:

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

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

Αποδεικνύεται ότι υπό την $H_0 : \rho = \rho_0$, όπου ρ_0 η αληθινή τιμή του πληθυσμιακού συντελεστή συσχέτισης ότι η στατιστική συνάρτηση $Z = \left(r_s - \rho_0\right)\sqrt{n-1} \underset{H_0}{\overset{\pi\rho o\sigma\varepsilon\gamma.}{\sim}} N\left(0,1\right)$. Το αποτέλεσμα αυτό χρησιμοποιείται για τον έλεγχο της υπόθεσης $H_0 : \rho = 0$.

Συντελεστής συσχέτισης του Kendall: Ο Kendall πρότεινε το στατιστικό:

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

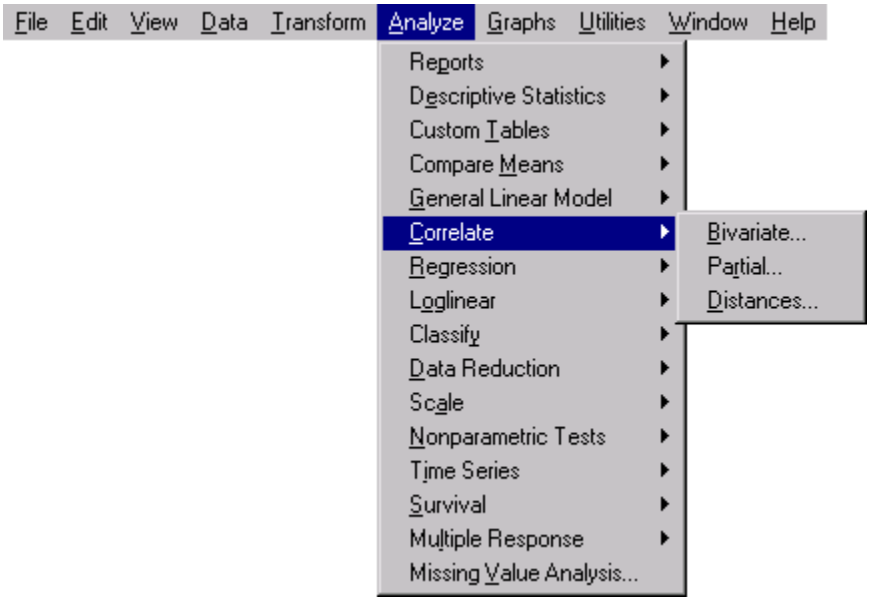
όπου n_c και n_d είναι το πλήθος των σύμφωνων και ασύμφωνων ζευγαριών, αντίστοιχα, δηλαδή των τιμών $\left(X_i, Y_i\right)$ και $\left(X_j, Y_j\right)$, $i \neq j$, $i, j = 1,...,n$, που το πρόσημο της διαφοράς $X_i - X_j$ είναι σύμφωνο, δεν είναι σύμφωνο, αντίστοιχα, με το πρόσημο της διαφοράς $Y_i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

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

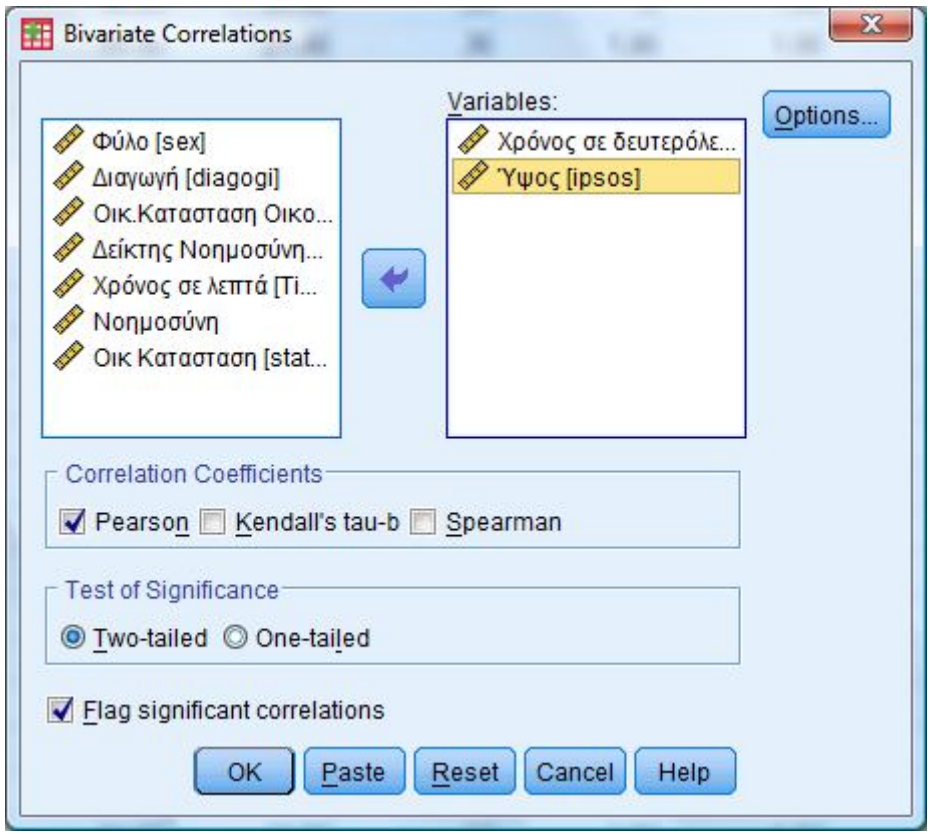
Σχόλιο: Βασικό πλεονέκτημα των συντελεστών συσχέτισης του Spearman και του Kendall είναι ότι μπορούν να χρησιμοποιηθούν και για διατάξιμες μεταβλητές, είναι ανθεκτικοί στην ύπαρξη ακραίων τιμών, και ως μη παραμετρικοί συντελεστές δεν απαιτούν καμία υπόθεση για τους πληθυσμούς. Από την άλλη μεριά βασικό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

		΄Υψος	Χρόνος σε δευτερόλεπτα
΄Υψος	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.

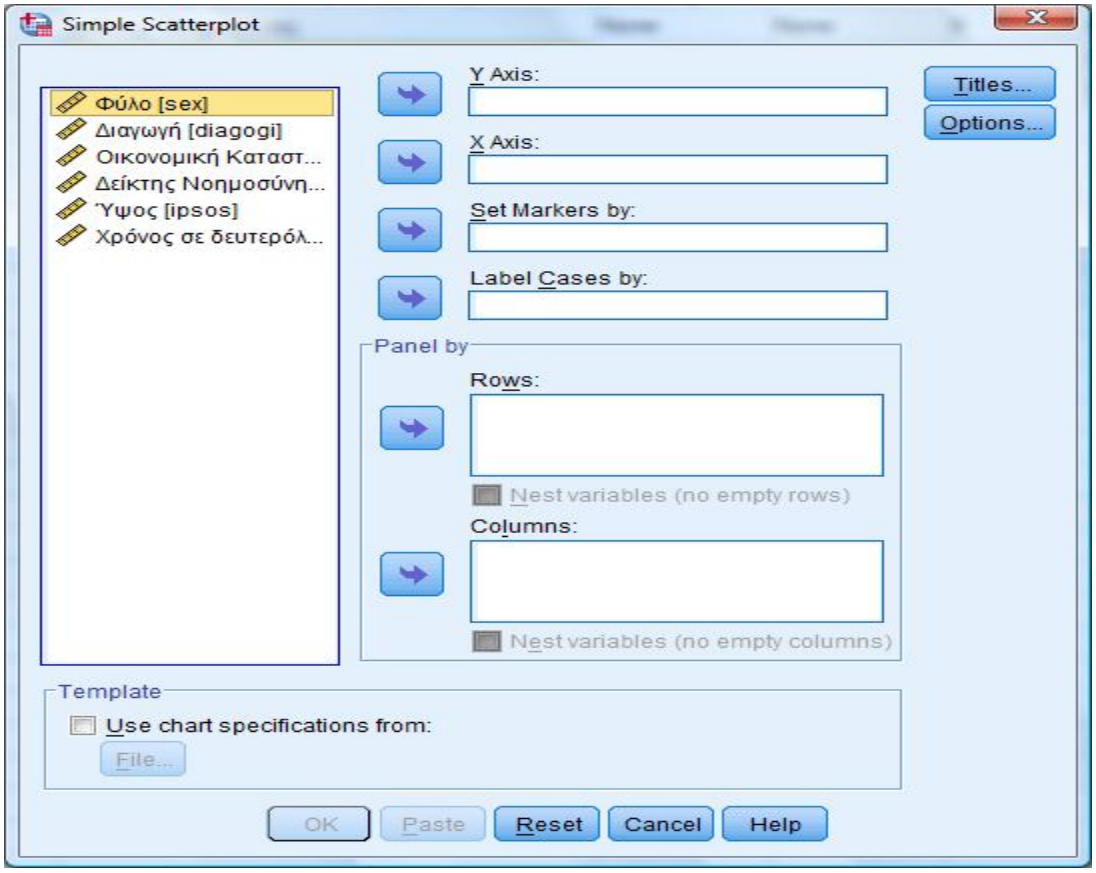
Για να αποκτήσουμε το διάγραμμα διασποράς μπορούμε να ακολουθήσουμε μία από τις παρακάτω δύο διαδικασίες:

Διαδικασία Scatter/Dot από Graphs→Legacy Dialogs

Από τη βασική ράβδο του λογισμικού επιλέγουμε

i. Graphs→ Legacy Dialogs→ Scatter/Dot.

ii. Στο νέο παράθυρο διαλόγου που προκύπτει μπορούμε να επιλέξουμε τον τύπο του Scatter/Dot. Έστω Simple Scatter. Προκύπτει το ακόλουθο παράθυρο διαλόγου:

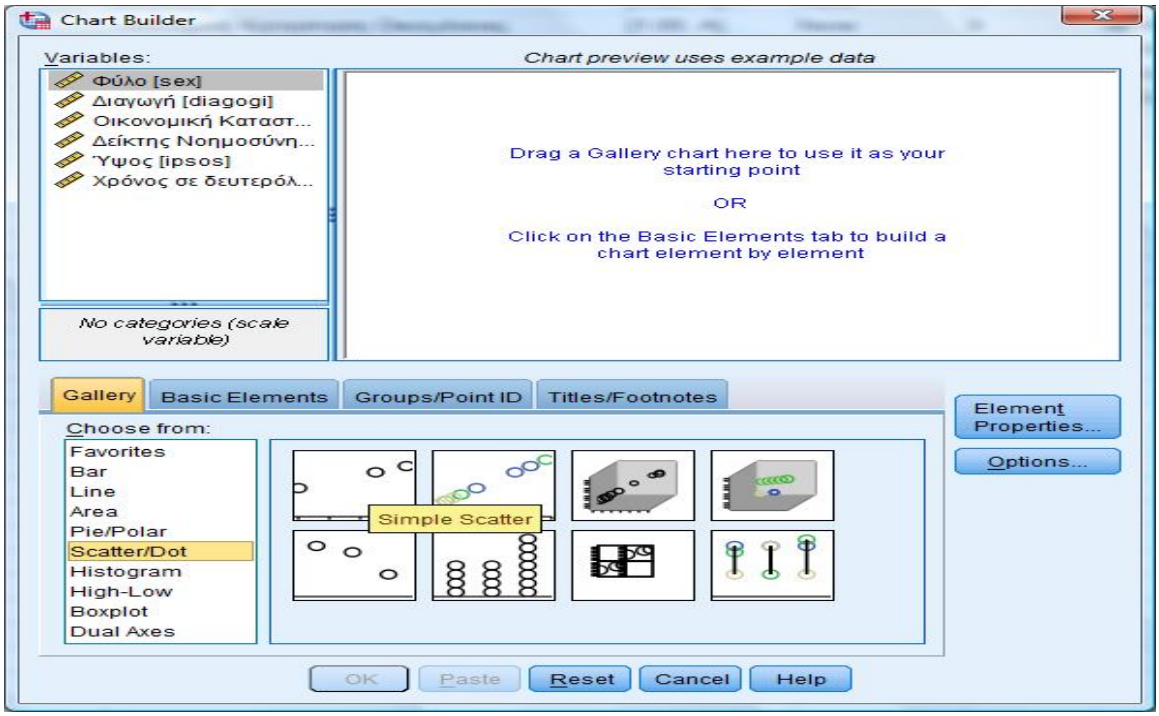


Μετακινούμε τη μεταβλητή (συνήθως την εξαρτημένη) στον κατακόρυφο άξονα του ορθογωνίου συστήματος αξόνων, ενώ στον οριζόντιο άξονα μετακινούμε την άλλη ποσοτική μεταβλητή (συνήθως την ανεξάρτητη). Επιπρόσθετα στο πλαίσιο Set Markers by μπορούμε να δηλώσουμε μία ποιοτική μεταβλητή ελέγχου έτσι ώστε στο διάγραμμα διασποράς να υπάρχει διάκριση των σημείων αντίστοιχη με τις κατηγορίες της ποιοτικής μεταβλητής. Έτσι για παράδειγμα μπορούμε να τοποθετήσουμε τη μεταβλητή Φύλο. 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.

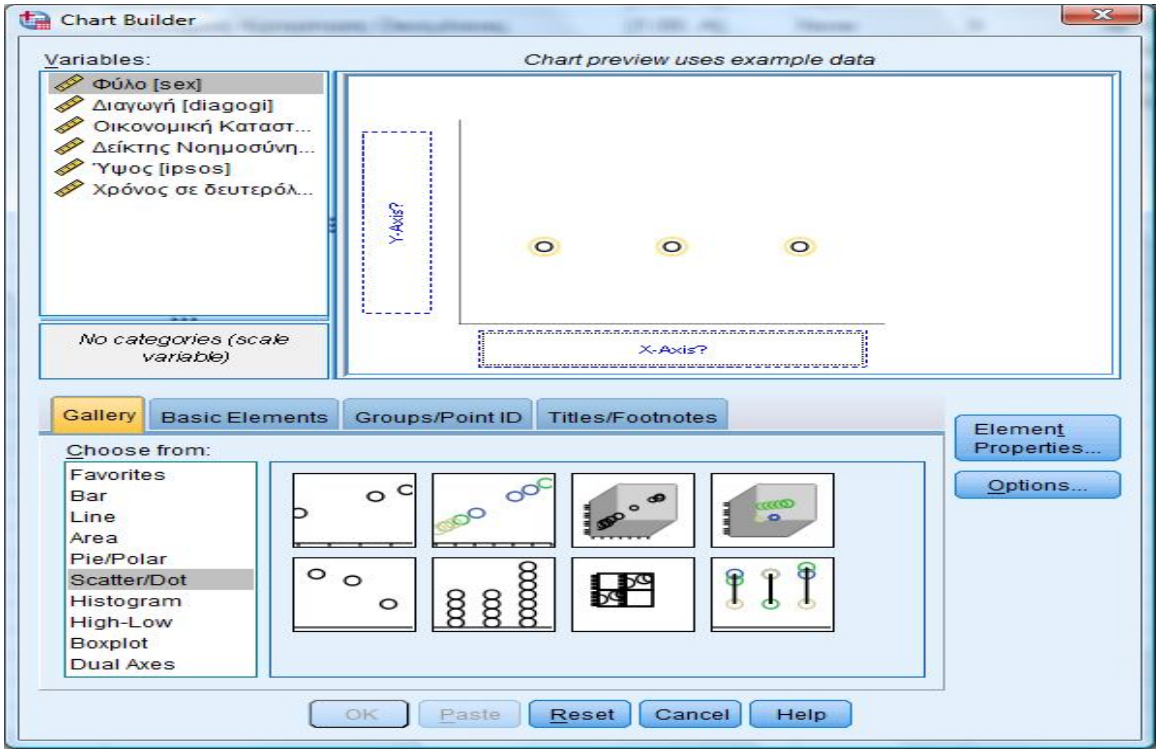
Διαδικασία Chart Builder

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

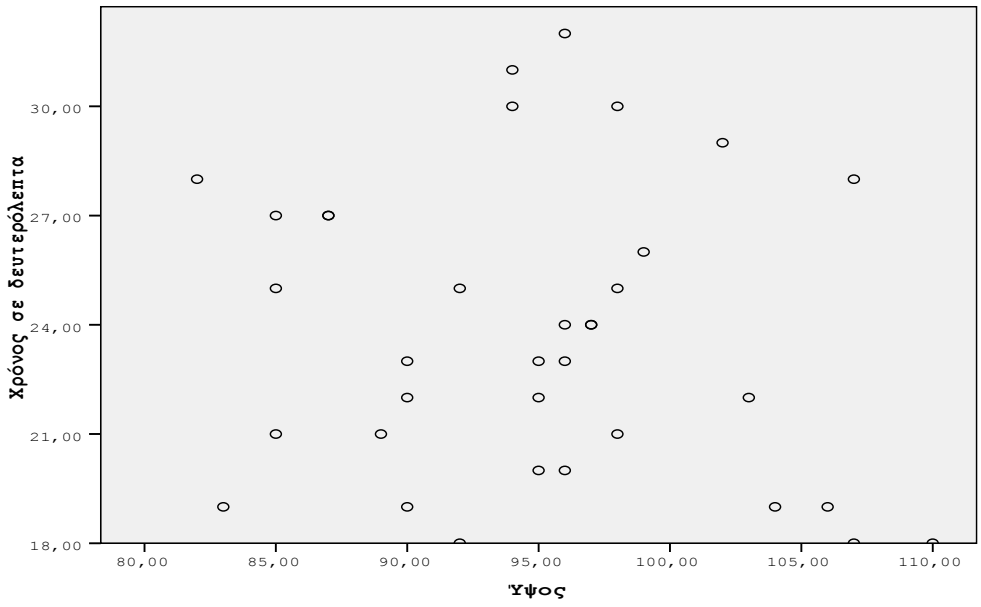
ii. Στο νέο παράθυρο διαλόγου που προκύπτει μπορούμε να επιλέξουμε τον τύπο του γραφήματος που θέλουμε να κατασκευάσουμε. Καθώς επιθυμούμε να κατασκευαστεί ένα διάγραμμα διασποράς επιλέγουμε Scatter/Dot. Έπειτα, όταν σκοπός μας είναι να κατασκευάσουμε ένα σύνηθες διάγραμμα διασποράς επιλέγουμε Simple 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). Για την καλύτερη παρουσίαση του γραφήματος δίνονται διάφορες δυνατότητες π.χ. στα πλαίσια Group/Point ID και Titles/Footnotes. Η παρουσίαση αυτών των επιλογών δεν είναι αντικείμενο και στόχος αυτών των διδακτικών σημειώσεων.



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

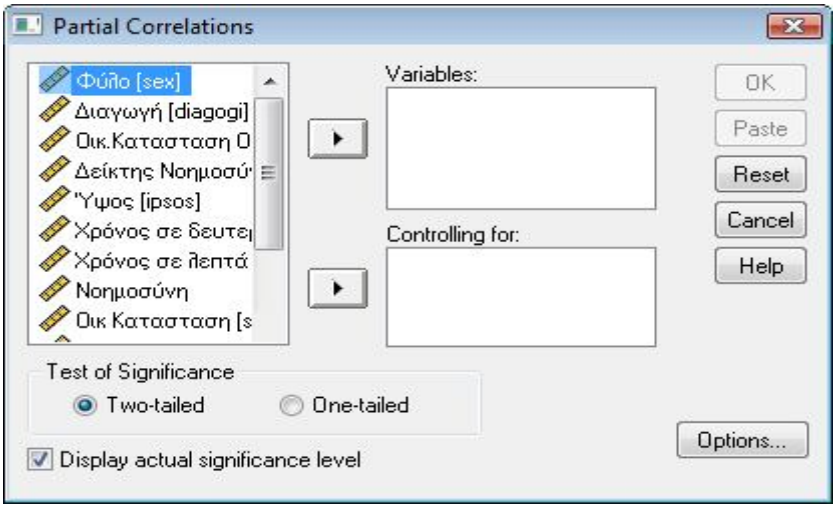


Μερικός συντελεστής συσχέτισης

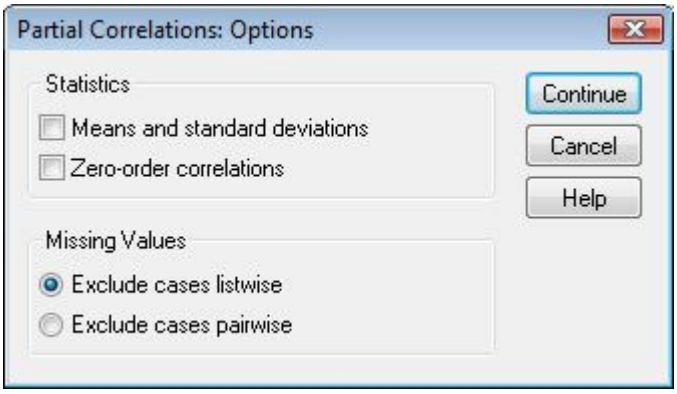
Κλείνοντας τούτη την παράγραφο θα ήταν ίσως παράλειψη να μην αναφέρουμε ότι όταν θέλουμε να εξετάσουμε την ένταση της γραμμικής σχέσης μεταξύ δύο μεταβλητών υπό την επίδραση μίας ή περισσοτέρων μεταβλητών ελέγχου χρησιμοποιούμε το μερικό συντελεστή συσχέτισης. Με τον μερικό συντελεστή συσχέτισης αυτό 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.

ΚΕΦΑΛΑΙΟ ΤΕΤΑΡΤΟ

Έλεγχος ότι η παράμετρος θέσης ενός πληθυσμού είναι ίση με δοθείσα γνωστή τιμή

Έστω ένα τυχαίο δείγμα $X_1,\ldots,X_n$ μεγέθους n από έναν πληθυσμό με μέση τιμή μ και διακύμανση σ^2, άγνωστη. Ενδιαφερόμαστε για τον έλεγχο, σε επίπεδο σημαντικότητας α, της μηδενικής υπόθεσης

$$H_0 : \mu = \mu_0,$$

με μ_0 μια σταθερά, ως προς μία εκ των

$$H_a : \mu > \mu_0, \quad H_a : \mu < \mu_0 , \; H_0 : \mu \neq \mu_0 .$$

Το παραπάνω πρόβλημα ελέγχεται υπό κάποιες υποθέσεις με τον παραμετρικό έλεγχο του t-test. Όταν κάποια από τις υποθέσεις αυτές δεν ικανοποιείται και δεν υπάρχει τρόπος διόρθωσης του προβλήματος, ο έλεγχος ανάγεται σε αυτόν ότι η πληθυσμιακή διάμεσος είναι ίση με τη δοθείσα τιμή με μεθόδους της μη παραμετρικής στατιστικής. Τα αποτελέσματα του τελευταίου ελέγχου γενικεύονται για τη μέση τιμή όταν τα δεδομένα είναι συμμετρικά.

Σχόλιο: Επισημαίνεται ότι οι μη παραμετρικές μέθοδοι είναι λιγότερο ισχυρές στο να «ανακαλύπτουν» στατιστικά σημαντικές σχέσεις συγκριτικά με τις ανάλογες, αντίστοιχες παραμετρικές.

Υπενθύμιση: Σε κάθε στατιστικό έλεγχο αποφασίζουμε στη βάση ενός στατιστικού για την αποδοχή ή την απόρριψη μίας υπόθεσης (της μηδενικής υπόθεσης όπως λέγεται και η οποία συμβολίζεται με H_0). Επομένως υπάρχει ο «κίνδυνος» είτε ο στατιστικός να απορρίπτει την προς έλεγχο μηδενική υπόθεση (να αποδέχεται την λεγόμενη εναλλακτική υπόθεση H_a), ενώ η H_0 είναι αληθής είτε ο στατιστικός να αποδέχεται την H_0, ενώ η H_a είναι αληθής. Στην πρώτη περίπτωση έχουμε το λεγόμενο **σφάλμα τύπου Ι**, ενώ στη δεύτερη το **σφάλμα τύπου ΙΙ**. Είναι τότε:

$$a = P\left(\text{σφάλμα τύπου I}\right) = P\left(\text{απορρίπτω } H_0 / H_0 \text{ αληθής}\right)$$

και

$$\beta = P(\text{σφάλμα τύπου II}) = P(\text{αποδέχομαι } H_0 / H_\alpha \text{ αληθής}).$$

Το επιθυμητό θα ήταν να επιτυγχάνεται η ταυτόχρονη ελαχιστοποίηση των α και β. Όμως κάτι τέτοιο είναι αδύνατο. Το πρόβλημα αυτό παρακάμπτεται, προκαθορίζοντας το α και ελαχιστοποιώντας το β ή ισοδύναμα μεγιστοποιώντας την **ισχύ του τεστ** $\gamma = 1 - \beta = P(\text{απορρίπτω } H_0 / H_\alpha \text{ αληθής})$. Το προκαθορισμένο α είναι γνωστό και ως **επίπεδο σημαντικότητας** και συνήθως επιλέγεται να είναι είτε 5% είτε 1%.

Στα στατιστικά πακέτα η απόφαση για την αποδοχή ή απόρριψη της υπόθεσης δεν γίνεται εξετάζοντας αν η τιμή του στατιστικού ανήκει στην **περιοχή απόρριψης** (γνωστή και ως **κρίσιμη περιοχή**), αλλά στη βάση των p-τιμών (p-value ή Sig.) Η **p-τιμή** ενός στατιστικού τεστ είναι η μικρότερη τιμή του επιπέδου σημαντικότητας για την οποία απορρίπτεται η μηδενική υπόθεση. Εύκολα προκύπτει τότε ότι **απορρίπτουμε την προς έλεγχο μηδενική υπόθεση αν η p-τιμή είναι μικρότερη από το προκαθορισμένο επίπεδο σημαντικότητας (συνήθως 0.05).**

4.1 Μεθοδολογία- Υλοποίηση στο S.P.S.S.

Η μεθοδολογία που θα χρησιμοποιηθεί για τη στατιστική ανάλυση του πιο πάνω προβλήματος εξαρτάται από το αν πληρούνται ή όχι κάποιες προϋποθέσεις, τις οποίες και πρέπει αρχικά να ελέγξει ο ερευνητής. Πιο συγκεκριμένα, ελέγχουμε

α) αν 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 μπορούμε να ισχυριστούμε ότι περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Ανάλογα με τα αποτελέσματα των παραπάνω ελέγχων προβαίνουμε στον παραμετρικό έλεγχο του t test ή στο μη παραμετρικό έλεγχο (προσημικό στατιστικό τεστ). Ο μη παραμετρικός έλεγχος ουσιαστικά ελέγχει αν η πληθυσμιακή διάμεσος είναι ίση με την προκαθορισμένη τιμή και τα αποτελέσματα όπως θα δούμε στη συνέχεια μπορούν να γενικευθούν υπό κάποια προϋπόθεση για τον ζητούμενο έλεγχο.

Από τα παραπάνω ίσως έγινε ήδη αντιληπτό ότι κομβικό σημείο για τον τρόπο διεξαγωγής του υπό μελέτη ελέγχου αποτελεί η διενέργεια των προκαταρκτικών ελέγχων α) και β), με βάση τα αποτελέσματα των οποίων θα αποφανθούμε αν θα προχωρήσουμε παραμετρικά ή μη παραμετρικά. Για το λόγο αυτό στη συνέχεια παρουσιάζονται όλα τα

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

1. Αρχικά ελέγχουμε αν υπάρχουν ακραίες τιμές στις διαθέσιμες δειγματικές τιμές. Αν το ποσοστό των ακραίων τιμών, που προκύπτει αφαιρώντας μία ακραία κάθε φορά με τη βοήθεια του θηκογράμματος, δε ξεπερνά το 10%, τότε προχωρούμε στο βήμα 2. Αν το ποσοστό των ακραίων τιμών ξεπερνά το 10%, τότε, αφού συμπεριλάβουμε τις δειγματικές παρατηρήσεις που προηγούμενα έχουν αποκλειστεί από την ανάλυση, 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 έλεγχος (βλέπε βήμα 4).

2. Στο βήμα 2, 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 έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή. Αν ο έλεγχος της κανονικότητας μας υποδεικνύει ότι η υπόθεση της κανονικότητας δεν απορρίπτεται (p-τιμή >α), τότε η ανάλυση θα συνεχιστεί με τον παραμετρικό έλεγχο του t τεστ (βλέπε βήμα 3). Αν η υπόθεση της κανονικότητας απορρίπτεται (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, εννοείται χωρίς να προκύπτει πρόβλημα ύπαρξης ακραίων τιμών, συνεχίζουμε την ανάλυση παραμετρικά (βήμα 3). Σε αντίθετη περίπτωση, αν το πλήθος των δειγματικών παρατηρήσεων (μη λαμβάνοντας υπόψη αυτές που έχουν αφαιρεθεί στο βήμα 1) είναι μεγάλο (συνήθως μεγαλύτερο ή ίσο του 30) κάνοντας χρήση του Κεντρικού Οριακού Θεωρήματος, προβαίνουμε στον παραμετρικό έλεγχο της υπό έλεγχο υπόθεσης (βλέπε βήμα 3), όπου η p-τιμή του ελέγχου και το διάστημα εμπιστοσύνης θα είναι προσεγγιστικά. Στην περίπτωση τώρα που 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. <u>Παραμετρικός έλεγχος t τεστ</u>: Οι κρίσιμες περιοχές μεγέθους α για τον έλεγχο της μηδενικής υπόθεσης

$$H_0 : \mu = \mu_0,$$

με μ_0 μια σταθερά, ως προς τις εναλλακτικές

$$H_a : \mu > \mu_0, \quad H_a : \mu < \mu_0, \; H_a : \mu \neq \mu_0,$$

είναι αντίστοιχα

$$t \geq t_{n-1,a}\,,\; t \leq -t_{n-1,a}\,,\; |t| \geq t_{n-1,a/2} \;\; \left(t \geq t_{n-1,a/2} \text{ ή } t \leq -t_{n-1,a/2}\right),$$

όπου $t = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$.

Επιπλέον το 100(1-α)% Δ.Ε. για τη μέση τιμή είναι

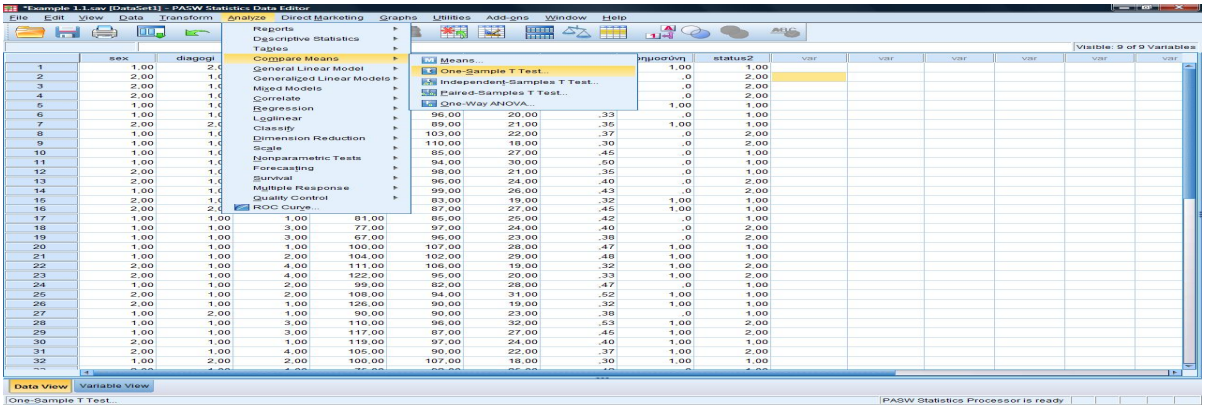
$$\left(\bar{X} - t_{n-1,a/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,a/2} \frac{S}{\sqrt{n}} \right).$$

Επισήμανση: Σε περίπτωση που έχει χρησιμοποιηθεί κάποιος μετασχηματισμός διόρθωσης του προβλήματος είτε της ύπαρξης πολλών ακραίων τιμών είτε της μη κανονικότητας, τότε όλα τα παραπάνω αναφέρονται στις μετασχηματισμένες τιμές και στο τροποποιημένο σε μέγεθος δείγμα. Ειδικότερα, αν έχει χρησιμοποιηθεί ο μετασχηματισμός του λογαρίθμου, θα προβούμε στον έλεγχο αν ο μέσος λογάριθμος δε διαφέρει στατιστικά σημαντικά από το λογάριθμο της δοθείσας γνωστής τιμής.

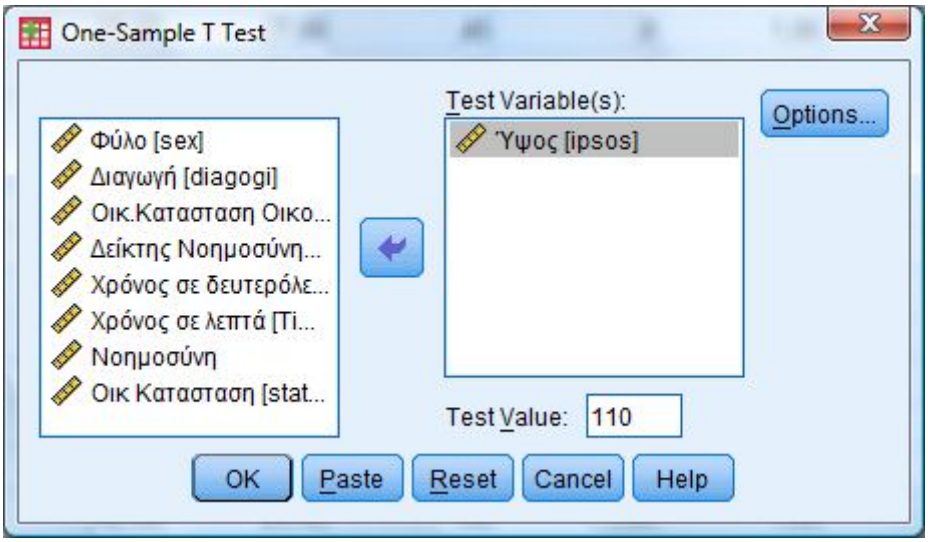
Υλοποίηση στο S.P.S.S.

Ο έλεγχος αυτός υλοποιείται ως εξής:

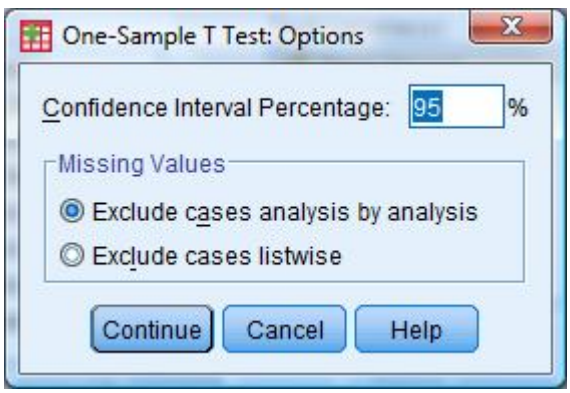
i. Analyze→Compare Means→One-Sample T Test.



ii. Στο παράθυρο διαλόγου που προκύπτει μετακινούμε στο πλαίσιο Test Variable τη μεταβλητή (στήλη του αρχείου δεδομένων, έστω το Ύψος) που καταγράφονται οι δειγματικές τιμές της υπό μελέτη μεταβλητής. Στο πλαίσιο Test Value εισάγουμε την τιμή ως προς την οποία θέλουμε να γίνει ο έλεγχος (έστω 110 cm). Προσοχή: Αν έχει χρησιμοποιηθεί κατάλληλος μετασχηματισμός διόρθωσης του προβλήματος των ακραίων τιμών ή της μη κανονικότητας θα πρέπει να τροποποιείται κατάλληλα και η αρχική τιμή προς έλεγχο.



iii. Από την επιλογή Options έχουμε τη δυνατότητα να καθορίσουμε τον τρόπο χειρισμού των ελλιπών τιμών καθώς και να προσδιορίσουμε το βαθμό εμπιστοσύνης του διαστήματος εμπιστοσύνης που θα κατασκευαστεί για τη διαφορά της καθορισμένης τιμής από την πληθυσμιακή μέση τιμή, δηλαδή για την ποσότητα $\mu - \mu_0$. Έτσι για παράδειγμα ζητάμε τον υπολογισμό του 95% διαστήματος εμπιστοσύνης για το $\mu - \mu_0$.



Ερμηνεία αποτελεσμάτων του S.P.S.S.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Ύψος	35	94,8571	7,18308	1,21416

One-Sample Test

	Test Value = 110					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Ύψος	-12,472	34	,000	-15,14286	-17,6103	-12,6754

Από τον πίνακα One-Sample Statistics πληροφορούμαστε ότι 35 παρατηρήσεις είναι διαθέσιμες στη μεταβλητή Ύψος. Δηλαδή, έχει καταγραφεί το ύψος 35 παιδιών. Το μέσο ύψος αυτών των παιδιών είναι 94,8571 εκατοστά, η δειγματική τυπική απόκλιση του ύψους είναι S =7.18308 εκατοστά και το τυπικό σφάλμα για τη μέση τιμή είναι $\frac{S}{\sqrt{n}}$ =1.21416.

Από τον δεύτερο πίνακα (One-Sample Test) το λογισμικό αρχικά στο πλαίσιο Test Value μας υπενθυμίζει ότι η τιμή ελέγχου είναι ίση με 110 (Test Value =110). Επιπλέον μας δίνει την τιμή του t στατιστικού για τον έλεγχο της υπόθεσης ότι το πληθυσμιακό μέσο ύψος είναι ίσο με 110 εκατοστά. Επιπλέον προκύπτει ότι το μέσο ύψος των παιδιών είναι στατιστικά σημαντικά διαφορετικό από 110 εκατοστά (t = -12.472, p-τιμή<0.001). Μάλιστα το μέσο ύψος των παιδιών είναι στατιστικά σημαντικά μικρότερο από 110 εκατοστά (καθώς από τη στήλη Mean Difference παρατηρούμε ότι η μέση διαφορά (Μέσο ύψος -110) είναι ίση με -15.14286 εκατοστά). Τέλος, το 95% διάστημα εμπιστοσύνης για τη διαφορά $\mu - 110$ είναι το (-17.6103, -12.6754) και επομένως ένα 95% διάστημα εμπιστοσύνης για το μέσο ύψος μ είναι το (92.3897,97.3246).

4. Τεστ του Wilcoxon (Wilcoxon Signed Rank Test) : Ο έλεγχος της μηδενικής υπόθεσης

$$H_0 : \mu = \mu_0,$$

με μ_0 μια σταθερά, ως προς μία εκ των $H_a : \mu > \mu_0$, $H_a : \mu < \mu_0$, $H_a : \mu \neq \mu_0$, <u>υπό την προϋπόθεση ότι ο πληθυσμός είναι συμμετρικός</u> γύρω από το μ_0, ανάγεται σε έλεγχο της διαμέσου και μεταξύ άλλων έχει προταθεί ο προσημικός έλεγχος καθώς και το τεστ του Wilcoxon για ένα δείγμα (για περισσότερες πληροφορίες βλέπε Παπαϊωάννου και Λουκάς (2002)).

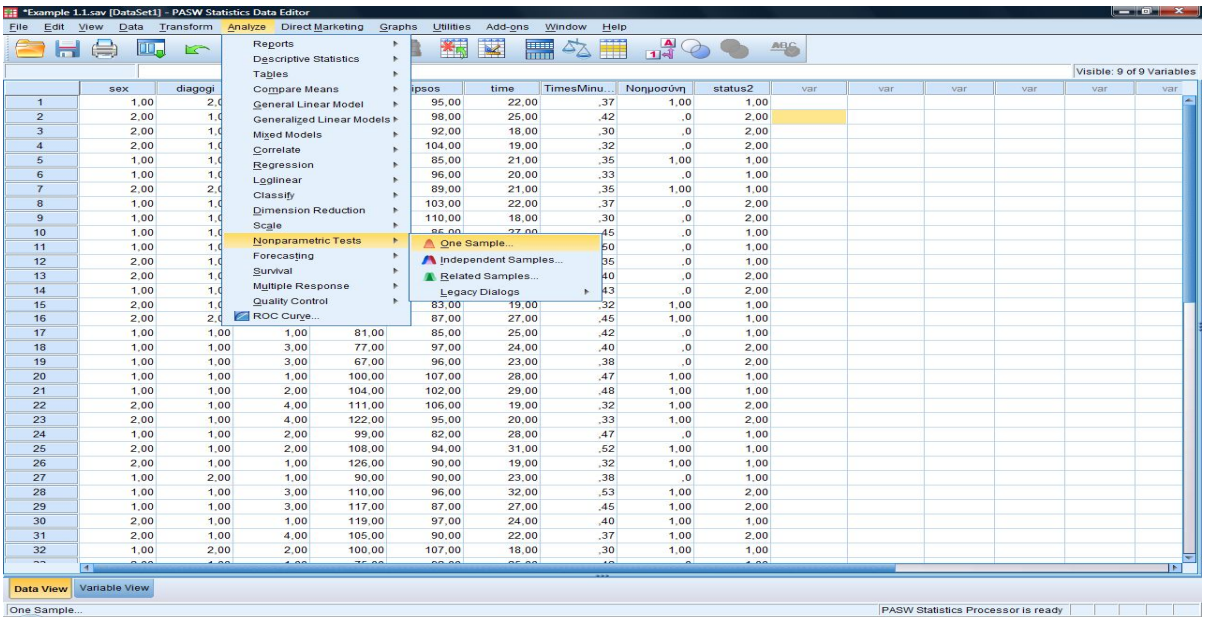
Επιγραμματικά αναφέρουμε ότι αν $D_i = X_i - m_0$, $i = 1,...,n$, είναι οι μη μηδενικές διαφορές και $d_1,...,d_c$, ο αριθμός των παρατηρήσεων σε καθεμία από τις c διαφορετικές απόλυτες διαφορές (σε αύξουσα τάξη μεγέθους), με $d_i \geq 1$, και $\sum_{i=1}^{c} d_i = \frac{n(n+1)}{2}$ τότε η προσεγγιστική κατανομή του στατιστικού T^+ που παριστάνει το άθροισμα των τάξεων που αντιστοιχούν στις θετικές διαφορές, υπό τη μηδενική υπόθεση, είναι

$$W = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{c} \frac{d_i\left(d_i^2 - 1\right)}{48}}} \underset{H_0}{\overset{\pi\rho o\sigma}{\sim}} N(0,1)$$

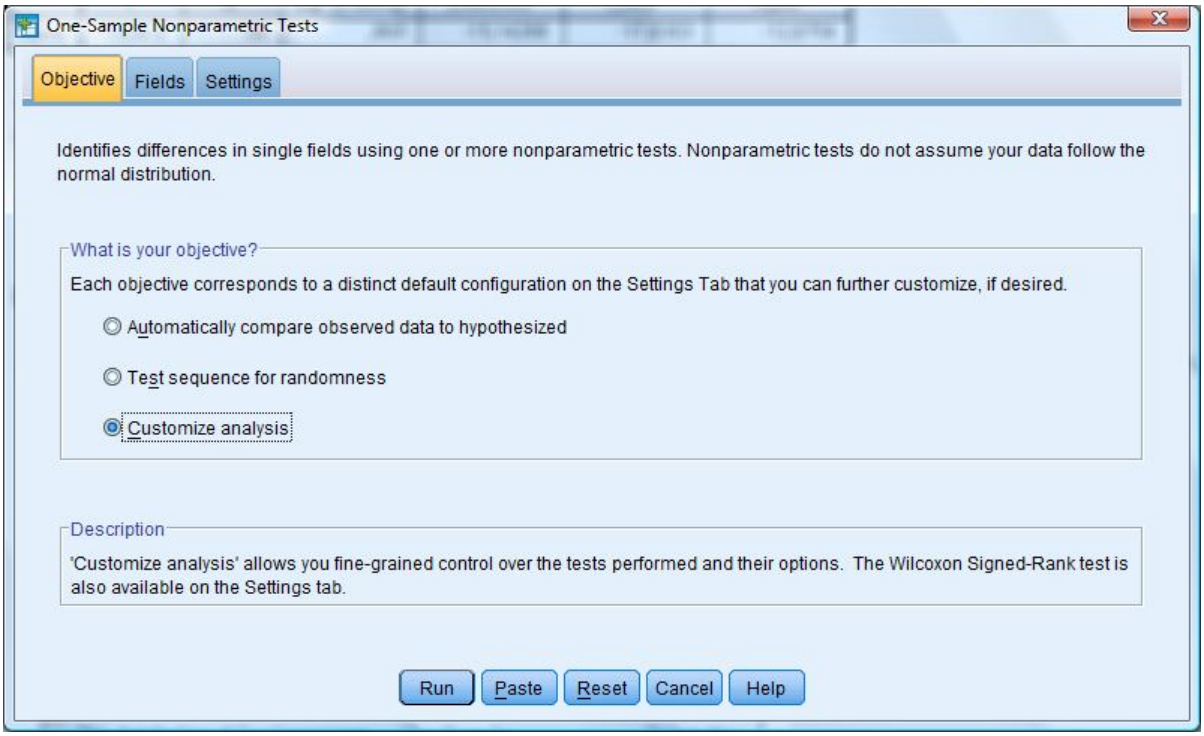
και αυτή η στατιστική συνάρτηση χρησιμοποιείται για τον έλεγχο της υπό μελέτης μηδενικής υπόθεσης.

Ο έλεγχος αυτός υλοποιείται ως εξής:

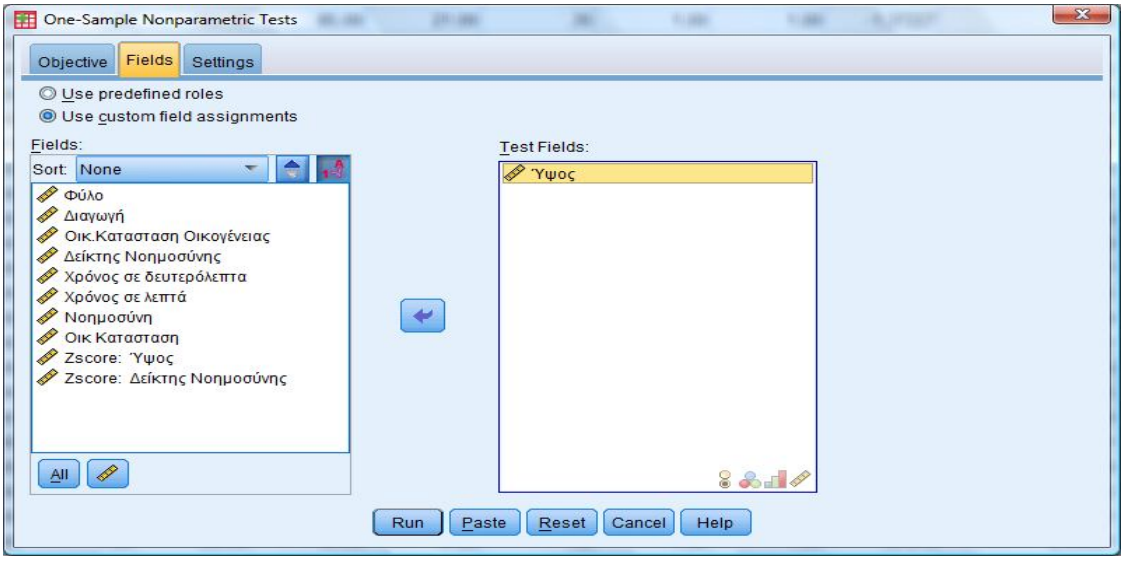
i. Analyze→Nonparametric Tests→One Sample.



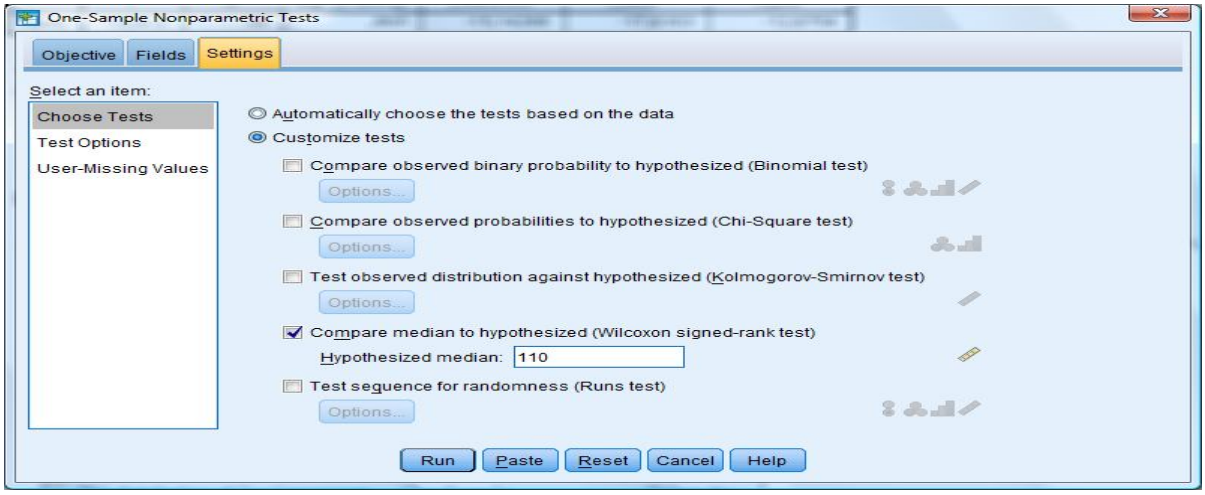
ii. Στο παράθυρο διαλόγου που προκύπτει επιλέγουμε στο πλαίσιο Objective την επιλογή Customize analysis, έτσι ώστε στη συνέχεια από τα πλαίσια Fields και Settings να καθορίσουμε τον έλεγχο τον οποίο θέλουμε να διενεργηθεί.



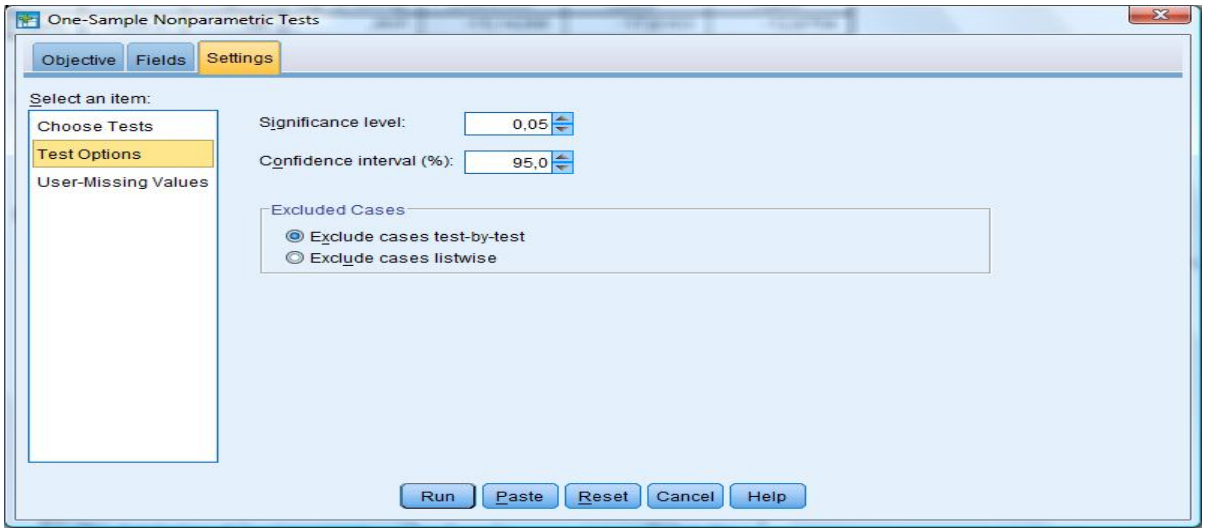
Στο πλαίσιο Fields τοποθετούμε στο πλαίσιο Test Fields τις μεταβλητές που θέλουμε να ελέγξουμε π.χ.



Στο πλαίσιο Settings και έχοντας ενεργοποιημένη την επιλογή Choose Tests και Customize tests επιλέγουμε τον κατάλληλο μη παραμετρικό έλεγχο, δηλαδή στη συγκεκριμένη περίπτωση την επιλογή Compare median to hypothesized (Wilcoxon signed-rank test) τοποθετώντας στο πλαίσιο Hypothesized median την προκαθορισμένη τιμή π.χ. 110.



Από την επιλογή Test Options μας δίνεται μεταξύ άλλων η δυνατότητα καθορισμού του επιπέδου σημαντικότητας και του βαθμού εμπιστοσύνης των Δ.Ε. Επιλέγοντας Run διενεργούνται οι επιλογές που ζητήσαμε.



Ερμηνεία αποτελεσμάτων του S.P.S.S.

Η υπόθεση ότι η πληθυσμιακή διάμεσος του ύψους ισούται με 110 cm απορρίπτεται καθώς η p-τιμή του One-Sample Wilcoxon Signed Ranks test είναι μικρότερη από 0.001. Το αποτέλεσμα αυτό για να μπορεί να γενικευθεί στην πληθυσμιακή μέση τιμή του ύψους θα πρέπει η δειγματική μέση τιμή να είναι κοντά στην δειγματική διάμεσο, κάτι που μπορεί να ελέγξει κάποιος μέσω της διαδικασίας Descriptive Statistics Explore.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of Ύψος equals 110.	One-Sample Wilcoxon Signed Ranks Test	.000	Reject the null hypothesis.

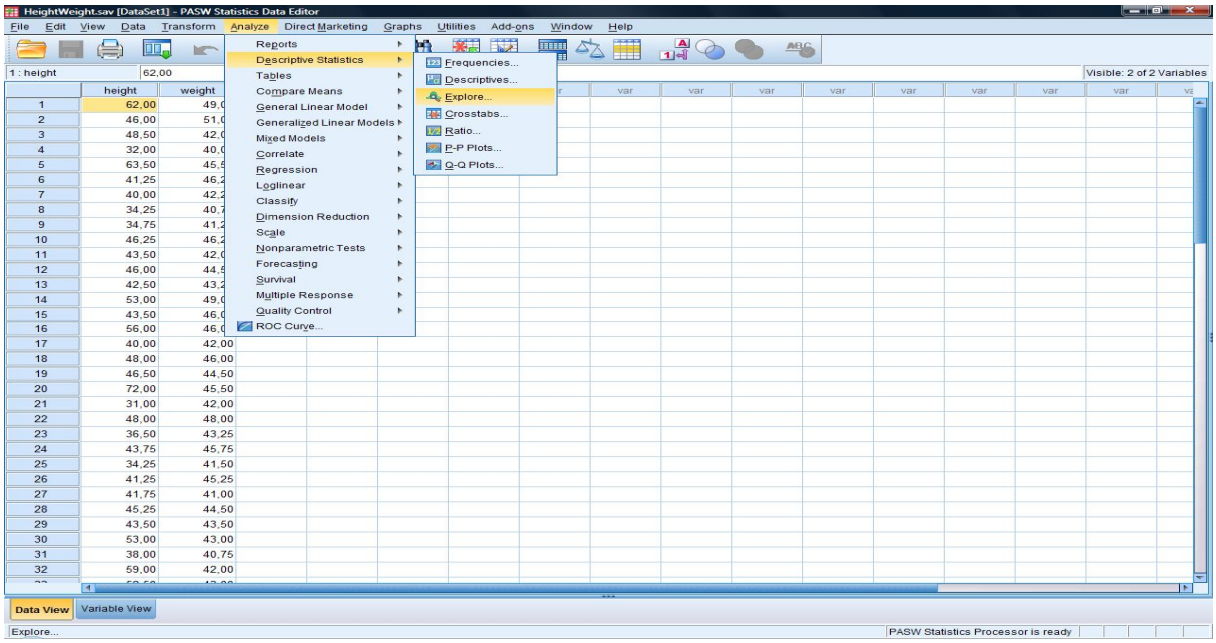
Asymptotic significances are displayed. The significance level is .05.

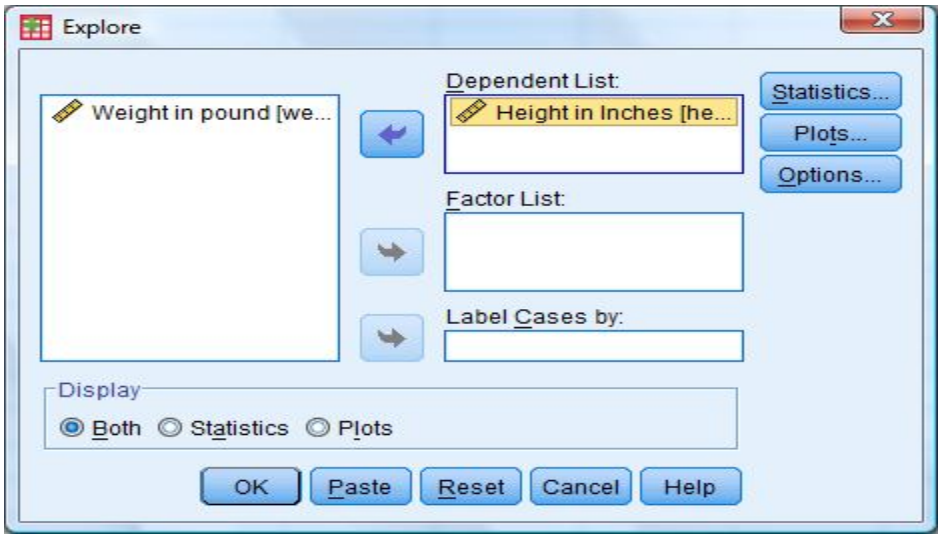
4.2 Παραδείγματα

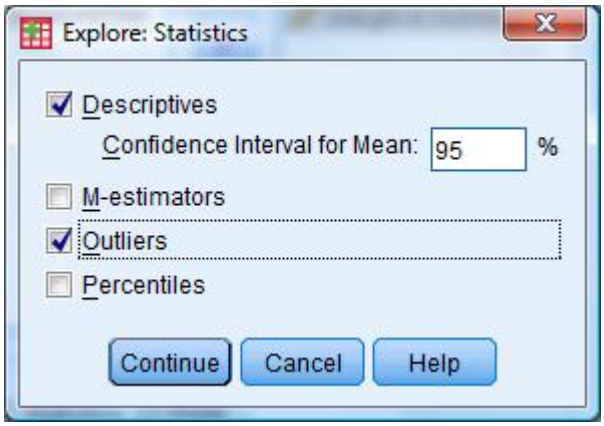
Παράδειγμα 1ο Στο αρχείο HeightWeight.sav* (βλέπε φάκελο DataII 1Mean t-test) καταγράφονται οι τιμές του βάρους και του ύψους (σε ίντσες) 73 τυχαία επιλεγμένων ατόμων από έναν πληθυσμό. Θέλουμε να ελέγξουμε, αν είναι εφικτό, αν το μέσο ύψος του πληθυσμού είναι στατιστικά σημαντικά διαφορετικό από τις 60 inches.

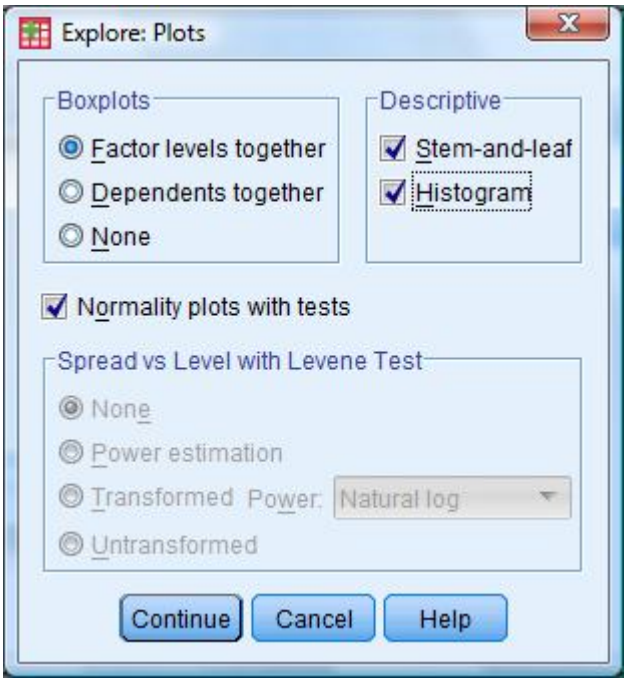
Υλοποίηση: Η υλοποίηση θα γίνει ακολουθώντας τη μεθοδολογία της παραγράφου 4.1.

Έλεγχος ακραίων τιμών. Ο έλεγχος για την ύπαρξη ακραίων τιμών γίνεται με το θηκόγραμμα. Παρατήρηση: ταυτόχρονα ζητάμε και στατιστικούς και γραφικούς τρόπους ελέγχου της κανονικότητας έτσι ώστε να μην επανερχόμαστε σε περίπτωση που δεν υπάρχει πρόβλημα ακραίων τιμών. Ακολουθούμε τα επόμενα βήματα μέσω της διαδικασίας Analyze Descriptive Statistics Explore.


HeightWeight.sav [DataSet1] - PASW 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
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
Data View
Variable View
Explore...
PASW Statistics Processor is ready

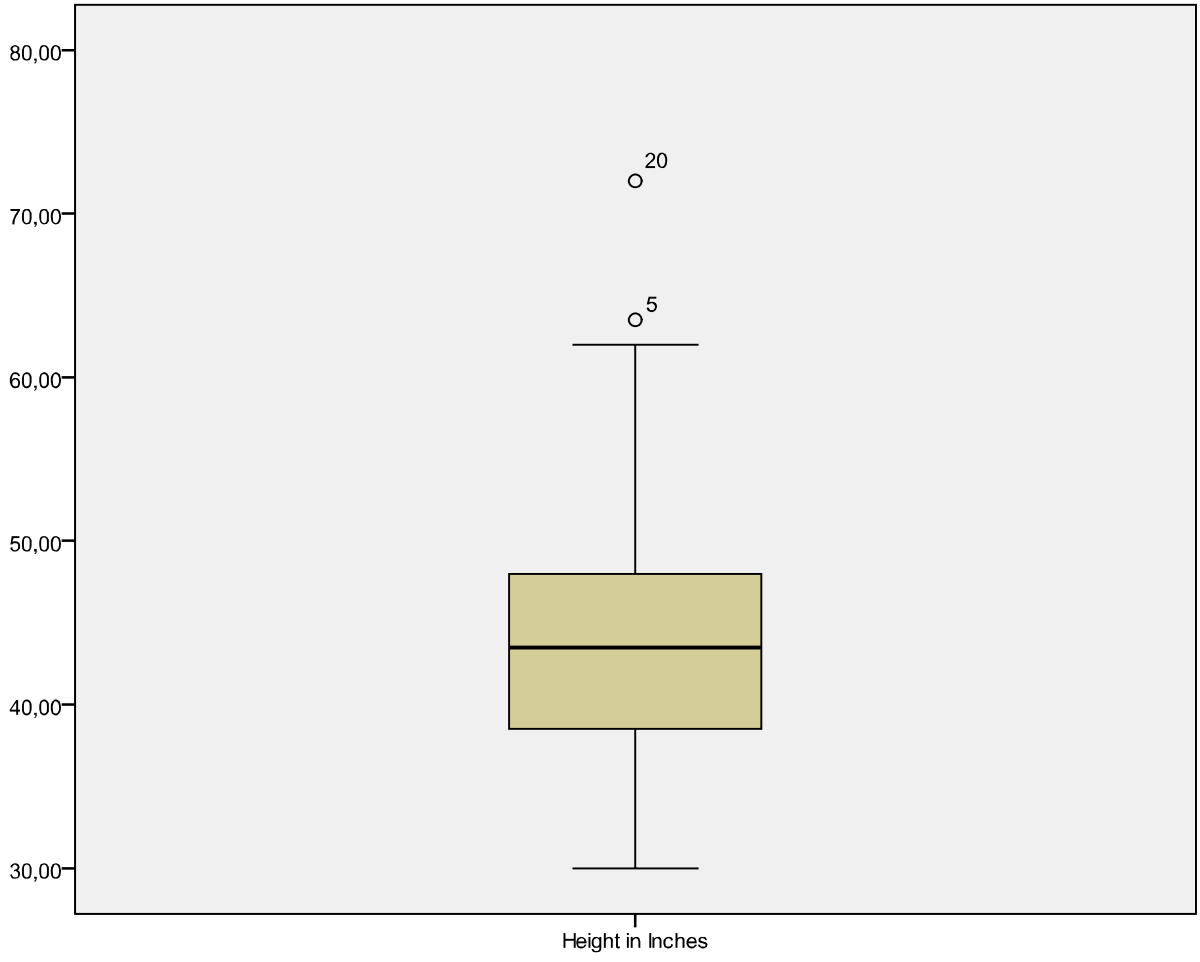

Explore
Weight in pound [we...
Dependent List:
Height in Inches [he...
Factor List:
Label Cases by:
Statistics...
Plots...
Options...
Display
Both
Statistics
Plots
OK
Paste
Reset
Cancel
Help


Explore: Statistics
Descriptives
Confidence Interval for Mean: 95 %
M-estimators
Outliers
Percentiles
Continue
Cancel
Help

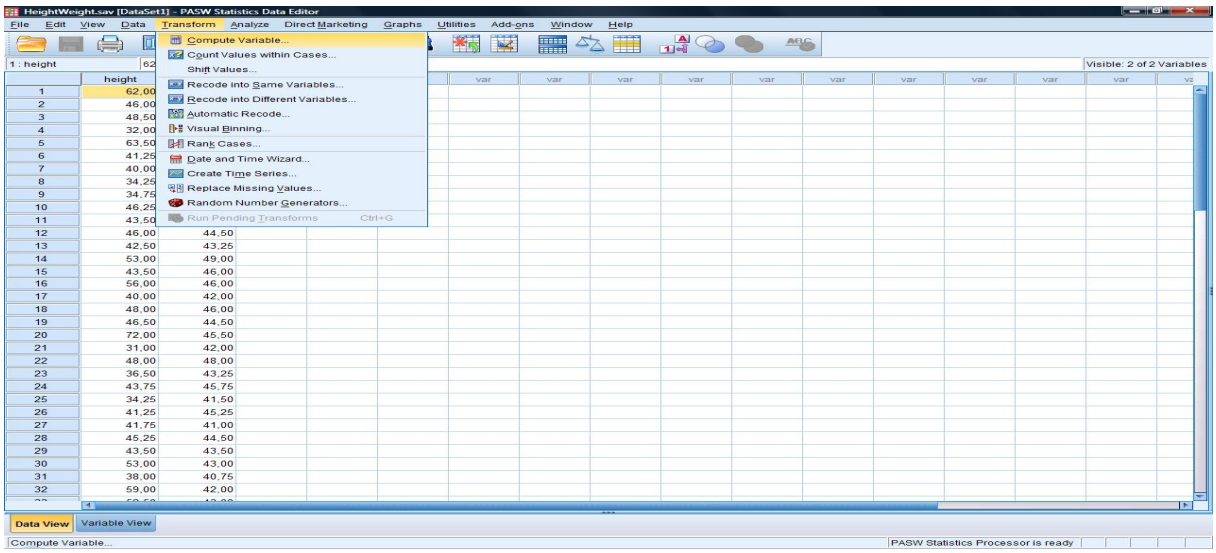


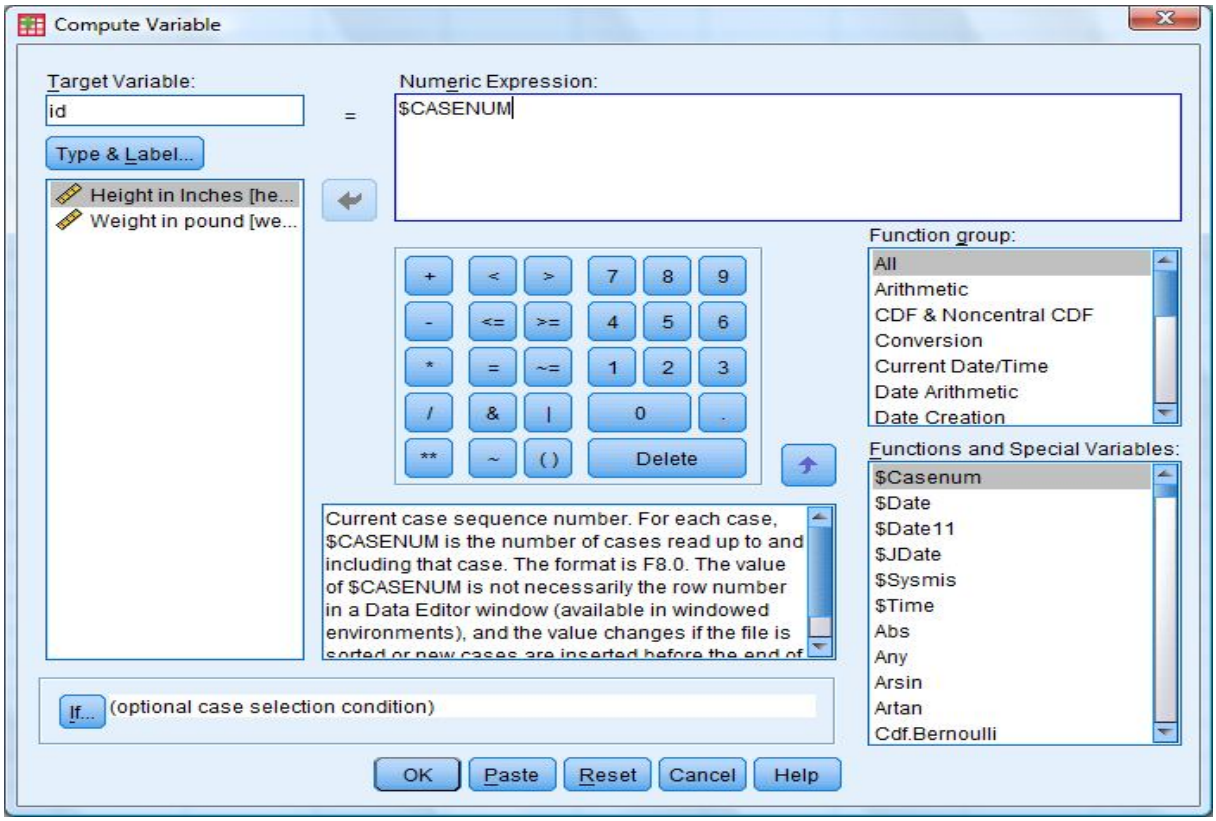
Προκύπτει μεταξύ άλλων το ακόλουθο θηκόγραμμα:

Θηκόγραμμα 1

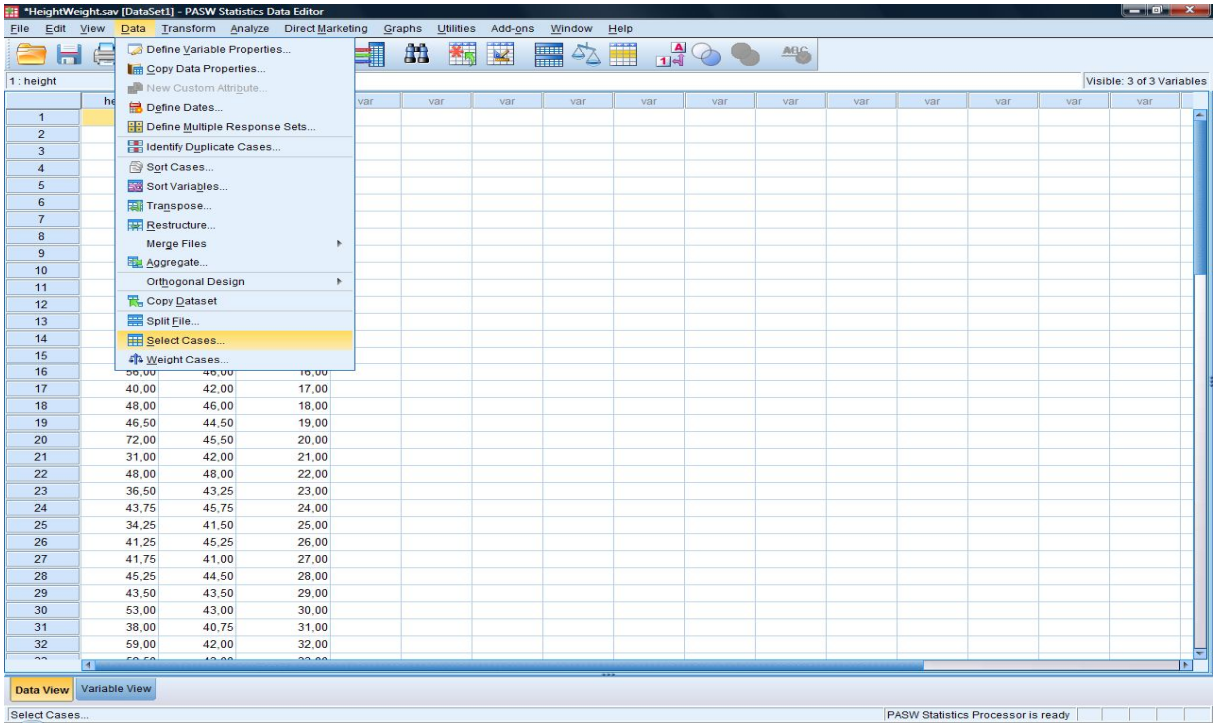


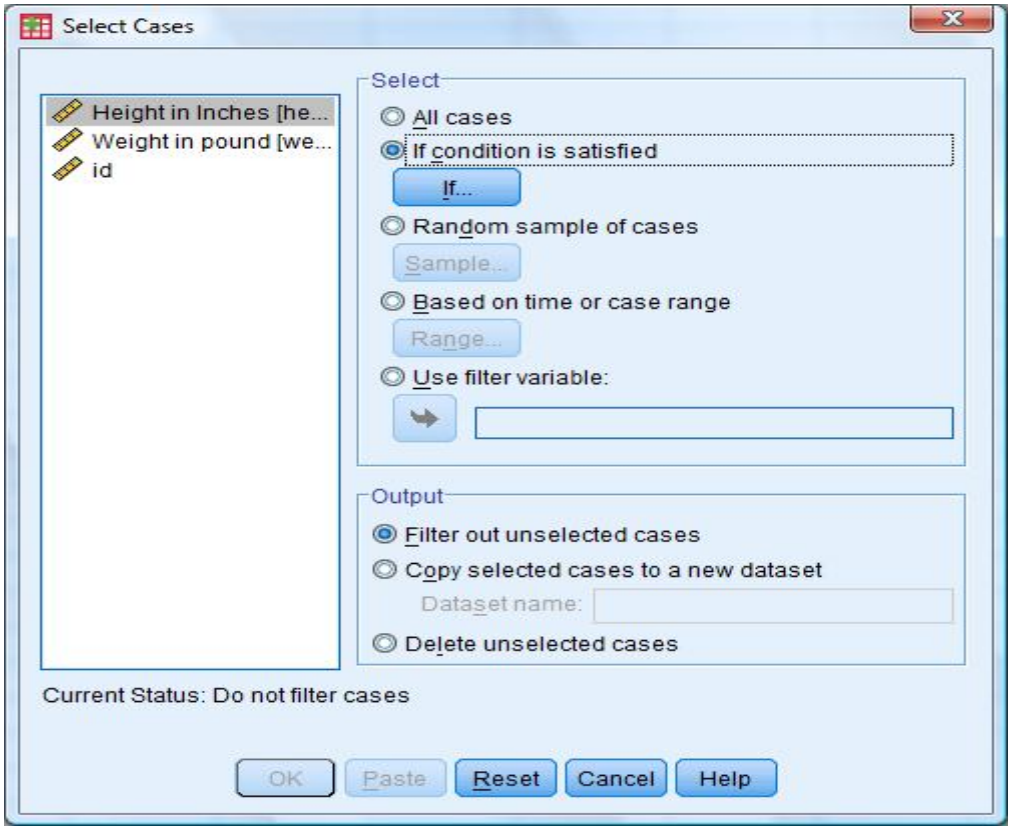
Στο σχήμα αυτό βλέπουμε ότι έχουμε μία τουλάχιστον ακραία τιμή, την παρατήρηση 20 την οποία και θα αποκλείσουμε από την περαιτέρω ανάλυση. Δημιουργούμε για αυτό το σκοπό μία νέα στήλη, μέσω της διαδικασίας Transform Compute Variable, που θα μας δίνει τον αύξοντα αριθμό των παρατηρήσεων,.

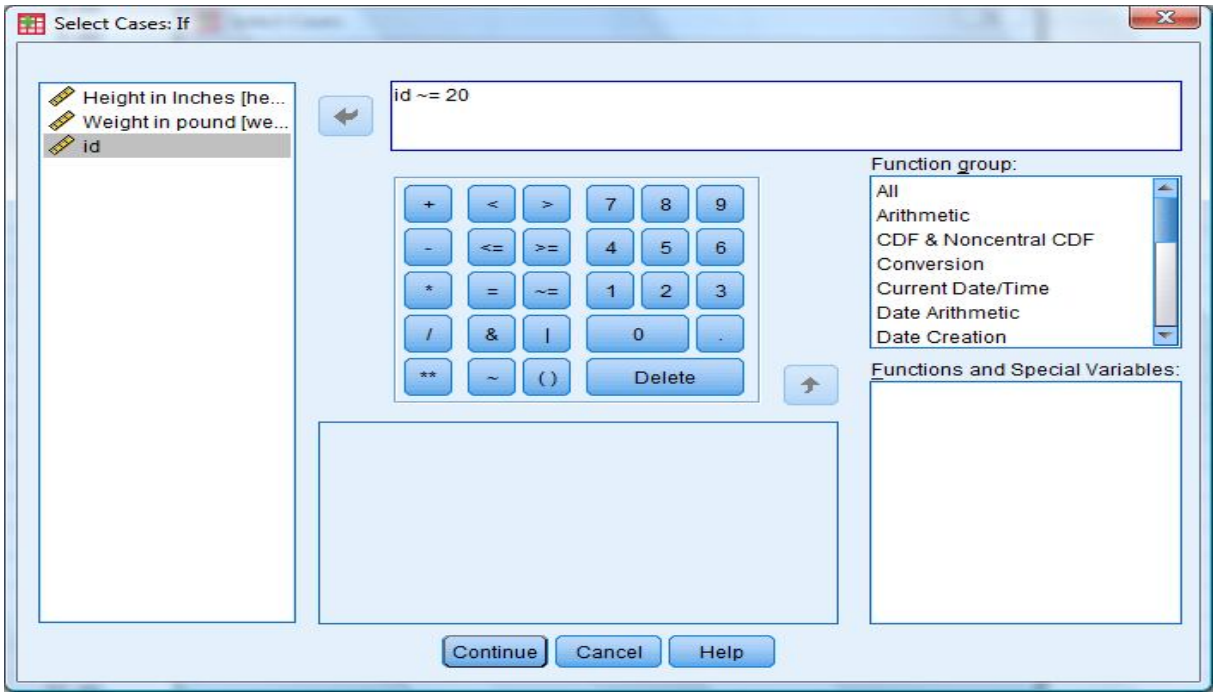




Με τη βοήθεια της στήλης αυτής αποκλείουμε από την περαιτέρω ανάλυση την παρατήρηση 20 μέσω της διαδικασίας Data → Select Cases ως εξής:

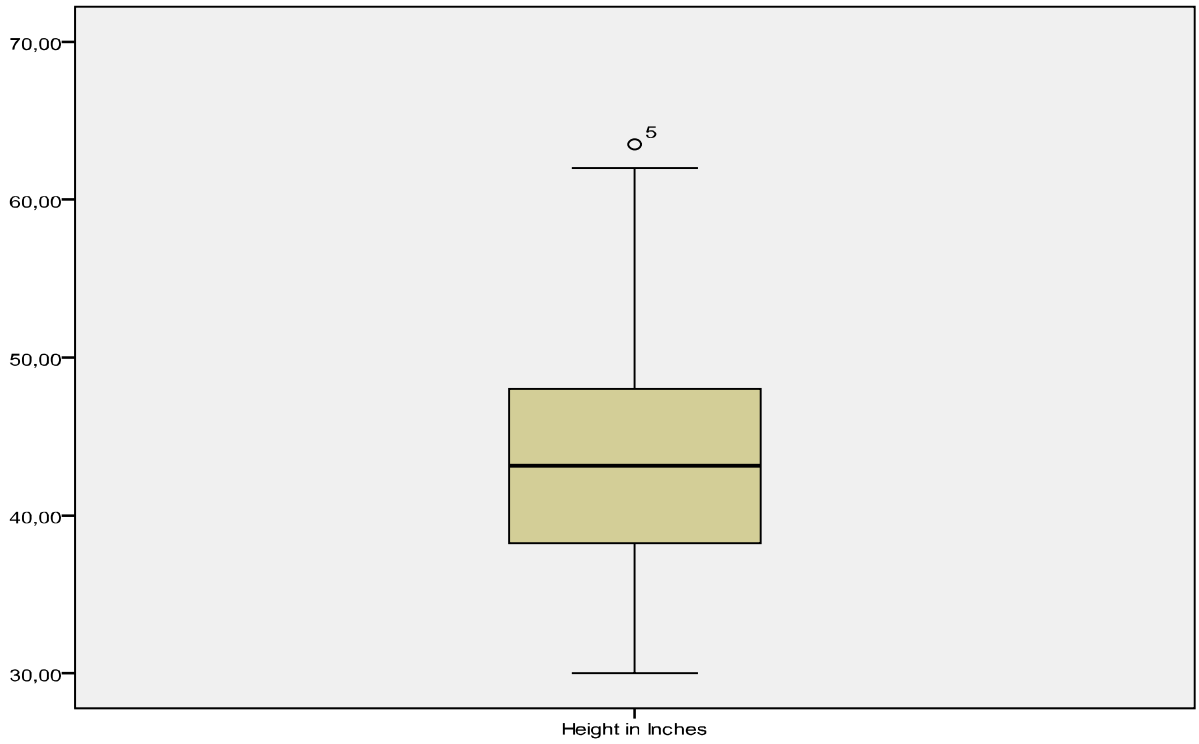






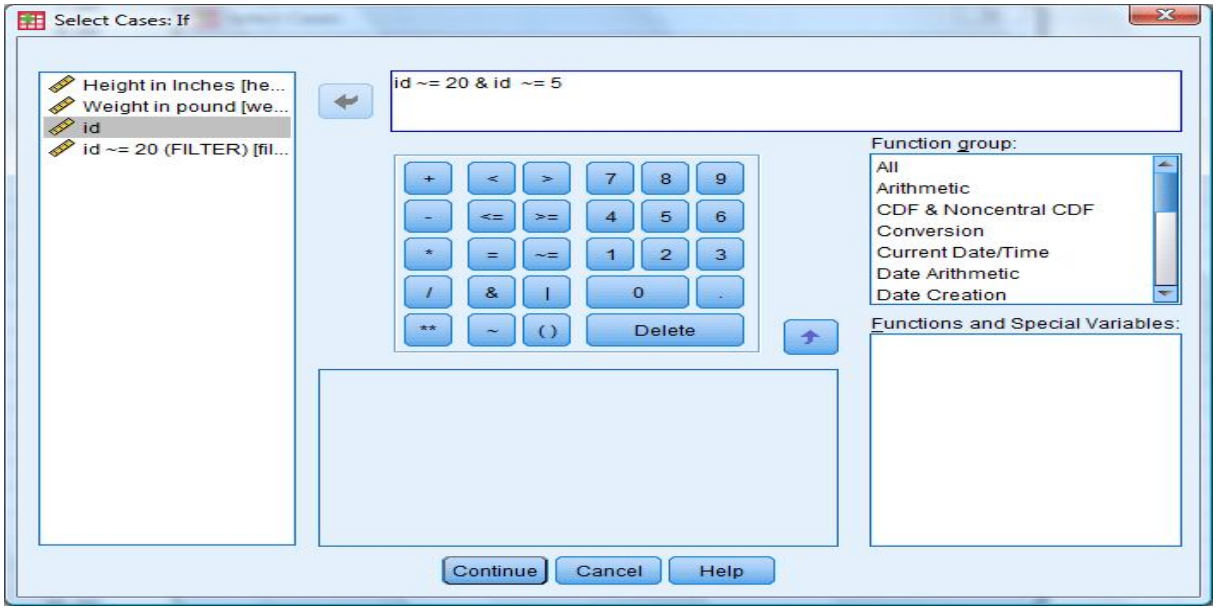
Έπειτα επαναλαμβάνεται ο έλεγχος των ακραίων τιμών και προκύπτει το ακόλουθο θηκόγραμμα:

Θηκόγραμμα 2



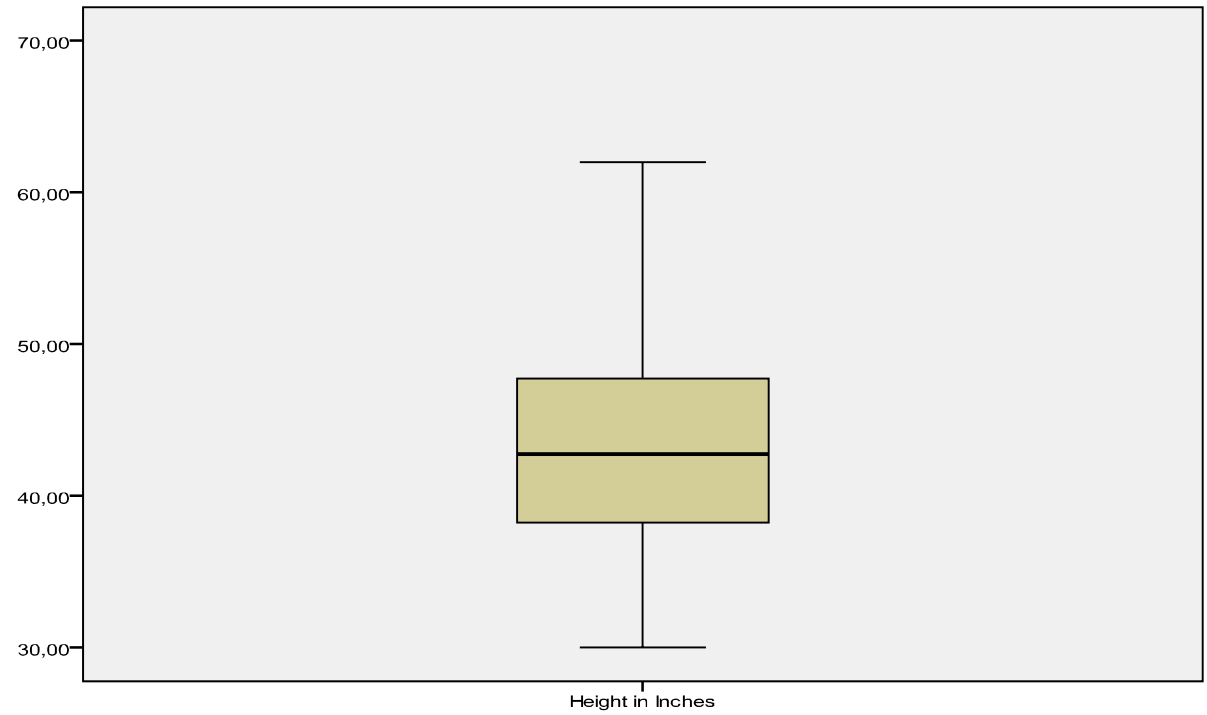
Επομένως καθώς είναι ακραία τιμή πρέπει να αποκλειστεί και η παρατήρηση 5 ($2^{η}$ ακραία σε σύνολο 73 παρατηρήσεων, ποσοστό μικρότερο του 10%) από την περαιτέρω ανάλυση.

Επιλέγουμε μέσω της διαδικασίας Data → Select Cases τα ακόλουθα:



Έπειτα επαναλαμβάνεται η διαδικασία ελέγχου ακραίων τιμών.

Θηκόγραμμα 3



Στην περίπτωση αυτή βλέπουμε ότι δεν υπάρχουν άλλες ακραίες τιμές.

Έλεγχος κανονικής κατανομής: Το αποτέλεσμα αυτού του ελέγχου είναι διαθέσιμο ήδη αφού στο προηγούμενο βήμα ζητάμε ταυτόχρονα και γραφικούς και στατιστικούς τρόπους ελέγχου της υπόθεσης της κανονικότητας.

Tests of Normality

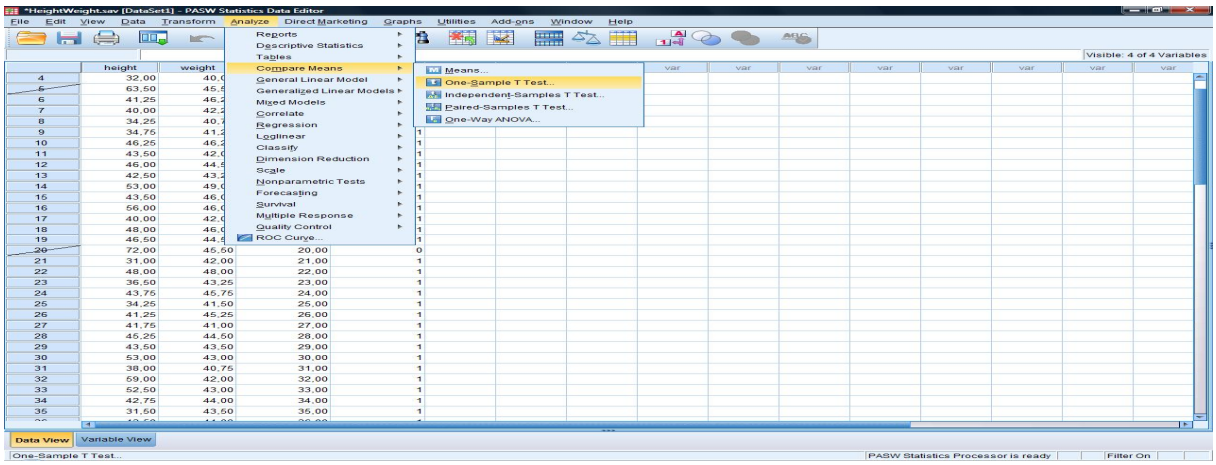
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Height in Inches	,078	71	,200*	,969	71	,079

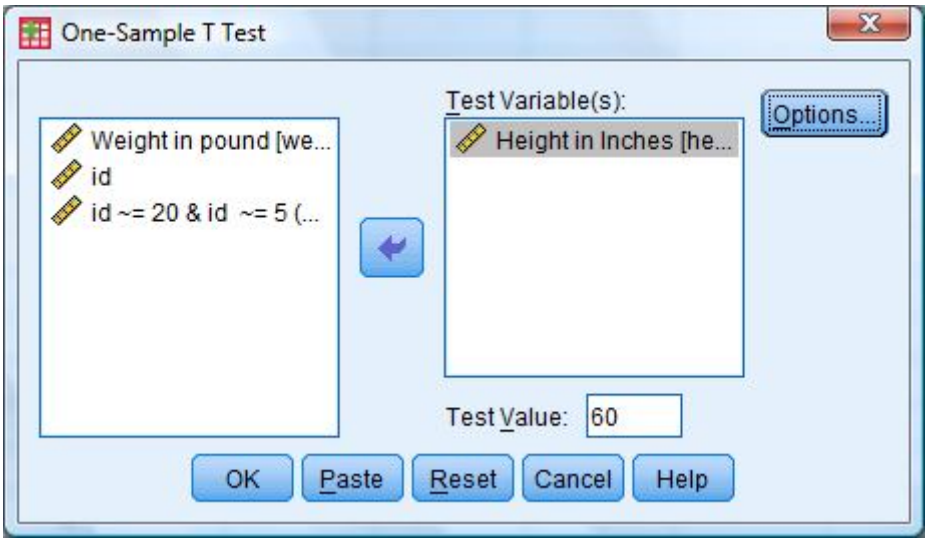
a. Lilliefors Significance Correction

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

Η τιμή που κοιτάζουμε στον πίνακα αυτό, είναι η η κρίσιμη πιθανότητα, (p-value) στη στήλη Sig του Shapiro Wilk. Επειδή η τιμή είναι μεγαλύτερη του 5% (δηλ. του α=0,05), λέμε ότι η υπόθεση ότι οι δειγματικές τιμές του ύψους προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή δεν μπορεί να απορριφθεί. Επομένως συμπεραίνουμε ότι θα ελέγξουμε την υπόθεση ότι το μέσο ύψος του πληθυσμού είναι ίσο με 60 ίντσες παραμετρικά.

Παραμετρικός έλεγχος t-Test Ο έλεγχος διεξάγεται μέσω της διαδικασίας Analyze→Compare Means→One-Sample T Test όπως αναλυτικά περιγράφηκε στην παράγραφο 4.1.





Στο παράθυρο των αποτελεσμάτων προκύπτουν τα ακόλουθα αποτελέσματα:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Height in Inches	71	43,2359	7,36278	,87380

One-Sample Test

	Test Value = 60					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Height in Inches	-19,185	70	,000	-16,76408	-18,5068	-15,0213

Κοιτάζουμε την κρίσιμη πιθανότητα, Sig. (2-tailed). Στην περίπτωσή μας p<0,001. Άρα η τιμή αυτή είναι μικρότερη από το 5% (0,000<0,05), συνεπώς το μέσο ύψος είναι στατιστικά σημαντικά διαφορετικό από τις 60 inches και καθώς το πρόσημο της μέσης διαφοράς του

μέσου ύψους από τις 60 ίντσες είναι αρνητικό (-16,764) συμπεραίνουμε ότι το μέσο ύψος του πληθυσμού είναι στατιστικά σημαντικά μικρότερο από την τιμή των 60 inches.

Από την παραπάνω ανάλυση προκύπτει η ακόλουθη αναφορά:

Αναφορά Θέλουμε να ελέγξουμε αν η μέση τιμή του ύψους των ατόμων του πληθυσμού από τον οποίο επιλέξαμε το δείγμα μας ισούται με 60 ίντσες. Το πρόβλημα αυτό είναι ένας έλεγχος για τη μέση τιμή ενός πληθυσμού. Θα ελέγξουμε αρχικά αν ικανοποιούνται οι παρακάτω υποθέσεις χρήσης του παραμετρικού αυτού ελέγχου.

1. Το δείγμα μας είναι τυχαίο.
2. Δεν υπάρχουν ακραίες τιμές στα δεδομένα μας που ξεπερνούν σε ποσοστό το 10%.
3. Τα δεδομένα μας ακολουθούν κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε το δείγμα μας και ικανοποιείται.

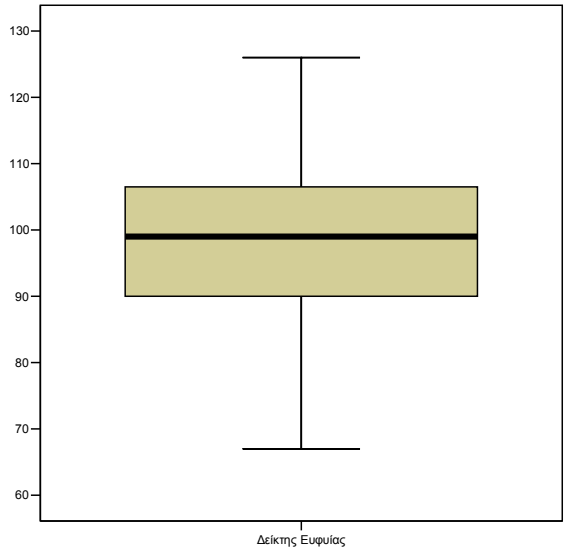
Ο έλεγχος των ακραίων τιμών έγινε με το θηκόγραμμα και έδειξε ότι υπάρχουν 2 ακραίες τιμές οι παρατηρήσεις με αύξοντα αριθμό 20 και 5 και τιμές του ύψους 72 και 63,5 ίντσες αντίστοιχα (βλέπε θηκογράμματα 1,2,3). Καθώς το ποσοστό των ακραίων τιμών (2/73*100%) δεν υπερβαίνει το 10% συνεχίζουμε την περαιτέρω ανάλυση έχοντας αποκλείσει τις δύο αυτές παρατηρήσεις. Από το τεστ των Shapiro-Wilk έχουμε ότι η υπόθεση ότι οι δειγματικές τιμές του ύψους προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την υπόθεση της κανονικής κατανομής δεν μπορεί να απορριφθεί (τιμή του τεστ 0.961, β.ε. 71, p=0,079).

Εφόσον ικανοποιούνται όλες οι προϋποθέσεις, μπορούμε να κάνουμε χρήση του παραμετρικού ελέγχου t-Test για τον έλεγχο της υπόθεσης ότι το μέσο ύψος του πληθυσμού είναι ίσο με 60 ίντσες. Από τον έλεγχο αυτό προκύπτει ότι το μέσο ύψος είναι στατιστικά σημαντικά διαφορετικό από τις 60 inches ($p<0,001$) και καθώς το πρόσημο της μέσης διαφοράς του μέσου ύψους από τις 60 ίντσες είναι αρνητικό (-16,764) συμπεραίνουμε ότι το μέσο ύψος του πληθυσμού είναι στατιστικά σημαντικά μικρότερο από την τιμή των 60 inches. Επιπλέον ένα 95% Δ.Ε. για το μέσο ύψος του πληθυσμού είναι το (60-18.5068,60-15.0213).

<u>Παράδειγμα 2ο</u> Στο αρχείο GeneralExample.sav* (βλέπε φάκελο DataII 1Mean t-test) καταγράφονται μεταξύ άλλων ο δείκτης ευφυΐας 35 ατόμων τυχαία επιλεγμένων από τον υπό μελέτη πληθυσμό. Θέλουμε να ελέγξουμε, αν είναι εφικτό, αν ο μέσος δείκτης ευφυΐας του πληθυσμού είναι στατιστικά σημαντικά διαφορετικός από τις 100 μονάδες.

Η υλοποίηση είναι ανάλογη του προηγούμενου παραδείγματος και παραλείπεται. Στη συνέχεια θα δοθεί η αναφορά των ακόλουθων αποτελεσμάτων

Θηκόγραμμα 1



Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
iq	,087	35	,200(*)	,986	35	,932

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

One-Sample Test

	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
iq	-1,250	34	,220	-3,057	-8,03	1,91

Αναφορά: Θέλουμε να ελέγξουμε αν η μέση τιμή του δείκτη ευφυΐας για τον πληθυσμό από τον οποίο επιλέξαμε το δείγμα μας ισούται με 100. Το πρόβλημα αυτό είναι ένας έλεγχος για τη μέση τιμή ενός πληθυσμού. Θα ελέγξουμε αρχικά αν ικανοποιούνται οι παρακάτω υποθέσεις χρήσης του παραμετρικού αυτού ελέγχου.

1. Το δείγμα μας είναι τυχαίο.
2. Δεν υπάρχουν ακραίες τιμές στα δεδομένα μας που ξεπερνούν σε ποσοστό το 10%.
3. Τα δεδομένα μας ακολουθούν κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε το δείγμα μας και ικανοποιείται.

Ο έλεγχος των ακραίων τιμών έγινε με το θηκόγραμμα και έδειξε ότι δεν υπάρχουν (βλέπε θηκόγραμμα 1) τέτοιες στις δειγματικές τιμές που καταγράφεται ο δείκτης ευφυΐας των 35 ατόμων. Καθώς δεν υπάρχουν ακραίες τιμές συνεχίζουμε την περαιτέρω ανάλυση ελέγχοντας την υπόθεση ότι οι διαθέσιμες δειγματικές παρατηρήσεις που καταγράφεται ο δείκτης ευφυΐας προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή. Από το τεστ των Shapiro-Wilk έχουμε ότι η υπόθεση ότι οι δειγματικές τιμές του δείκτη ευφυΐας προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την υπόθεση της κανονικής κατανομής δεν μπορεί να απορριφθεί (τιμή του τεστ 0.986, β.ε. 35, p=0,932).

Εφόσον ικανοποιούνται όλες οι προϋποθέσεις, μπορούμε να κάνουμε χρήση του παραμετρικού ελέγχου t-Test για τον έλεγχο της υπόθεσης ότι ο μέσος δείκτης ευφυΐας του πληθυσμού είναι ίσος με 100. Από τον έλεγχο αυτό προκύπτει ότι ο μέσο δείκτης ευφυΐας δε διαφέρει στατιστικά σημαντικά από το 100 (p=0,220), ενώ ένα 95% Δ.Ε. για το μέσο δείκτη ευφυΐας είναι το (100-8.03,100+1.91).

<u>Παράδειγμα 3^{ο}</u> Στο αρχείο HeightWeight15.sav* (βλέπε φάκελο DataII 2Mean t-test) καταγράφονται οι τιμές του βάρους (σε λίμπρες) και του ύψους (σε ίντσες) 39 τυχαία επιλεγμένων ατόμων από έναν πληθυσμό. Θέλουμε να ελέγξουμε, αν είναι εφικτό, αν το μέσο βάρος του πληθυσμού είναι στατιστικά σημαντικά διαφορετικό από τις 70 λίμπρες.

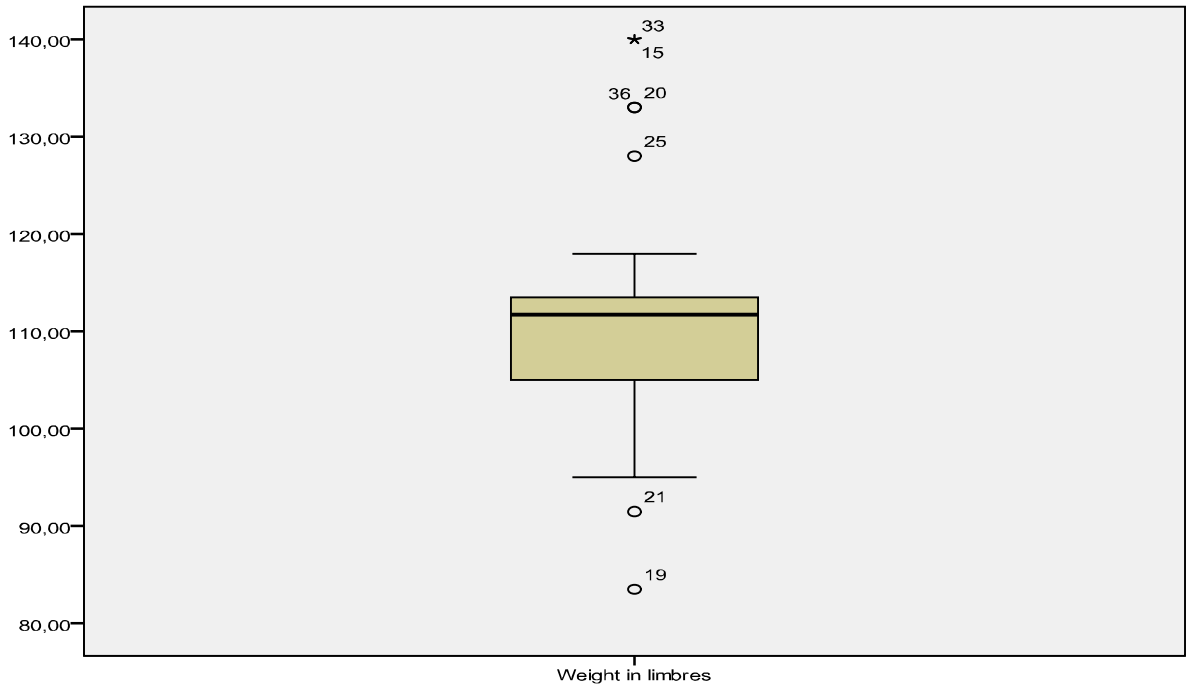
Υλοποίηση-Αποτελέσματα:

Έλεγχος ακραίων τιμών. Ο έλεγχος για την ύπαρξη ακραίων τιμών γίνεται με το θηκόγραμμα, όπως αναλυτικά περιγράφηκε στο Παράδειγμα 1, μέσω της διαδικασίας Analyze Descriptive Statistics Explore.

Θηκόγραμμα 1

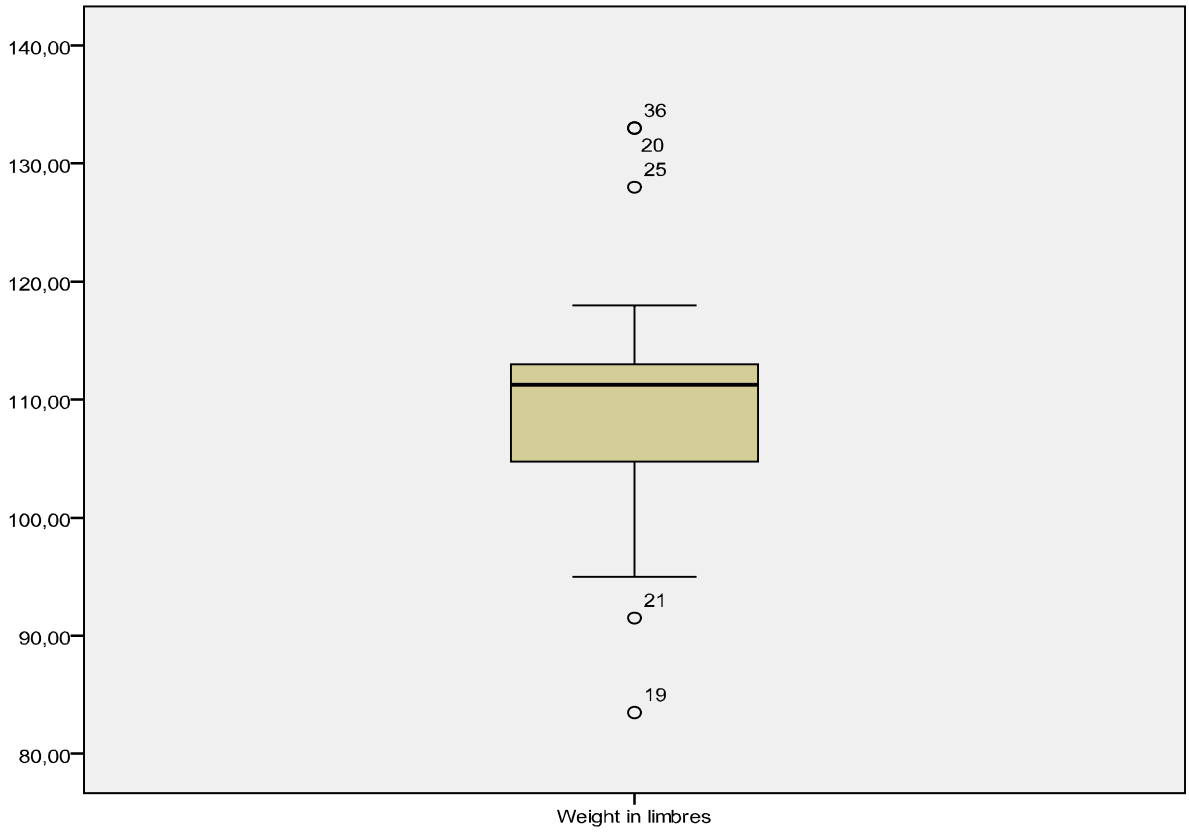
Καθώς η παρατήρηση 35 είναι εκείνη που είναι πιο απομακρυσμένη από τους φράκτες (whiskers) θα είναι αυτή που αρχικά αποκλείουμε από την περαιτέρω ανάλυση με τον τρόπο που περιγράφηκε στο Παράδειγμα 1. Έπειτα επαναλαμβάνουμε τον έλεγχο ύπαρξης ακραίων τιμών. Από το νέο θηκόγραμμα προκύπτει ότι οι παρατηρήσεις με αύξοντα αριθμό 33 και 15 είναι επίσης ακραίες. Καθώς το ποσοστό των ακραίων παρατηρήσεων δεν ξεπέρασε ακόμη το 10% των διαθέσιμων παρατηρήσεων τις αποκλείουμε και προβαίνουμε σε επανέλεγχο για την ύπαρξη ακραίων τιμών.

Θηκόγραμμα 2



Το νέο θηκόγραμμα που προκύπτει είναι το:

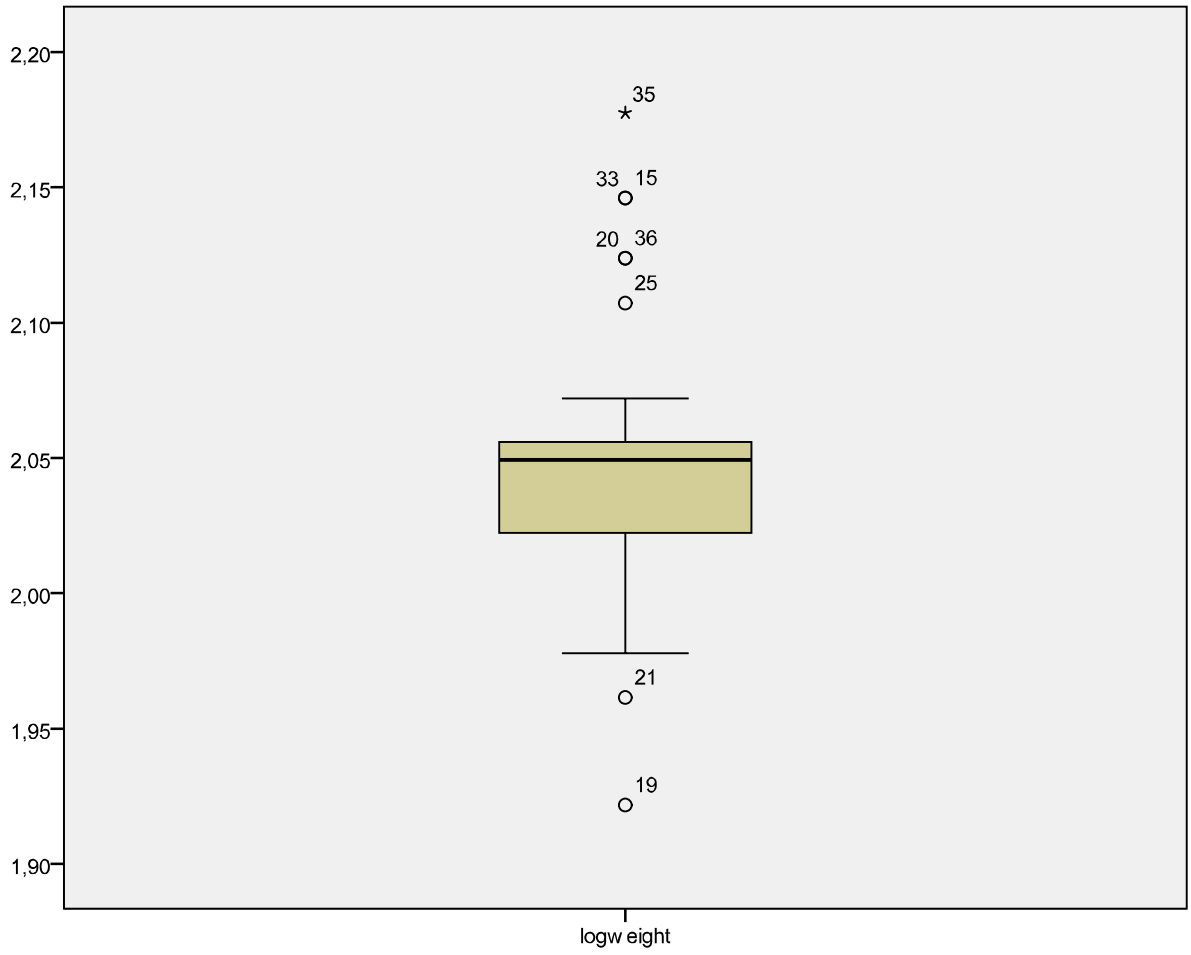
Θηκόγραμμα 3



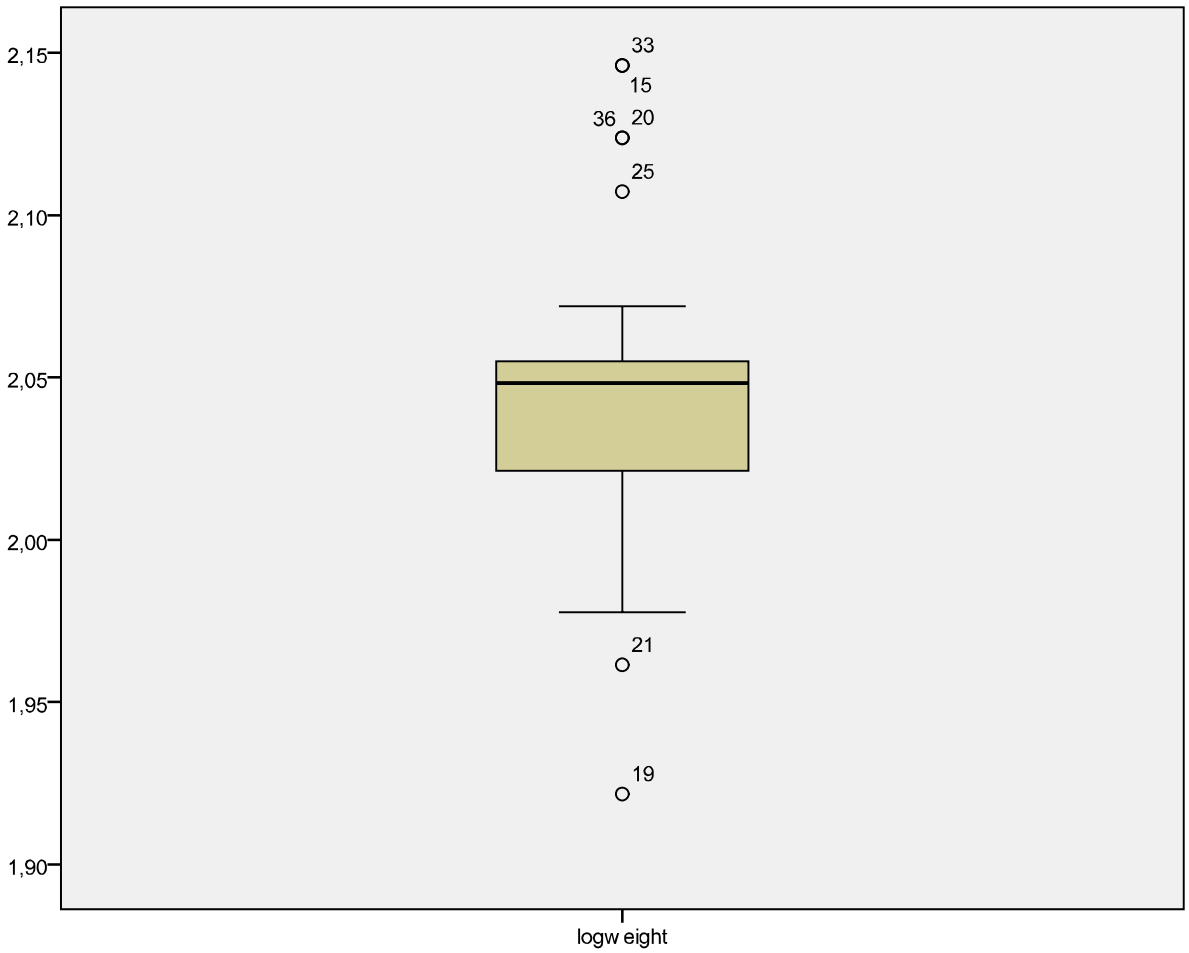
Επομένως προκύπτει ότι οι παρατηρήσεις με αύξοντα αριθμό 36 και 20 είναι ακραίες. Επομένως πλέον έχουμε διαπιστώσει ότι υπάρχουν τουλάχιστον 5 ακραίες τιμές οι παρατηρήσεις με αύξοντα αριθμό 35,33,15,36,20 και το ποσοστό τους ξεπερνά το 10% (καθώς 5/39*100%>10%). Αφού τις επαναφέρουμε όλες τις παρατηρήσεις θα εξετάσουμε αν ο μετασχηματισμός του λογαρίθμου διορθώνει το πρόβλημα. Προσοχή: Πάντοτε πριν μετασχηματίσουμε τα δεδομένα μας κάνοντας χρήση της συνάρτησης του λογαρίθμου θα πρέπει λόγω ορισμού της να ελέγχουμε αν περιέχονται στα δεδομένα μη θετικές τιμές. Σε μία τέτοια περίπτωση αν X είναι η μεταβλητή που θα μετασχηματιστεί και α η μικρότερη μη θετική τιμή τότε προτείνεται ο μετασχηματισμός $\log\left(X+|a|+1\right)$.

Μετασχηματίζοντας τα δεδομένα και μέσω της διαδικασίας Explore προκύπτουν τα ακόλουθα θηκογράμματα με παρόμοιο τρόπο αποκλείοντας μία-μία τις ακραίες τιμές.

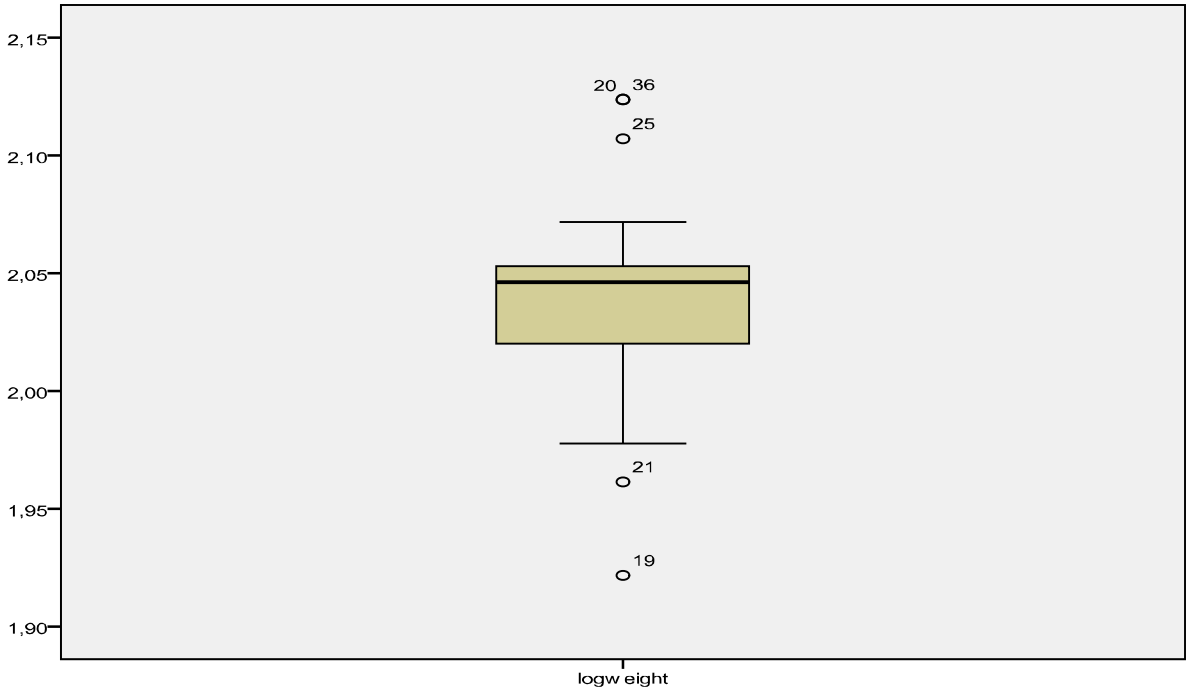
Θηκόγραμμα 4



Θηκόγραμμα 5



Θηκόγραμμα 6



Επομένως ο μετασχηματισμός του λογαρίθμου δε διορθώνει το πρόβλημα καθώς το ποσοστό αυτών στα μετασχηματισμένα δεδομένα είναι μεγαλύτερο του 10%.

Άρα θα προβούμε στον μη παραμετρικό έλεγχο ότι η πληθυσμιακή διάμεσος του βάρους είναι 70 λίμπρες μέσω της διαδικασίας Analyze→Nonparametric Tests→One Sample που αναλυτικά περιγράφηκε προηγούμενα, χρησιμοποιώντας προφανώς όλες τις παρατηρήσεις και τα αρχικά δεδομένα.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of Weight in limbres equals 70.	One-Sample Wilcoxon Signed Ranks Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Καθώς ο παραπάνω πίνακας μας πληροφορεί αν η πληθυσμιακή διάμεσος του βάρους είναι στατιστικά σημαντικά διαφορετική από 70 λίμπρες, από τη διαδικασία Analyze Descriptive Statistics Explore θα διαπιστώσουμε αν τα αποτελέσματα που αφορούν την πληθυσμιακή διάμεσο μπορούν να γενικευτούν στην πληθυσμιακή μέση τιμή, εξετάζοντας αν τα δεδομένα είναι συμμετρικά.

Descriptives

			Statistic	Std. Error
Weight in limbres	Mean		112,2051	2,10084
	95% Confidence Interval for Mean	Lower Bound	107,9522	
		Upper Bound	116,4581	
	5% Trimmed Mean		111,7400	
	Median		112,0000	
	Variance		172,128	
	Std. Deviation		13,11975	
	Minimum		83,50	
	Maximum		150,50	
	Range		67,00	
	Interquartile Range		9,00	
	Skewness		,933	,378
	Kurtosis		1,754	,741

Αναφορά: Θέλουμε να ελέγξουμε αν το μέσο βάρος του πληθυσμού είναι 70 λίμπρες. Το πρόβλημα αυτό είναι ένας έλεγχος για τη μέση τιμή ενός πληθυσμού. Θα ελέγξουμε αρχικά αν ικανοποιούνται οι παρακάτω υποθέσεις χρήσης του παραμετρικού αυτού ελέγχου.

1. Το δείγμα μας είναι τυχαίο
2. Δεν υπάρχουν ακραίες τιμές στα δεδομένα μας που ξεπερνούν σε ποσοστό το 10%.
3. Τα δεδομένα μας ακολουθούν κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε το δείγμα μας και ικανοποιείται.

Ο έλεγχος των ακραίων τιμών έδειξε ότι έχουμε μεγάλο αριθμό ακραίων παρατηρήσεων που ξεπερνούν σε ποσοστό το 10%. Συγκεκριμένα υπάρχουν τουλάχιστον 5 ακραίες παρατηρήσεις, οι δειγματικές παρατηρήσεις με αύξοντα αριθμό 35,33,15,20,36 (βλέπε θηκογράμματα 1,2,3). Ο μετασχηματισμός του λογαρίθμου (που πραγματοποιείται επαναφέροντας όλες τις παρατηρήσεις που είχαν αποκλειστεί) δε διορθώνει το πρόβλημα των ακραίων τιμών (βλέπε θηκογράμματα 4,5,6). Συγκεκριμένα ακραίες παρατηρήσεις, με την σειρά που εμφανίσθηκαν, ήταν οι δειγματικές παρατηρήσεις με αύξοντα αριθμό 35,33,15,20,36. Για το λόγο αυτό θα καταφύγουμε στον μη παραμετρικό έλεγχο της υπόθεσης ότι η πληθυσμιακή διάμεσος του βάρους ισούται με 70 λίμπρες.

Επειδή η κρίσιμη πιθανότητα είναι $p<0,001$ συμπεραίνουμε ότι η διάμεσος του βάρος του πληθυσμού είναι στατιστικά σημαντικά διαφορετική από 70 λίμπρες. Καθώς η δειγματική μέση τιμή του βάρους είναι 112,2051 λίμπρες ενώ η δειγματική διάμεσος είναι 112, τα αποτελέσματα γενικεύονται για τη μέση τιμή και ειδικότερα συμπεραίνουμε ότι το μέσο βάρος του πληθυσμού είναι στατιστικά σημαντικά μεγαλύτερο από 70 λίμπρες.

<u>Παράδειγμα 4ο</u> Στο αρχείο TVAdv.sav* καταγράφεται το ποσό που δαπανούν 21 τυχαία επιλεγμένες εταιρείες ενός πληθυσμού σε διαφημίσεις. Να ελεγχτεί, αν είναι εφικτό, αν οι μέσες δαπάνες των εταιρειών διαφέρουν στατιστικά σημαντικά από τα 185 εκατομμύρια δολάρια (τα δεδομένα δίνονται σε εκατομμύρια δολάρια).

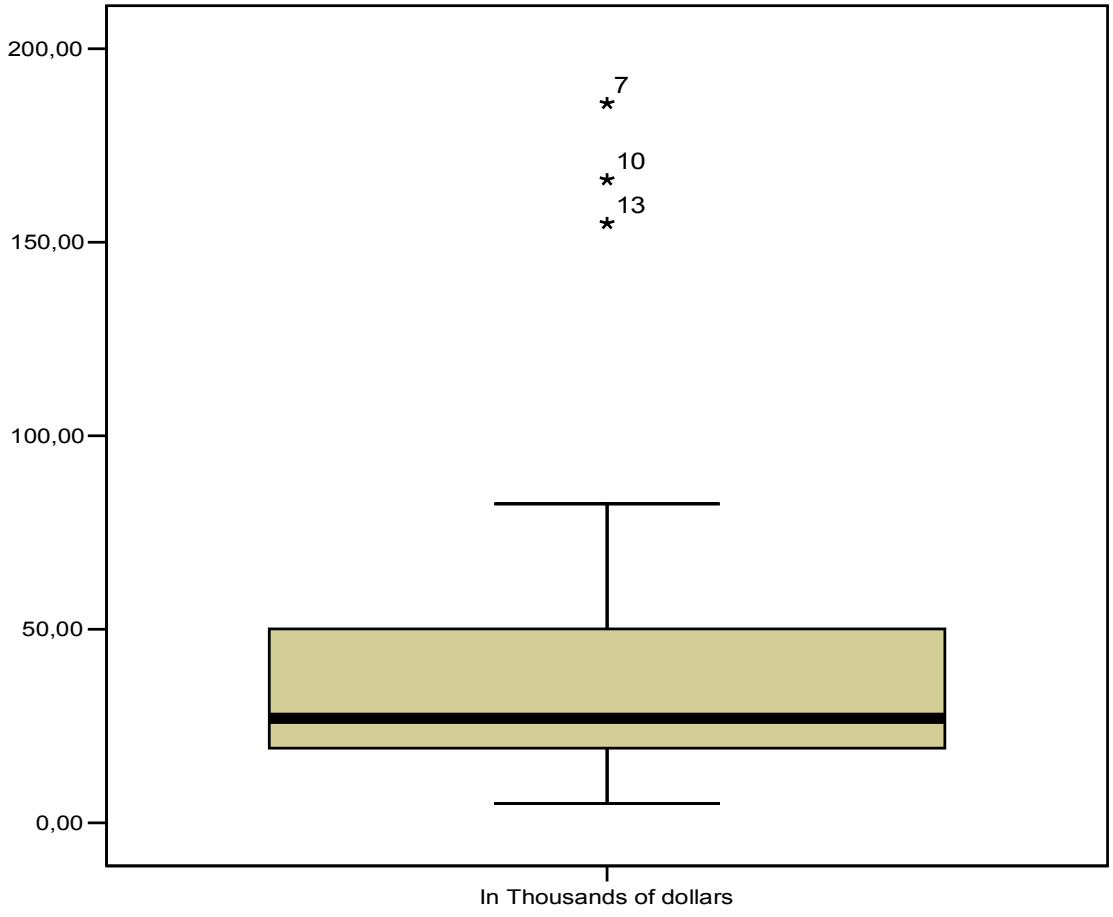
Αναφορά: Θέλουμε να ελέγξουμε αν κατά μέσο όρο το ποσό που δαπανούν οι εταιρείες ισούται με 185 εκατομμύρια δολάρια. Το πρόβλημα αυτό είναι ένας έλεγχος για τη μέση τιμή ενός πληθυσμού. Για να το ελέγξουμε θα χρησιμοποιήσουμε το t-Test για έναν πληθυσμό εφόσον ικανοποιούνται οι εξής προϋποθέσεις:

1. Το δείγμα μας είναι τυχαίο
2. Δεν υπάρχουν ακραίες τιμές στα δεδομένα μας.
3. Τα δεδομένα μας ακολουθούν κανονική κατανομή.

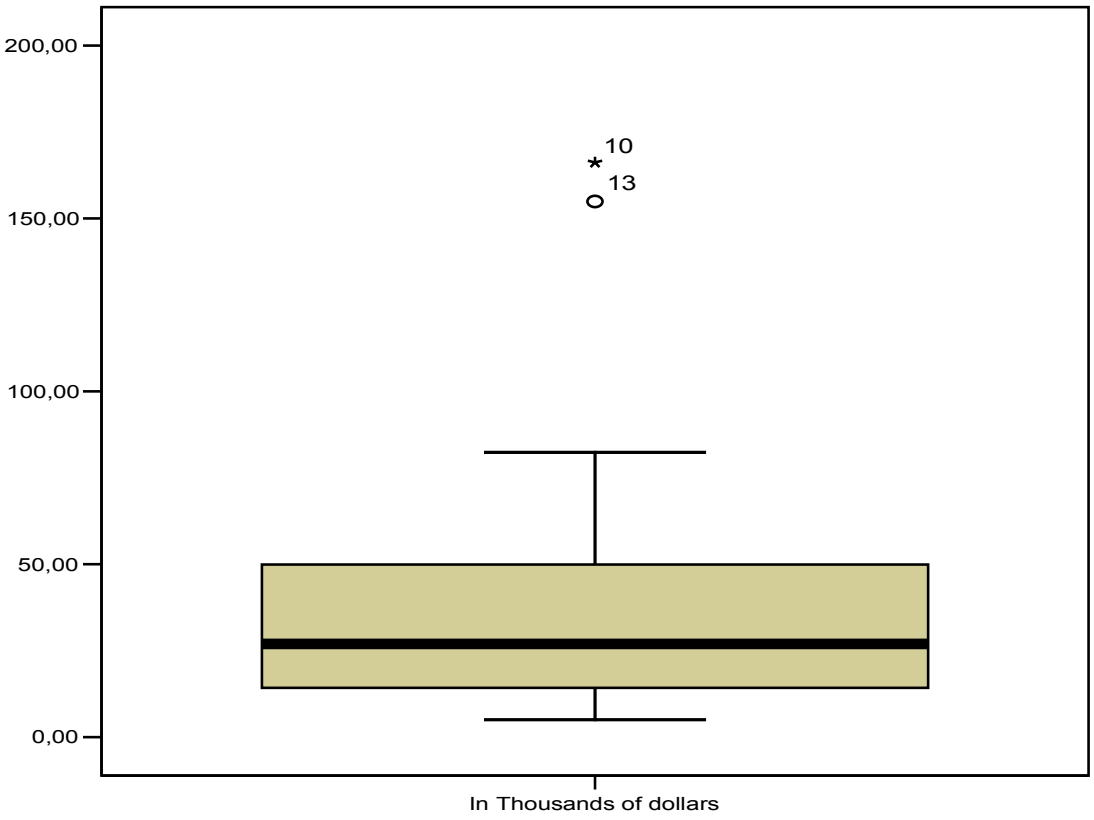
Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε το δείγμα μας και ικανοποιείται.

Ο έλεγχος των ακραίων τιμών έδειξε ότι ο αριθμός των ακραίων παρατηρήσεων υπερβαίνει το 10% του μεγέθους του δείγματος (3/21*100%>10%), καθώς υπάρχουν τουλάχιστον 3 ακραίες τιμές οι παρατηρήσεις με αύξοντα αριθμό 7,10,13 και τιμές στις δαπάνες 185.9, 166.20 και 154.90 αντίστοιχα (βλέπε θηκογράμματα 1,2,3).

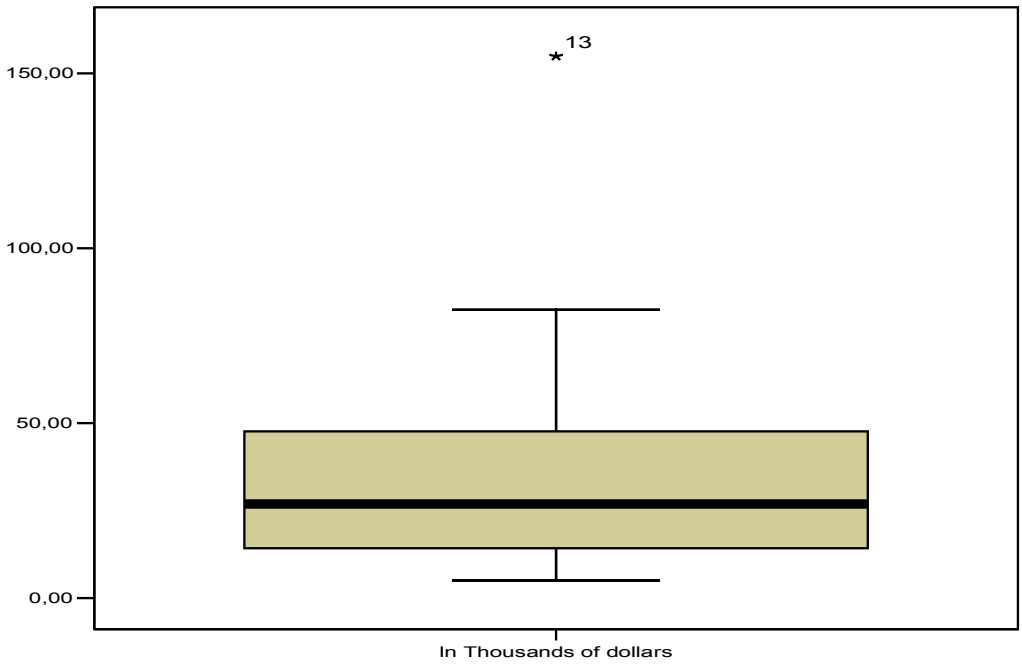
Θηκόγραμμα 1



Θηκόγραμμα 2



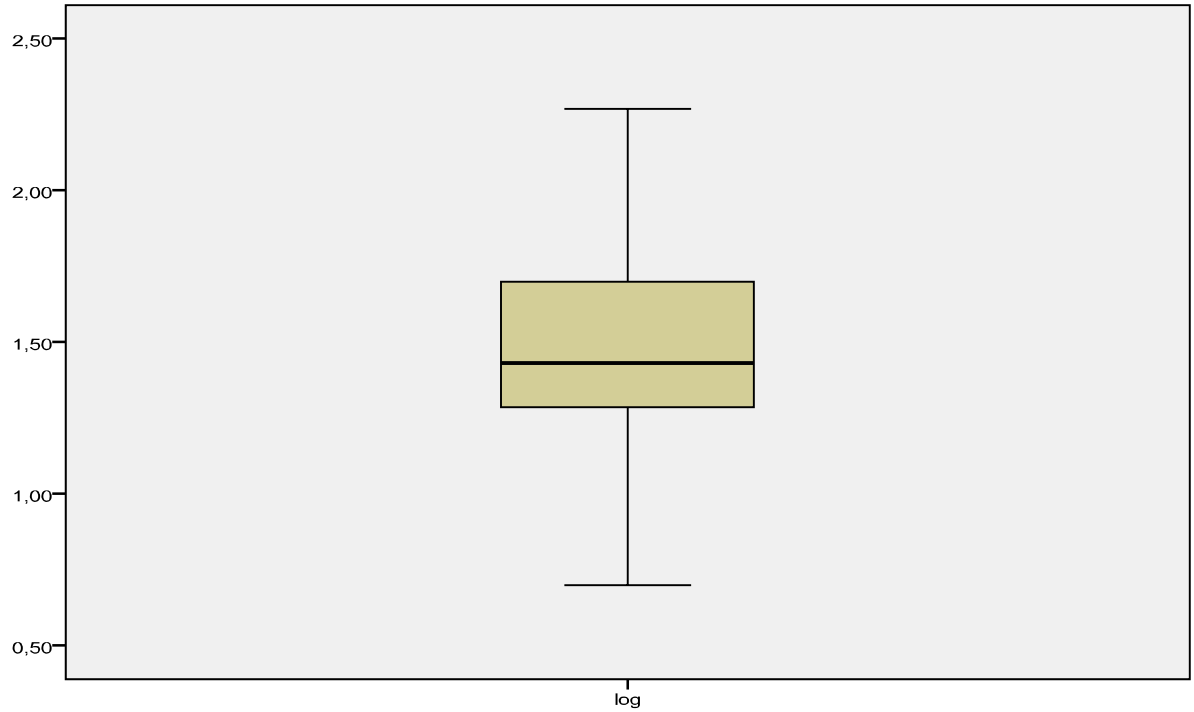
Θηκόγραμμα 3



Για τον λόγο αυτό εξετάζουμε αν ο μετασχηματισμός του λογαρίθμου θα διορθώσει το πρόβλημα αφού πρώτα επαναφέρουμε τις ακραίες παρατηρήσεις που πρωτύτερα έχουν αποκλειστεί. Ο μετασχηματισμός του λογαρίθμου διορθώνει το πρόβλημα καθώς από το

θηκόγραμμα που προέκυψε (βλέπε θηκόγραμμα 4) συμπεραίνουμε ότι δεν υπάρχουν ακραίες παρατηρήσεις στις δειγματικές παρατηρήσεις του λογαρίθμου των δαπανών των εταιρειών.

Θηκόγραμμα 4



Καθώς δεν υπάρχουν ακραίες τιμές στις δειγματικές παρατηρήσεις του λογαρίθμου των δαπανών συνεχίζουμε την περαιτέρω ανάλυση ελέγχοντας την υπόθεση ότι οι διαθέσιμες δειγματικές παρατηρήσεις που καταγράφεται ο λογάριθμος των δαπανών προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή. Από το τεστ των Shapiro-Wilk έχουμε ότι η υπόθεση ότι οι δειγματικές τιμές των δαπανών για διαφημίσεις των εταιρειών προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την υπόθεση της κανονικής κατανομής δε μπορεί να απορριφθεί (p=0,381).

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
log	,107	21	,200*	,953	21	,381

a. Lilliefors Significance Correction

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

Εφόσον ικανοποιούνται όλες οι προϋποθέσεις, μπορούμε να κάνουμε χρήση του παραμετρικού ελέγχου t-Test για τον έλεγχο της υπόθεσης ότι ο μέσος λογάριθμος των δαπανών είναι ίσος με το δεκαδικό λογάριθμο του 185, δηλαδή με 2.267. Από τον έλεγχο αυτό προκύπτει ότι ο μέσος λογάριθμος των δαπανών διαφέρει στατιστικά σημαντικά από το 2.267 ($p<0.001$), ενώ ένα 95% Δ.Ε. για το μέσο λογάριθμο των δαπανών είναι το (2.267-1,0082, 2.267-0,5768).

<u>Παράδειγμα 5^{ο}</u> Στο αρχείο HeightWeight15.sav* καταγράφονται οι τιμές του βάρους και του ύψους (σε ίντσες) τυχαία επιλεγμένων ατόμων από έναν πληθυσμό. Θέλουμε να ελέγξουμε, αν είναι εφικτό, αν το μέσο ύψος του πληθυσμού είναι στατιστικά σημαντικά διαφορετικό από τις 65 inches.

Η υλοποίηση παραλείπεται και αφήνεται ως άσκηση ενώ δίδεται μία συνοπτική αναφορά.

Συνοπτική Αναφορά*:* Θέλουμε να ελέγξουμε αν η μέση τιμή του ύψους για τον πληθυσμό, από τον οποίο επιλέξαμε το δείγμα μας, ισούται με 65 ίντσες. Το πρόβλημα αυτό είναι ένας έλεγχος για τη μέση τιμή ενός πληθυσμού. Για να το ελέγξουμε θα χρησιμοποιήσουμε το t-Test για έναν πληθυσμό εφόσον ικανοποιούνται οι εξής προϋποθέσεις:

1. Το δείγμα μας είναι τυχαίο
2. Δεν υπάρχουν ακραίες τιμές στα δεδομένα μας.
3. Τα δεδομένα μας ακολουθούν κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε το δείγμα μας και ικανοποιείται.

Ο έλεγχος των ακραίων τιμών έγινε με το θηκόγραμμα και έδειξε ότι δεν υπάρχουν ακραίες τιμές.

Ο έλεγχος της κανονικής κατανομής με το τεστ των Shapiro-Wilk έδειξε ότι η υπόθεση αυτή απορρίπτεται για επίπεδο σημαντικότητας 5%, ενώ δεν μπορεί να απορριφθεί για επίπεδο σημαντικότητας 1% ($p=0,037$). Συνέπεια αυτού είναι στη συνεχεία να χρησιμοποιούμε το 1% σαν επίπεδο σημαντικότητας, καθώς θέλουμε να αποφύγουμε τόσο το μετασχηματισμό των δεδομένων όσο και τη χρήση του Κ.Ο.Θ.

Εφόσον ικανοποιούνται όλες οι προϋποθέσεις, μπορούμε να κάνουμε χρήση του t-Test. Επειδή p=0,020 το συμπέρασμα από τη χρήση του τεστ αυτού είναι ότι η υπόθεση πως το μέσο ύψος του πληθυσμού είναι στατιστικά σημαντικά ίσο με 65 ίντσες, δε μπορεί να απορριφθεί για επίπεδο σημαντικότητας 1%.

ΚΕΦΑΛΑΙΟ ΠΕΜΠΤΟ

Έλεγχος για τις παραμέτρους θέσης δύο πληθυσμών με ανεξάρτητα δείγματα

Θέλοντας να εξετάσουμε τις μέσες τιμές δύο πληθυσμών πρέπει να διακρίνουμε κατά τα γνωστά από τη θεωρία δύο περιπτώσεις ανάλογα με το αν τα δείγματα είναι ανεξάρτητα ή εξαρτημένα. Στο κεφάλαιο αυτό θα ασχοληθούμε με την περίπτωση που τα δείγματα είναι ανεξάρτητα.

Σχόλιο: Ο καθορισμός των δύο δειγμάτων στην περίπτωση ανεξάρτητων δειγμάτων στο λογισμικό γίνεται με τη βοήθεια μίας ποιοτικής μεταβλητής που διαχωρίζεται σε δύο κατηγορίες (π.χ. άνδρας-γυναίκα) και χρησιμεύει για τον καθορισμό των δύο υποθετικών πληθυσμών. Από την άλλη μεριά, ο καθορισμός των δύο εξαρτημένων δειγμάτων στο λογισμικό γίνεται με τη βοήθεια δύο στηλών. Στη μία καταγράφονται για παράδειγμα οι τιμές της υπό μελέτης ποσοτικής μεταβλητής πριν την εφαρμογή της μεθόδου, ενώ στην άλλη οι τιμές της μετά την εφαρμογή της μεθόδου. Λεπτομέρειες σχετικά με τα εξαρτημένα δείγματα θα δοθούν στο επόμενο κεφάλαιο.

Έστω ένα τυχαίο δείγμα $X_1,\ldots,X_n$ μεγέθους n από έναν πληθυσμό με μέση τιμή μ_1 και διακύμανση σ_1^2, άγνωστη. Επιπλέον έστω ένα τυχαίο δείγμα $Y_1,\ldots,Y_m$ μεγέθους m από έναν πληθυσμό με μέση τιμή μ_2 και διακύμανση σ_2^2, άγνωστη. Επιπρόσθετα υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Ενδιαφερόμαστε για τον έλεγχο, σε επίπεδο σημαντικότητας α, της μηδενικής υπόθεσης

$$H_0 : \mu_1 = \mu_2,$$

ως προς μία εκ των

$$H_a : \mu_1 > \mu_2, \quad H_a : \mu_1 < \mu_2, \; H_a : \mu_1 \neq \mu_2.$$

Το παραπάνω πρόβλημα ελέγχεται υπό κάποιες υποθέσεις με τον παραμετρικό έλεγχο του t-test. Όταν κάποια από τις υποθέσεις αυτές δεν ικανοποιείται και δεν υπάρχει τρόπος διόρθωσης του προβλήματος ο έλεγχος ανάγεται σε αυτόν ότι οι πληθυσμιακές

διάμεσοι είναι ίσες. Τα αποτελέσματα του τελευταίου ελέγχου γενικεύονται για τον δοθέν έλεγχο όταν τα δεδομένα είναι συμμετρικά.

5.1 Μεθοδολογία-Υλοποίηση στο S.P.S.S.

Η μεθοδολογία που θα χρησιμοποιηθεί για τη στατιστική ανάλυση ενός τέτοιου προβλήματος εξαρτάται από το αν πληρούνται ή όχι κάποιες προϋποθέσεις, τις οποίες και πρέπει αρχικά να ελέγξει ο ερευνητής. Πιο συγκεκριμένα, ελέγχουμε
α) αν το ποσοστό των ακραίων τιμών στις διαθέσιμες δειγματικές παρατηρήσεις από καθένα από τους δύο το πλήθος πληθυσμούς ξεπερνά το 10% αυτών, και
β) αν οι πληθυσμοί από τους οποίους λαμβάνονται τα τυχαία δείγματα μπορούμε να ισχυριστούμε ότι περιγράφονται ικανοποιητικά από την κανονική κατανομή.

Ανάλογα με τα αποτελέσματα των παραπάνω ελέγχων προβαίνουμε στον παραμετρικό έλεγχο του t test ή στο μη παραμετρικό έλεγχο (Wilcoxon-Mann-Whitney).

Από τα παραπάνω ίσως έγινε ήδη αντιληπτό ότι κομβικό σημείο για τον τρόπο διεξαγωγής του υπό μελέτη ελέγχου αποτελεί η διενέργεια των προκαταρκτικών ελέγχων α) και β), με βάση τα αποτελέσματα των οποίων θα αποφανθούμε αν θα προχωρήσουμε παραμετρικά ή μη παραμετρικά. Για το λόγο αυτό στη συνέχεια παρουσιάζονται όλα τα πιθανά αποτελέσματα των α) και β), τα διάφορα βήματα της ανάλυσης και οι αποφάσεις στις οποίες οδηγούμαστε.

1. Αρχικά ελέγχουμε αν υπάρχουν ακραίες τιμές στις διαθέσιμες δειγματικές τιμές καθενός από τους 2 το πλήθος πληθυσμούς. Αν το ποσοστό των ακραίων τιμών σε καθένα από τα δύο δείγματα δε ξεπερνά το 10%, τότε προχωρούμε στο βήμα 2. Αν το ποσοστό των ακραίων τιμών σε κάποιο από τα δύο δείγματα 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 έλεγχος (βλέπε βήμα 4).
2. Στο βήμα 2, χρησιμοποιώντας το τεστ των Shapiro-Wilk καθώς και γραφικούς τρόπους, ελέγχουμε αν οι διαθέσιμες δειγματικές παρατηρήσεις (είτε οι αρχικές είτε οι μετασχηματισμένες) καθενός από τους δύο πληθυσμούς προέρχονται από έναν πληθυσμό που περιγράφεται 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 τον παραμετρικό έλεγχο του t τεστ (βλέπε βήμα 3). Αν

η υπόθεση της κανονικότητας απορρίπτεται για έναν ή και για τους δύο υπό εξέταση πληθυσμούς (τεστ Shapiro-Wilk, p-τιμή <α), τότε ελέγχουμε αν το πρόβλημα της μη κανονικότητας διορθώνεται μετασχηματίζοντας κατάλληλα 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 και των δύο πληθυσμών, συνεχίζουμε την ανάλυση παραμετρικά (βήμα 3). Σε αντίθετη περίπτωση, αν το πλήθος των δειγματικών παρατηρήσεων (μη λαμβάνοντας υπόψη αυτές που έχουν αφαιρεθεί στο βήμα 1) εκείνου του πληθυσμού που δεν περιγράφεται από την κανονική κατανομή είναι μεγάλο (συνήθως μεγαλύτερο ή ίσο του 30) κάνοντας χρήση του Κεντρικού Οριακού Θεωρήματος, προβαίνουμε στον παραμετρικό έλεγχο της υπό έλεγχο υπόθεσης (βλέπε βήμα 3). Τότε η κρίσιμη πιθανότητα του ελέγχου και το διάστημα εμπιστοσύνης θα είναι προσεγγιστικά. Στην περίπτωση τώρα που το πρόβλημα της μη κανονικότητας, κάποιου ή και των δύο πληθυσμών δε διορθώνεται (τεστ Shapiro-Wilk, p-τιμή <α), και ταυτόχρονα το πλήθος των δειγματικών παρατηρήσεων από αυτόν τον πληθυσμό ή από αυτούς τους πληθυσμούς ανάλογα (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. Παραμετρικός έλεγχος t τεστ: Η στατιστική συνάρτηση που θα χρησιμοποιηθεί και οι κρίσιμες περιοχές για την υπό έλεγχο μηδενική υπόθεση καθορίζονται στη βάση της ισότητας ή μη των δύο πληθυσμιακών διακυμάνσεων.

ι) Ειδικότερα, αν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων δεν απορρίπτεται (τεστ του Levene, p-τιμή>α), χρησιμοποιείται η στατιστική συνάρτηση

$$t = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \overset{H_0}{\sim} t_{n+m-2},$$

όπου $\bar{X}$ και $\bar{Y}$ οι δειγματικές μέσες τιμές και $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$, με S_1^2, S_2^2 τις δειγματικές διακυμάνσεις. Οι κρίσιμες περιοχές του ελέγχου είναι: $t \geq t_{n+m-2,a}$, $t \leq -t_{n+m-2,a}$ και $|t| \geq t_{n+m-2,a/2}$, για τον έλεγχο της $H_0 : \mu_1 = \mu_2$, ως προς τις εναλλακτικές $H_a : \mu_1 > \mu_2$, $H_a : \mu_1 < \mu_2$, $H_a : \mu_1 \neq \mu_2$, αντίστοιχα. Επιπλέον το 100(1-α)% Δ.Ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ είναι

$$\left(\bar{X} - \bar{Y} - t_{n+m-2,a/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X} - \bar{Y} + t_{n+m-2,a/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right).$$

ιι) Αν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων απορρίπτεται (τεστ του Levene, p-τιμή<α), χρησιμοποιείται η στατιστική συνάρτηση (γνωστό ως τεστ του Welch)

$$t = \frac{\bar{X} - \bar{Y}}{S} \overset{H_0}{\sim} t_v \ ,$$

όπου $S^2 = \frac{S_1^2}{n} + \frac{S_2^2}{m}$, και $v = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$, όπου $c = \frac{S_1^2}{nS^2}$.

Οι κρίσιμες περιοχές του ελέγχου είναι: $t \geq t_{v,a}$, $t \leq -t_{v,a}$ και $|t| \geq t_{v,a/2}$, για τον έλεγχο της $H_0 : \mu_1 = \mu_2$, ως προς τις εναλλακτικές $H_a : \mu_1 > \mu_2$, $H_a : \mu_1 < \mu_2$, $H_a : \mu_1 \neq \mu_2$, αντίστοιχα. Επιπλέον το 100(1-α)% Δ.Ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ είναι

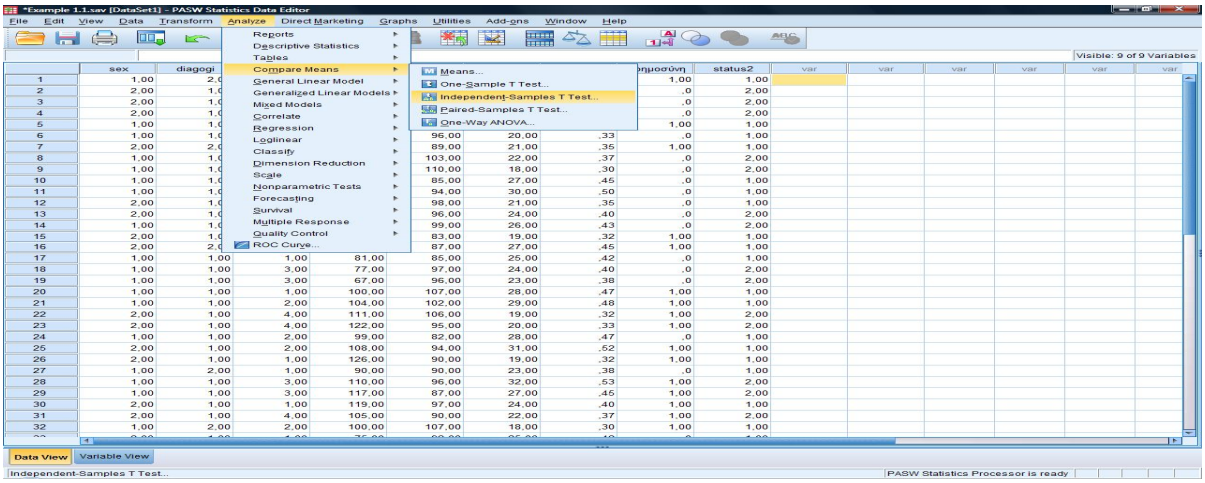
$$\left(\bar{X} - \bar{Y} - t_{v,a/2}S, \bar{X} - \bar{Y} + t_{v,a/2}S\right).$$

Επισήμανση: Σε περίπτωση που έχει χρησιμοποιηθεί κάποιος μετασχηματισμός διόρθωσης του προβλήματος είτε λόγω της ύπαρξης πολλών ακραίων τιμών είτε λόγω της απόκλισης από την κανονικότητα, τότε όλα τα παραπάνω αναφέρονται στις μετασχηματισμένες τιμές και στο τροποποιημένο σε μέγεθος δείγμα. Ειδικότερα, αν έχει χρησιμοποιηθεί ο μετασχηματισμός του λογαρίθμου, θα προβούμε στον έλεγχο αν ο μέσος λογάριθμος του ενός πληθυσμού δε διαφέρει στατιστικά σημαντικά από το μέσο λογάριθμο του άλλου.

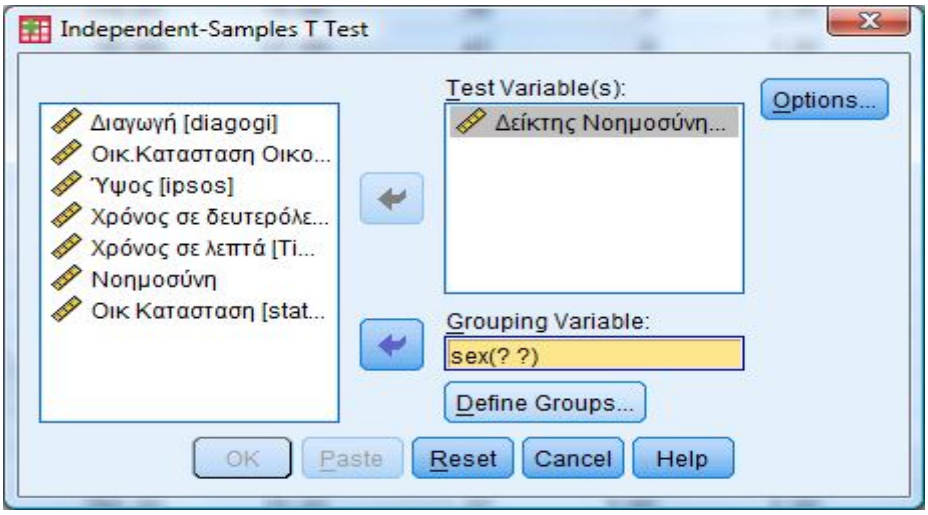
Υλοποίηση στο S.P.S.S.

Υλοποιείται από τη βασική ράβδο του λογισμικού ακολουθώντας τη διαδικασία

i. Analyze→Compare Means→Independent-Samples T Test.



ii. Στο νέο παράθυρο διαλόγου που προκύπτει διαλέγουμε τη μεταβλητή (ποσοτική) που παριστά το χαρακτηριστικό που μας ενδιαφέρει να μελετήσουμε και τη μετακινούμε στο πλαίσιο Test Variable(s).

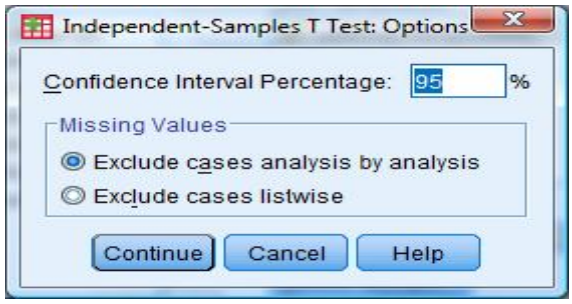


Στο πλαίσιο Grouping Variable καθορίζουμε τη μεταβλητή (ποιοτική) που διαχωρίζει τα δύο δείγματα.

Για παράδειγμα αν θέλουμε να μελετήσουμε αν υπάρχει στατιστικά σημαντική διαφορά στο μέσο δείκτη νοημοσύνης αγοριών και κοριτσιών, τοποθετούμε στα πλαίσια Test Variable και Grouping Variable, τις μεταβλητές «Δείκτης Νοημοσύνης» και «Sex», αντίστοιχα.

iii. Στη συνέχεια από το πλαίσιο Define Groups προσδιορίζουμε τον τρόπο διαχωρισμού των δύο δειγμάτων είτε από το πλαίσιο Use specified values είτε από το πλαίσιο Cut point. Αν χρησιμοποιήσουμε την πρώτη επιλογή δηλώνουμε την τιμή της ποιοτικής μεταβλητής που καθορίστηκε στο πλαίσιο Grouping Variable για καθένα από τα Group, λαμβάνοντας υπόψη ότι κάθε άλλη τιμή της θα εξαιρεθεί από την ανάλυση. Αν χρησιμοποιηθεί η δεύτερη επιλογή, η πρώτη κατηγορία θα αποτελείται από όλες εκείνες τις περιπτώσεις που αντιστοιχούν σε τιμές μικρότερες του αριθμού που δηλώσαμε (κατά αυτόν τον τρόπο μπορούμε να δηλώσουμε και ποσοτική μεταβλητή που με τη χρήση του Cut point ουσιαστικά μετατρέπεται σε ποιοτική).

iv. Από την επιλογή Options έχουμε τη δυνατότητα να καθορίσουμε τον τρόπο χειρισμού των ελλιπών τιμών καθώς και να προσδιορίσουμε το βαθμό εμπιστοσύνης του διαστήματος εμπιστοσύνης που θα κατασκευαστεί για τη διαφορά των μέσων τιμών.



Ερμηνεία αποτελεσμάτων του S.P.S.S

Από τον πίνακα Group Statistics το λογισμικό μας πληροφορεί ότι είναι διαθέσιμες 19 και 16 αντίστοιχα παρατηρήσεις για αγόρια και κορίτσια. Ο μέσος δείκτης νοημοσύνης των 19 αγοριών και 16 κοριτσιών είναι 93.4211 και 101.125 αντίστοιχα. Παρατηρούμε ότι ο μέσος δείκτης νοημοσύνης των κοριτσιών είναι μεγαλύτερος. Μένει να διαπιστώσουμε αν είναι στατιστικά σημαντικά μεγαλύτερος. Επιπλέον, στον πίνακα αυτό μας δίνονται οι τυπικές αποκλίσεις και το τυπικό σφάλμα της μέσης τιμής του δείκτη νοημοσύνης ως προς το φύλο.

Group Statistics

	Φύλο	N	Mean	Std. Deviation	Std. Error Mean
Δείκτης Νοημοσύνης	Αγόρι	19	93,4211	13,82154	3,17088
	Κορίτσι	16	101,1250	14,51379	3,62845

Στον πίνακα Independent Samples Test υλοποιείται ο έλεγχος της υπόθεσης ότι δε διαφέρει ο μέσος δείκτης νοημοσύνης των αγοριών από των κοριτσιών. Η στατιστική συνάρτηση ελέγχου καθορίζεται από την απόρριψη ή όχι της υπόθεσης της ισότητας των πληθυσμιακών διακυμάνσεων. Η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων δεν απορρίπτεται (τεστ του Levene, τιμή του F στατιστικού τεστ =0.079, p-τιμή=0.781>0.05). Ο μέσος δείκτης νοημοσύνης αγοριών και κοριτσιών δε διαφέρει στατιστικά σημαντικά (t στατιστικό, τιμή=-1.606, β.ε.=33, p-τιμή=0.118>0.05). Τέλος, ένα 95% διάστημα εμπιστοσύνης για τη μέση διαφορά του δείκτη νοημοσύνης αγοριών και κοριτσιών είναι (βλέπε 95% Confidence Interval of the Difference) είναι το (-17.46552, 2.05763).

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Δείκτης Νοημοσύνης	Equal variances assumed	,079	,781	-1,606	33	,118	-7,70395	4,79799	-17,46552	2,05763
	Equal variances not assumed			-1,599	31,399	,120	-7,70395	4,81872	-17,52674	2,11885

4. Μη παραμετρικός έλεγχος Wilcoxon-Mann-Whitney: Θέλοντας να εφαρμόσουμε το Mann-Whitney test, το οποίο είναι ισοδύναμο με το τεστ του Wilcoxon, για τον έλεγχο ότι οι δύο πληθυσμοί δε διαφέρουν ως προς την παράμετρο θέσης τα δύο δείγματα αναμιγνύονται και διατάσσονται κατά αύξουσα σειρά. Επιπλέον, υπολογίζονται οι τάξεις (ranks). Σημειώνεται ότι στην περίπτωση δεσμών οι τάξεις προκύπτουν ως ο μέσος όρος των τάξεων που θα έπαιρναν οι παρατηρήσεις αυτές αν δε διέφεραν μεταξύ τους. Αν οι πληθυσμοί είναι ίδιοι ως προς την παράμετρο θέσης, οι τάξεις θα πρέπει να είναι τυχαία αναμεμειγμένες στα δύο δείγματα. Έστω U_X (U_Y ανάλογα) ο αριθμός των φορών που μία παρατήρηση x ακολουθεί μία παρατήρηση y (ο αριθμός των φορών που μία παρατήρηση y ακολουθεί μία παρατήρηση x, ανάλογα). Τότε προκύπτει ότι: $U_X = \sum_{i=1}^{n} R_i(X_i) - \frac{n(n+1)}{2}$, και $U_Y = \sum_{j=1}^{m} R_j(Y_j) - \frac{m(m+1)}{2}$, όπου $R_i(X_i)$, $R_j(Y_j)$ οι τάξεις των $X_1,\ldots,X_n$ και $Y_1,\ldots,Y_m$. Το Mann-Whitney U στατιστικό ορίζεται από τη σχέση: $U = \min(U_X, U_Y)$. Αποδεικνύεται ότι για τιμές των n, m μεγαλύτερες ή ίσες του οκτώ το Mann-Whitney U στατιστικό ακολουθεί προσεγγιστικά μία κανονική κατανομή. Ειδικότερα, είναι γνωστό ότι προσεγγιστικά ισχύει ότι:

$$Z = \frac{U - \frac{nm}{2}}{\sqrt{\frac{nm(m+n+1)}{2}}} \overset{H_0}{\sim} N(0,1),$$

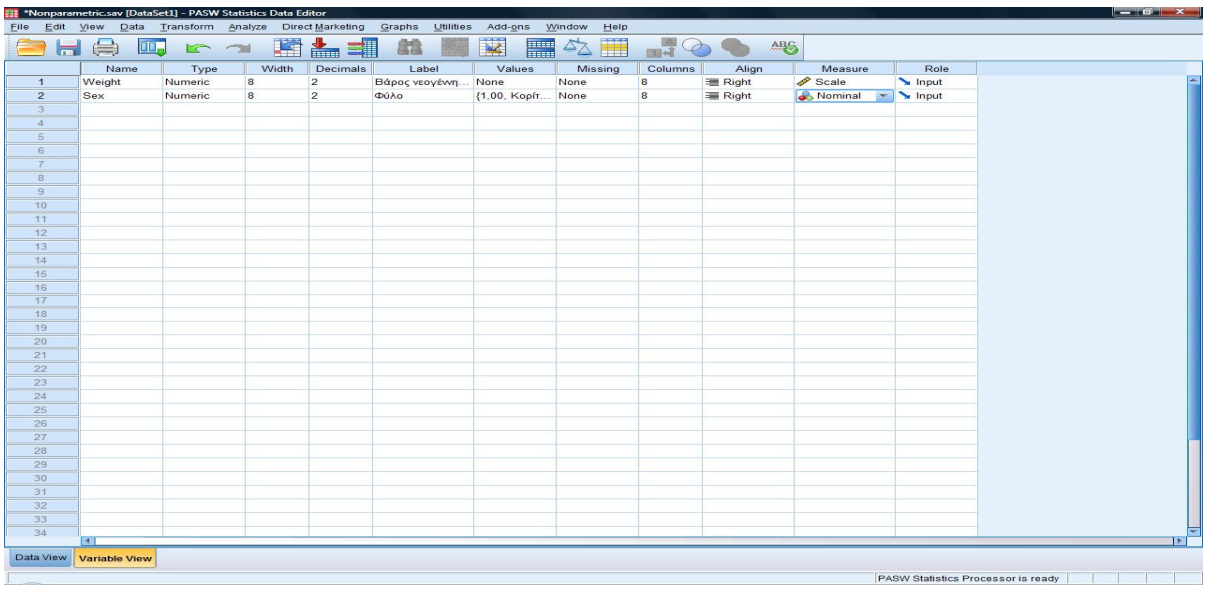
και οι κρίσιμες περιοχές για τους ελέγχους είναι: $Z \geq z_a$, $Z \leq -z_a$ και $|Z| \geq z_{a/2}$, αντίστοιχα.

Υλοποίηση στο S.P.S.S.

Στον παρακάτω πίνακα δεδομένων, που προέρχεται από το άρθρο του Peter K. Dunn (1999), δίνεται το βάρος ενός νεογέννητου σε γραμμάρια και το φύλο του (1=κορίτσι 2=αγόρι). Θέλουμε να ελέγξουμε αν το μέσο βάρος των νεογέννητων αγοριών διαφέρει από το μέσο βάρος των νεογέννητων κοριτσιών, με το τεστ Wilcoxon-Mann-Whitney (χωρίς να εξεταστεί αν είναι απαραίτητη η χρήση του).

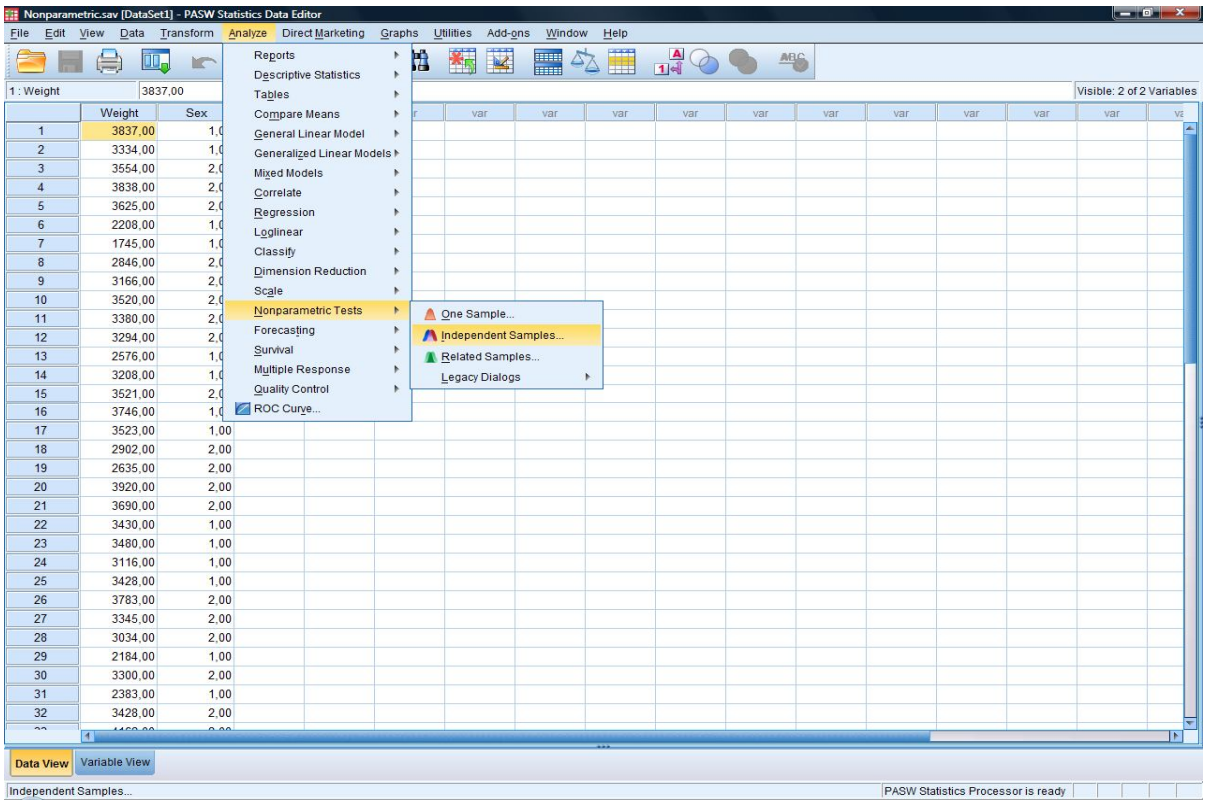
Φύλο	Βάρος	Φύλο	Βάρος	Φύλο	Βάρος
1	3837	2	2635	1	3500
1	3334	2	3920	2	3736
2	3554	2	3690	2	3370
2	3838	1	3430	2	2121
2	3625	1	3480	2	3150
1	2208	1	3116	1	3866
1	1745	1	3428	1	3542
2	2846	2	3783	1	3278
2	3166	2	3345	2	3402
2	3520	2	3034	2	2902
2	3380	1	2184	2	3406
2	3294	2	3300	1	3523
1	2576	1	2383	2	3630
1	3208	2	3428	1	3746
2	3521	2	4162		

Πριν προχωρήσουμε στην ανάλυση καθορίζουμε στο πλαίσιο Variable View ποια μεταβλητή είναι συνεχής και ποια είναι ονομαστική.

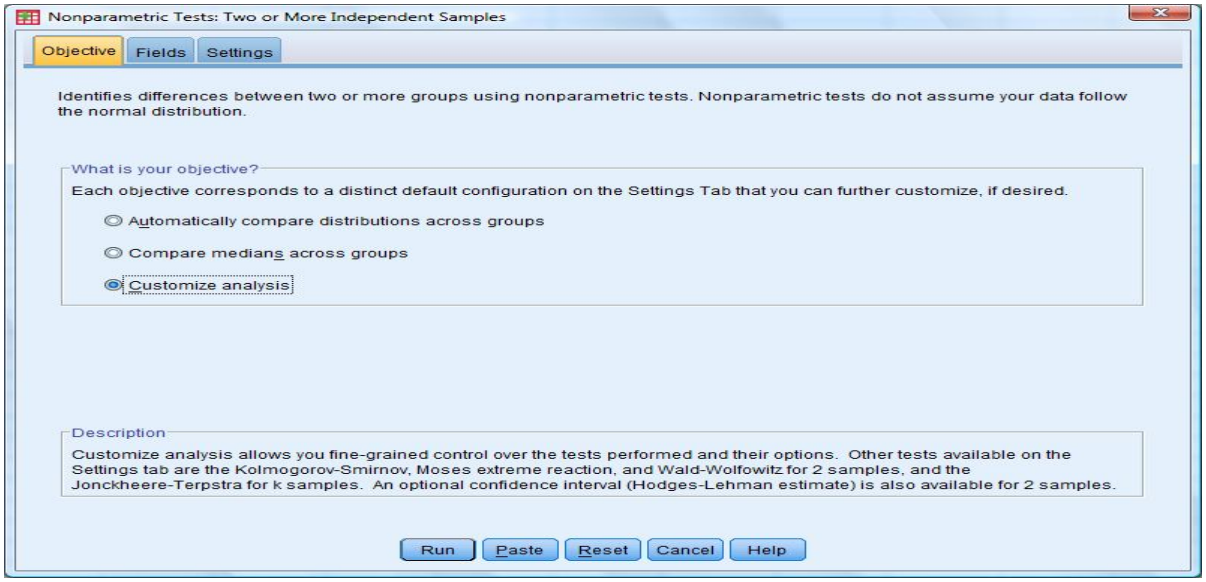


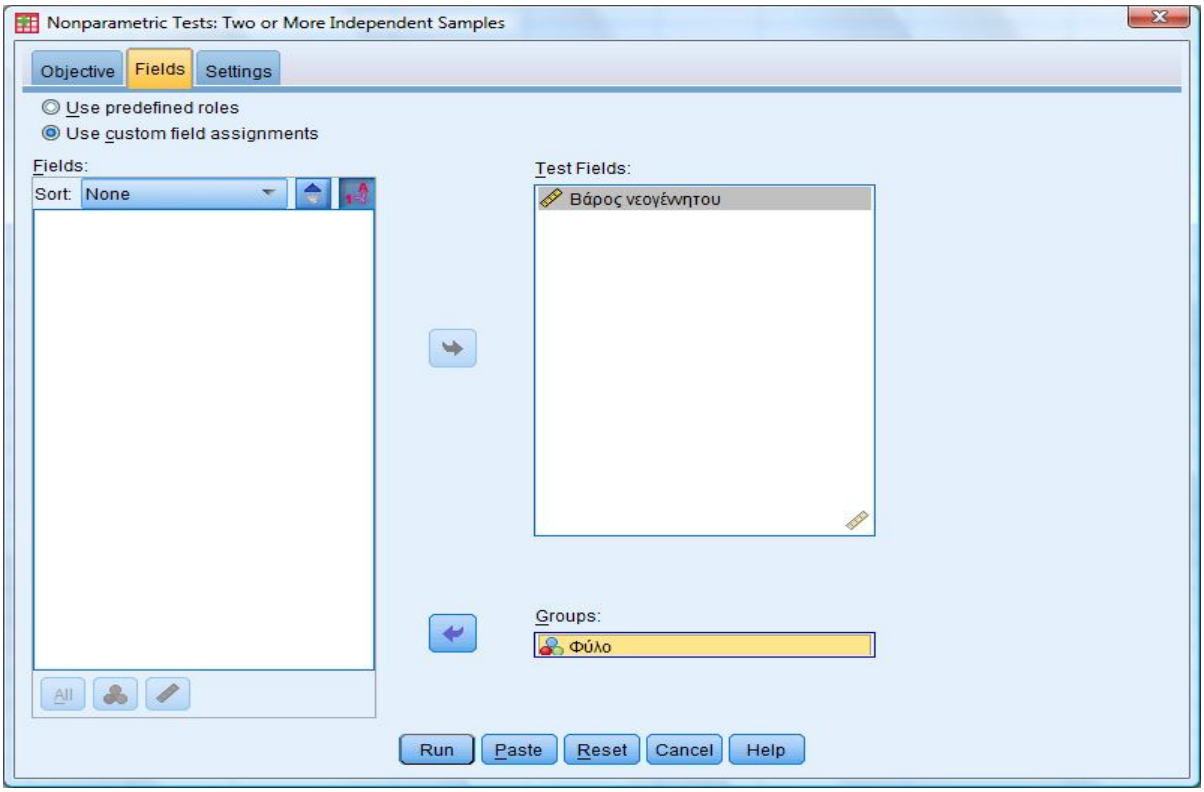
Έπειτα από το κύριο μενού επιλέγουμε

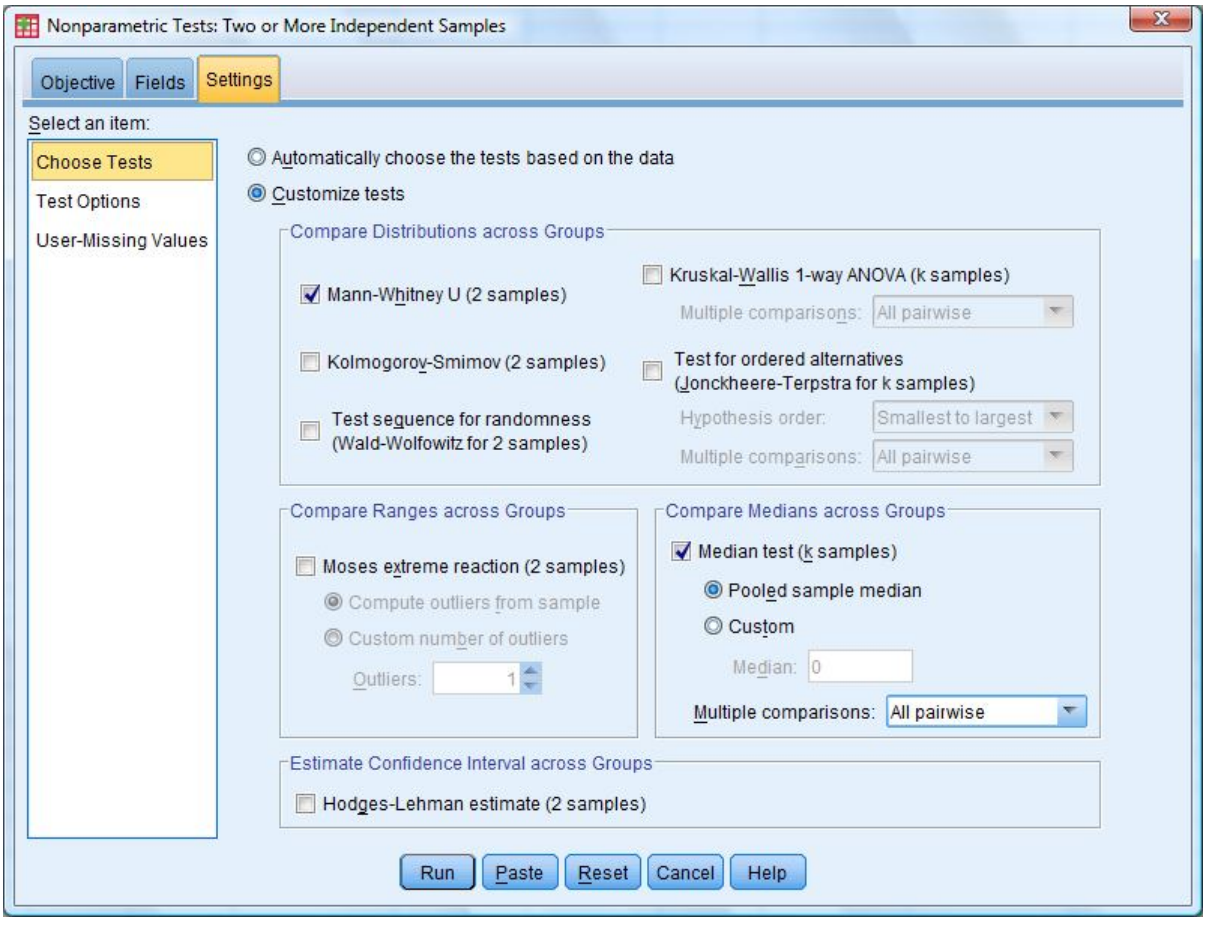
i. Analyze→Nonparametric Tests→Independent Samples.



Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε στο πλαίσιο Objective την επιλογή Customize analysis, έτσι ώστε στη συνέχεια από τα πλαίσια Fields και Settings να καθορίσουμε τον έλεγχο τον οποίο θέλουμε να διενεργηθεί όπως φαίνεται στα σχήματα που ακολουθούν.




Nonparametric Tests: Two or More Independent Samples
Objective
Fields
Settings
Use predefined roles
Use custom field assignments
Fields:
Sort: None
Test Fields:
Βάρος νεογέννητου
Groups:
Φύλο
All
Run
Paste
Reset
Cancel
Help


Nonparametric Tests: Two or More Independent Samples
Objective
Fields
Settings
Select an item:
Choose Tests
Test Options
User-Missing Values
Automatically choose the tests based on the data
Customize tests
Compare Distributions across Groups
Mann-Whitney U (2 samples)
Kruskal-Wallis 1-way ANOVA (k samples)
Multiple comparisons: All pairwise
Kolmogorov-Smimov (2 samples)
Test for ordered alternatives (Jonckheere-Terpstra for k samples)
Test sequence for randomness (Wald-Wolfowitz for 2 samples)
Hypothesis order: Smallest to largest
Multiple comparisons: All pairwise
Compare Ranges across Groups
Moses extreme reaction (2 samples)
Compute outliers from sample
Custom number of outliers
Outliers: 1
Compare Medians across Groups
Median test (k samples)
Pooled sample median
Custom
Median: 0
Multiple comparisons: All pairwise
Estimate Confidence Interval across Groups
Hodges-Lehman estimate (2 samples)
Run
Paste
Reset
Cancel
Help

Δηλαδή τοποθετούμε στο πλαίσιο Test Field(s) την υπό μελέτη ποσοτική μεταβλητή (έστω το Βάρος νεογέννητου), ενώ στο πλαίσιο Groups την ποιοτική μεταβλητή (έστω το Φύλο), η οποία μας διαχωρίζει τους δύο πληθυσμούς.

Στο παράθυρο διαλόγου Settings μας δίνεται η δυνατότητα να διαλέξουμε τον τύπο του μη παραμετρικού ελέγχου που θέλουμε να διενεργηθεί. Επιλέγουμε το πλαίσιο Mann-Whitney U και την επιλογή Median Test (k samples). Το πρώτο στατιστικό τεστ όπως μας πληροφορεί το Help του στατιστικού πακέτου ελέγχει την υπόθεση ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό ενώ το δεύτερο την υπόθεση της ισότητας των πληθυσμιακών διαμέσων.

Ερμηνεία αποτελεσμάτων

Από τον παρακάτω πίνακα προκύπτει ότι δεν υπάρχουν στατιστικά σημαντικές διαφορές στη διάμεσο του βάρους νεογέννητων αγοριών και κοριτσιών, καθώς είναι p-τιμή=0,759>0,05. Το ερώτημα τώρα είναι πότε τα αποτελέσματα αυτά γενικεύονται για το μέσο βάρος;

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The medians of Βάρος νεογέννητου are the same across categories of Φύλο.	Independent-Samples Median Test	,759	Retain the null hypothesis.
2	The distribution of Βάρος νεογέννητου is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,346	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

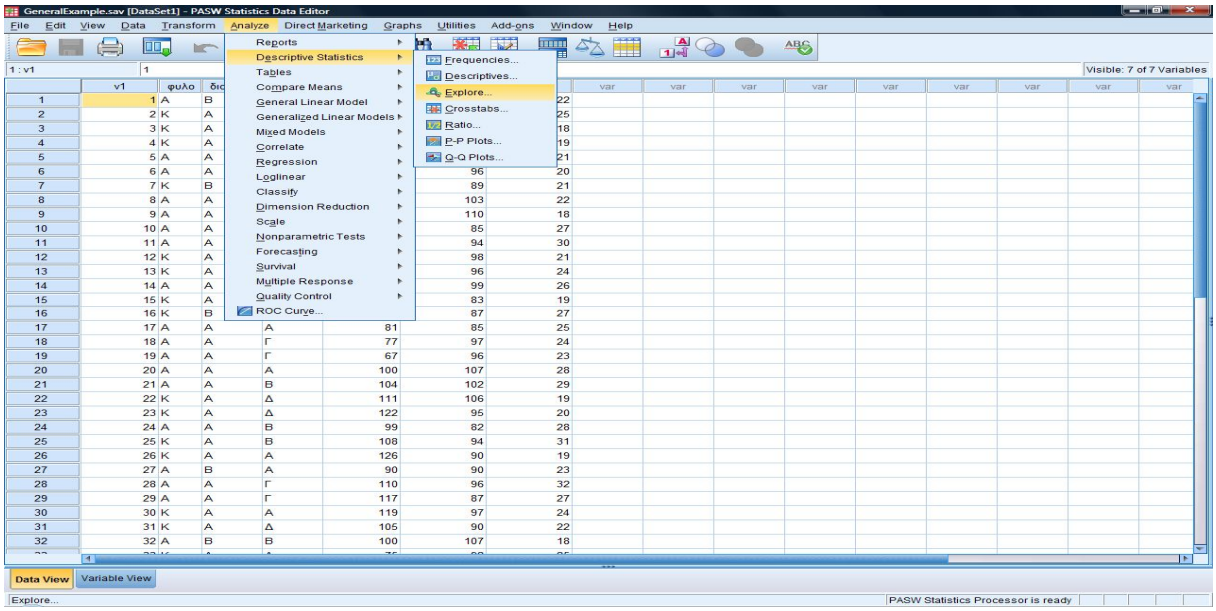
5.2 Παραδείγματα

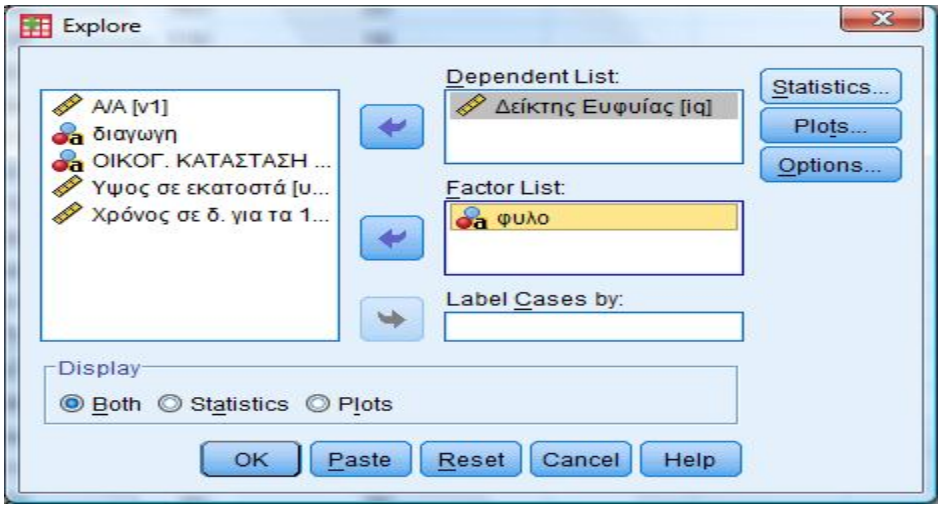
Παράδειγμα 1ο Χρησιμοποιώντας τα δεδομένα του αρχείου GeneralExample.sav* θέλουμε να ελέγξουμε, αν είναι εφικτό, αν υπάρχει στατιστικά σημαντική διαφορά στο μέσο δείκτη ευφυΐας αγοριών και κοριτσιών.

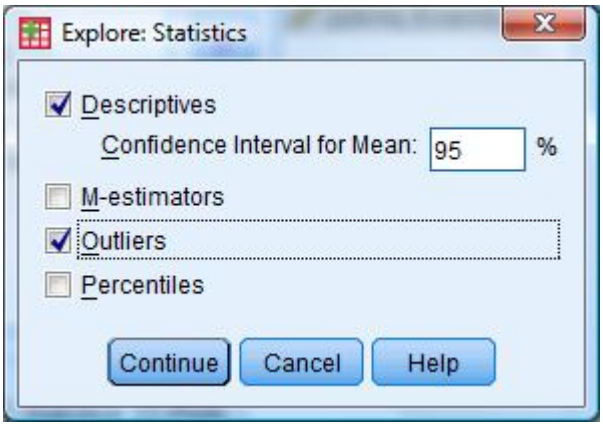
Υλοποίηση:

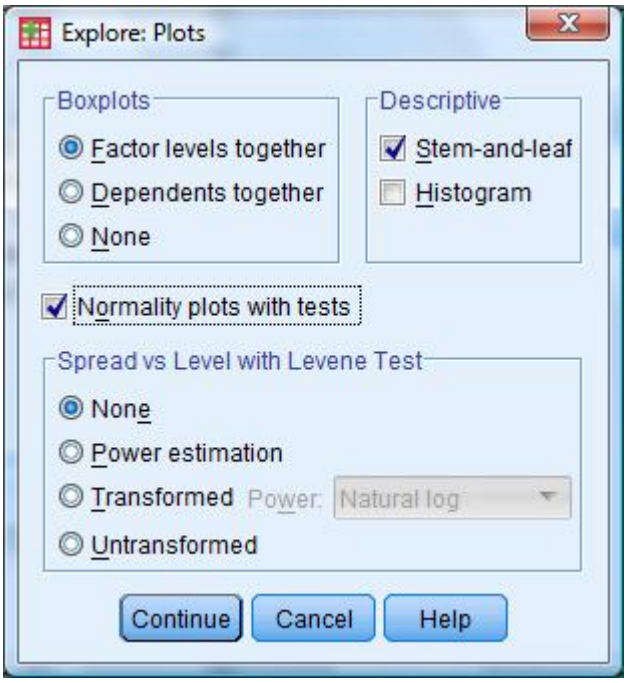
Έλεγχος Ακραίων τιμών: Αρχικά θα ελέγξουμε την ύπαρξη ακραίων τιμών στις δειγματικές τιμές που καταγράφεται ο δείκτης ευφυΐας αγοριών και κοριτσιών και

ταυτόχρονα για να μην επανερχόμαστε θα ζητούμε και γραφικούς-στατιστικούς τρόπους ελέγχου της κανονικότητας. Για το σκοπό αυτό ακολουθούμε τα παρακάτω βήματα (διαδικασία Explore):

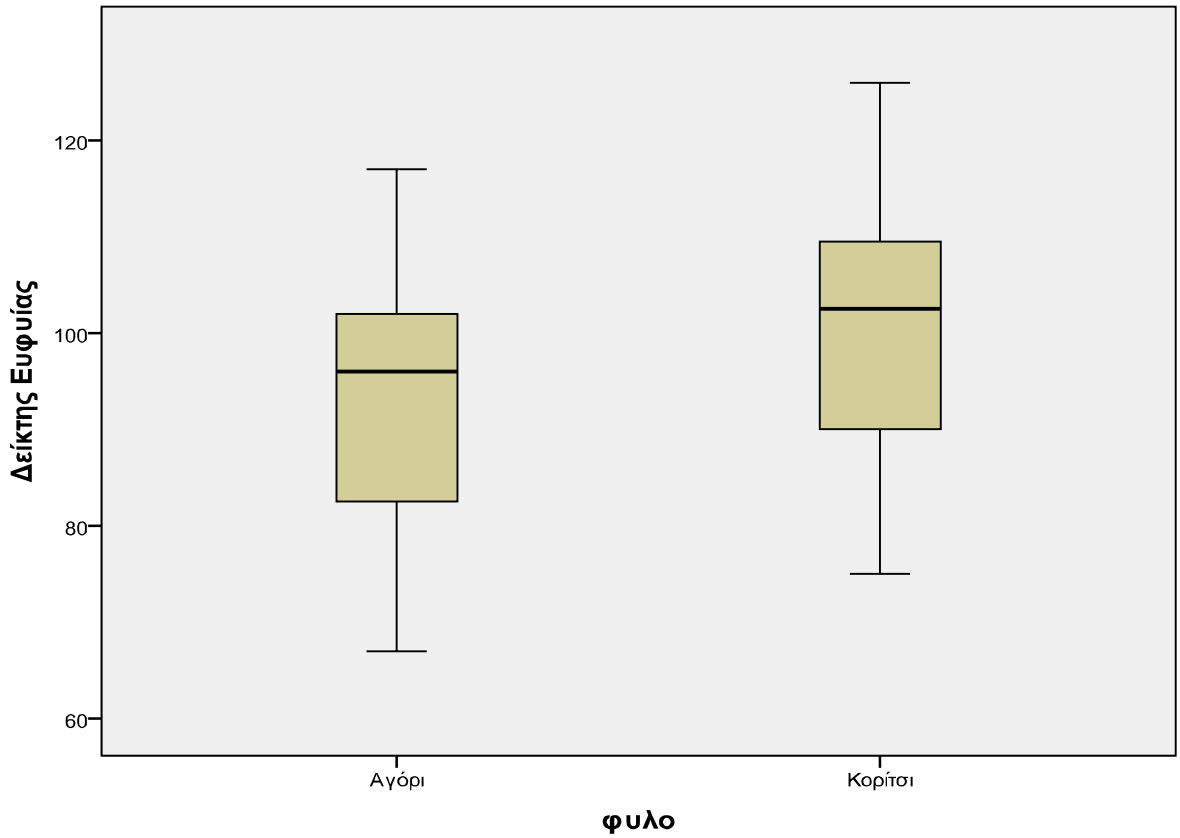





Explore: Statistics
Descriptives
Confidence Interval for Mean: 95 %
M-estimators
Outliers
Percentiles
Continue
Cancel
Help


Explore: Plots
Boxplots
Factor levels together
Dependents together
None
Descriptive
Stem-and-leaf
Histogram
Normality plots with tests
Spread vs Level with Levene Test
None
Power estimation
Transformed Power: Natural log
Untransformed
Continue
Cancel
Help

Σχήμα 1: Θηκόγραμματα 1 και 2



Από τα παραπάνω θηκογράμματα προκύπτει ότι δεν υπάρχουν ακραίες τιμές στις δειγματικές τιμές του δείκτη ευφυΐας αγοριών και κοριτσιών.

Προσοχή: Αν υπάρχουν ακραίες τιμές τις αποκλείουμε μία μία για κάθε «ομάδα», ξεκινώντας από την πιο απομακρυσμένη της ομάδας. **Το ποσοστό 10% δεν το υπολογίζουμε στο σύνολο των παρατηρήσεων αλλά στον αριθμό των παρατηρήσεων εντός κάθε ομάδας.**

Έπειτα ελέγχουμε αν οι δειγματικές τιμές του δείκτη ευφυΐας αγοριών και κοριτσιών προέρχονται από κανονικούς πληθυσμούς (τεστ Shapiro Wilk έχει ήδη ζητηθεί η υλοποίηση του στο προηγούμενο βήμα)

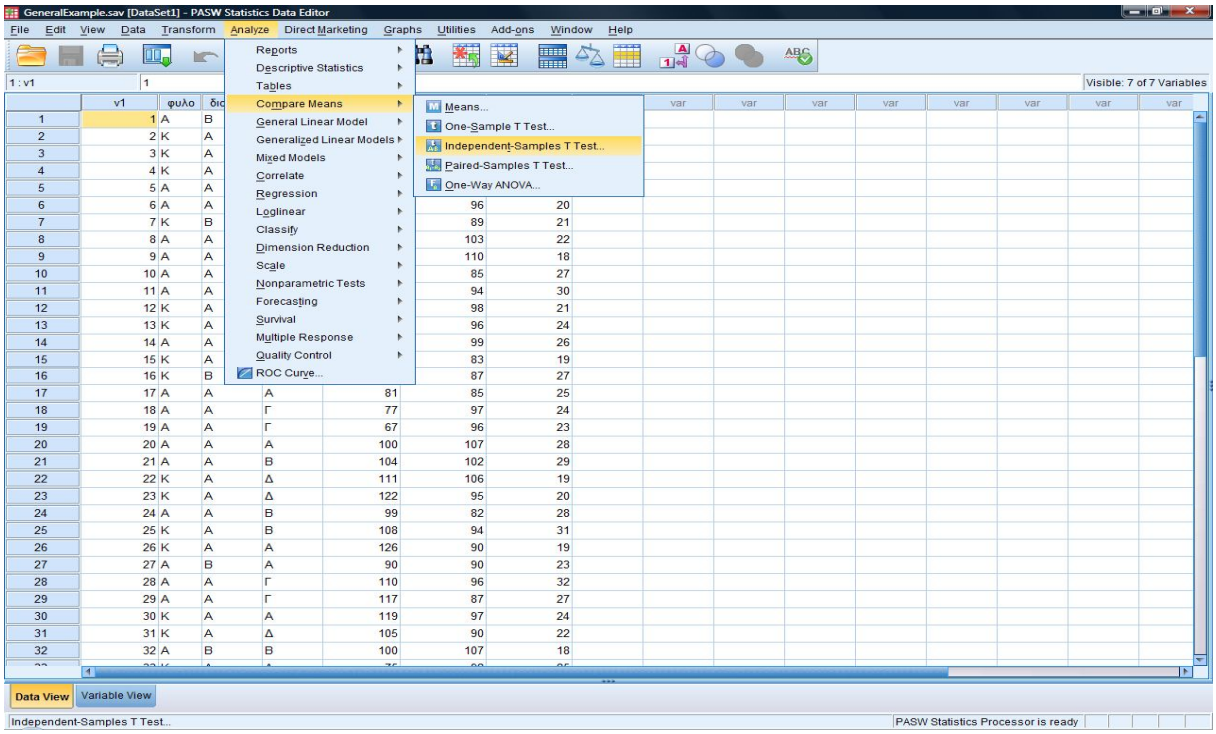
Tests of Normality

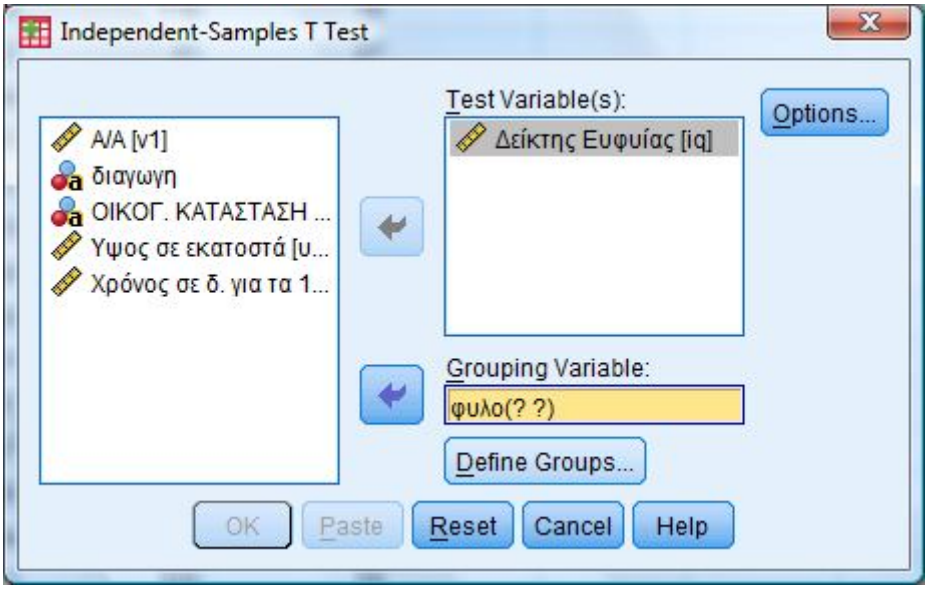
	φυλο	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Δείκτης Ευφυίας	Αγόρι	,130	19	,200*	,969	19	,764
	Κορίτσι	,105	16	,200*	,974	16	,893

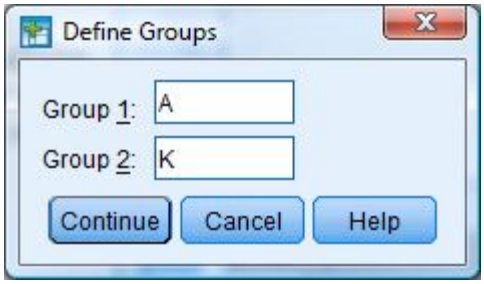
a. Lilliefors Significance Correction

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

Η υπόθεση της κανονικότητας δεν απορρίπτεται καθώς από τον πίνακα αυτό και από την στήλη Shapiro-Wilk Sig. βλέπουμε ότι η κρίσιμη πιθανότητα (p-value) για την ομάδα των αγοριών είναι 0,764 ενώ για την ομάδα των κοριτσιών 0,893. Επειδή και οι δυο αυτές τιμές είναι μεγαλύτερες από το 0,05 (5%) συμπεραίνουμε ότι η υπόθεση ότι τα δύο δείγματα προέρχονται από πληθυσμούς που περιγράφονται ικανοποιητικά από την κανονική κατανομή δεν μπορεί να απορριφθεί. Επομένως θα προβούμε σε παραμετρικό έλεγχο δύο μέσων τιμών: (από τη θεωρία γνωρίζουμε ότι ποια μορφή του t τεστ θα χρησιμοποιηθεί καθορίζεται από την ικανοποίηση ή όχι της ισότητας των πληθυσμιακών διακυμάνσεων. Η υπόθεση αυτή ελέγχεται από το στατιστικό τεστ του Levene).







Group Statistics

	φυλο	N	Mean	Std. Deviation	Std. Error Mean
Δείκτης Ευφυίας	Αγόρι	19	93,42	13,822	3,171
	Κορίτσι	16	101,13	14,514	3,628

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Δείκτης Ευφυίας	Equal variances assumed	,079	,781	-1,606	33	,118	-7,704	4,798	-17,466	2,058
	Equal variances not assumed			-1,599	31,399	,120	-7,704	4,819	-17,527	2,119

Η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων δεν απορρίπτεται (τεστ Levene p-τιμή=0,781). Ο μέσος δείκτης ευφυΐας αγοριών και κοριτσιών δε διαφέρει στατιστικά σημαντικά (p-τιμή=0,118>0.05). **Προσοχή**: Αν δεν ισχύει η ισότητα των πληθυσμιακών διακυμάνσεων τότε οδηγούμαστε στα αποτελέσματα του t τεστ της γραμμής Equal variances not assumed.

Αναφορά: Θέλουμε να ελέγξουμε αν ο μέσος δείκτης ευφυΐας αγοριών και κοριτσιών δε διαφέρει στατιστικά σημαντικά. Επομένως πρόκειται για έναν έλεγχο της ισότητας δύο μέσων τιμών με ανεξάρτητα τυχαία δείγματα. Για να μπορούμε να αποφανθούμε χρησιμοποιώντας τον παραμετρικό έλεγχο του t-test με ανεξάρτητα δείγματα θα πρέπει να πληρούνται οι ακόλουθες υποθέσεις:

1. Τα δείγματά μας να είναι τυχαία επιλεγμένα
2. Να μην υπάρχουν ακραίες τιμές στα δειγματικά δεδομένα κάθε πληθυσμού που να ξεπερνούν σε ποσοστό το 10%.
3. Κάθε πληθυσμός να περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε τα δείγματά μας και ικανοποιείται.

Αρχικά ελέγχουμε την ύπαρξη ακραίων τιμών στις δειγματικές παρατηρήσεις που καταγράφεται ο δείκτης ευφυΐας αγοριών και κοριτσιών αντίστοιχα. Ο έλεγχος της ύπαρξης ακραίων τιμών στο δείγμα των 19 ανδρών και στο δείγμα των 16 γυναικών αποδεικνύει ότι

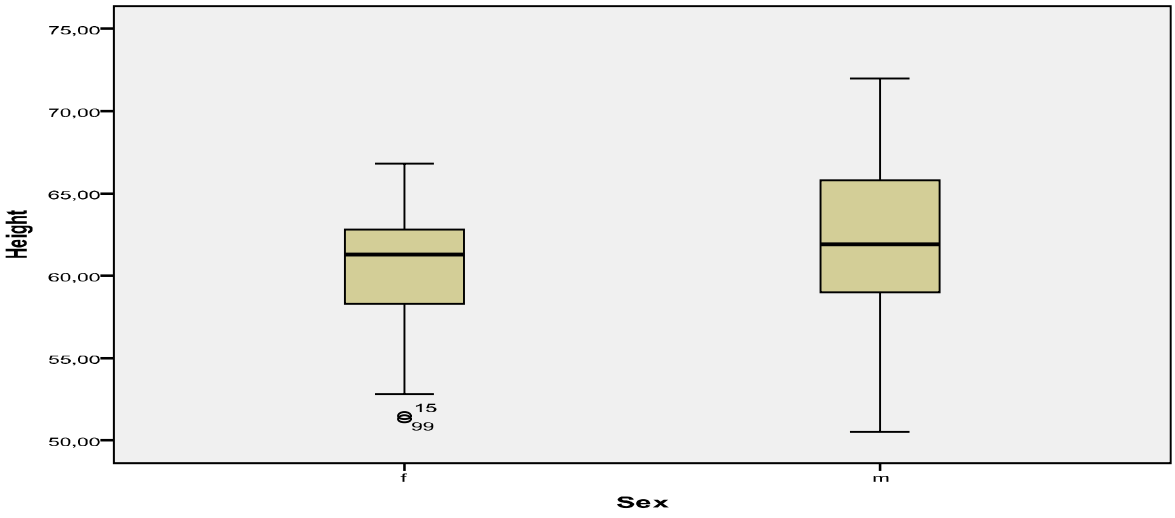
δεν υπάρχουν ακραίες τιμές (βλέπε θηκογράμματα 1,2 στο σχήμα 1). Έπειτα ελέγχουμε την υπόθεση ότι τα δύο δείγματα προέρχονται από κανονικούς πληθυσμούς. Προκύπτει ότι ικανοποιείται τόσο η υπόθεση ότι οι δειγματικές τιμές του δείκτη ευφυΐας των αγοριών προέρχονται από κανονικό πληθυσμό (Shapiro Wilk, p-τιμή=0,764) όσο και η υπόθεση ότι οι δειγματικές τιμές του δείκτη ευφυΐας των κοριτσιών προέρχονται από κανονικό πληθυσμό (Shapiro Wilk, p-τιμή=0,893)

Επομένως θα χρησιμοποιήσουμε τον παραμετρικό έλεγχο του τεστ για να ελέγξουμε την υπόθεση της ισότητας των μέσου δείκτη ευφυΐας αγοριών και κοριτσιών. Από τη θεωρία γνωρίζουμε ότι ποια μορφή του t τεστ θα χρησιμοποιηθεί καθορίζεται από την ικανοποίηση ή όχι της ισότητας των πληθυσμιακών διακυμάνσεων. Η υπόθεση της ισότητας των διακυμάνσεων δεν απορρίπτεται (τεστ του Levene, F=0,079, p-value=0,781). Ο μέσος δείκτης ευφυΐας των αγοριών δεν διαφέρει στατιστικά σημαντικά από το μέσο δείκτη ευφυΐας των κοριτσιών (t=-1,606, df=33, p=0,118>0,05).

<u>Παράδειγμα 2ο</u> Με βάση τα δεδομένα του αρχείου HeightWeight1.sav να εξετασθεί, αν είναι εφικτό, αν υπάρχει στατιστικά σημαντική διαφορά στο μέσο ύψος μεταξύ αγοριών και κοριτσιών.

Όπως πριν προκύπτουν τα ακόλουθα

Σχήμα 1: Θηκόγραμματα 1 και 2

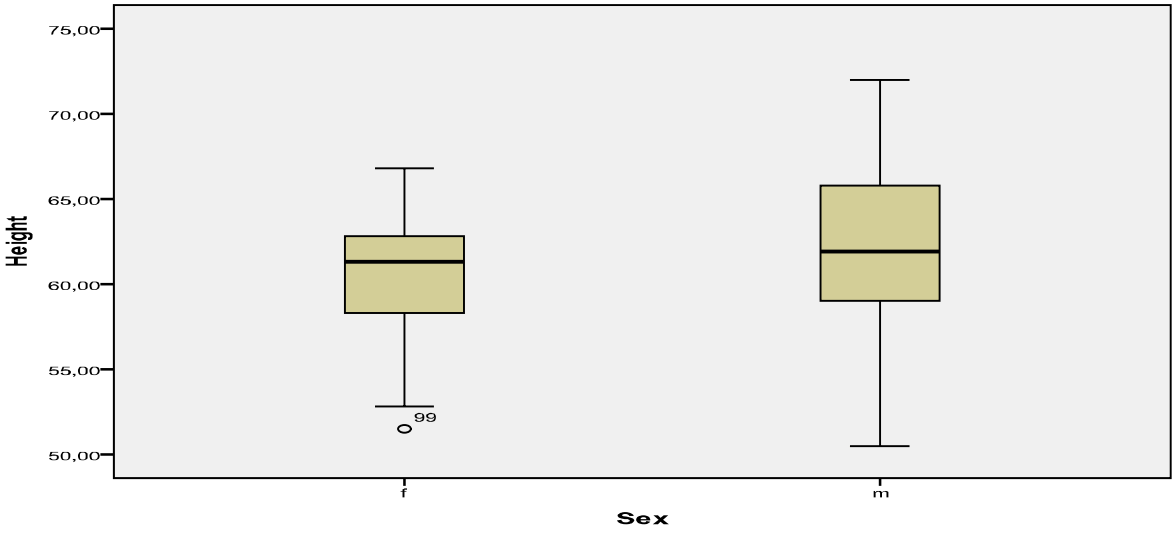


Από το θηκόγραμμα βλέπουμε ότι η ομάδα των αγοριών (m) δεν έχει ακραίες παρατηρήσεις, ενώ η ομάδα των κοριτσιών (f) έχει τουλάχιστον μία ακραία τιμή (βλ. αριστερά σχήμα). Αποκλείουμε την πιο απομακρυσμένη, που είναι η παρατήρηση 15 (προκύπτει από τον πίνακα Extreme Values). Την αποκλείουμε με τη γνωστή διαδικασία μέσω της επιλογής Data

Select Cases και επαναλαμβάνουμε τον έλεγχο ύπαρξης ακραίων τιμών στις υπόλοιπες δειγματικές τιμές.

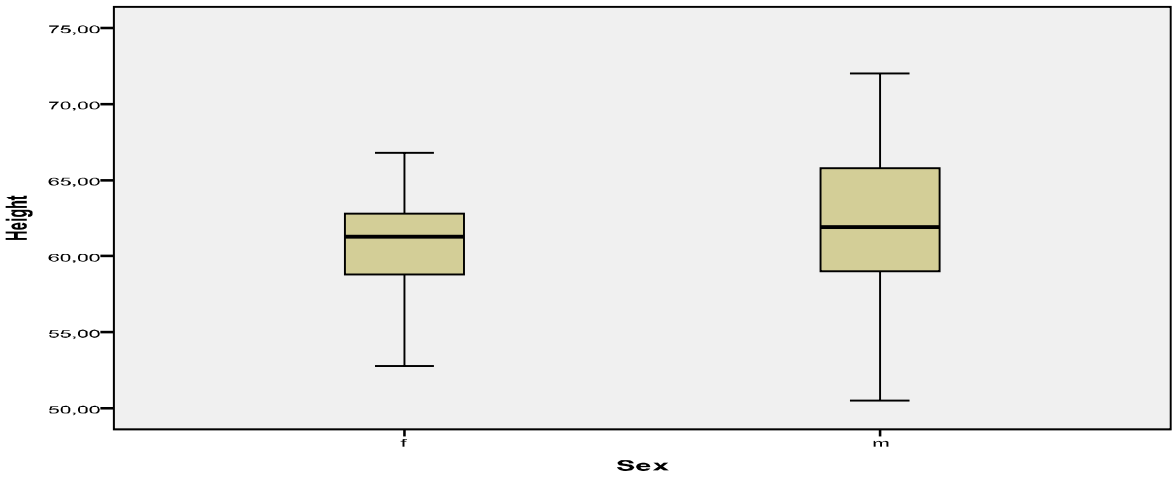
Προσοχή: Οι ακραίες τιμές αποκλείονται μία μία εντός κάθε δείγματος και μέχρι να ξεπεράσουμε σε αριθμό το 10% των διαθέσιμων παρατηρήσεων.

Σχήμα 2: Θηκόγραμματα 3 και 4



Η παρατήρηση 99 είναι ακραία. Αν διαγράψουμε την παρατήρηση αυτή και επαναλάβουμε τον έλεγχο θα δούμε ότι δεν υπάρχουν άλλες ακραίες παρατηρήσεις.

Σχήμα 3: Θηκόγραμματα 5 και 6



Συνολικά δηλαδή έχουμε για το δείγμα των κοριτσιών 2 ακραίες στις 111 διαθέσιμες, επομένως δεν έχουμε ξεπεράσει το 10%.

Υπόθεση κανονικότητας:

Από τον πίνακα που ακολουθεί προκύπτει ότι η υπόθεση της κανονικής κατανομής και για τους δύο πληθυσμούς δεν μπορεί να απορριφθεί.

Tests of Normality

	Sex	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Height	f	,108	109	,003	,979	109	,090
	m	,069	126	,200*	,988	126	,374

a. Lilliefors Significance Correction

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

Παραμετρικός έλεγχος

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Height	f	109	60,6936	3,14899	,30162
	m	126	62,1032	4,27669	,38100

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Height	Equal variances assumed	13,851	,000	-2,839	233	,005	-1,40960	,49653	-2,38786	-,43133
	Equal variances not assumed			-2,901	227,4	,004	-1,40960	,48594	-2,36711	-,45208

Η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων απορρίπτεται (τεστ Levene p-τιμή<0,001). Το μέσο ύψος αγοριών και κοριτσιών διαφέρει στατιστικά σημαντικά (p-τιμή=0,004<0,05) και μάλιστα το μέσο ύψος των αγοριών είναι στατιστικά σημαντικά μεγαλύτερο (το καταλαβαίνουμε είτε από τον πίνακα Group Statistics είτε από πρόσημο του t τεστ ή/και της μέσης διαφοράς είτε από το 95% δ.ε. για τη μέση διαφορά).

Αναφορά: Θέλουμε να ελέγξουμε αν το μέσο ύψος αγοριών και κοριτσιών διαφέρει στατιστικά σημαντικά. Επομένως πρόκειται για έναν έλεγχο της ισότητας δύο μέσων τιμών με ανεξάρτητα δείγματα. Για να μπορούμε να αποφανθούμε χρησιμοποιώντας τον παραμετρικό έλεγχο του t-test με ανεξάρτητα δείγματα θα πρέπει να πληρούνται οι ακόλουθες υποθέσεις:

1. Τα δείγματά μας να είναι τυχαία επιλεγμένα
2. Να μην υπάρχουν ακραίες τιμές στα δειγματικά δεδομένα κάθε πληθυσμού που να ξεπερνούν σε ποσοστό το 10%.
3. Κάθε πληθυσμός να περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε τα δείγματά μας και ικανοποιείται.

Αρχικά ελέγχουμε την ύπαρξη ακραίων τιμών στις δειγματικές παρατηρήσεις που καταγράφεται το ύψος αγοριών και κοριτσιών αντίστοιχα. Ο έλεγχος της ύπαρξης ακραίων τιμών στις δειγματικές τιμές του ύψους των 111 γυναικών έδειξε ότι υπάρχουν δύο ακραίες τιμές (ποσοστό αριθμού ακραίων τιμών <10%) οι παρατηρήσεις με αύξοντα αριθμό 15 και 99 και ύψος 51.3 και 51.5 αντίστοιχα, ενώ ο έλεγχος της ύπαρξης ακραίων τιμών στις δειγματικές τιμές του ύψους των 128 ανδρών έδειξε ότι δεν υπάρχουν ακραίες τιμές (βλέπε θηκογράμματα 1,2,3,4 στα σχήματα 1,2). Έπειτα καθώς το ποσοστό των ακραίων τιμών δεν ξεπερνά το 10% και αφού τις αποκλείσουμε από την περαιτέρω ανάλυση ελέγχουμε την υπόθεση ότι τα δύο δείγματα προέρχονται από κανονικούς πληθυσμούς. Προκύπτει ότι ικανοποιείται τόσο η υπόθεση ότι οι δειγματικές τιμές του ύψους των αγοριών προέρχονται από κανονικό πληθυσμό (Shapiro Wilk, p-τιμή=0,374) όσο και η υπόθεση ότι οι δειγματικές τιμές του ύψους των κοριτσιών προέρχονται από κανονικό πληθυσμό (Shapiro Wilk, p-τιμή=0,090)

Επομένως θα χρησιμοποιήσουμε τον παραμετρικό έλεγχο του τεστ για να ελέγξουμε την υπόθεση της ισότητας του μέσου ύψους αγοριών και κοριτσιών. Από τη θεωρία γνωρίζουμε ότι ποια μορφή του t τεστ θα χρησιμοποιηθεί καθορίζεται από την ικανοποίηση ή όχι της ισότητας των πληθυσμιακών διακυμάνσεων. Η υπόθεση της ισότητας των διακυμάνσεων απορρίπτεται (τεστ του Levene, F=13,851, p-value<0,001). Το μέσο ύψος αγοριών και κοριτσιών διαφέρει στατιστικά σημαντικά (p-τιμή=0,004<0.05) και μάλιστα το μέσο ύψος των αγοριών είναι στατιστικά σημαντικά μεγαλύτερο (δειγματική μέση τιμή ύψους αγοριών και κοριτσιών 62,1032 και 60,6936 αντίστοιχα) και ένα 95% Δ.Ε. για τη διαφορά: μέσο ύψους κοριτσιών-μέσο ύψος αγοριών είναι το (-2.36711,-0.45208).

<u>Παράδειγμα 3ο</u> Στο αρχείο CompanyProfit.sav* καταγράφονται τα κέρδη (Profit), σε χιλιάδες δολάρια, για δύο τύπους εταιρειών, τις Φαρμακευτικές (Pharmac) και τις εταιρείες υπολογιστών (Computer). Οι εταιρείες αυτές έχουν επιλεγεί με τυχαίο τρόπο από τις αντίστοιχες εταιρείες που δραστηριοποιούνται στην Ελλάδα. Να εξετασθεί, αν είναι εφικτό, αν το μέσο κέρδος των φαρμακευτικών εταιρειών διαφέρει στατιστικά σημαντικά από το αντίστοιχο των εταιρειών υπολογιστών.

Παράθεση Αποτελεσμάτων-Ανάλυση

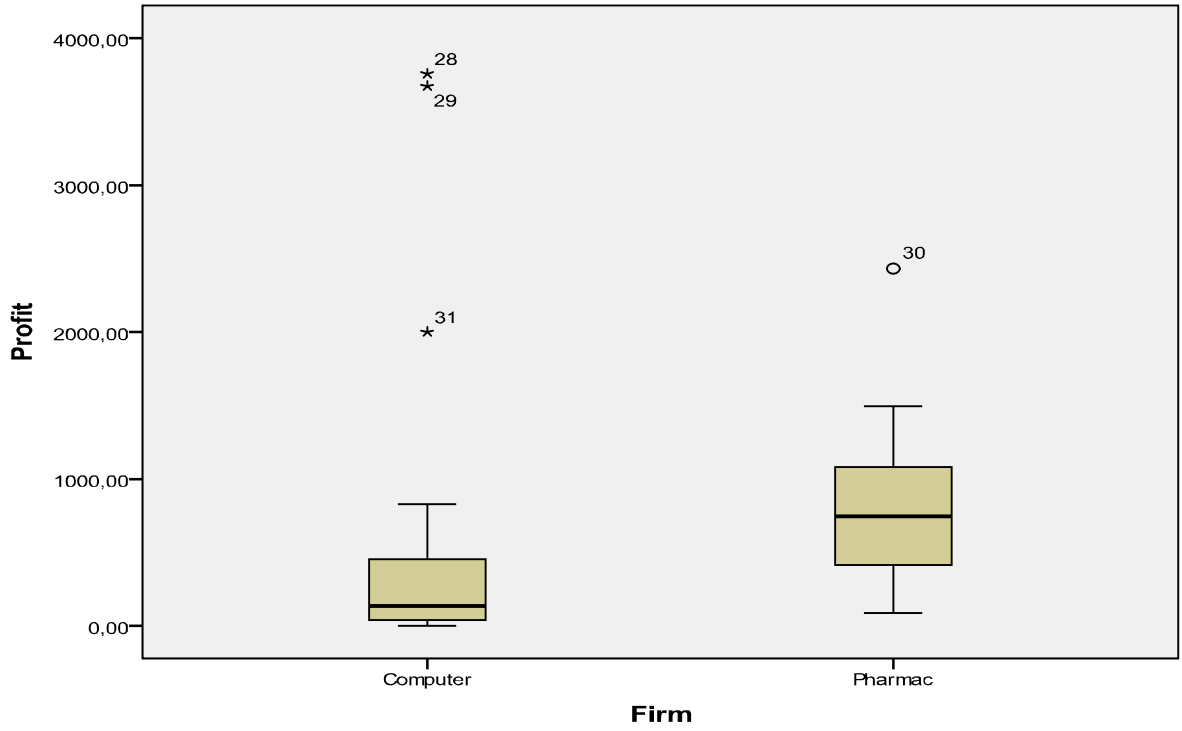
Αρχικά παρατηρούμε ότι έχουμε διαθέσιμες 18 και 13 παρατηρήσεις αντίστοιχα που αφορούν τα κέρδη για εταιρείες υπολογιστών και φαρμακευτικές αντίστοιχα.

Case Processing Summary

Firm		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Profit	Computer	18	100,0%	0	,0%	18	100,0%
	Pharmac	13	100,0%	0	,0%	13	100,0%

Έλεγχος ακραίων τιμών: Από το θηκόγραμμα που έπεται προκύπτει αρχικά ότι στις δειγματικές τιμές του κέρδους των εταιρειών πληροφορικής υπάρχει τουλάχιστον μία ακραία τιμή η παρατήρηση με αύξοντα αριθμό 28. Επίσης προκύπτει ότι στις δειγματικές τιμές του κέρδους των φαρμακευτικών εταιρειών υπάρχει τουλάχιστον μία ακραία τιμή η παρατήρηση με αύξοντα αριθμό 30. Οι παρατηρήσεις αυτές αποκλείονται από την περαιτέρω ανάλυση και διενεργείται πάλι έλεγχος ακραίων τιμών**.**

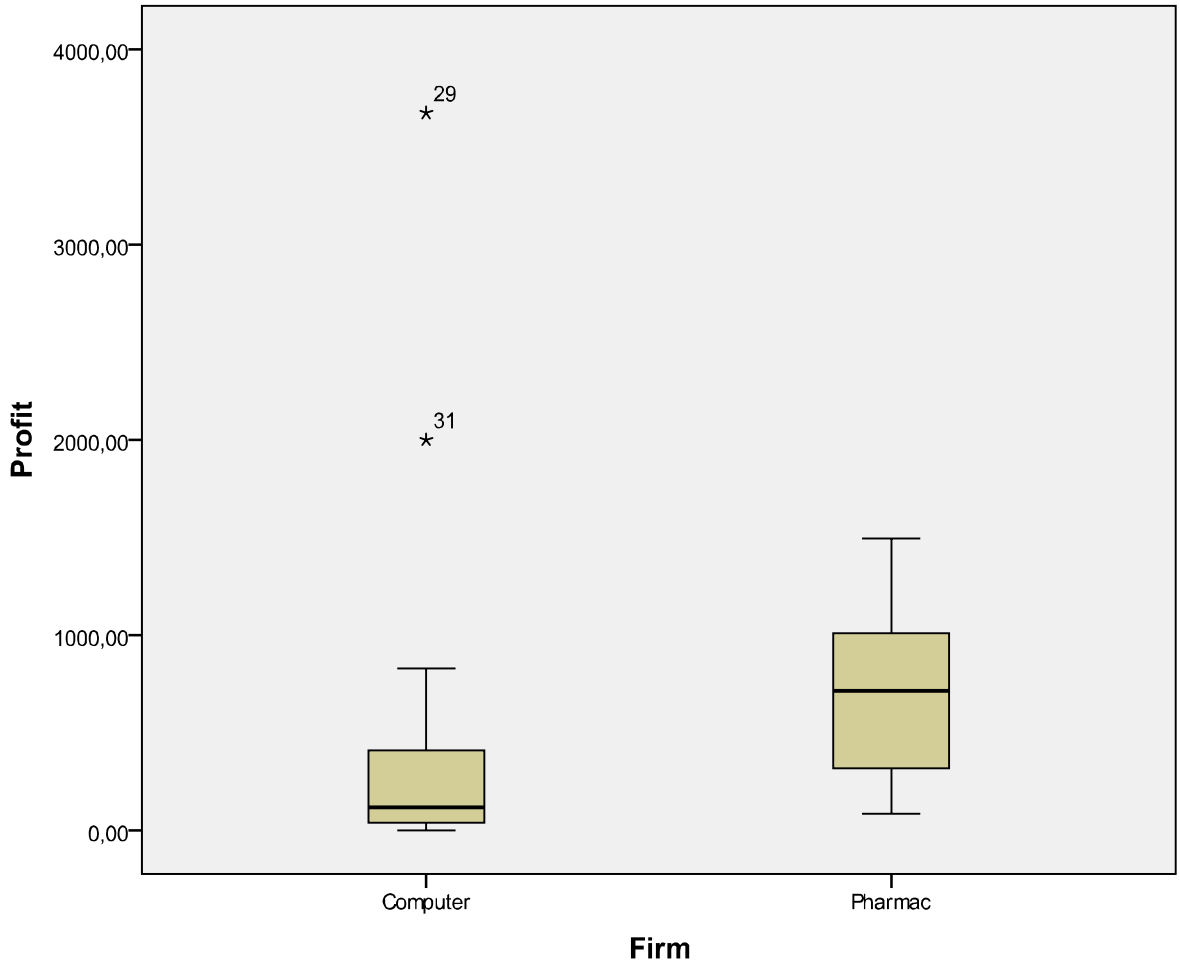
Σχήμα 1: Θηκόγραμματα 1 και 2



Από το θηκόγραμμα που ακολουθεί προκύπτει ότι υπάρχει τουλάχιστον μία ακόμη ακραία τιμή στις δειγματικές τιμές του κέρδους των εταιρειών πληροφορικής η παρατήρηση με

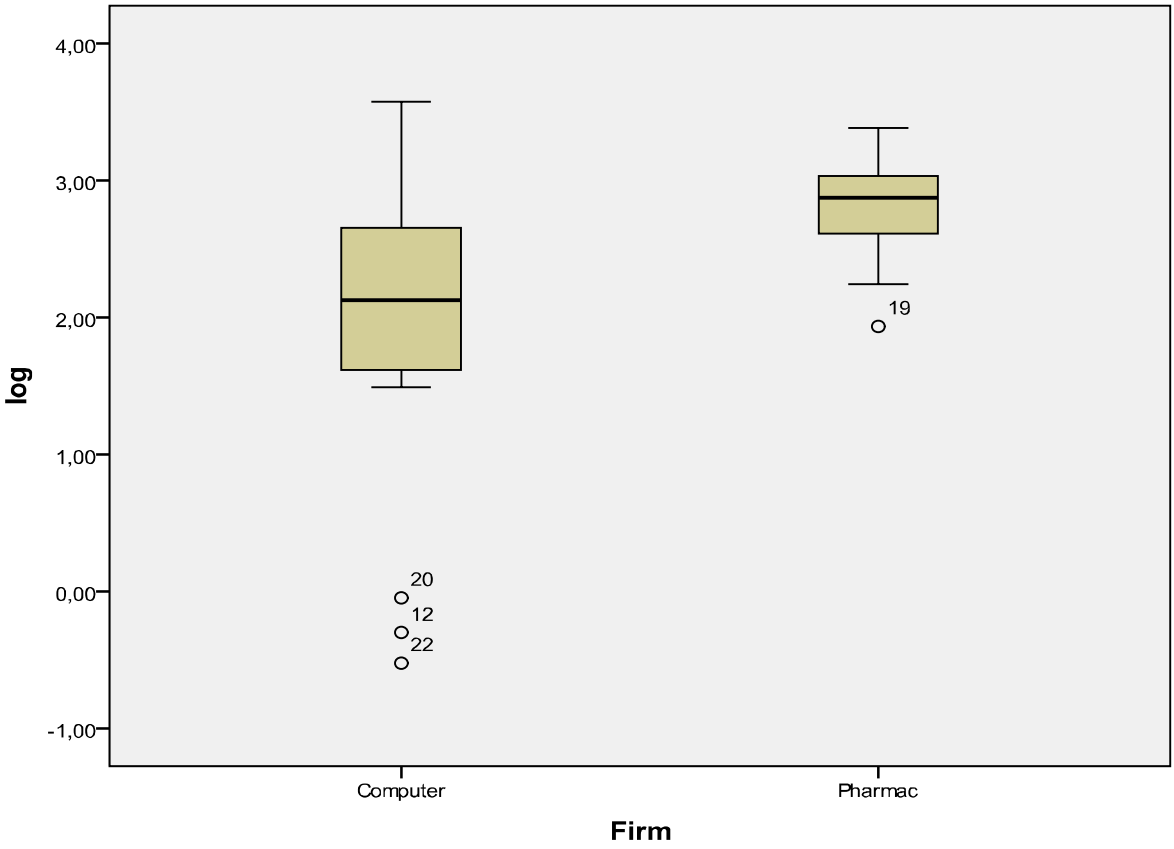
αύξοντα αριθμό 29, ενώ δεν προκύπτει άλλη ακραία τιμή στις δειγματικές τιμές του κέρδους των φαρμακευτικών εταιρειών. Επομένως συνολικά στις δειγματικές τιμές του κέρδους των εταιρειών πληροφορικής υπάρχουν τουλάχιστον δύο ακραίες τιμές στις διαθέσιμες 18 και επομένως ο αριθμός των ακραίων τιμών εντός αυτού του δείγματος ξεπερνά το 10%.

Σχήμα 2: Θηκόγραμματα 3 και 4

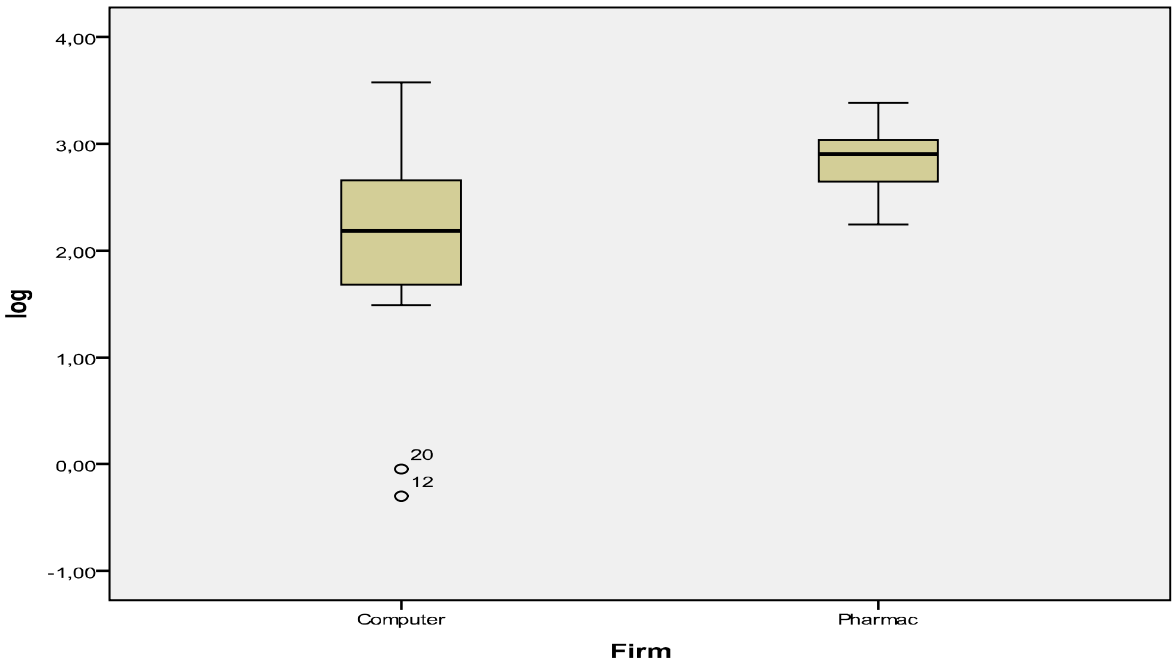


Άρα επαναφέρουμε όλες τις τιμές, προχωράμε σε λογαριθμικό μετασχηματισμό των δεδομένων μας για την διόρθωση του προβλήματος των ακραίων παρατηρήσεων (πάλι επισημαίνεται ότι θα πρέπει να είμαστε προσεκτικοί στην ύπαρξη μη θετικών τιμών στην υπό μετασχηματισμό μεταβλητή). Διεξάγεται στη συνέχεια έλεγχος ύπαρξης ακραίων τιμών στις δειγματικές τιμές του λογαρίθμου του κέρδους των εταιρειών πληροφορικής και των εταιριών των σχετικών με το εμπόριο φαρμάκων. Προκύπτουν τότε κατά σειρά και με τη γνωστή διαδικασία τα ακόλουθα θηκογράμματα:

Σχήμα 3: Θηκόγραμματα 5 και 6

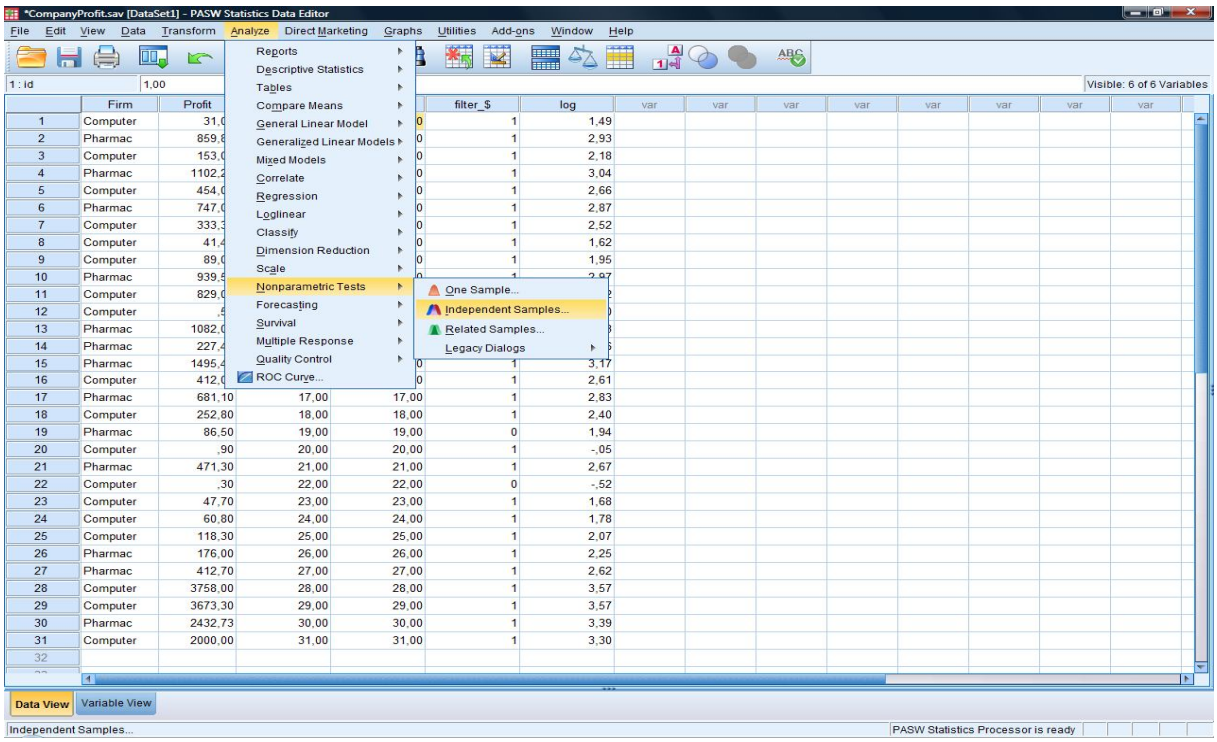


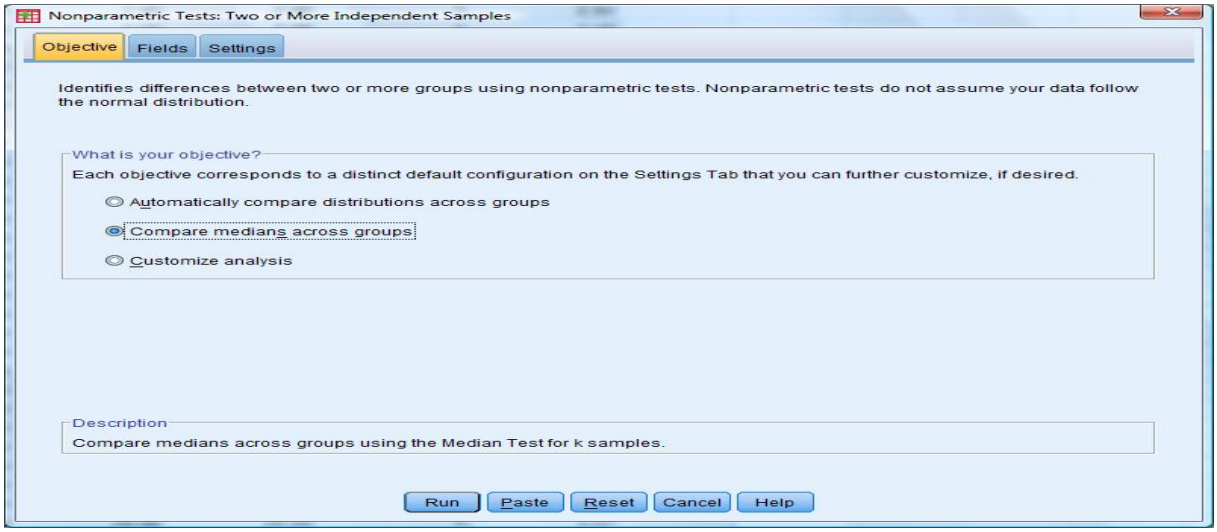
Σχήμα 4: Θηκόγραμματα 7 και 8

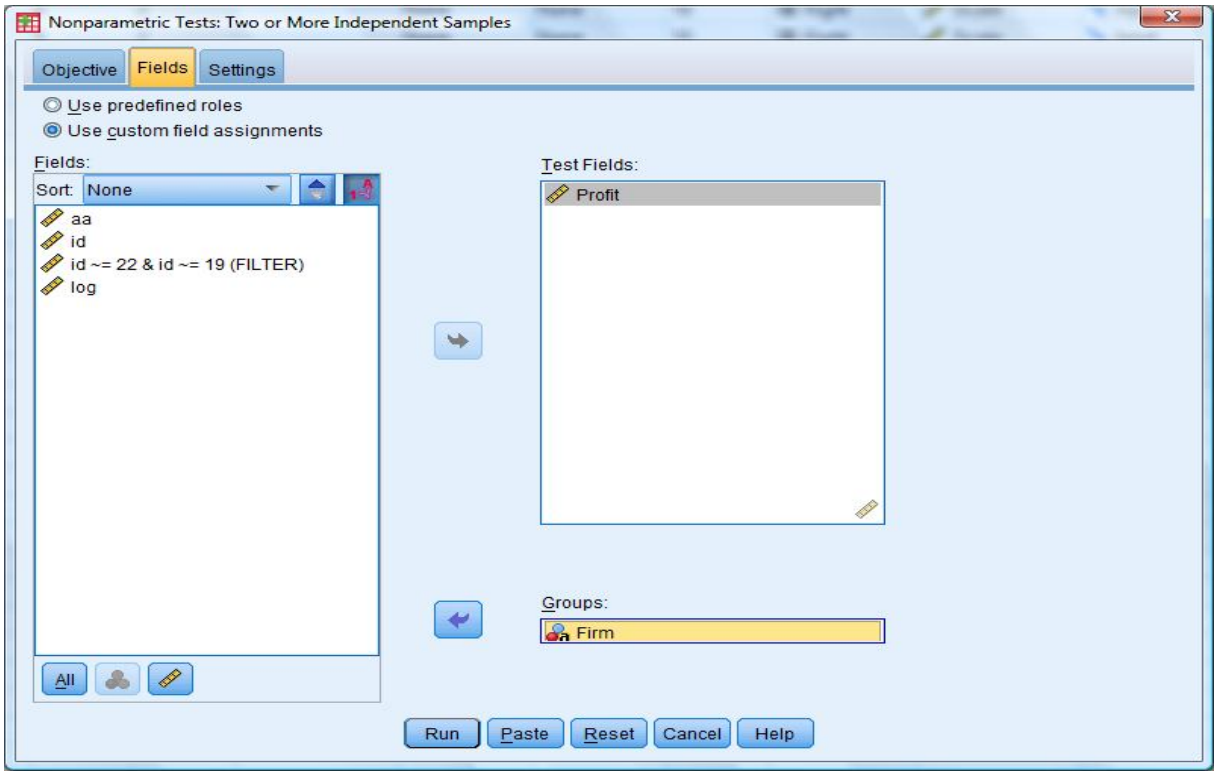


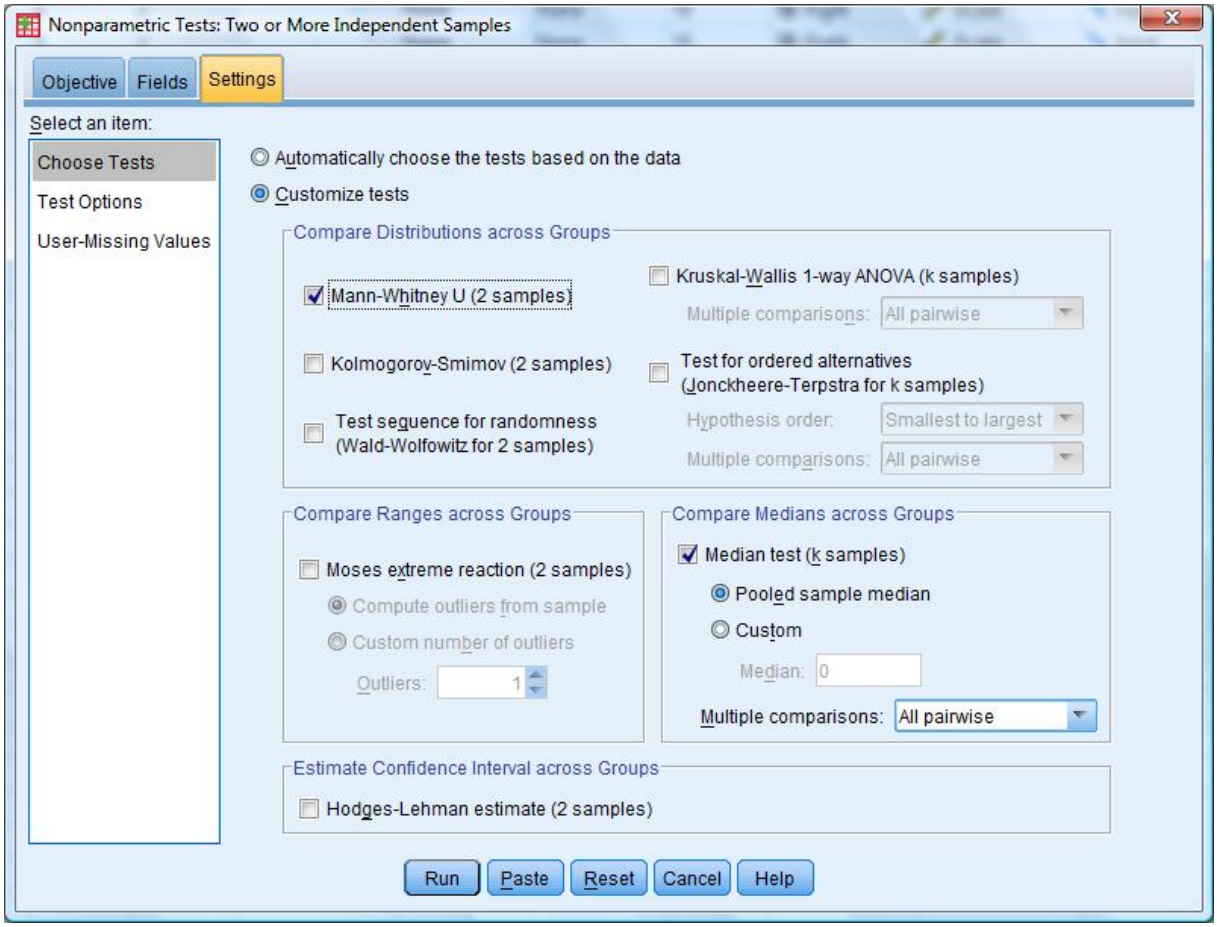
Επομένως ο μετασχηματισμός του λογαρίθμου δε διορθώνει το πρόβλημα της ύπαρξης ακραίων τιμών. Αφού επαναφέρουμε όλες τις παρατηρήσεις προβαίνουμε σε μη παραμετρικό έλεγχο της υπόθεσης της ισότητας των πληθυσμιακών διαμέσων του κέρδους των δύο τύπων εταιρειών και θα εξεταστεί έπειτα κατά πόσο τα αποτελέσματα αυτού γενικεύονται στις πληθυσμιακές μέσες τιμές.

Μη παραμετρικός έλεγχος (αφού επαναφέρουμε όλες τις παρατηρήσεις και ακολουθώντας τη διαδικασία Nonparametric Tests Independent Samples).






Nonparametric Tests: Two or More Independent Samples
Objective
Fields
Settings
Use predefined roles
Use custom field assignments
Fields:
Sort: None
aa
id
id ~= 22 & id ~= 19 (FILTER)
log
All
Test Fields:
Profit
Groups:
Firm
Run
Paste
Reset
Cancel
Help


Nonparametric Tests: Two or More Independent Samples
Objective
Fields
Settings
Select an item:
Choose Tests
Test Options
User-Missing Values
Automatically choose the tests based on the data
Customize tests
Compare Distributions across Groups
Mann-Whitney U (2 samples)
Kruskal-Wallis 1-way ANOVA (k samples)
Multiple comparisons: All pairwise
Kolmogorov-Smimov (2 samples)
Test for ordered alternatives (Jonckheere-Terpstra for k samples)
Test sequence for randomness (Wald-Wolfowitz for 2 samples)
Hypothesis order: Smallest to largest
Multiple comparisons: All pairwise
Compare Ranges across Groups
Moses extreme reaction (2 samples)
Compute outliers from sample
Custom number of outliers
Outliers: 1
Compare Medians across Groups
Median test (k samples)
Pooled sample median
Custom
Median: 0
Multiple comparisons: All pairwise
Estimate Confidence Interval across Groups
Hodges-Lehman estimate (2 samples)
Run
Paste
Reset
Cancel
Help

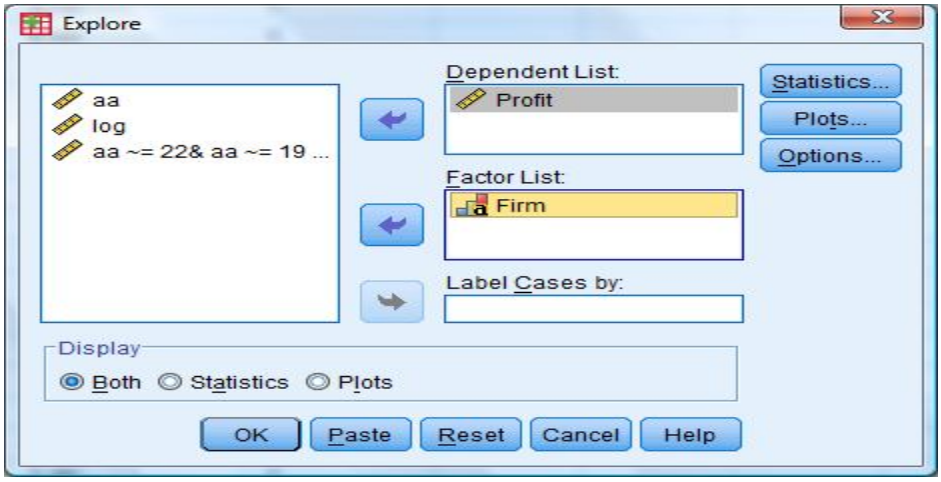
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The medians of Profit are the same across categories of Firm.	Independent-Samples Median Test	,019	Reject the null hypothesis.
2	The distribution of Profit is the same across categories of Firm.	Independent-Samples Mann-Whitney U Test	,025[1]	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1]Exact significance is displayed for this test.

Η κρίσιμη πιθανότητα που μας έδωσε το τεστ αυτό είναι p=0,019. Σαν συμπέρασμα έχουμε ότι για επίπεδο σημαντικότητας 5% η διάμεσος του κέρδους των φαρμακευτικών εταιρειών διαφέρει στατιστικά σημαντικά από αυτό των εταιρειών των υπολογιστών, ενώ για επίπεδο σημαντικότητας 1% δε διαφέρουν στατιστικά σημαντικά. Για να μπορούν να γενικευτούν τα αποτελέσματα για το μέσο κέρδος θα πρέπει η δειγματική μέση τιμή και η αντίστοιχη διάμεσος να είναι αρκετά κοντά. Για να το εξετάσουμε αυτό επιλέγουμε μέσω της διαδικασίας Analyze Descriptive Statistics Explore τα ακόλουθα:



Από τον πίνακα που προκύπτει έχουμε ότι για τις εταιρείες πληροφορικής η δειγματική μέση τιμή και διάμεσος του κέρδους είναι αντίστοιχα 680,8500 και 135,6500 αντίστοιχα ενώ οι αντίστοιχες τιμές για τις φαρμακευτικές εταιρείες είναι 824,1254 και 747. Άρα τα αποτελέσματα δεν γενικεύονται στις πληθυσμιακές μέσες τιμές, καθώς τα δεδομένα δεν είναι συμμετρικά.

Descriptives

	Firm			Statistic	Std. Error
Profit	Computer	Mean		680,8500	283,32610
		95% Confidence Interval for Mean	Lower Bound	83,0842	
			Upper Bound	1278,6158	
		5% Trimmed Mean		547,7056	
		Median		135,6500	
		Variance		1444926,267	
		Std. Deviation		1202,05086	
		Minimum		,30	
		Maximum		3758,00	
		Range		3757,70	
		Interquartile Range		508,95	
		Skewness		2,117	,536
		Kurtosis		3,395	1,038
	Pharmac	Mean		824,1254	176,08687
		95% Confidence Interval for Mean	Lower Bound	440,4651	
			Upper Bound	1207,7857	
		5% Trimmed Mean		775,7376	
		Median		747,0000	
		Variance		403085,608	
		Std. Deviation		634,89023	
		Minimum		86,50	
		Maximum		2432,73	
		Range		2346,23	
		Interquartile Range		772,05	
		Skewness		1,348	,616
		Kurtosis		2,482	1,191

Αναφορά: Θέλουμε να ελέγξουμε αν το μέσο κέρδος εταιρειών πληροφορικής και φαρμακευτικών εταιρειών διαφέρει στατιστικά σημαντικά. Επομένως πρόκειται για έναν έλεγχο της ισότητας δύο μέσων τιμών με ανεξάρτητα δείγματα. Για να μπορούμε να αποφανθούμε χρησιμοποιώντας τον παραμετρικό έλεγχο του t-test με ανεξάρτητα δείγματα θα πρέπει να πληρούνται οι ακόλουθες υποθέσεις:

1. Τα δείγματά μας να είναι τυχαία επιλεγμένα
2. Να μην υπάρχουν ακραίες τιμές στα δειγματικά δεδομένα κάθε πληθυσμού που να ξεπερνούν σε ποσοστό το 10%.
3. Κάθε πληθυσμός να περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε τα δείγματά μας και ικανοποιείται.

Αρχικά ελέγχουμε την ύπαρξη ακραίων τιμών στις δειγματικές παρατηρήσεις που καταγράφεται το ύψος αγοριών και κοριτσιών αντίστοιχα. Ο έλεγχος της ύπαρξης ακραίων τιμών στις δειγματικές τιμές του κέρδους των 18 εταιρειών πληροφορικής έδειξε ότι υπάρχουν τουλάχιστον δύο ακραίες τιμές (ποσοστό αριθμού ακραίων τιμών >10%) οι παρατηρήσεις με αύξοντα αριθμό 28 και 29 και κέρδος 3758 και 3673.3 αντίστοιχα, ενώ ο έλεγχος της ύπαρξης ακραίων τιμών στις δειγματικές τιμές του κέρδους των 13 φαρμακευτικών εταιρειών έδειξε ότι υπάρχει μία ακραία τιμή η παρατήρηση με αύξοντα αριθμό 30 και τιμή 2432,73 (βλέπε θηκογράμματα 1,2,3,4 στα σχήματα 1,2). Έπειτα καθώς το ποσοστό των ακραίων τιμών ξεπερνά το 10% εξετάζουμε αν ο μετασχηματισμός του λογαρίθμου διορθώνει το πρόβλημα. Ο μετασχηματισμός του λογαρίθμου δε διορθώνει το πρόβλημα καθώς ο έλεγχος της ύπαρξης ακραίων τιμών στις δειγματικές τιμές του λογαρίθμου του κέρδους των 18 εταιρειών πληροφορικής έδειξε ότι υπάρχουν τουλάχιστον δύο ακραίες τιμές (ποσοστό αριθμού ακραίων τιμών >10%) οι παρατηρήσεις με αύξοντα αριθμό 22 και 12, ενώ ο έλεγχος της ύπαρξης ακραίων τιμών στις δειγματικές τιμές του λογαρίθμου του κέρδους των 13 φαρμακευτικών εταιρειών έδειξε ότι υπάρχει μία ακραία τιμή η παρατήρηση με αύξοντα αριθμό 19 (βλέπε θηκογράμματα 1,2,3,4 στα σχήματα 3,4). Επομένως αφού επαναφέρουμε όλες τις δειγματικές παρατηρήσεις θα χρησιμοποιήσουμε τον μη παραμετρικό έλεγχο για να ελέγξουμε την υπόθεση της ισότητας των πληθυσμιακών διαμέσων του κέρδους των δύο τύπων εταιρειών. Σαν συμπέρασμα έχουμε ότι για επίπεδο σημαντικότητας 5% η διάμεσος του κέρδους των φαρμακευτικών εταιρειών διαφέρει από αυτή των εταιρειών των υπολογιστών (Median Test, p-τιμή=0.007, δειγματικές τιμές της διαμέσου του κέρδους 135,65 και 747 αντίστοιχα). Για να μπορούν να γενικευτούν τα αποτελέσματα για το μέσο κέρδος θα πρέπει η δειγματική μέση τιμή και η αντίστοιχη διάμεσος να είναι αρκετά κοντά. Κάτι τέτοιο όμως δεν συμβαίνει καθώς για τις εταιρείες πληροφορικής η δειγματική μέση τιμή και διάμεσος του κέρδους είναι αντίστοιχα 680,8500 και 135,6500 αντίστοιχα ενώ οι αντίστοιχες τιμές για τις φαρμακευτικές εταιρείες είναι 824,1254 και 747. Άρα τα αποτελέσματα δεν γενικεύονται στις πληθυσμιακές μέσες τιμές.

ΚΕΦΑΛΑΙΟ ΕΚΤΟ

Έλεγχος για τις παραμέτρους θέσης δύο πληθυσμών με εξαρτημένα δείγματα

Στο κεφάλαιο αυτό θα ασχοληθούμε με τον έλεγχο της υπόθεσης της ισότητας δύο μέσων τιμών με εξαρτημένα δείγματα. Εξαρτημένα δείγματα εμφανίζονται συνήθως στις ακόλουθες περιπτώσεις:

α) Σε πειράματα, μελέτες των οποίων ο σκοπός είναι η διερεύνηση της αποτελεσματικότητας μίας θεραπείας. Για το λόγο αυτό οι τιμές μίας ή περισσοτέρων μεταβλητών καταγράφονται στην ίδια πειραματική μονάδα πριν και μετά την εφαρμογή της μεθόδου.

β) Στην περίπτωση των διδύμων.

γ) Όταν θεωρούμε πειραματικές μονάδες που μοιάζουν σε όλα τα υπόλοιπα χαρακτηριστικά πλην αυτού που θέλουμε να μελετήσουμε (ταιριαστά δεδομένα).

Στη βάση των εξαρτημένων δειγμάτων θα ασχοληθούμε με το ακόλουθο πρόβλημα:

Έστω ένα τυχαίο δείγμα $X_1,\ldots,X_n$ μεγέθους n από έναν πληθυσμό με μέση τιμή μ_1 και διακύμανση σ_1^2. Επιπλέον έστω ένα τυχαίο δείγμα $Y_1,\ldots,Y_n$ μεγέθους n από έναν πληθυσμό με μέση τιμή μ_2 και διακύμανση σ_2^2. Επιπρόσθετα υποθέτουμε ότι τα δύο δείγματα είναι εξαρτημένα. Ενδιαφερόμαστε για τον έλεγχο, σε επίπεδο σημαντικότητας α, της μηδενικής υπόθεσης

$$H_0 : \mu_1 = \mu_2,$$

ως προς μία εκ των

$$H_a : \mu_1 > \mu_2, \quad H_a : \mu_1 < \mu_2, \; H_a : \mu_1 \neq \mu_2 .$$

Το πρώτο βήμα για τη μελέτη του προβλήματος είναι η δημιουργία των διαφορών $D_i = X_i - Y_i,\ i = 1,...,n$.

Το παραπάνω πρόβλημα ελέγχεται υπό κάποιες υποθέσεις με τον παραμετρικό έλεγχο του t-test. Όταν κάποια από τις υποθέσεις αυτές δεν ικανοποιείται και δεν υπάρχει τρόπος διόρθωσης του προβλήματος ο έλεγχος ανάγεται σε αυτόν ότι οι πληθυσμιακές

διάμεσοι είναι ίσες. Τα αποτελέσματα του τελευταίου ελέγχου γενικεύονται για τον δοθέν έλεγχο όταν τα δεδομένα είναι συμμετρικά.

6.1 Μεθοδολογία-Υλοποίηση στο S.P.S.S.

Η μεθοδολογία που θα χρησιμοποιηθεί για τη στατιστική ανάλυση ενός τέτοιου προβλήματος εξαρτάται από το αν πληρούνται ή όχι κάποιες προϋποθέσεις, τις οποίες και πρέπει αρχικά να ελέγξει ο ερευνητής. Πιο συγκεκριμένα, ελέγχουμε

α) αν το ποσοστό των ακραίων τιμών στις διαθέσιμες παρατηρήσεις $D_i = X_i - Y_i,\ i = 1,...,n$, ξεπερνά το 10% αυτών, και

β) αν ο πληθυσμός από τον οποίο λαμβάνεται το τυχαίο δείγμα $D_i = X_i - Y_i,\ i = 1,...,n$, μπορούμε να ισχυριστούμε ότι περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Ανάλογα με τα αποτελέσματα των παραπάνω ελέγχων προβαίνουμε στον παραμετρικό ή στο μη παραμετρικό έλεγχο. Για το λόγο αυτό στη συνέχεια παρουσιάζονται όλα τα πιθανά αποτελέσματα των α) και β), τα διάφορα βήματα της ανάλυσης και οι αποφάσεις στις οποίες οδηγούμαστε.

1. Αρχικά ελέγχουμε αν υπάρχουν ακραίες τιμές στις διαθέσιμες δειγματικές τιμές D_i . Αν το ποσοστό των ακραίων τιμών δε ξεπερνά το 10%, τότε προχωρούμε στο επόμενο βήμα. 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 έλεγχος (βλέπε βήμα 4).
2. Στο βήμα 2, χρησιμοποιώντας το τεστ των Shapiro-Wilk καθώς και γραφικούς τρόπους, ελέγχουμε αν οι διαθέσιμες δειγματικές παρατηρήσεις D_i (είτε οι αρχικές είτε οι μετασχηματισμένες του βήματος 1) μπορούν να θεωρηθούν ότι προέρχονται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή. Αν ο έλεγχος της κανονικότητας μας υποδεικνύει ότι η υπόθεση της κανονικότητας δεν απορρίπτεται (p-τιμή >α), τότε η ανάλυση θα συνεχιστεί με τον παραμετρικό έλεγχο (βλέπε βήμα 3). Αν η υπόθεση της κανονικότητας απορρίπτεται (τεστ Shapiro-Wilk, p-τιμή <α), τότε ελέγχουμε αν το πρόβλημα της μη κανονικότητας διορθώνεται μετασχηματίζοντας κατάλληλα 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 (βήμα 3). Σε αντίθετη περίπτωση, αν το πλήθος των δειγματικών παρατηρήσεων D_i, μη λαμβάνοντας υπόψη αυτές που έχουν αφαιρεθεί στο βήμα 1, είναι μεγάλο (συνήθως μεγαλύτερο ή ίσο του 30), κάνοντας χρήση του Κεντρικού Οριακού Θεωρήματος, προβαίνουμε στον παραμετρικό έλεγχο της υπό έλεγχο υπόθεσης (βλέπε βήμα 3), όπου η p-τιμή του ελέγχου και το διάστημα εμπιστοσύνης θα είναι προσεγγιστικά. Στην περίπτωση τώρα που το πρόβλημα της μη κανονικότητας δε διορθώνεται (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. <u>Παραμετρικός έλεγχος- Τ τεστ συγκρίσεως ζευγών</u>: Χρησιμοποιούμε τη στατιστική συνάρτηση $t = \dfrac{\bar{D}}{S_D/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$, όπου $\bar{D}$ και S_D η μέση τιμή και τυπική απόκλιση του δείγματος των διαφορών $D_i = X_i - Y_i,\ i = 1,2,...,n$. Οι κρίσιμες περιοχές του ελέγχου είναι αντίστοιχα $t \geq t_{n-1,a}$, $t \leq -t_{n-1,a}$, $|t| \geq t_{n-1,a/2}$ $\left(t \geq t_{n-1,a/2}\ \text{ή}\ t \leq -t_{n-1,a/2}\right)$. Επιπλέον το 100(1-α)% Δ.Ε. για την $\mu_1 - \mu_2$ είναι:

$$\left(\bar{D} - t_{n-1,a/2}\frac{S_D}{\sqrt{n}}, \bar{D} + t_{n-1,a/2}\frac{S_D}{\sqrt{n}}\right).$$

<u>Επισήμανση</u>: Σε περίπτωση που έχει χρησιμοποιηθεί κάποιος μετασχηματισμός διόρθωσης του προβλήματος είτε της ύπαρξης πολλών ακραίων τιμών είτε της μη κανονικότητας, τότε όλα τα παραπάνω αναφέρονται στις μετασχηματισμένες τιμές και στο τροποποιημένο σε μέγεθος δείγμα. Ειδικότερα, αν έχει χρησιμοποιηθεί ο μετασχηματισμός του λογαρίθμου, θα προβούμε στον έλεγχο αν ο μέσος λογάριθμος του ενός πληθυσμού δε διαφέρει στατιστικά σημαντικά από το μέσο λογάριθμο του άλλου.

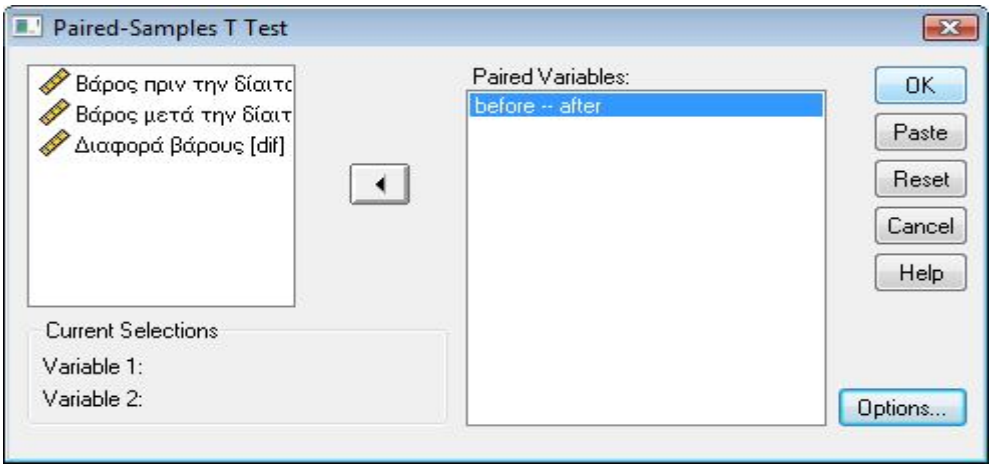
<u>Υλοποίηση στο S.P.S.S.</u>

Για τη διεξαγωγή του t τεστ συγκρίσεως ζευγών, από το κύριο μενού του λογισμικού επιλέγουμε

i. Analyze→Compare Means→Paired-Samples T Test.

ii. Στο νέο παράθυρο διαλόγου που προκύπτει διαλέγουμε αρχικά τη μεταβλητή που καταγράφει π.χ. τις τιμές του βάρους πριν την εφαρμογή της δίαιτας και η οποία θα χαρακτηρισθεί ως Variable 1. Έπειτα επιλέγουμε εκείνη που καταγράφει τις τιμές του

βάρους μετά την εφαρμογή της δίαιτα και η οποία θα πάει στο πλαίσιο Variable2. Στη συνέχεια μετακινούμε το ζεύγος των μεταβλητών στο πλαίσιο Paired Variables.



iii. Από την επιλογή Options έχουμε τη δυνατότητα να καθορίσουμε τον τρόπο χειρισμού των ελλιπών τιμών καθώς και να προσδιορίσουμε το βαθμό εμπιστοσύνης του διαστήματος εμπιστοσύνης που θα κατασκευαστεί για τη μέση απώλεια βάρους.

<u>Σχόλιο</u>: Εναλλακτικά ο έλεγχος της υπόθεσης $H_0 : \mu_D = d_0$ θα μπορούσε να διενεργηθεί μέσω της διαδικασίας Compare Means One Sample για τη μεταβλητή όπου καταγράφονται οι δειγματικές τιμές της διαφοράς και με Test Value την τιμή d_0.

4. <u>Μη παραμετρικός έλεγχος-Wilcoxon</u>: Σύμφωνα με αυτόν θεωρούμε προς έλεγχο την ισοδύναμη μηδενική υπόθεση ότι οι διαφορές $D_i = X_i - Y_i,\ i = 1,...,n$, προέρχονται από μία συμμετρική περί το μηδέν κατανομή. Για την υλοποίηση του αρχικά τοποθετούμε τις διαφορές $D_i = X_i - Y_i,\ i = 1,...,n$, σε αύξουσα τάξη μεγέθους μη λαμβάνοντας υπόψη το πρόσημο. Έπειτα αντικαθιστούμε την j κατά σειρά μεγέθους διαφορά με $+j$ ή $-j$ ανάλογα με το αν η συγκεκριμένη διαφορά είναι θετική ή αρνητική αντίστοιχα. Αν υπάρχουν μηδενικές διαφορές απορρίπτονται από τη μελέτη με ανάλογη μείωση του μεγέθους του δείγματος. Έπειτα υπολογίζουμε το στατιστικό $T = \min\left(T^+, T^-\right)$, όπου T^+ και T^- αντίστοιχα είναι το άθροισμα των θετικών και αρνητικών τάξεων αντίστοιχα. Αποδεικνύεται τότε ότι για $n \geq 8$, $Z = \dfrac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \underset{H_0}{\overset{\pi\rho\sigma.}{\sim}} N(0,1)$, και ο έλεγχος γίνεται κατά τα γνωστά

χρησιμοποιώντας αυτήν τη στατιστική συνάρτηση, με τις p-τιμές να προσδιορίζονται κατά ανάλογο τρόπο με το Z τεστ.

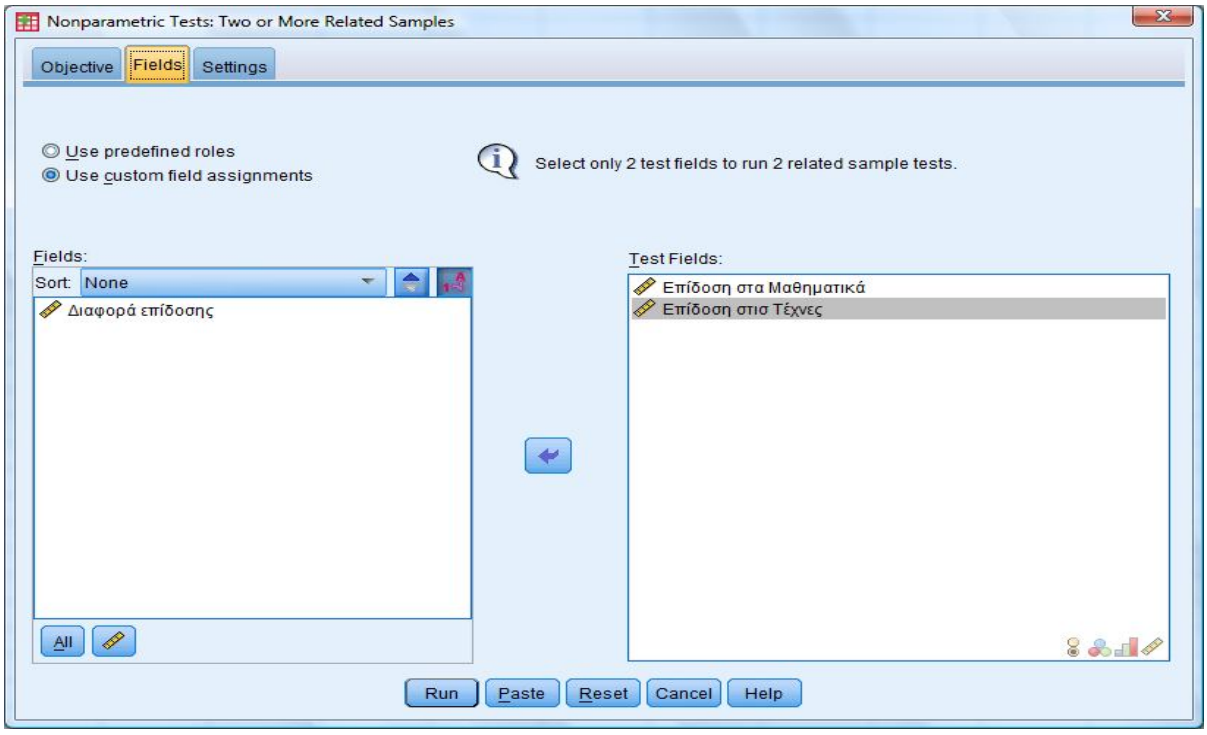
Σχόλιο: Για τον έλεγχο της υπό μελέτης μηδενικής υπόθεσης έχει προταθεί στη βιβλιογραφία και το προσημικό τεστ για συγκρίσεις κατά ζεύγη. Όμως, το τεστ του Wilcoxon είναι αποτελεσματικότερο για αυτό και δεν αναφέρθηκε το προσημικό τεστ.

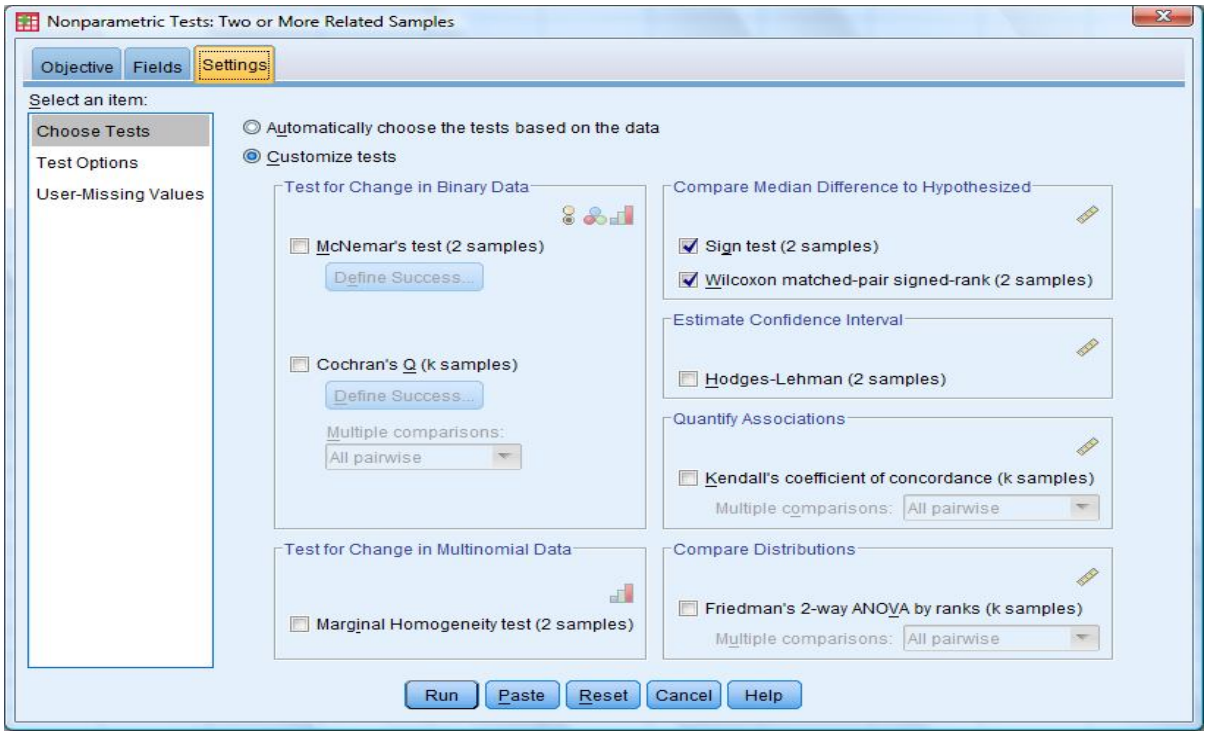
Υλοποίηση στο S.P.S.S.

Από το κύριο μενού επιλέγουμε

i. Analyze→Nonparametric Tests→Related Samples.

ii. Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε στο πλαίσιο Objective την επιλογή Customize analysis, έτσι ώστε στη συνέχεια από τα πλαίσια Fields και Settings να καθορίσουμε τον έλεγχο τον οποίο θέλουμε να διενεργηθεί όπως φαίνεται στα σχήματα που ακολουθούν.





Στο παράθυρο διαλόγου Settings επομένως επιλέγουμε τον προσημικό έλεγχο (Sign test) και τον έλεγχο των Mann-Whiney.

6.2 Παραδείγματα

Παράδειγμα 1ο

Στον πίνακα που ακολουθεί (βλέπε Ζωγράφος, 2003, σελ. 206) δίνεται το βάρος σε κιλά 9 γυναικών πριν και μετά την εφαρμογή μίας δίαιτας αδυνατίσματος τεσσάρων εβδομάδων

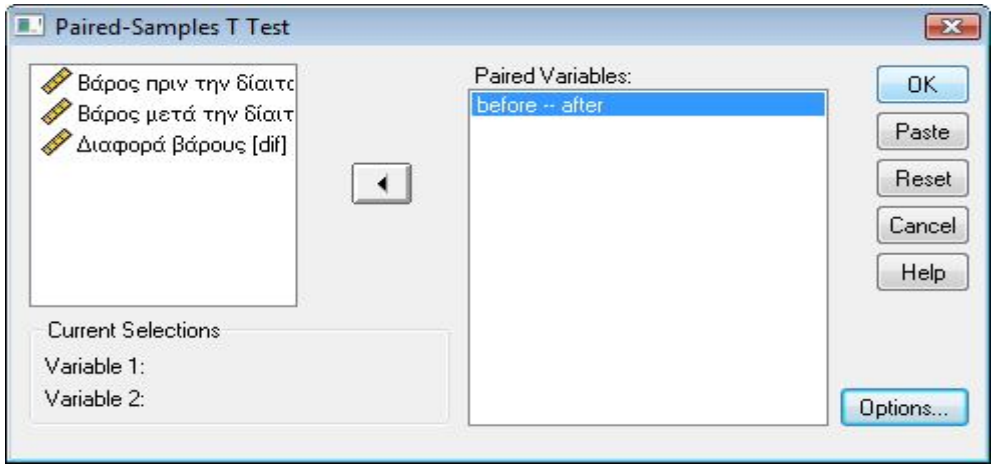
Πριν: 67 81 57 68 67 69 66 77 54
Μετά: 59 69 56 59 63 66 64 71 54.

Με βάση τα δεδομένα αυτά να ελεγχθεί αν είναι εφικτό ο ισχυρισμός ότι η δίαιτα είναι αποτελεσματική.

Επειδή πρόκειται για μετρήσεις του βάρους στα ίδια άτομα πριν και μετά την εφαρμογή μίας δίαιτας τα δύο δείγματα είναι εξαρτημένα. Επομένως, αρχικά ελέγχουμε αν

δεν υπάρχουν ακραίες τιμές στις δειγματικές τιμές της απώλεια βάρους και αν αυτές μπορούμε να ισχυριστούμε ότι προέρχονται από κανονικό πληθυσμό (δημιουργία στήλης διαφορών μέσω της διαδικασίας Transform Compute). Προκύπτει από το θηκόγραμμα ότι δεν υπάρχουν ακραίες τιμές και χρησιμοποιώντας το στατιστικό τεστ των Shapiro-Wilk προκύπτει ότι η υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί (p-τιμή=0,727>0.05).

Από τη διεξαγωγή του t τεστ συγκρίσεως ζευγών, που επιτυγχάνεται μέσω της διαδικασίας Analyze→Compare Means→Paired-Samples T Test όπως φαίνεται στο σχήμα που ακολουθεί προκύπτουν τα αποτελέσματα που παρατίθενται και ερμηνεύονται



Ερμηνεία αποτελεσμάτων

Από τον πίνακα Paired Samples Statistics έχουμε ότι το μέσο βάρος των 9 ατόμων πριν την δίαιτα ήταν 67.33 κιλά με τυπική απόκλιση 8.44097 και τυπικό σφάλμα για τη μέση τιμή 2.81366. Οι αντίστοιχες ποσότητες μετά τη διενέργεια της δίαιτας είναι 62.33 κιλά, 5.78792 και 1.92931, αντίστοιχα. Παρατηρούμε δηλαδή μία μείωση του βάρους κατά 5 κιλά. Θα εξεταστεί στη συνέχεια αν αυτή είναι στατιστικά σημαντική.

Στον πίνακα Paired Samples Correlations έχουμε ότι το βάρος πριν και το βάρος μετά την δίαιτα είναι συσχετισμένα (υψηλή στατιστικά σημαντική θετική γραμμική συσχέτιση).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Βάρος πριν την δίαιτα	67,3333	9	8,44097	2,81366
	Βάρος μετά την δίαιτα	62,3333	9	5,78792	1,92931

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Βάρος πριν την δίαιτα & Βάρος μετά την δίαιτα	9	,906	,001

Τα σημαντικότερα όμως αποτελέσματα δίνονται στον πίνακα Paired Samples Test. Προκύπτει ότι η μέση απώλεια βάρους είναι 5 κιλά (δίνεται στη στήλη Mean), ενώ η τυπική απόκλιση των διαφορών και το τυπικό σφάλμα για τη μέση τιμή είναι 4.03113 και 1.34371 κιλά αντίστοιχα. Επιπλέον μας δίνεται το 95% διάστημα εμπιστοσύνης για τη μέση διαφορά (1.90140,8.09860), καθώς και η τιμή του t τεστ για τον έλεγχο ότι δεν υπάρχουν στατιστικά σημαντικές διαφορές στο μέσο βάρος των ατόμων πριν και μετά τη δίαιτα. Από την p-τιμή του ελέγχου συμπεραίνουμε ότι υπάρχουν στατιστικά σημαντικές διαφορές στο μέσο βάρος πριν και μετά τη δίαιτα και αφού η μέση διαφορά είναι 5 καταλαβαίνουμε ότι η δίαιτα επιφέρει στατιστικά σημαντική μείωση του βάρους.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Βάρος πριν την δίαιτα - Βάρος μετά την δίαιτα	5	4,03113	1,34371	1,90140	8,09860	3,721	8	,006

Η αναφορά αφήνεται ως άσκηση.

Παράδειγμα 2^{ο}

Παρακάτω παρατίθενται οι επιδόσεις 15 μαθητών στα Μαθηματικά και στις Καλές Τέχνες (βλέπε Παπαΐωάννου και Φερεντίνος, 2000, σελ. 263). Να ελεγχθεί με επίπεδο σημαντικότητας 5% αν υπάρχει στατιστικά σημαντική διαφορά στις επιδόσεις των μαθητών στα δύο γνωστικά αντικείμενα.

Μαθηματικά: 22 37 36 38 42 58 58 60 62 65 66 56 66 67 62

Καλές Τέχνες: 53 68 42 49 51 65 51 71 55 74 68 64 67 73 65.

Πρόκειται για μετρήσεις της επίδοσης σε δύο γνωστικά αντικείμενα και στους ίδιους μαθητές. Επομένως γίνεται εύκολα κατανοητό ότι έχουμε δύο εξαρτημένα δείγματα.

Επειδή πρόκειται για μετρήσεις της επίδοσης στα ίδια άτομα τα δύο δείγματα είναι εξαρτημένα. Επομένως, αρχικά ελέγχουμε αν δεν υπάρχουν ακραίες τιμές στις δειγματικές τιμές της διαφοράς της επίδοσης και αν αυτές μπορούμε να ισχυριστούμε ότι προέρχονται από κανονικό πληθυσμό. Αρχικά σχηματίζουμε στο S.P.S.S. τη στήλη των διαφορών των επιδόσεων στα Μαθηματικά και στις Καλές τέχνες (έστω κωδική ονομασία diafora και Label Διαφορά Επίδοσης). Το πρώτο βήμα της ανάλυσης είναι ο έλεγχος της ύπαρξης ακραίων τιμών και της κανονικότητας των διαφορών. **Διαπιστώνεται (αφήνεται ως άσκηση) ότι υπάρχουν ακραίες τιμές σε ποσοστό μεγαλύτερο του 10% και το πρόβλημα δε διορθώνεται με το μετασχηματισμό του λογαρίθμου, επομένως καταφεύγουμε σε μη παραμετρικούς τρόπους ελέγχου**. Ακολουθώντας τα βήματα που αναλυτικά περιγράφηκαν στην παράγραφο 6.1 προκύπτουν τα ακόλουθα αποτελέσματα

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Επίδοση στισ Τέχνες and Επίδοση στα Μαθηματικά equals 0.	Related-Samples Sign Test	.007[1]	Reject the null hypothesis.
2	The median of differences between Επίδοση στα Μαθηματικά and Επίδοση στισ Τέχνες equals 0.	Related-Samples Wilcoxon Signed Ranks Test	.009	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1] Exact significance is displayed for this test.

Από την p-τιμή του στατιστικού τεστ του Wilcoxon συμπεραίνουμε ότι υπάρχουν στατιστικά σημαντικές διαφορές στην επίδοση στα Μαθηματικά και στις Τέχνες (p-τιμή=0.009<0.05). Επιπλέον από τον πίνακα που ακολουθεί (και αποκτήθηκε μέσω της διαδικασίας Analyze Descriptive Statistics Explore) προκύπτει ότι η πληθυσμιακή διάμεσος της επίδοσης στις τέχνες είναι στατιστικά σημαντικά μεγαλύτερη και τα αποτελέσματα δεν μπορούν να γενικευτούν στις μέσες τιμές.

Descriptives

			Statistic	Std. Error
Επίδοση στα Μαθηματικά	Mean		53,0000	3,65148
	95% Confidence Interval for Mean	Lower Bound	45,1683	
		Upper Bound	60,8317	
	5% Trimmed Mean		53,9444	
	Median		58,0000	
	Variance		200,000	
	Std. Deviation		14,14214	
	Minimum		22,00	
	Maximum		67,00	
	Range		45,00	
	Interquartile Range		27,00	
	Skewness		-,949	,580
	Kurtosis		-,278	1,121
Επίδοση στις Τέχνες	Mean		61,0667	2,57546
	95% Confidence Interval for Mean	Lower Bound	55,5428	
		Upper Bound	66,5905	

	5% Trimmed Mean	61,4074	
	Median	65,0000	
	Variance	99,495	
	Std. Deviation	9,97473	
	Minimum	42,00	
	Maximum	74,00	
	Range	32,00	
	Interquartile Range	17,00	
	Skewness	-,466	,580
	Kurtosis	-1,062	1,121

<u>Παράδειγμα 3ο</u> Αρχείο Teaching.sav *

Στο αρχείο αυτό υπάρχει η βαθμολογία 60 μαθητών οι οποίοι επιλεχθήκαν τυχαία από το σύνολο των μαθητών μια πόλης. Σκοπός της μελέτης αυτής ήταν να εξετάσει, αν μια καινούργια μέθοδος διδασκαλίας ενός μαθήματος βελτιώνει την απόδοση των μαθητών. Pretest είναι η βαθμολογία των μαθητών στο συγκεκριμένο μάθημα πριν την διδασκαλία με την νέα μέθοδο και Posttest είναι η βαθμολογία των ίδιων μαθητών στο συγκεκριμένο μάθημα μετά την διδασκαλία της νέας μεθόδου. Να εξετασθεί η αποτελεσματικότητα της νέας μεθόδου διδασκαλίας.

Η υλοποίηση αφήνεται ως άσκηση, ενώ δίνεται η αναφορά.

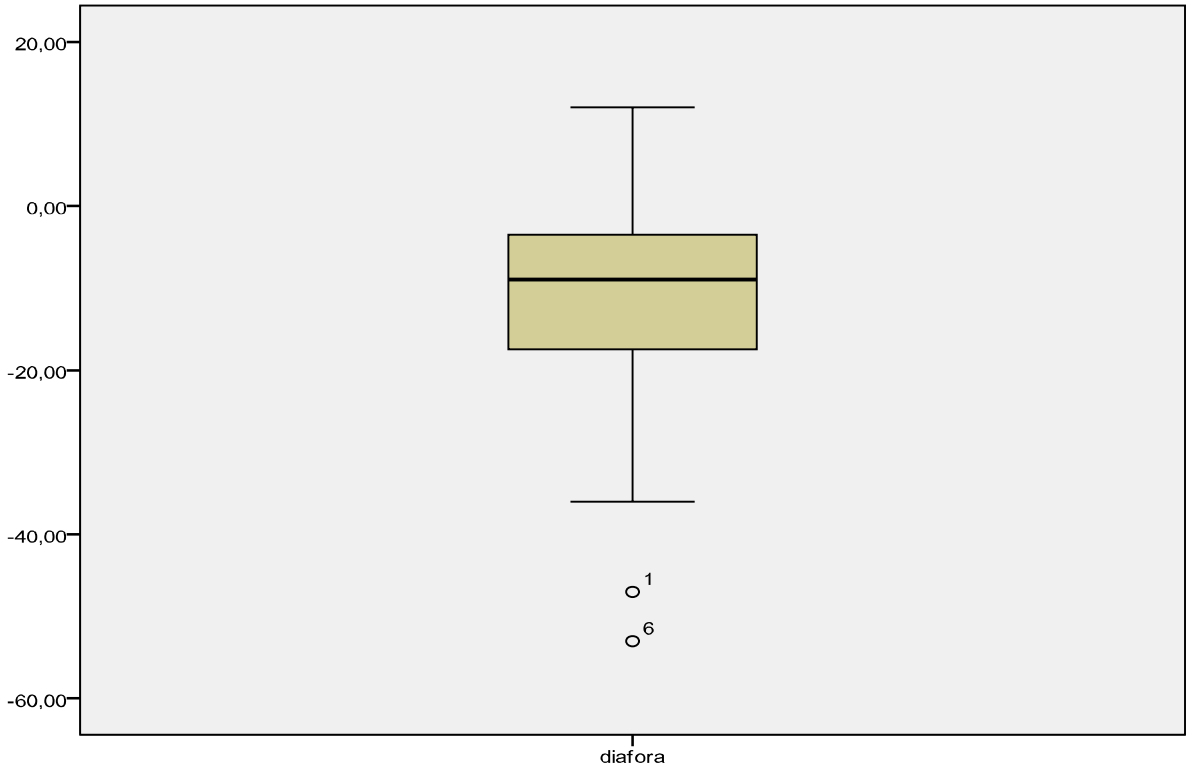
Αναφορά:

Στο πρόβλημα αυτό θέλουμε να εξετάσουμε την αποτελεσματικότητα της νέας μεθόδου διδασκαλίας σε σχέση με την υπάρχουσα μέθοδο. Το πρόβλημα μας είναι ένα πρόβλημα ελέγχου ισότητας των μέσων τιμών δυο πληθυσμών. Επειδή όμως οι μετρήσεις μας, και για τις δυο μεθόδους, γίνονται στις ίδιες πειραματικές μονάδες συμπεραίνουμε ότι τα δείγματα μας δεν είναι ανεξάρτητα. Το πείραμα αυτό είναι της μορφής ΠΡΙΝ-ΜΕΤΑ. Για να απαντήσουμε στο αρχικό μας ερώτημα θα πρέπει να σχηματίσουμε τις διαφορές π. χ. pretest-posttest και έτσι το πρόβλημά μας μετατρέπεται σε ένα πρόβλημα ελέγχου για την μέση τιμή ενός πληθυσμού. Πιο συγκεκριμένα για το αν μέση τιμή της διαφοράς στη βαθμολογία πριν και μετά τη μέθοδο διδασκαλίας είναι ίση με μηδέν (0). Για να κάνουμε χρήση του t-τεστ για έναν πληθυσμό θα πρέπει, για το δείγμα μας, να ικανοποιούνται οι επόμενες προϋποθέσεις:

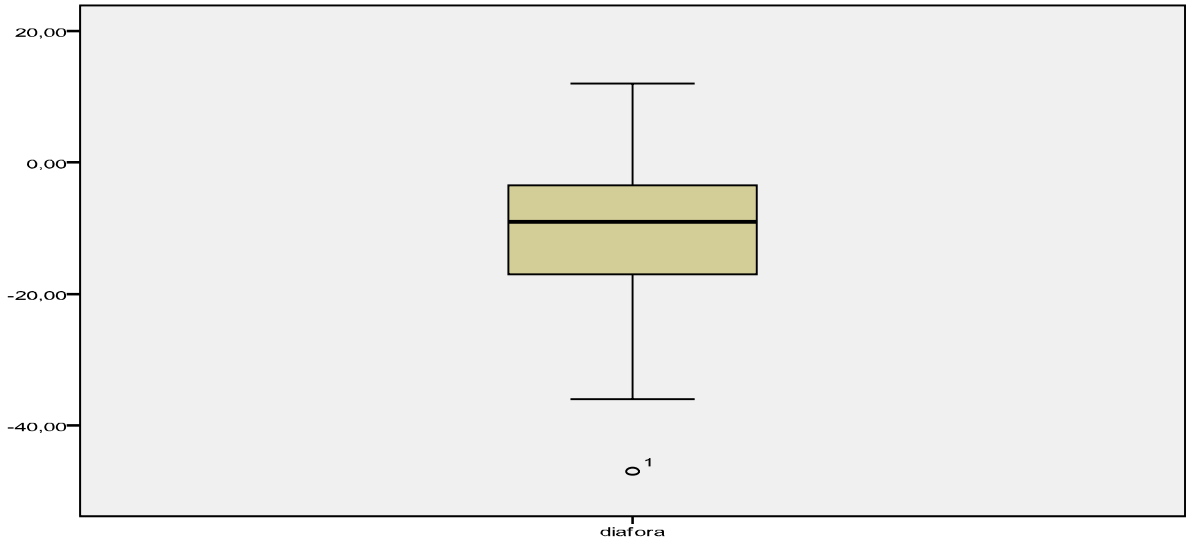
1. Να είναι τυχαίο.
2. Να μην έχει ακραίες τιμές σε ποσοστό μεγαλύτερο του 10%.
3. Να προέρχεται από πληθυσμό που περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Από το θηκόγραμμα προέκυψε ότι υπάρχουν δύο ακραίες παρατηρήσεις στις δειγματικές τιμές των διαφορών στην βαθμολογία πριν και μετά την εφαρμογή της μεθόδου διδασκαλίας οι παρατηρήσεις με αύξοντα αριθμό 6 και 1 με τιμές -53 και -47 αντίστοιχα (βλέπε θηκογράμματα 1,2,3). Οι παρατηρήσεις αυτές αποκλείονται από την περαιτέρω ανάλυση, επειδή ο συνολικός τους αριθμός δεν υπερβαίνει το 10% των παρατηρήσεων, 2/60*100%<10%).

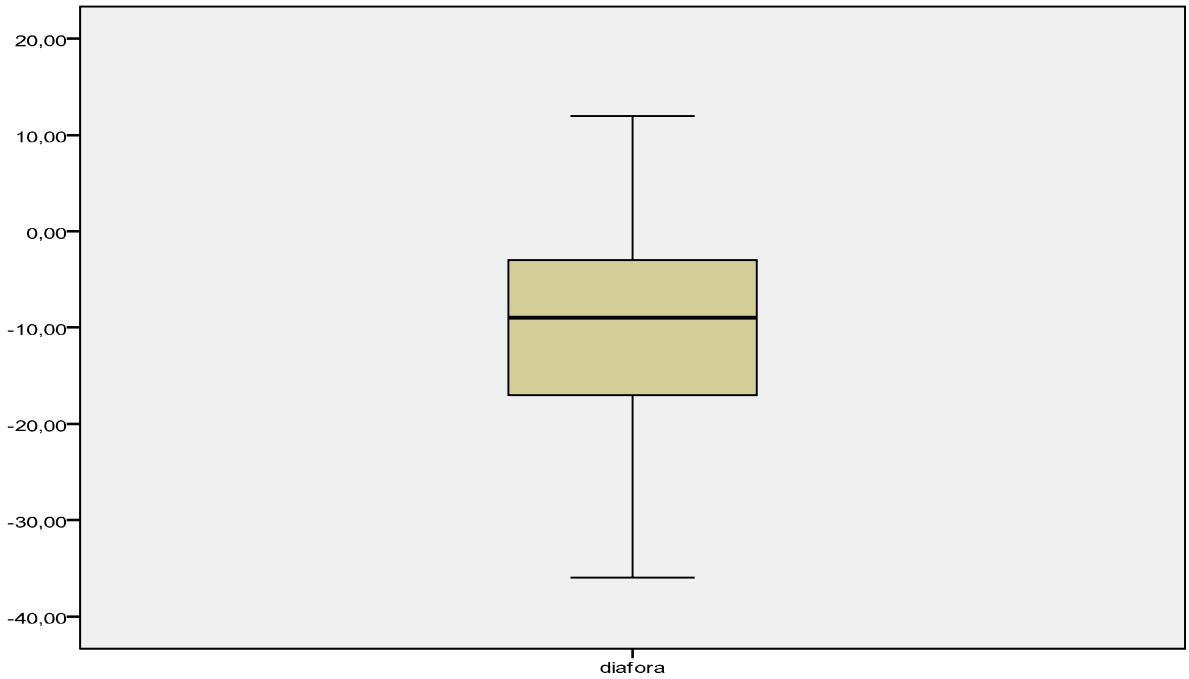
Θηκόγραμμα 1



Θηκόγραμμα 2



Θηκόγραμμα 3



Στη συνέχεια ελέγχουμε αν οι 58 δειγματικές παρατηρήσεις της διαφοράς της βαθμολογίας προέρχονται από κανονικό πληθυσμό. Η κρίσιμη πιθανότητα του τεστ των Shapiro-Wilk είναι p=0,021. Αυτό σημαίνει ότι η υπόθεση της κανονικής κατανομής με επίπεδο σημαντικότητας 5% θα πρέπει να απορριφθεί, ενώ με επίπεδο σημαντικότητας 1% δεν μπορεί να απορριφθεί. Αποτέλεσμα αυτού ήταν να καθορίσουμε το επίπεδο σημαντικότητας στο 1% σε όσα έπονται και θα χρησιμοποιηθεί για τον έλεγχο της υπό μελέτης υπόθεσης ο παραμετρικός έλεγχος του t-τεστ.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
diafora	,158	58	,001	,951	58	,021

a. Lilliefors Significance Correction

Σαν συμπέρασμα από την υλοποίηση του t-τεστ προέκυψε ότι: οι δύο μέθοδοι διδασκαλίας, η παλαιά και η καινούργια, διαφέρουν στατιστικά σημαντικά μεταξύ τους ($p<0,001$). Πιο συγκεκριμένα η μέση απόδοση της καινούργιας μεθόδου είναι κατά 10,5172 βαθμούς καλύτερη από την παλαιά. Ένα 99% διάστημα εμπιστοσύνης για την διαφορά protest-posttest είναι το (-14,0858 , -6,9487).

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
diafora	58	-10,5172	10,19845	1,33912

One-Sample Test

	Test Value = 0					
					99% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
diafora	-7,854	57	,000	-10,51724	-14,0858	-6,9487

ΚΕΦΑΛΑΙΟ ΕΒΔΟΜΟ

Έλεγχος για τις παραμέτρους θέσης περισσοτέρων των δύο πληθυσμών με ανεξάρτητα δείγματα

Έστω $Y_{j1},\ldots,Y_{jn_j}$, j το πλήθος $j=1,...,k$, $k\geq 2$ τυχαία ανεξάρτητα δείγματα μεγέθους n_j από έναν πληθυσμό με μέση τιμή μ_j και διακύμανση σ_j^2, άγνωστη. Ενδιαφερόμαστε για τον έλεγχο, σε επίπεδο σημαντικότητας α, της μηδενικής υπόθεσης

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k,$$

ως προς την εναλλακτική $H_a : \mu_i \neq \mu_j$, για τουλάχιστον ένα ζεύγος με $i \neq j, i, j = 1,...,k$, $k > 2$.

Επομένως, το ενδιαφέρον τώρα επικεντρώνεται στον έλεγχο ότι οι μέσες τιμές της εξαρτημένης μεταβλητής δύο ή περισσότερων ομάδων (επιπέδων του παράγοντα) δε διαφέρουν στατιστικά σημαντικά. Επομένως, γίνεται αντιληπτό ότι αποτελεί γενίκευση του προβλήματος της σύγκρισης των μέσων τιμών δύο πληθυσμών σε περισσότερους από δύο ανεξάρτητους πληθυσμούς (με ανεξάρτητα δείγματα).

Το παραπάνω πρόβλημα ελέγχεται υπό κάποιες υποθέσεις με τον παραμετρικό έλεγχο του t-test. Όταν κάποια από τις υποθέσεις αυτές δεν ικανοποιείται και δεν υπάρχει τρόπος διόρθωσης του προβλήματος ο έλεγχος ανάγεται σε αυτόν ότι οι πληθυσμιακές διάμεσοι είναι ίσες μεταξύ τους. Τα αποτελέσματα του τελευταίου ελέγχου γενικεύονται για τον δοθέν έλεγχο όταν τα δεδομένα είναι συμμετρικά.

7.1 Μεθοδολογία-Υλοποίηση στο S.P.S.S.

Η μεθοδολογία που θα χρησιμοποιηθεί για τη στατιστική ανάλυση ενός τέτοιου προβλήματος εξαρτάται από το αν πληρούνται ή όχι κάποιες προϋποθέσεις, τις οποίες και πρέπει αρχικά να ελέγξει ο ερευνητής. Πιο συγκεκριμένα, ελέγχουμε

α) αν το ποσοστό των ακραίων τιμών στις διαθέσιμες δειγματικές παρατηρήσεις από καθένα από τους k το πλήθος πληθυσμούς ξεπερνά το 10% αυτών, και

β) αν οι πληθυσμοί από τους οποίους λαμβάνονται τα τυχαία δείγματα μπορούμε να ισχυριστούμε ότι περιγράφονται ικανοποιητικά από την κανονική κατανομή.

Ανάλογα με τα αποτελέσματα των παραπάνω ελέγχων προβαίνουμε σε παραμετρικό έλεγχο ή σε μη παραμετρικό έλεγχο. Στη συνέχεια παρουσιάζονται όλα τα πιθανά αποτελέσματα των α) και β), τα διάφορα βήματα της ανάλυσης και οι αποφάσεις στις οποίες οδηγούμαστε.

1. Αρχικά ελέγχουμε αν υπάρχουν ακραίες τιμές στις διαθέσιμες δειγματικές τιμές σε καθένα από τους k σε πλήθος πληθυσμούς. Αν το ποσοστό των ακραίων τιμών σε καθένα από αυτά τα δείγματα δε ξεπερνά το 10%, τότε προχωρούμε στο βήμα 2. Αν το ποσοστό των ακραίων τιμών σε κάποιο από αυτά τα δείγματα 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 έλεγχος (βλέπε βήμα 4).

2. Στο βήμα 2, χρησιμοποιώντας το τεστ των Shapiro-Wilk καθώς και γραφικούς τρόπους, ελέγχουμε αν οι διαθέσιμες δειγματικές παρατηρήσεις από καθένα από τους k πληθυσμούς (είτε οι αρχικές είτε οι μετασχηματισμένες που έχουν προκύψει από το βήμα 1) προέρχονται από έναν πληθυσμό που περιγράφεται 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 τον παραμετρικό έλεγχο (βλέπε βήμα 3). Αν η υπόθεση της κανονικότητας απορρίπτεται για έναν από αυτούς ή και για τους k υπό εξέταση πληθυσμούς (τεστ Shapiro-Wilk, p-τιμή<α), τότε ελέγχουμε αν το πρόβλημα της μη κανονικότητας διορθώνεται μετασχηματίζοντας κατάλληλα 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 όλων των πληθυσμών, συνεχίζουμε την ανάλυση παραμετρικά (βήμα 3). Σε αντίθετη περίπτωση, αν το πλήθος των δειγματικών παρατηρήσεων (μη λαμβάνοντας υπόψη αυτές που έχουν αφαιρεθεί στο βήμα 1) εκείνου ή εκείνων των πληθυσμών που δεν περιγράφονται 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), όπου τα αποτελέσματα θα είναι προσεγγιστικά. Στην περίπτωση τώρα που το πρόβλημα της μη κανονικότητας κάποιου ή και των k πληθυσμών δε διορθώνεται (τεστ Shapiro-Wilk, p-τιμή <α), και ταυτόχρονα το πλήθος των δειγματικών παρατηρήσεων (μη λαμβάνοντας

υπόψη αυτές που έχουν αφαιρεθεί στο βήμα 1) από αυτόν τον πληθυσμό ή από αυτούς τους πληθυσμούς ανάλογα είναι μικρό (συνήθως μικρότερο του 30), συνεχίζεται η περαιτέρω ανάλυση μη παραμετρικά (βήμα 4).

3. Παραμετρικός έλεγχος: Σε αυτήν την περίπτωση η περαιτέρω ανάλυση επηρεάζεται από το αποτέλεσμα του έλεγχου της ισότητας των k το πλήθος πληθυσμιακών διακυμάνσεων.

ι) Ειδικότερα, αν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων δεν απορρίπτεται (τεστ του Levene, p-τιμή>α), τότε ο έλεγχος της υπόθεσης $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, γίνεται με τη βοήθεια του F στατιστικού του πίνακα ΑΝΑΔΙΑ, όπου

$$F = \frac{MS_{tr}}{MS_{res}} \overset{H_0}{\sim} F_{k-1,n-k},$$

με

$$MS_{tr} = \frac{SS_{tr}}{k-1} = \frac{\sum_{j=1}^{k} n_j \left(\overline{y}_{j.} - \overline{y}_{..} \right)^2}{k-1},$$

και

$$MS_{res} = \frac{SS_{res}}{n-k} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} n_j \left(y_{ji} - \overline{y}_{..} \right)^2 - \sum_{j=1}^{k} n_j \left(\overline{y}_{j.} - \overline{y}_{..} \right)^2}{n-k},$$

όπου $n = \sum_{j=1}^{k} n_j$, $\overline{y}_{j.} = \frac{\sum_{i=1}^{n_j} y_{ji}}{n_j}$, $\overline{y}_{..} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}}{n}$. Η κρίσιμη περιοχή του ελέγχου με επίπεδο σημαντικότητας a δίνεται από τη σχέση: $F \geq F_{k-1,n-k,a}$.

ii) Αν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων απορρίπτεται (τεστ του Levene, p-τιμή<α), τότε ο έλεγχος της υπόθεσης $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, γίνεται με το στατιστικό τεστ των Brown-Forsythe ή του Welch.

Αν χρησιμοποιώντας είτε το F στατιστικό του πίνακα ΑΝΑΔΙΑ είτε το στατιστικό τεστ των Brown-Forsythe ή του Welch, συμπεράνουμε ότι υπάρχουν στατιστικά σημαντικές διαφορές της μέσης τιμής της ποσοτικής μεταβλητής ως προς τις διάφορες κατηγορίες της ποιοτικής μεταβλητής (p-τιμή<α), τότε προκύπτει εύλογα το ερώτημα ποιο μ_i διαφέρει στατιστικά σημαντικά από τα υπόλοιπα. Για το σκοπό αυτό θα πρέπει να γίνουν οι

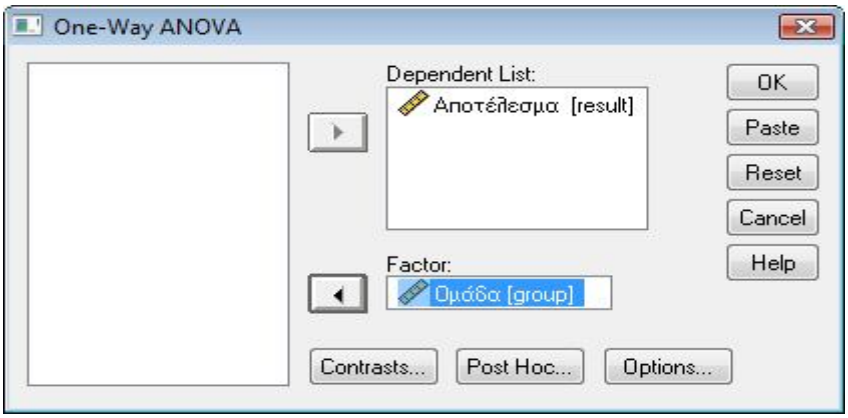
$\binom{k}{2} = \frac{k(k-1)}{2}$ το πλήθος έλεγχοι της μορφής: $H_0 : \mu_i = \mu_j, \quad i, j = 1, ..., k, \quad i \neq j$. Το συνηθέστερο λάθος που γίνεται είναι να πραγματοποιηθούν οι παραπάνω $\binom{k}{2}$ έλεγχοι χρησιμοποιώντας το t τεστ. Η χρήση του t τεστ έχει ως συνέπεια την αύξηση της πιθανότητας εσφαλμένης απόρριψης κάποιας από τις $\binom{k}{2}$ υποθέσεις. Η πιθανότητα εσφαλμένης απόρριψης δεν είναι πλέον ίση με το επίπεδο σημαντικότητας a κάθε ελέγχου, αλλά περίπου ίση με $1-(1-a)^{\binom{k}{2}}$ (βλέπε Καρακώστας (2002)). Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες μεθοδολογίες που είναι γνωστές ως Πολλαπλές Συγκρίσεις (Multiple Comparisons ή Post hoc tests). Άλλες από αυτές χρησιμοποιούνται όταν δεν έχει απορριφθεί η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων και άλλες στην περίπτωση που έχει απορριφθεί η υπόθεσης της ομοσκεδαστικότητας των πληθυσμών. Στην πρώτη περίπτωση ανήκουν μεταξύ άλλων, η ελάχιστη σημαντική διαφορά (έχουν ασκηθεί κριτικές ότι το επίπεδο σημαντικότητας αυξάνει όσο αυξάνει ο αριθμός των ομάδων k), η μέθοδος του Bonferroni (προτιμάται όταν το k είναι μικρό, επιτυγχάνει να διατηρεί το επίπεδο σημαντικότητας μικρότερο του a), του Sidak, του Scheffe (κατάλληλη μέθοδος όχι μόνο για σύγκριση ζευγών αλλά και περισσότερων). Στη δεύτερη περίπτωση, χρησιμοποιούνται οι μεθοδολογίες των Tamhane's T2 (συντηρητικό τεστ), Dunnett's T3, Games-Howell (τείνει να δίνει στατιστικά σημαντικές διαφορές) και Dunnett's C.

<u>Υλοποίηση στο S.P.S.S. (αρχείο school.sav)</u>

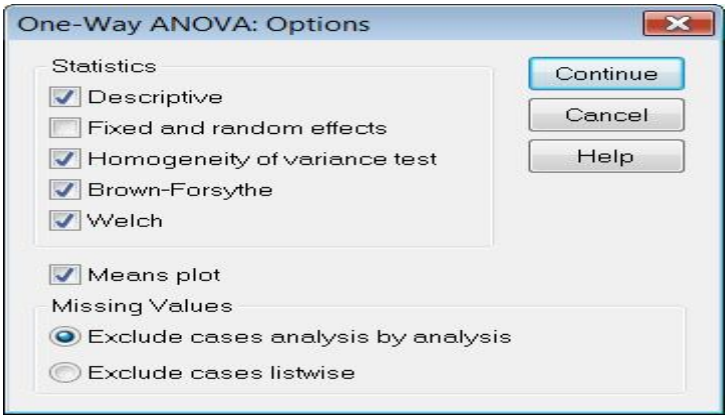
Επιλέγουμε από το αρχικό παράθυρο του S.P.S.S.

i. Analyze → Compare Means→ One-Way Anova.

ii. Στο παράθυρο διαλόγου που προκύπτει στο πλαίσιο Dependent List τοποθετούμε την εξαρτημένη ποσοτική μεταβλητή (έστω Αποτέλεσμα), ενώ στο πλαίσιο Factor τον παράγοντα που πιθανώς επηρεάζει την εξαρτημένη μεταβλητή (έστω Ομάδα).



Πατώντας το πλαίσιο Options από το πλαίσιο Statistics επιλέγουμε τα ακόλουθα:



- Descriptive. Δίνονται στο Output πληροφορίες για κάθε εξαρτημένη μεταβλητή, η οποία έχει δηλωθεί στο πλαίσιο Dependent List, ως προς τα διάφορα επίπεδα του παράγοντα που έχει δηλωθεί στο πλαίσιο Factor. Οι πληροφορίες αυτές αφορούν περιγραφικά μέτρα, όπως το πλήθος των πειραματικών μονάδων, τη μέση τιμή, την τυπική απόκλιση, το τυπικό σφάλμα για τη μέση τιμή, τη μέγιστη και την ελάχιστη τιμή. Επιπλέον υπολογίζεται το 95% διάστημα εμπιστοσύνης για τη μέση τιμή κάθε εξαρτημένης μεταβλητής για κάθε επίπεδο του παράγοντα
- Homogeneity of variance test. Υπολογίζει το στατιστικό του Levene για τον έλεγχο τις ισότητας των διακυμάνσεων. Επισημαίνεται ότι ο έλεγχος αυτός δεν είναι ανεξάρτητος της υπόθεσης της κανονικότητας.
- Brown-Forsythe και Welch. Υπολογίζει το στατιστικό των Brown-Forsythe και του Welch αντίστοιχα για τον έλεγχο της ισότητας των μέσων τιμών. Τα στατιστικά αυτά είναι καταλληλότερα από το F στατιστικό του πίνακα ΑΝΑΔΙΑ, όταν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων απορρίπτεται.

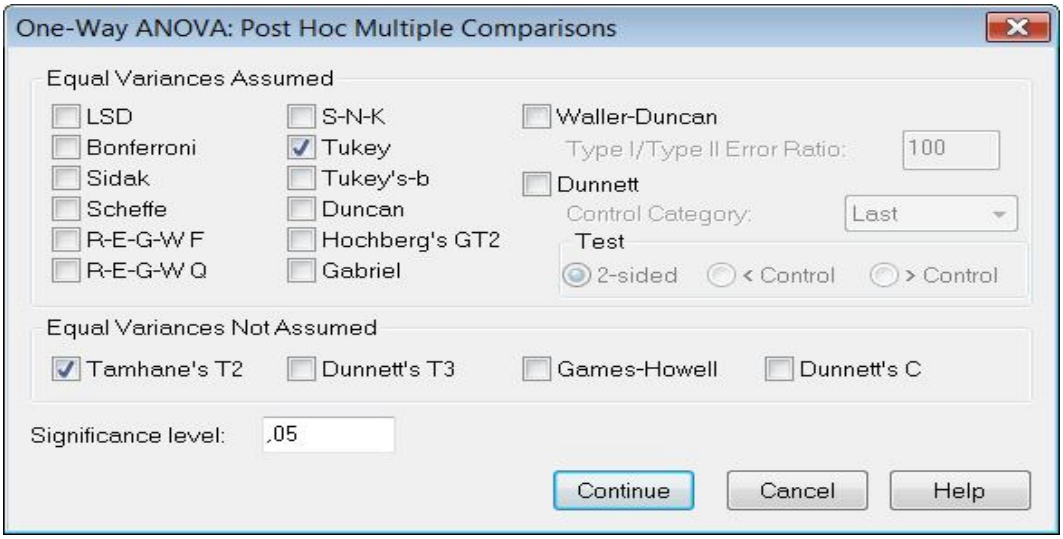
- Επιπλέον, επιλέγουμε το πλαίσιο Means plot προκειμένου να μας δώσει το λογισμικό ένα γράφημα των μέσων τιμών της εξαρτημένης μεταβλητής ως προς κάθε παράγοντα (για το σχηματισμό μίας εικόνας για τις διαφορές μεταξύ των επιπέδων).

Επιλέγουμε το πλαίσιο Continue και επανερχόμαστε στο αρχικό παράθυρο διαλόγου One-Way Anova. Παρατηρούμε ότι είναι διαθέσιμα και άλλα δύο πλαίσια τα Contrasts και Post Hoc.

Από την επιλογή Post Hoc Multiple Comparisons έχουμε ένα πλήθος μεθόδων πολλαπλών συγκρίσεων που άλλες εφαρμόζονται στην περίπτωση που ισχύει η ισότητα των πληθυσμιακών διακυμάνσεων (Equal Variances Assumed) και άλλες όταν είναι άνισες (Not Equal Variances Assumed).

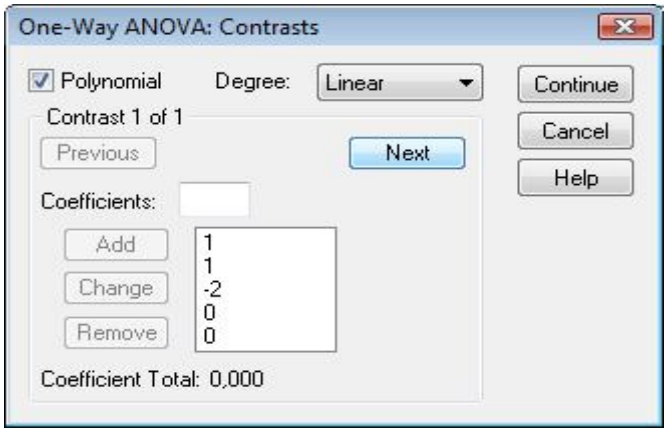
Δεν θα δώσουμε λεπτομέρειες για κάθε έλεγχο. Απλά αναφέρουμε ότι οι πιο δημοφιλείς από αυτούς τους ελέγχους είναι οι: LSD (ελάχιστη σημαντική διαφορά), Bonferroni, Tukey, Scheffe, Tamhane, Dunnett με επίπεδο σημαντικότητας που καθορίζεται στο πλαίσιο Significance level (συνήθως είναι 0.05 ή 0.01).

Ας υποθέσουμε ότι επιλέγουμε τα ακόλουθα:



Πολλές φορές ο ερευνητής δεν ενδιαφέρεται μόνο για τη σύγκριση μεταξύ δύο οποιωνδήποτε μέσων τιμών, αλλά είναι πιθανό να ενδιαφέρεται για τη σύγκριση περισσοτέρων των δύο. Στο πλαίσιο αυτό ενδιαφέρεται για τον έλεγχο κάποιας γραμμικής σχέσης της μορφής $\sum_{i=1}^{k} c_i \mu_i = 0$, όπου μ_i, $i = 1,...,k$, $k \geq 2$, είναι η πληθυσμιακή μέση τιμή του $i = 1,...,k$, $k \geq 2$, πληθυσμού και c_i κατάλληλοι συντελεστές τέτοιοι ώστε να ικανοποιείται η σχέση: $\sum_{i=1}^{k} c_i = 0$, $k \geq 2$. Τότε λέμε ότι πρόκειται για τον έλεγχο μίας

γραμμικής αντίθεσης. Το λογισμικό του S.P.S.S. μας δίνει τη δυνατότητα για τη διενέργεια τέτοιων ελέγχων από το παράθυρο διαλόγου One-Way Anova και την επιλογή Contrasts με την εισαγωγή των κατάλληλων συντελεστών. Στο πλαίσιο Degree διατηρούμε την επιλογή Linear. Έπειτα από το πλαίσιο Coefficients καθορίζουμε τους συντελεστές της γραμμικής αντίθεσης. Η σειρά των συντελεστών αυτών είναι σημαντική. Τοποθετούνται κατά αύξουσα σειρά σε αντιστοιχία με τις τιμές της ποιοτικής μεταβλητής. Έστω ότι θέλουμε να ελέγξουμε για παράδειγμα τις δύο πρώτες ομάδες με την τρίτη. Γίνεται αντιληπτό ότι οι συντελεστές θα είναι 1 1 -2 0 0.



Ερμηνεία αποτελεσμάτων

Στον πίνακα Descriptives, μας δίνονται για την Επίδοση ως προς τις πέντε διαφορετικές μεθόδους διδασκαλίας, το πλήθος των πειραματικών μονάδων (N), η μέση τιμή (Mean), η τυπική απόκλιση (Std. Deviation), το τυπικό σφάλμα για τη μέση τιμή (Std. Error), η ελάχιστη (Min) και η μέγιστη τιμή (Max). Επιπλέον υπολογίζεται το 95% διάστημα εμπιστοσύνης για τη μέση επίδοση ως προς τις πέντε μεθόδους διδασκαλίας (95% Confidence Interval for Mean). Παρατηρούμε ότι η μέση επίδοση των μαθητών της τρίτης ομάδας φαίνεται να είναι μεγαλύτερη-καλύτερη, ενώ αυτή των μαθητών της πέμπτης ομάδας δείχνει να είναι η χειρότερη. Μένει να επιβεβαιωθούν, στη συνέχεια, και στατιστικά αυτές οι αρχικές παρατηρήσεις.

Descriptives

Αποτέλεσμα

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
Ομάδα Α	9	19,6667	4,21307	1,40436	16,4282	22,9051	14	24
Ομάδα Β	9	18,3333	3,57071	1,19024	15,5886	21,0780	13	23
Ομάδα Γ	9	27,4444	2,45515	,81838	25,5572	29,3316	23,	30
Ομάδα Δ	9	23,4444	3,08671	1,02890	21,0718	25,8171	19,	28
Ομάδα Ε	9	16,1111	3,62093	1,20698	13,3278	18,8944	10	21
Total	45	21,0000	5,21362	,77720	19,4337	22,5663	10	30

Από τον πίνακα Test of Homogeneity of Variances μας δίνεται η τιμή, οι βαθμοί ελευθερίας και η p-τιμή του στατιστικού τεστ του Levene για τον έλεγχο της υπόθεσης των ίσων διακυμάνσεων. Συμπεραίνουμε ότι η υπόθεση της ισότητας των διακυμάνσεων δε μπορεί να απορριφθεί (p-τιμή=0.201>0.05). Επομένως πρέπει να χρησιμοποιήσουμε το F τεστ από τον πίνακα ΑΝΑΔΙΑ (άρα είναι λανθασμένο να χρησιμοποιήσουμε τα τεστ των Brown-Forsythe και Welch) για τον έλεγχο της υπόθεσης ότι δε διαφέρουν οι μέσες τιμές της επίδοσης ως προς τις 5 διαφορετικές μεθοδολογίες διδασκαλίας.

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
1,569	4	40	,201

Στον πίνακα ANOVA μας δίνεται ο πίνακας ΑΝΑΔΙΑ και όλες οι πληροφορίες που περιέχονται σε αυτόν. Από την τιμή και την p-τιμή του F στατιστικού τεστ προκύπτει ότι η υπόθεση της ισότητας των μέσων τιμών απορρίπτεται (p-τιμή<0.001).

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	722,667	4	180,667	15,268	,000
Within Groups	473,333	40	11,833		
Total	1196,000	44			

Άρα υπάρχει στατιστικά σημαντική διαφορά στη μέση επίδοση των μαθητών ως προς τις 5 μεθόδους διδασκαλίας. Για να διαπιστώσουμε ποιες μέθοδοι διδασκαλίας διαφέρουν

στατιστικά σημαντικά μεταξύ τους θα αποφανθούμε χρησιμοποιώντας τα αποτελέσματα των πολλαπλών συγκρίσεων με τη μέθοδο του Tukey (είναι μία από τις μεθόδους πολλαπλών συγκρίσεων που ισχύει όταν η υπόθεση της ομοσκεδαστικότητας των πληθυσμών δεν απορρίπτεται).

Με μία πρώτη ματιά κάποιος ίσως απογοητευτεί καθώς υπάρχει πληθώρα αριθμών στον πίνακα των πολλαπλών συγκρίσεων (πίνακας Multiple Comparisons). Όμως, η ερμηνεία αυτών των αποτελεσμάτων είναι εύκολη. Στις δύο πρώτες στήλες κάθε πίνακα μας δηλώνεται ποιος πληθυσμός συγκρίνεται με ποιον και αυτό γίνεται για κάθε συνδυασμό των πληθυσμών. Το αποτέλεσμα της σύγκρισης κάθε συνδυασμού φαίνεται στην αντίστοιχη γραμμή. Έτσι, θέλοντας να συγκρίνουμε την Ομάδα Γ με την Ομάδα Ε με την μέθοδο π.χ. του Tukey θα ερμηνεύσουμε τα αποτελέσματα της γραμμής με **bold** γραφή.

Παρατήρηση (βλέπε Καρακώστας, 2002, σελ. 116)

Πολλές φορές οι πολλαπλές συγκρίσεις οδηγούν σε μη συμβατά αποτελέσματα. Για παράδειγμα μπορούμε να πάρουμε ότι το Α δε διαφέρει στατιστικά σημαντικά από το Β, το Α είναι σημαντικά καλύτερο από το Γ και το Β δε διαφέρει σημαντικά από το Γ.

Μεταξύ άλλων παρατηρούμε ότι η μέση διαφορά στην επίδοση των μαθητών των ομάδων Γ και Ε είναι ίση με 11.3333, με τυπικό σφάλμα 1.6261. Δηλαδή η μέση επίδοση των μαθητών της τρίτης ομάδας είναι υψηλότερη κατά 11.3333 βαθμούς. Επιπλέον δίπλα σε αυτή την τιμή υπάρχει ένα αστεράκι (*). Αυτό μας υποδεικνύει (βλέπε υποσημείωση του πίνακα) ότι υπάρχει στατιστικά σημαντική διαφορά στη μέση επίδοση των μαθητών που ακολουθούν τον τρόπο διδασκαλίας Γ και αυτών της Ε. Η επίδοση αυτών της Γ ομάδας είναι στατιστικά σημαντικά μεγαλύτερη (λαμβάνοντας υπόψη και το αποτέλεσμα από τη στήλη Mean Difference). Τέλος, στη στήλη 95% Confidence Interval μας δίνεται το 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών.

Multiple Comparisons

Dependent Variable: Αποτέλεσμα

Tukey HSD

(I) Ομάδα	(J) Ομάδα	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower	Upper
Ομάδα Α	Ομάδα Β	1,33333	1,62161	,922	Lower	Upper
	Ομάδα Γ	-7,77778(*)	1,62161	,000	-12,4093	-3,1463
	Ομάδα Δ	-3,77778	1,62161	,157	-8,4093	,8537
	Ομάδα Ε	3,55556	1,62161	,203	-1,0759	8,1870
Ομάδα Β	Ομάδα Α	-1,33333	1,62161	,922	-5,9648	3,2981
	Ομάδα Γ	-9,11111(*)	1,62161	,000	-13,7426	-4,4796
	Ομάδα Δ	-5,11111(*)	1,62161	,024	-9,7426	-,4796
	Ομάδα Ε	2,22222	1,62161	,650	-2,4093	6,8537
Ομάδα Γ	Ομάδα Α	7,77778(*)	1,62161	,000	3,1463	12,4093
	Ομάδα Β	9,11111(*)	1,62161	,000	4,4796	13,7426
	Ομάδα Δ	4,00000	1,62161	,119	-,6315	8,6315
	Ομάδα Ε	**11,33333(*)**	**1,62161**	**,000**	**6,7019**	**15,9648**
Ομάδα Δ	Ομάδα Α	3,77778	1,62161	,157	-,8537	8,4093
	Ομάδα Β	5,11111(*)	1,62161	,024	,4796	9,7426
	Ομάδα Γ	-4,00000	1,62161	,119	-8,6315	,6315
	Ομάδα Ε	7,33333(*)	1,62161	,000	2,7019	11,9648
Ομάδα Ε	Ομάδα Α	-3,55556	1,62161	,203	-8,1870	1,0759
	Ομάδα Β	-2,22222	1,62161	,650	-6,8537	2,4093
	Ομάδα Γ	-11,33333(*)	1,62161	,000	-15,9648	-6,7019
	Ομάδα Δ	-7,33333(*)	1,62161	,000	-11,9648	-2,7019

* The mean difference is significant at the .05 level.

Επιπρόσθετα, από τον πίνακα Homogeneous Subsets μας δίνονται οι «πιθανές» ομογενείς ομάδες μέσων μέσω της μεθόδου του Tukey. Έτσι, για το παράδειγμά μας έχουμε τις ομάδες των {Ε, Β, Α}, των {Α, Δ} και {Δ, Γ}. Από τις αντίστοιχες p-τιμές συμπεραίνουμε ότι οι ομάδες αυτές είναι ομογενείς (δηλαδή η υπόθεση της ισότητας των μέσων τιμών εντός αυτών δεν απορρίπτεται).

Homogeneous Subsets

Αποτέλεσμα

	Ομάδα	N	Subset for alpha = .05		
		1	2	3	1
Tukey HSD(a)	Ομάδα Ε	9	16,1111		
	Ομάδα Β	9	18,3333		
	Ομάδα Α	9	19,6667	19,6667	
	Ομάδα Δ	9		23,4444	23,4444
	Ομάδα Γ	9			27,4444
	Sig.		,203	,157	,119

Means for groups in homogeneous subsets are displayed.

a Uses Harmonic Mean Sample Size = 9,000.

Στον πίνακα Contrast Coefficients του Output το λογισμικό μας δίνει τους συντελεστές κάθε γραμμικής αντίθεσης που σχηματίσαμε.

Contrast Coefficients

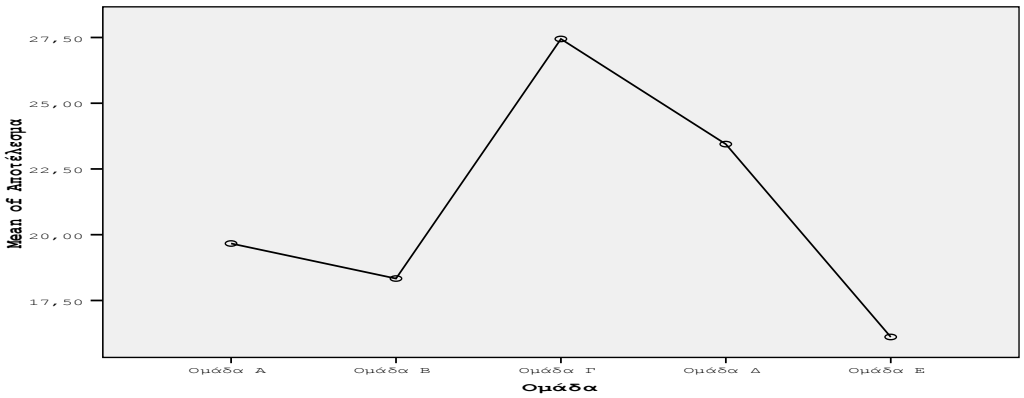
	Ομάδα				
Contrast	Ομάδα Α	Ομάδα Β	Ομάδα Γ	Ομάδα Δ	Ομάδα Ε
1	1	1	-2	0	0

Στον πίνακα Contrast Tests δίνονται τα αποτελέσματα για τους ελέγχους των γραμμικών αντιθέσεων που ζητήσαμε, ανάλογα με το αν ισχύει ή όχι η υπόθεση της ισότητας των διακυμάνσεων. Λόγω του αποτελέσματος του τεστ Levene έχουμε ότι ισχύει επομένως περιοριζόμαστε στα αποτελέσματα του πλαισίου Assume Equal Variances. Για κάθε γραμμική αντίθεση έχουμε στη στήλη Value of Contrast μία εκτίμηση και το αντίστοιχο τυπικό σφάλμα. Τέλος έχουμε την τιμή του t στατιστικού, τους βαθμούς ελευθερίας και την αντίστοιχη p-τιμή. Συμπεραίνουμε ότι η ομάδα Γ διαφέρει στατιστικά σημαντικά από τις ομάδες Α και Β.

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Αποτέλεσμα	Assume equal variances	1	-16,8889	2,80872	-6,013	40	,000
	Does not assume equal variances	1	-16,8889	2,46331	-6,856	22,530	,000

Τέλος, μας δίνεται το γράφημα των μέσων τιμών. Η εικόνα αυτή πολλές φορές μπορεί να μας ξεγελάσει, καθώς θα πρέπει να είμαστε προσεκτικοί στην κλίμακα μέτρησης που χρησιμοποιείται στον κατακόρυφο άξονα. Επιπλέον, λαμβάνοντας υπόψη ότι δε μας δίνεται καμία πρόσθετη και ουσιαστική πληροφορία σε σχέση με αυτές της στήλης Mean του πίνακα Descriptives προτείνεται η αποφυγή δημιουργίας της.



4. Μη παραμετρικός έλεγχος: Σε αυτή την περίπτωση το πρόβλημα ελέγχου αντιμετωπίζεται με το μη παραμετρικό έλεγχο των Kruskal and Wallis (1952), το οποίο αποτελεί επέκταση του τεστ των Mann-Whitney σε περισσότερους από δύο πληθυσμούς. Αποδεικνύεται ότι το στατιστικό

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left(R_i - \frac{n_i(n+1)}{2} \right)^2 ,$$

όπου $R_i = \sum_{j=1}^{n_i} R(X_{ij})$, $i = 1,...,k$, $k \geq 3$, με $R(X_{ij})$ $i = 1,...,k$, $j = 1,...,n_i$, $k \geq 3$, οι τάξεις των διαθέσιμων δειγματικών τιμών των k δειγμάτων στο σύνολο των n παρατηρήσεων, στην περίπτωση μη ύπαρξης δεσμών ακολουθεί προσεγγιστικά υπό τη μηδενική υπόθεση X^2 κατανομή με $k-1$ βαθμούς ελευθερίας και η μηδενική υπόθεση απορρίπτεται αν $KW \geq X^2_{k-1,1-a}$.

Σχόλιο: Σε περίπτωση ύπαρξης δεσμών ανάμεσα στις δειγματικές παρατηρήσεις οι Kruskal-Wallis πρότειναν το στατιστικό: $KW^* = \frac{KW}{1 - \frac{\sum (d_i^3 - d_i)}{n^3 - n}}$, όπου KW είναι το σύνηθες στατιστικό των Kruskal-Wallis υπολογισμένο χρησιμοποιώντας τα midranks και d_i είναι ο αριθμός των δεσμών στο i δείγμα.

Παρατήρηση Το S.P.S.S. διεξάγει δύο ελέγχους για το παραπάνω πρόβλημα. Το τεστ των διαμέσων (median test) που ελέγχει την ισότητα των πληθυσμιακών διαμέσων και το τεστ των Kruskal-Wallis test που ουσιαστικά ελέγχει αν τα δείγματα προέρχονται από τον ίδιο πληθυσμό.

<u>Υλοποίηση στο S.P.S.S.</u> (Kruskal and Wallis (1952))

Σε ένα εργοστάσιο τρεις μηχανές χρησιμοποιούνται για την παραγωγή δοχείων εμφιαλώσεως. Κατά τη διάρκεια μιας εργάσιμης εβδομάδας καταγράφεται ο αριθμός των δοχείων που κατασκευάστηκαν από κάθε μηχανή και τα αποτελέσματα παρατίθενται στον πίνακα που ακολουθεί, με την επιπλέον επισήμανση ότι κάποιες μέρες δεν παρήχθησαν δοχεία από κάποιες μηχανές λόγω ότι είχαν τεθεί εκτός λειτουργίας.

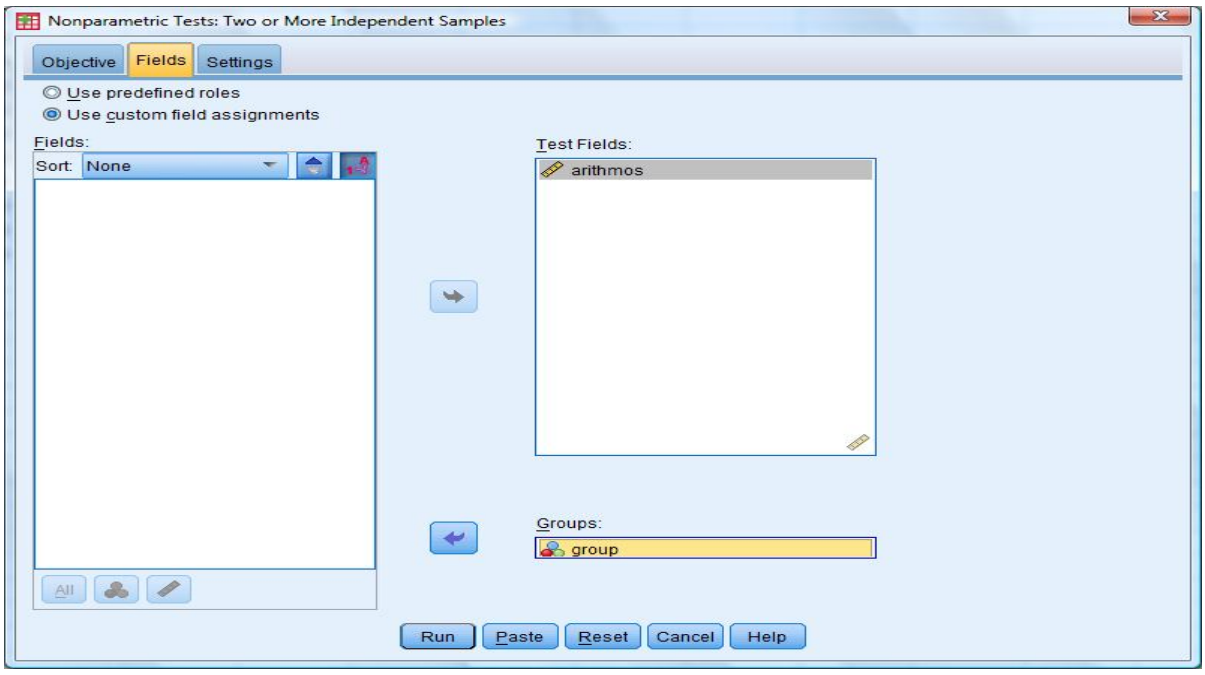
Μηχανή 1	340	345	330	342	338
Μηχανή 2	339	333	344		
Μηχανή 3	347	343	349	355	

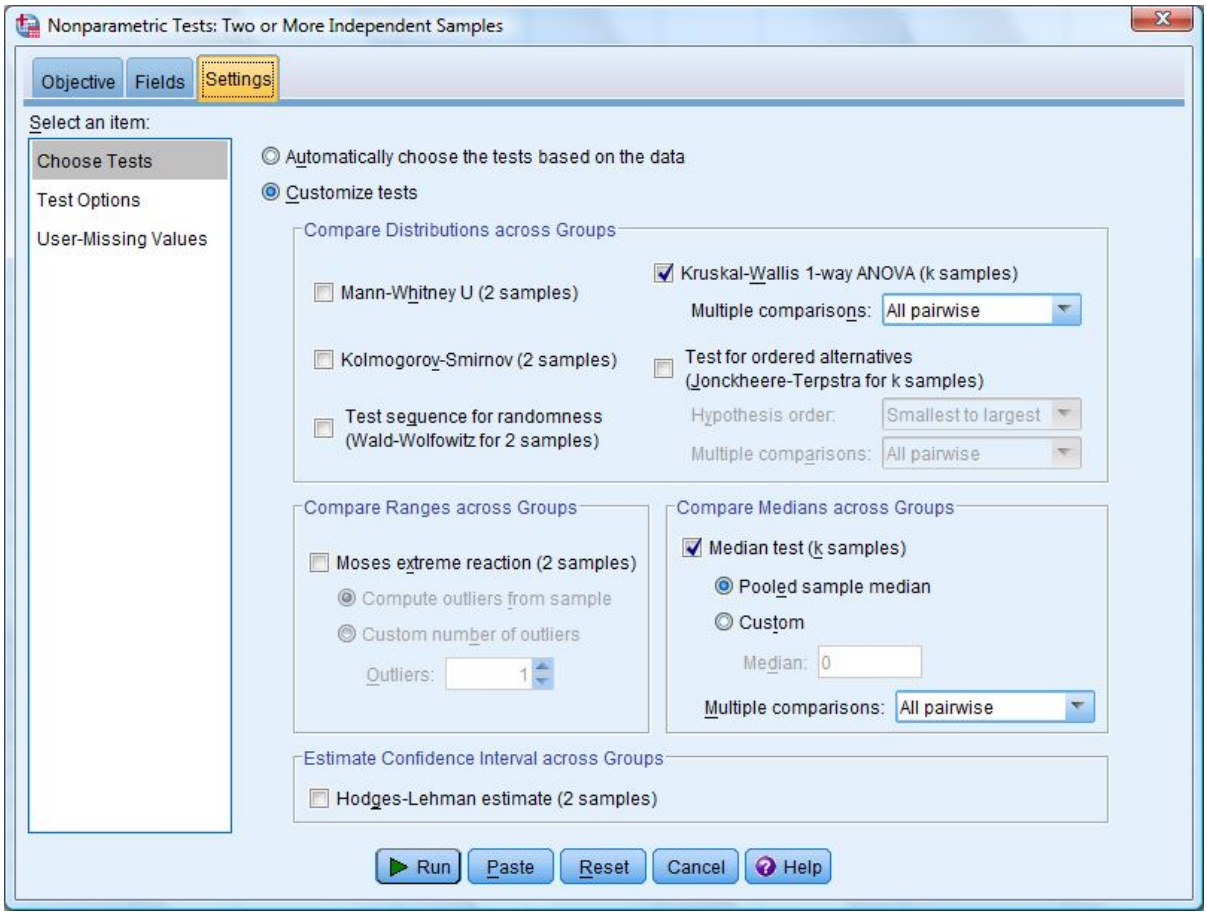
Θέλουμε να ελέγξουμε αν υπάρχει στατιστικά σημαντική διαφορά ως προς την πληθυσμιακή διάμεσο του αριθμού των δοχείων που παράγουν οι τρεις μηχανές. (Σχόλιο: να διαπιστώσετε ότι όντως χρειάζεται μη παραμετρικός έλεγχος)

Από το κύριο μενού επιλέγουμε

i. Analyze→Nonparametric Tests→ Independent Samples.

ii. Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε στο πλαίσιο Objective την επιλογή Customize analysis, έτσι ώστε στη συνέχεια από τα πλαίσια Fields και Settings να καθορίσουμε τον έλεγχο τον οποίο θέλουμε να διενεργηθεί.





Στο πλαίσιο Test Fields τοποθετούμε την υπό μελέτη ποσοτική μεταβλητή, ενώ στο πλαίσιο Groups την ποιοτική μεταβλητή η οποία μας διαχωρίζει τους k πληθυσμούς. Από το πλαίσιο Settings επιλέγουμε Customize Tests και έπειτα Kruskal-Wallis 1-way Anova και Median test (k samples)

Ερμηνεία αποτελεσμάτων

Συμπεραίνουμε με το Independent Samples Median test ότι δεν υπάρχει στατιστικά σημαντική διαφορά στην πληθυσμιακή διάμεσο του αριθμού των δοχείων που παρήχθησαν ως προς τις 3 μηχανές. Από την άλλη μεριά χρησιμοποιώντας το Kruskal Wallis test προκύπτει ότι η κατανομή του αριθμού των δοχείων δεν διαφοροποιείται στατιστικά σημαντικά στις 3 ομάδες. Τα αποτελέσματα για να γενικευτούν στις μέσες τιμές θα πρέπει να είναι συμμετρικοί οι πληθυσμοί από όπου προέρχονται τα δεδομένα μας. (Σχόλιο: στο συγκεκριμένο παράδειγμα οι πληθυσμοί μπορούν να θεωρηθούν συμμετρικοί. Μπορείτε να το διαπιστώσετε; Επιπλέον ποιο το αποτέλεσμα του Median Test). Συνοψίζοντας, θα πρέπει

να είμαστε επιφυλακτικοί για τα συμπεράσματά μας μιας και οι p-τιμές είναι πολύ κοντά στο 5% και τα αποτελέσματα αντικρουόμενα.

Hypothesis Test Summary

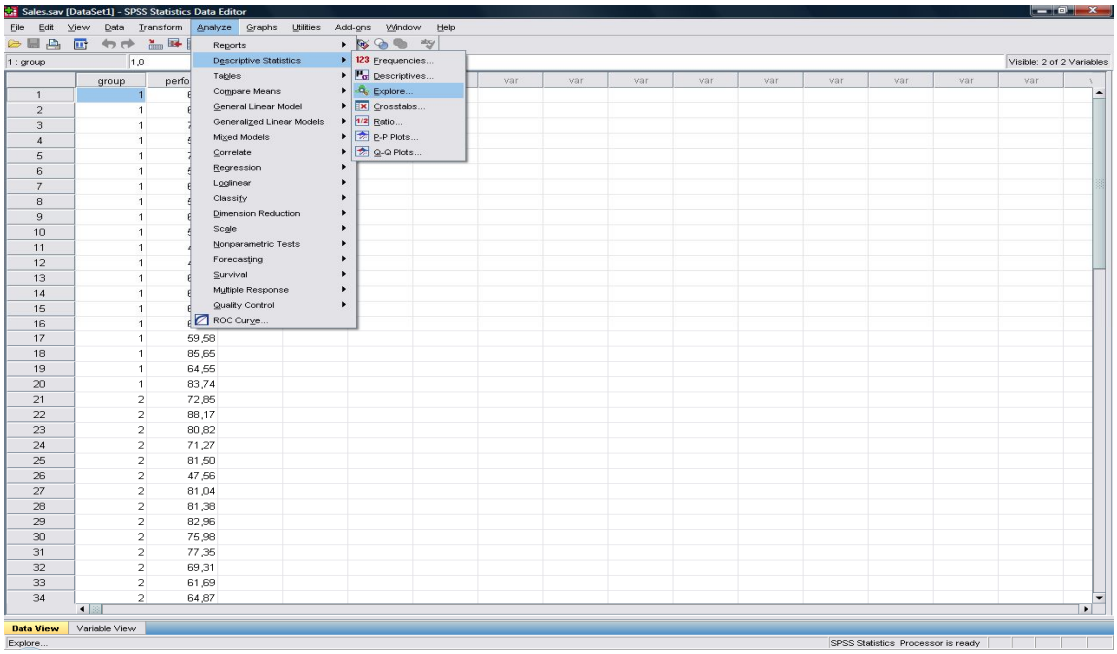
	Null Hypothesis	Test	Sig.	Decision
1	The medians of arithmos are the same across categories of group.	Independent-Samples Median Test	,047	Reject the null hypothesis.
2	The distribution of arithmos is the same across categories of group.	Independent-Samples Kruskal-Wallis Test	,059	Retain the null hypothesis.

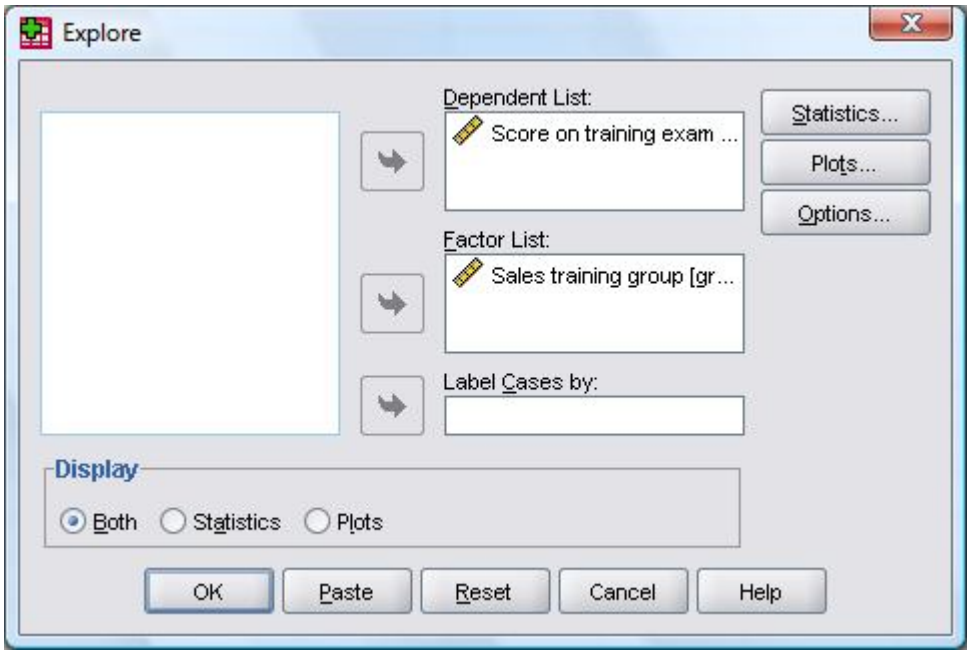
Asymptotic significances are displayed. The significance level is ,05.

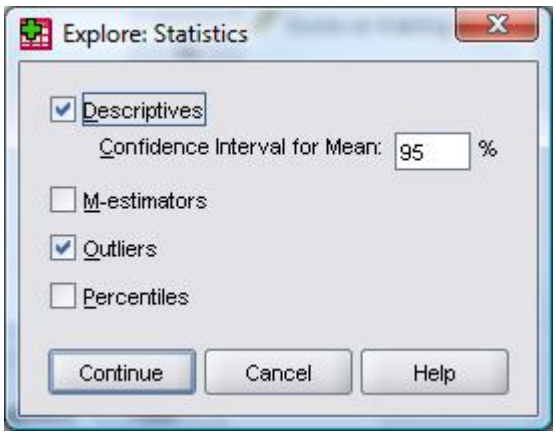
7.2 Παραδείγματα

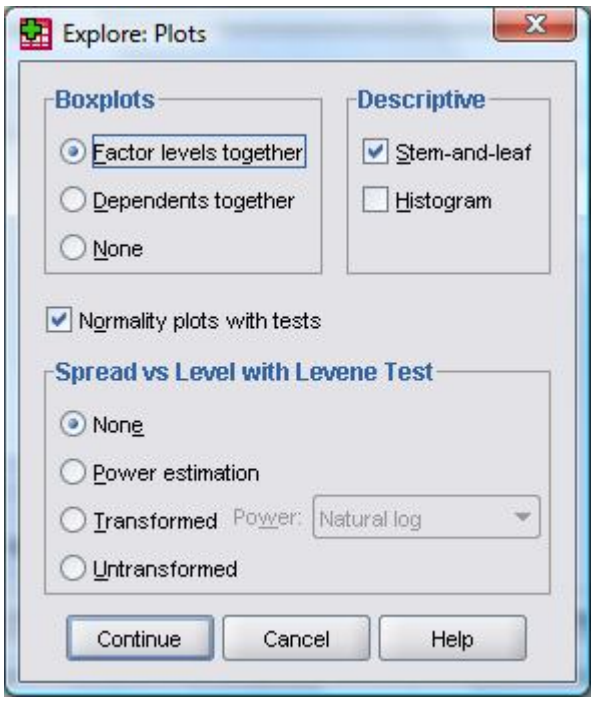
<u>Παράδειγμα 1ο</u> Στο αρχείο Sales.sav * καταγράφονται οι αποδόσεις (στήλη Perform) 60 τυχαία επιλεγμένων πωλητών μιας εταιρείας οι οποίοι έχουν χωριστεί σε τρεις ομάδες (Group). Θέλουμε να ελέγξουμε αν είναι εφικτό υπάρχει στατιστικά σημαντική διαφορά στη μέση απόδοσή των πωλητών ανάλογα με την ομάδα που ανήκουν.

Αρχικά θα ελέγξουμε την ύπαρξη ακραίων τιμών στις δειγματικές τιμές που καταγράφεται η επίδοση των 3 ομάδων πωλητών. Για το σκοπό αυτό ακολουθούμε τα βήματα:



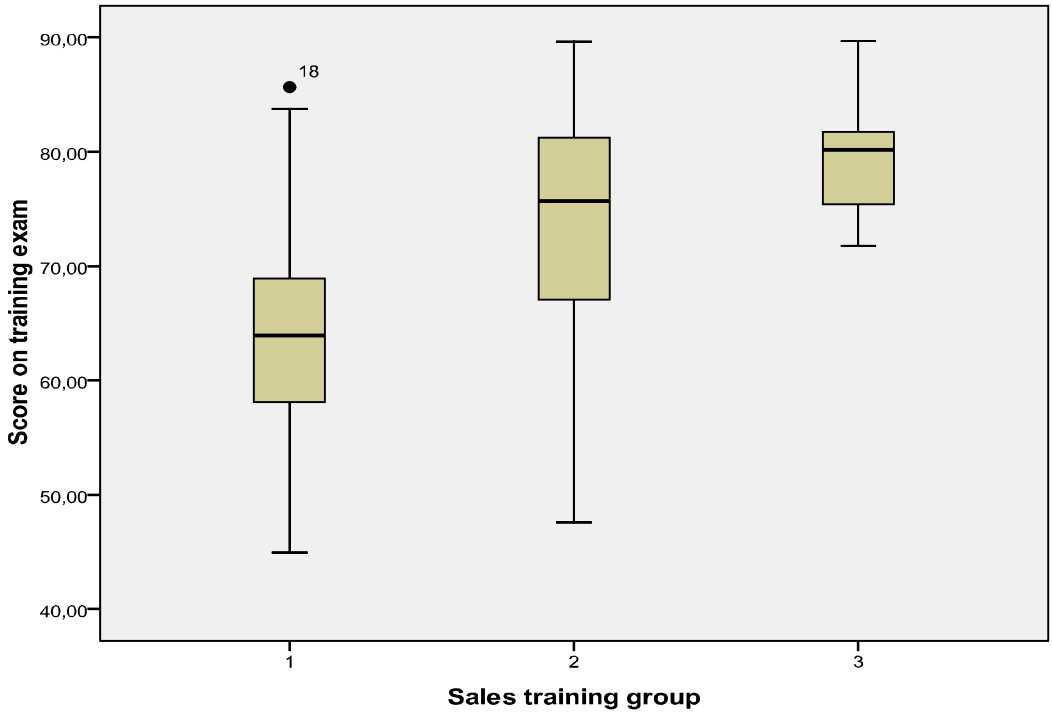

Explore
Dependent List:
Score on training exam ...
Factor List:
Sales training group [gr...
Label Cases by:
Statistics...
Plots...
Options...
Display
Both
Statistics
Plots
OK
Paste
Reset
Cancel
Help


Explore: Statistics
Descriptives
Confidence Interval for Mean:
95
%
M-estimators
Outliers
Percentiles
Continue
Cancel
Help

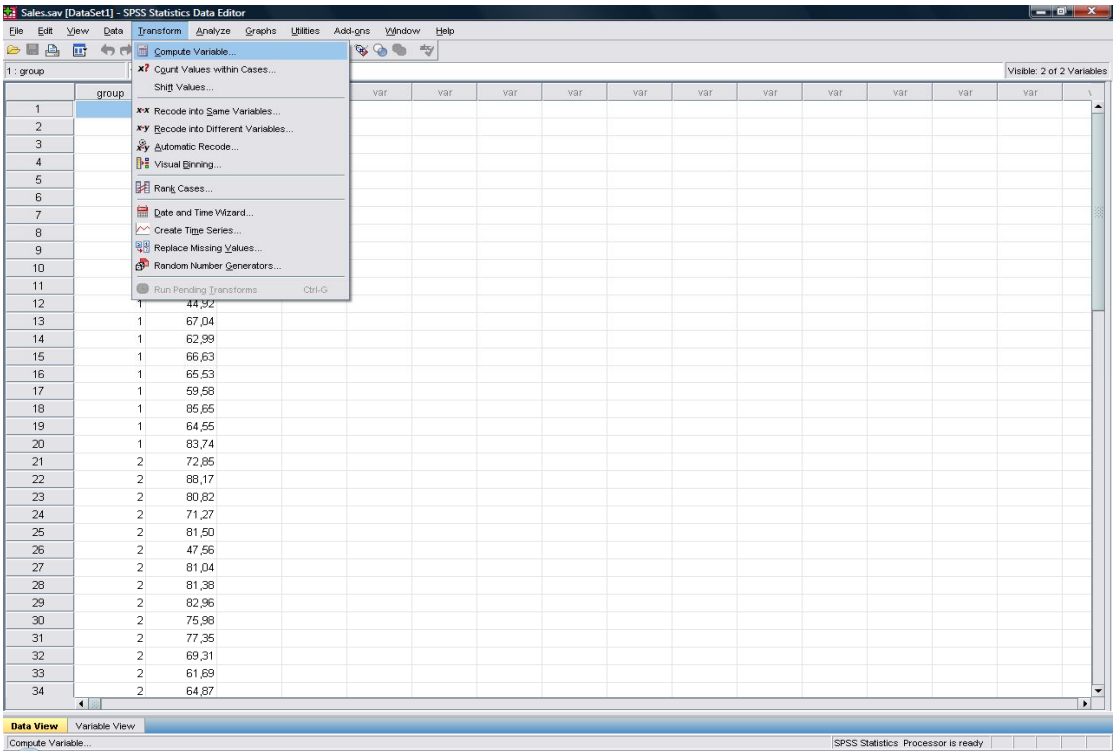

Explore: Plots
Boxplots
Factor levels together
Dependents together
None
Descriptive
Stem-and-leaf
Histogram
Normality plots with tests
Spread vs Level with Levene Test
None
Power estimation
Transformed
Power:
Natural log
Untransformed
Continue
Cancel
Help

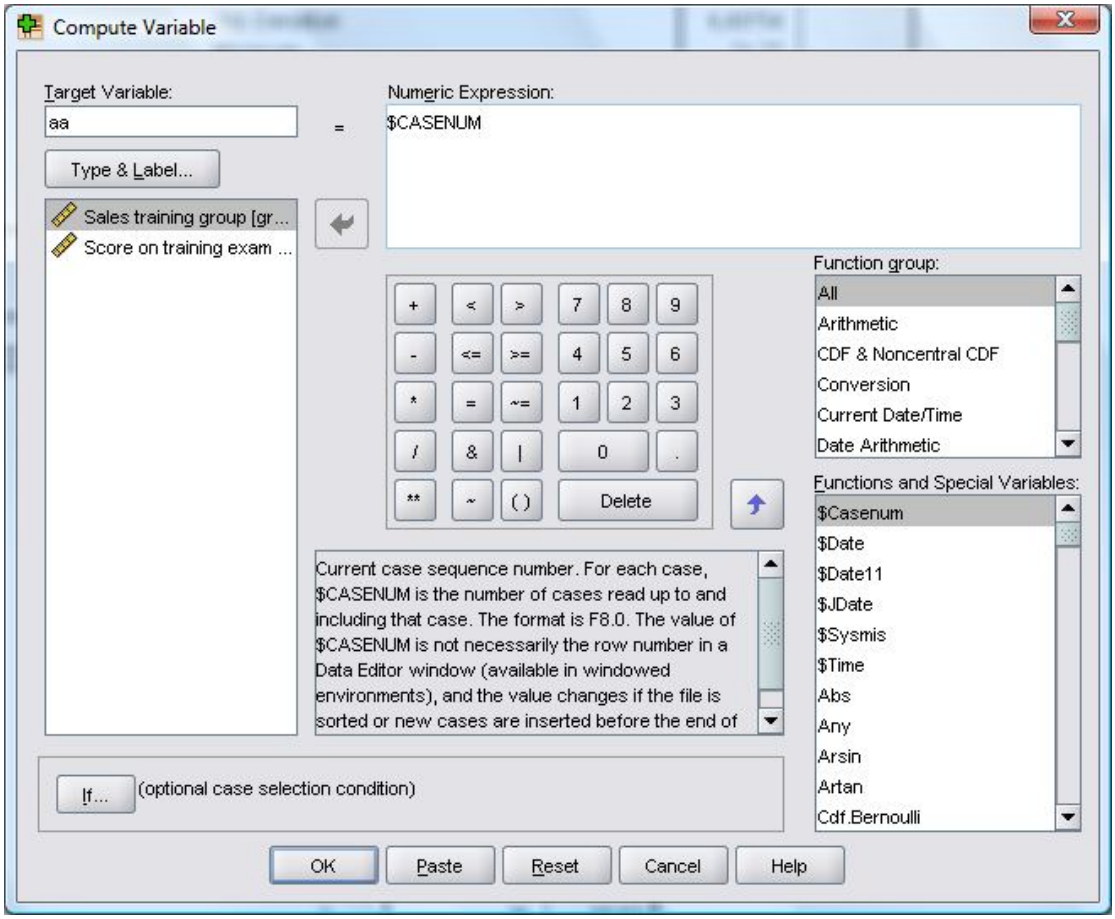
Παρατηρούμε, από το θηκόγραμμα που ακολουθεί, ότι πρέπει να αποκλειστεί από την περαιτέρω ανάλυση η παρατήρηση με α/α 18, λόγω του ότι είναι ακραία. Η παρατήρηση αυτή προέρχεται από την πρώτη ομάδα πωλητών και έχει τιμή απόδοσης 85,65.

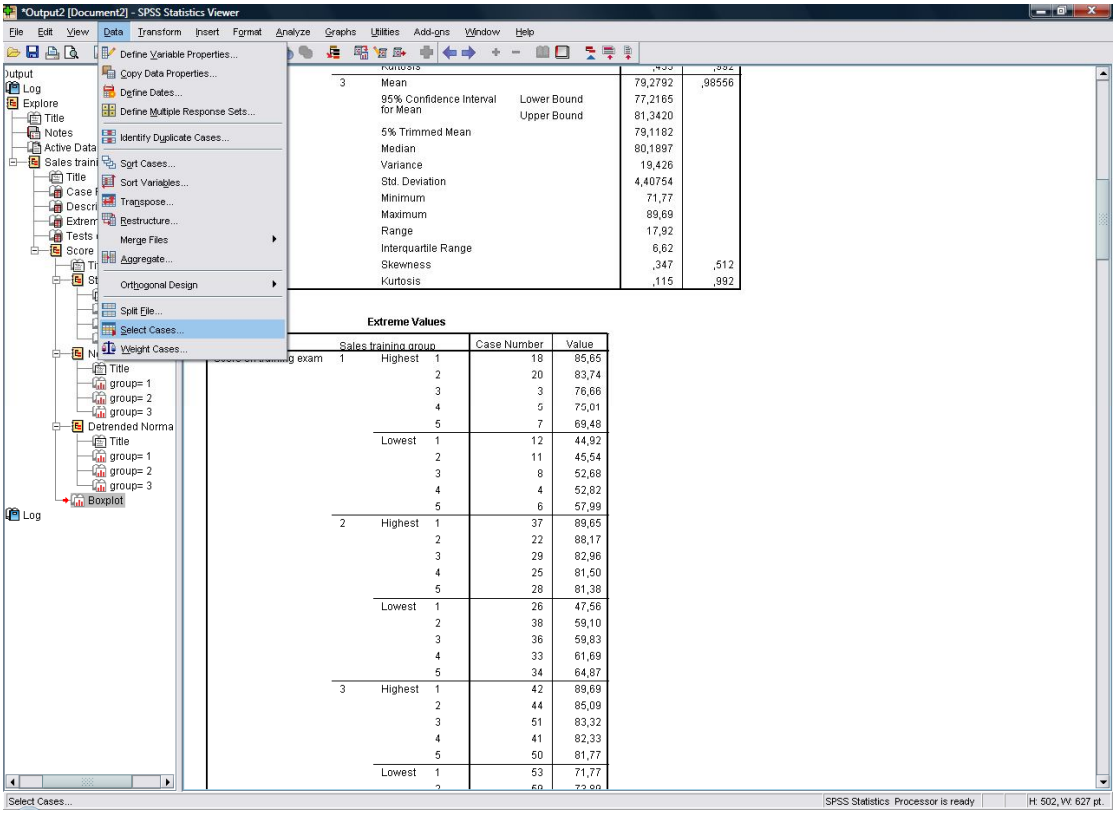
Σχήμα 1 Θηκογράμματα 1, 2 και 3

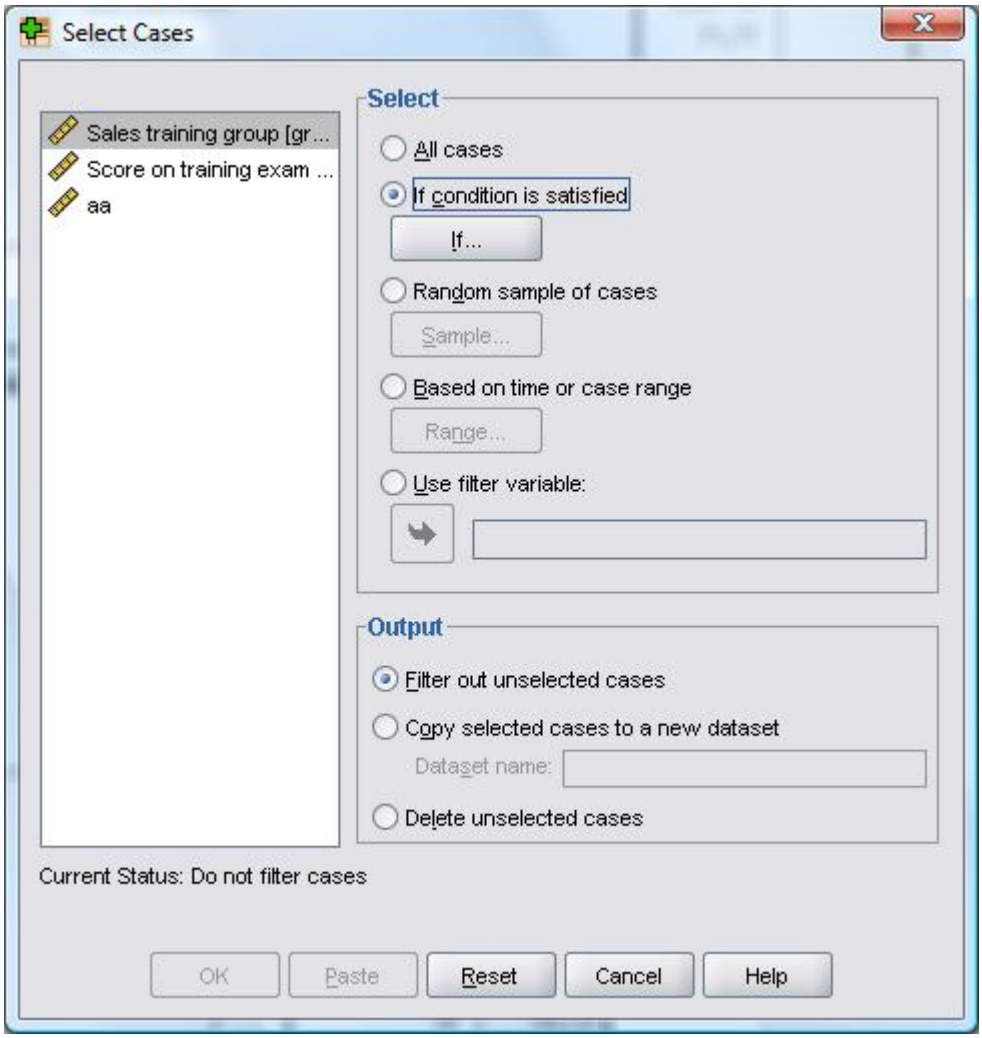


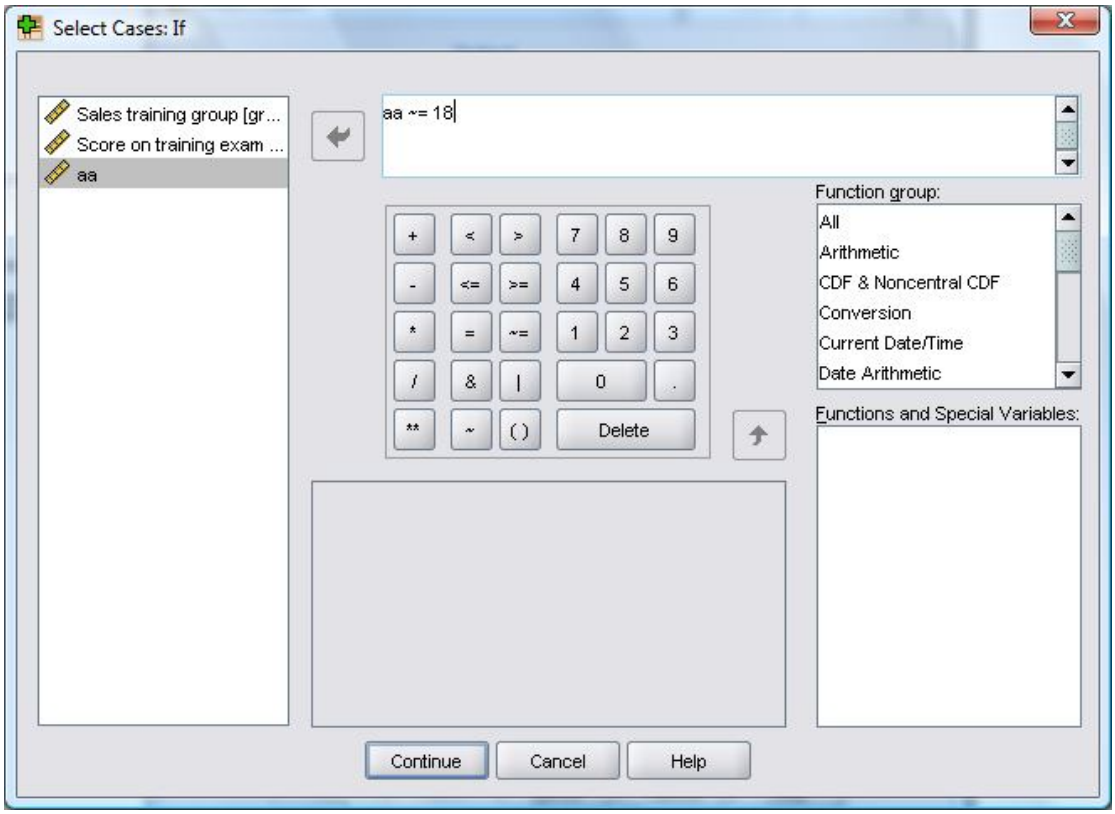
Δημιουργούμε κατά τα γνωστά μία νέα μεταβλητή με τον αύξοντα αριθμό παρατήρησης ως βοηθητική για τον αποκλεισμό της παρατήρησης 18 από την περαιτέρω ανάλυση.


Sales.sav [DataSet1] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Compute Variable...
Count Values within Cases...
Shift Values...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Visual Binning...
Rank Cases...
Date and Time Wizard...
Create Time Series...
Replace Missing Values...
Random Number Generators...
Run Pending Transforms Ctrl-G
Visible: 2 of 2 Variables
group
Data View
Variable View
Compute Variable...
SPSS Statistics Processor is ready

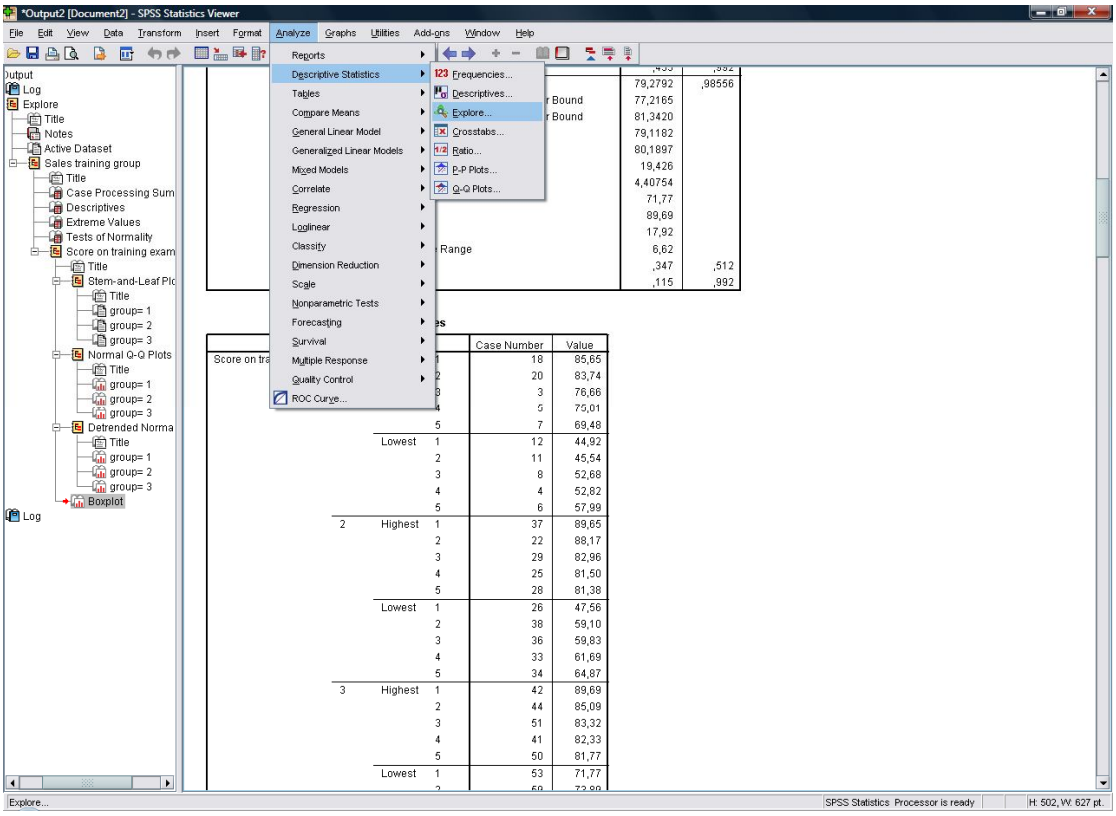

Compute Variable
Target Variable:
aa
=
Numeric Expression:
$CASENUM
Type & Label...
Sales training group [gr...
Score on training exam ...
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) Delete
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
$Casenum
$Date
$Date11
$JDate
$Sysmis
$Time
Abs
Any
Arsin
Artan
Cdf.Bernoulli
Current case sequence number. For each case, $CASENUM is the number of cases read up to and including that case. The format is F8.0. The value of $CASENUM is not necessarily the row number in a Data Editor window (available in windowed environments), and the value changes if the file is sorted or new cases are inserted before the end of
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help

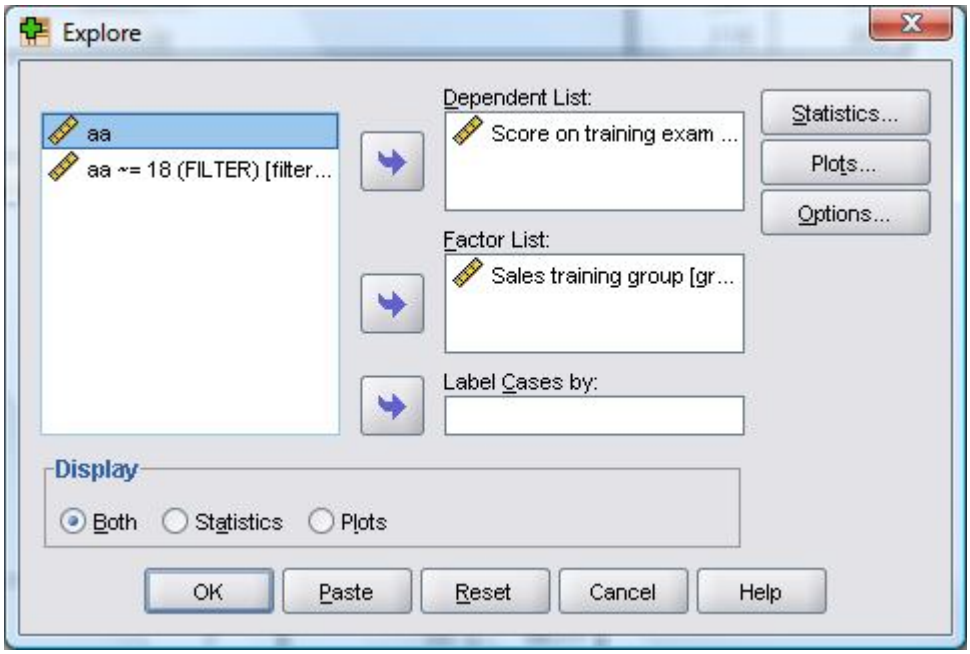

*Output2 [Document2] - SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Define Variable Properties...
Copy Data Properties...
Define Dates...
Define Multiple Response Sets...
Identify Duplicate Cases...
Sort Cases...
Sort Variables...
Transpose...
Restructure...
Merge Files
Aggregate...
Orthogonal Design
Split File...
Select Cases...
Weight Cases...
Mean 79,2792 ,98556
95% Confidence Interval for Mean Lower Bound 77,2165
Upper Bound 81,3420
5% Trimmed Mean 79,1182
Median 80,1897
Variance 19,426
Std. Deviation 4,40754
Minimum 71,77
Maximum 89,69
Range 17,92
Interquartile Range 6,62
Skewness ,347 ,512
Kurtosis ,115 ,992
Extreme Values
Sales training group
Case Number
Value
Highest 1 18 85,65
2 20 83,74
3 3 76,66
4 5 75,01
5 7 69,48
Lowest 1 12 44,92
2 11 45,54
3 8 52,68
4 4 52,82
5 6 57,99
2 Highest 1 37 89,65
2 22 88,17
3 29 82,96
4 25 81,50
5 28 81,38
Lowest 1 26 47,56
2 38 59,10
3 36 59,83
4 33 61,69
5 34 64,87
3 Highest 1 42 89,69
2 44 85,09
3 51 83,32
4 41 82,33
5 50 81,77
Lowest 1 53 71,77
Select Cases...
SPSS Statistics Processor is ready
H: 502, W: 627 pt.


Select Cases
Sales training group [gr...
Score on training exam ...
aa
Select
All cases
If condition is satisfied
If...
Random sample of cases
Sample...
Based on time or case range
Range...
Use filter variable:
Output
Filter out unselected cases
Copy selected cases to a new dataset
Dataset name:
Delete unselected cases
Current Status: Do not filter cases
OK
Paste
Reset
Cancel
Help

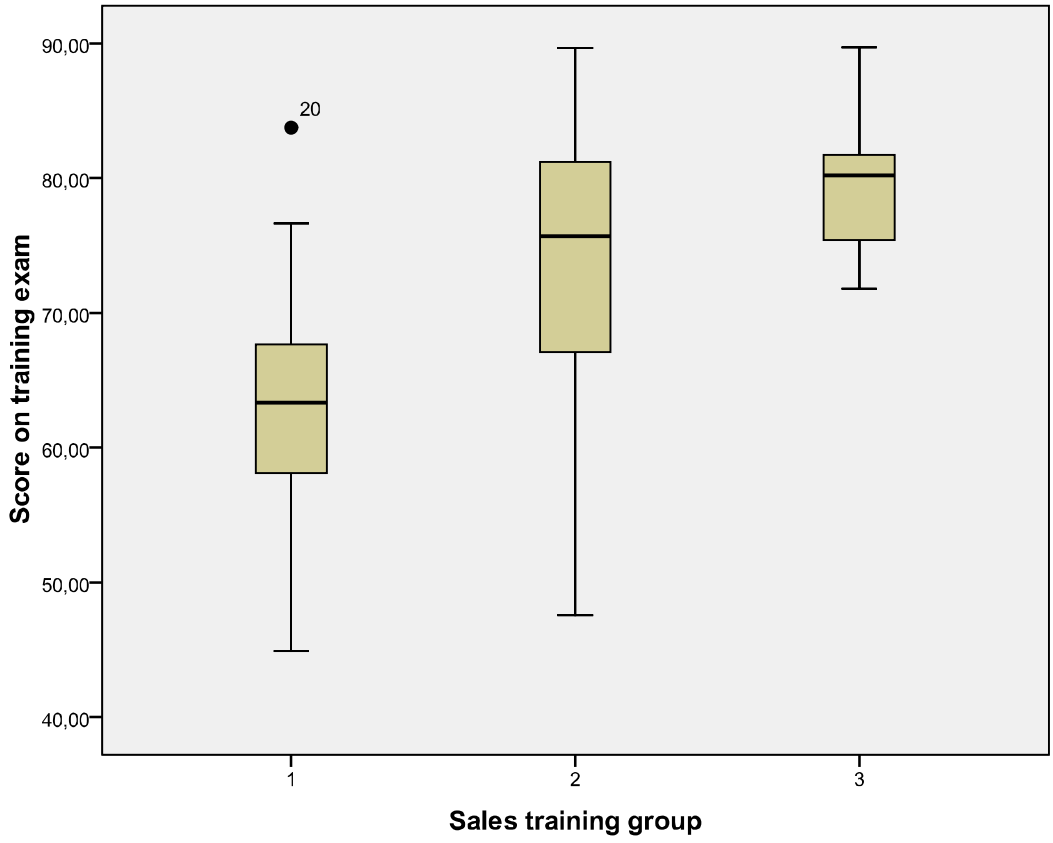


Στη συνέχεια προβαίνουμε πάλι σε έλεγχο ύπαρξης ακραίων τιμών.



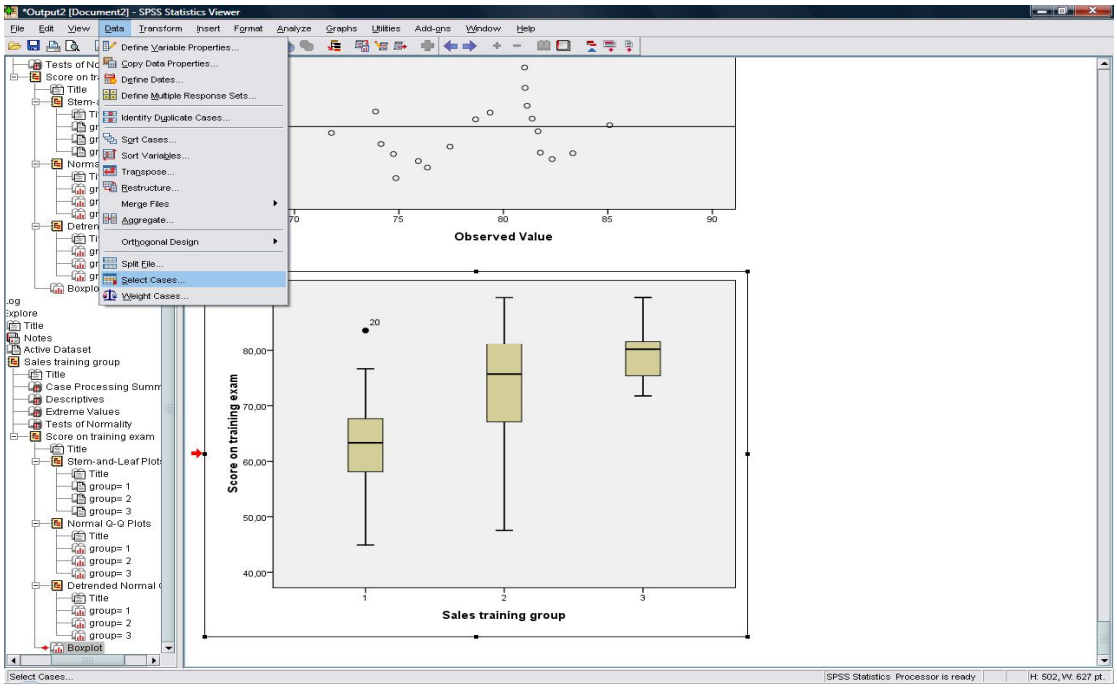


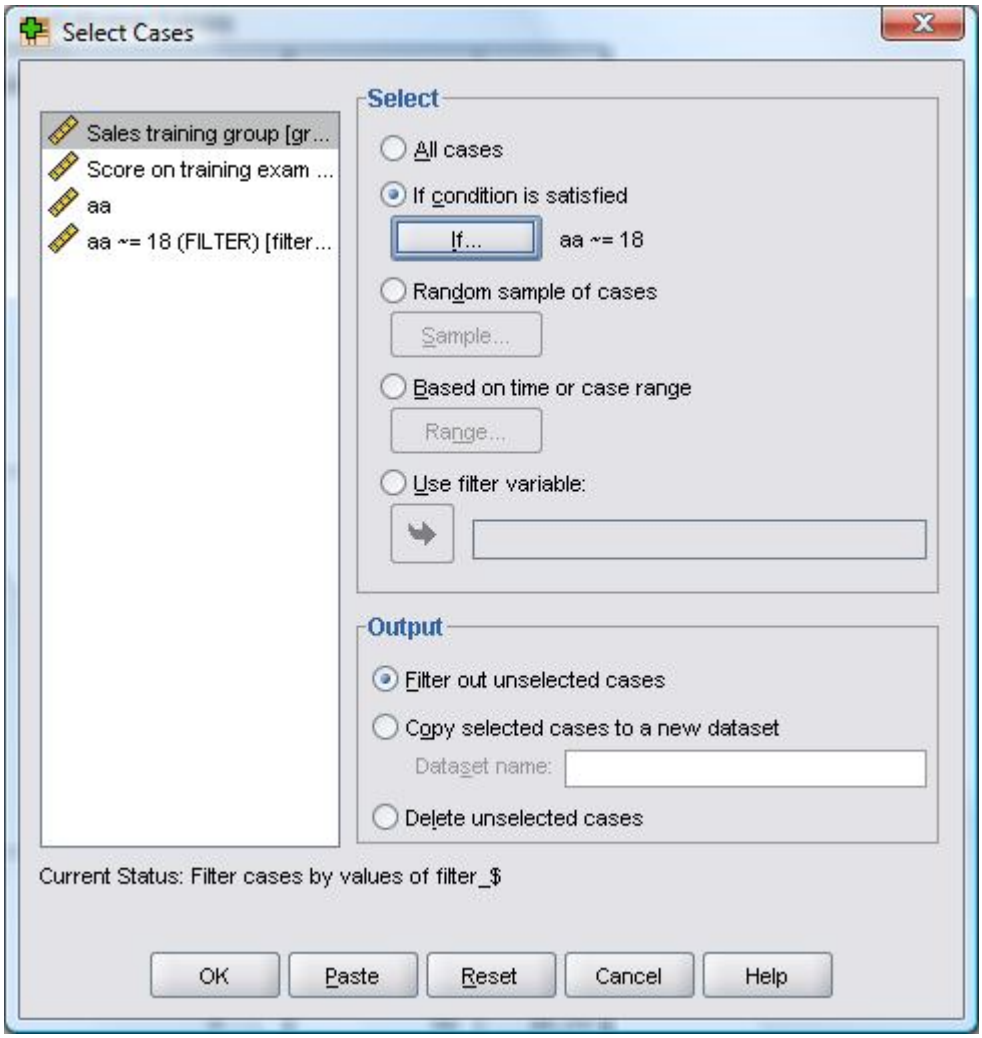
Σχήμα 2 Θηκογράμματα 4, 5 και 6

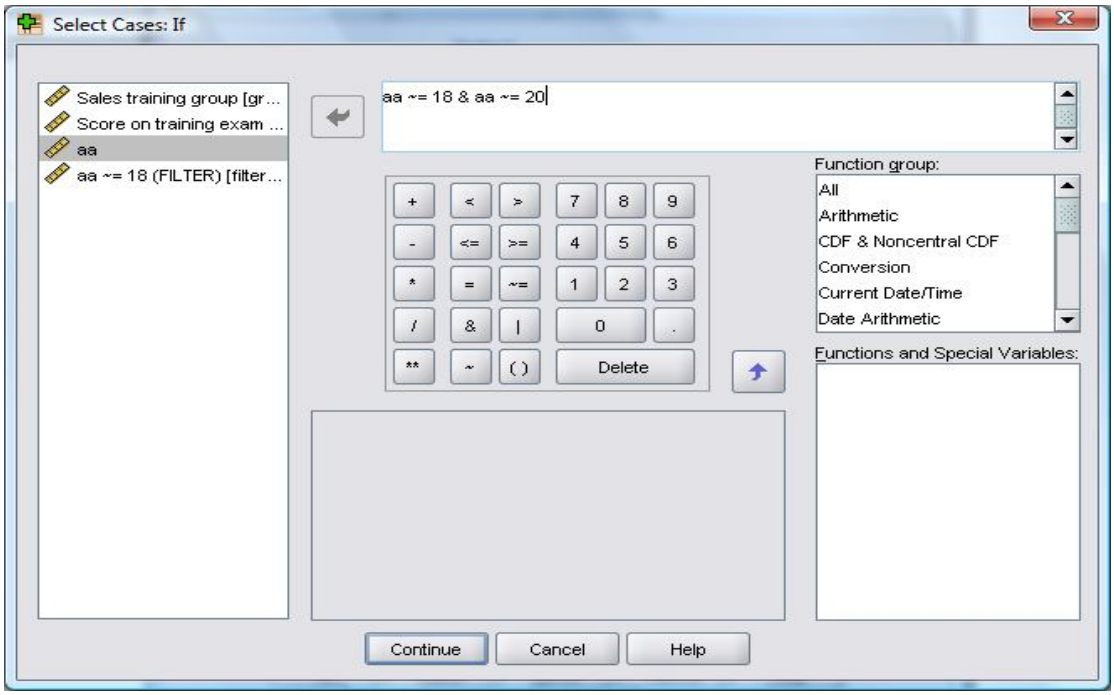


Θα αποκλειστεί επομένως και η παρατήρηση με αύξοντα αριθμό 20 (το ποσοστό των ακραίων είναι τώρα 2/20* 100%=10%) με τιμή 83,74. Για το σκοπό αυτό επιλέγουμε τα

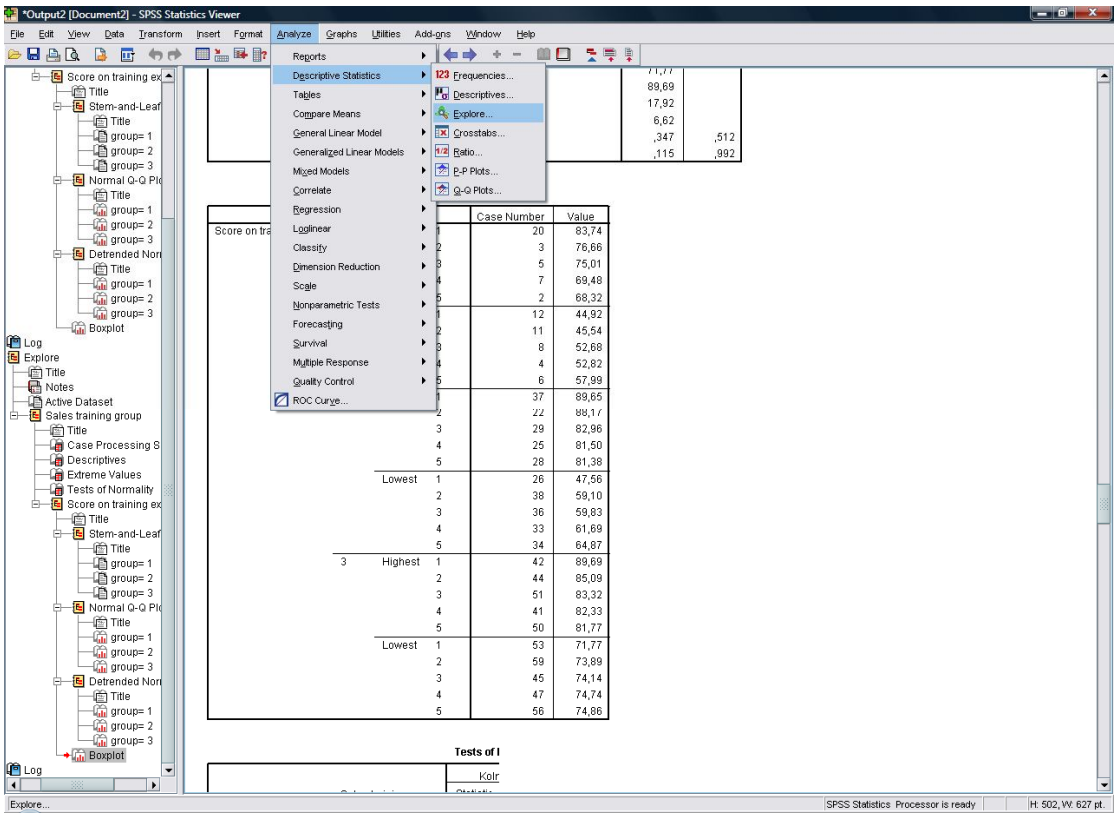
ακόλουθα μέσω της διαδικασίας Data Select Cases. Παρατήρηση: Η 20 δεν εμφανιζόταν πριν ως πιθανή ακραία.

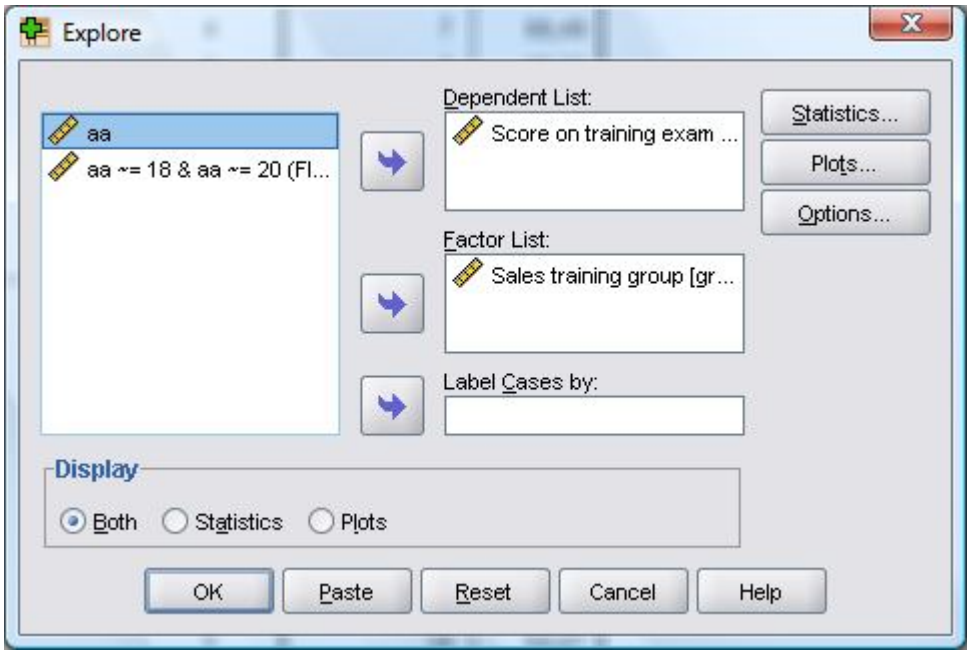




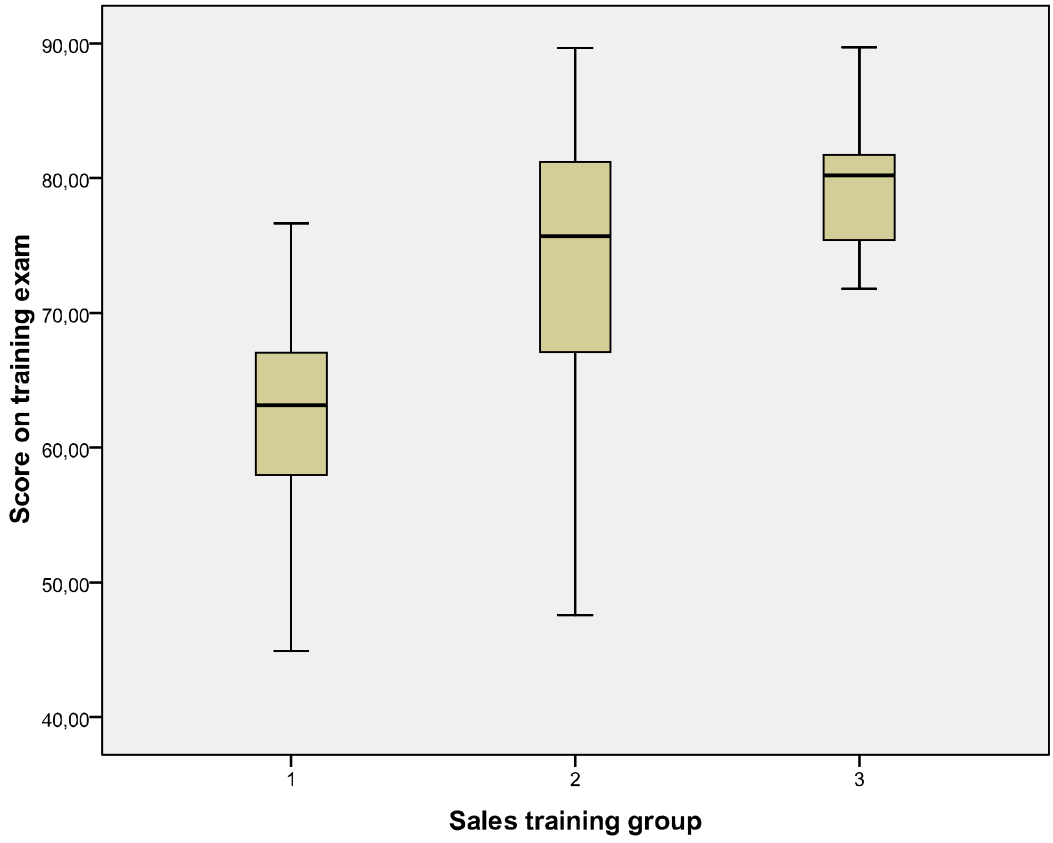


Επαναλαμβάνουμε τον έλεγχο ύπαρξης ακραίων τιμών





Σχήμα 3 Θηκογράμματα 7, 8 και 9



Συνοψίζοντας, από τα παραπάνω θηκογράμματα προκύπτει ότι υπάρχουν δύο ακραίες τιμές στις δειγματικές τιμές της επίδοσης της πρώτης ομάδας, ενώ δεν υπάρχουν ακραίες τιμές

στις δειγματικές τιμές της επίδοσης των υπολοίπων ομάδων. Καθώς το ποσοστό των ακραίων τιμών εντός της πρώτης ομάδας δεν ξεπερνά το 10% (2/10*100%=10%) συνεχίζουμε την περαιτέρω ανάλυση, έχοντας αποκλείσει από την περαιτέρω ανάλυση της προαναφερθείσες δειγματικές τιμές.

Προσοχή: Τις ακραίες τιμές τις αποκλείουμε μία μία για κάθε «ομάδα», ξεκινώντας από την πιο απομακρυσμένη της ομάδας. Το ποσοστό 10% δεν το υπολογίζουμε στο σύνολο των παρατηρήσεων αλλά στον αριθμό των παρατηρήσεων εντός κάθε ομάδας.

Έπειτα ελέγχουμε αν οι δειγματικές τιμές της επίδοσης των τριών ομάδων προέρχονται από κανονικούς πληθυσμούς (τεστ Shapiro Wilk έχει ήδη ζητηθεί η υλοποίηση του στο προηγούμενο βήμα).

Tests of Normality

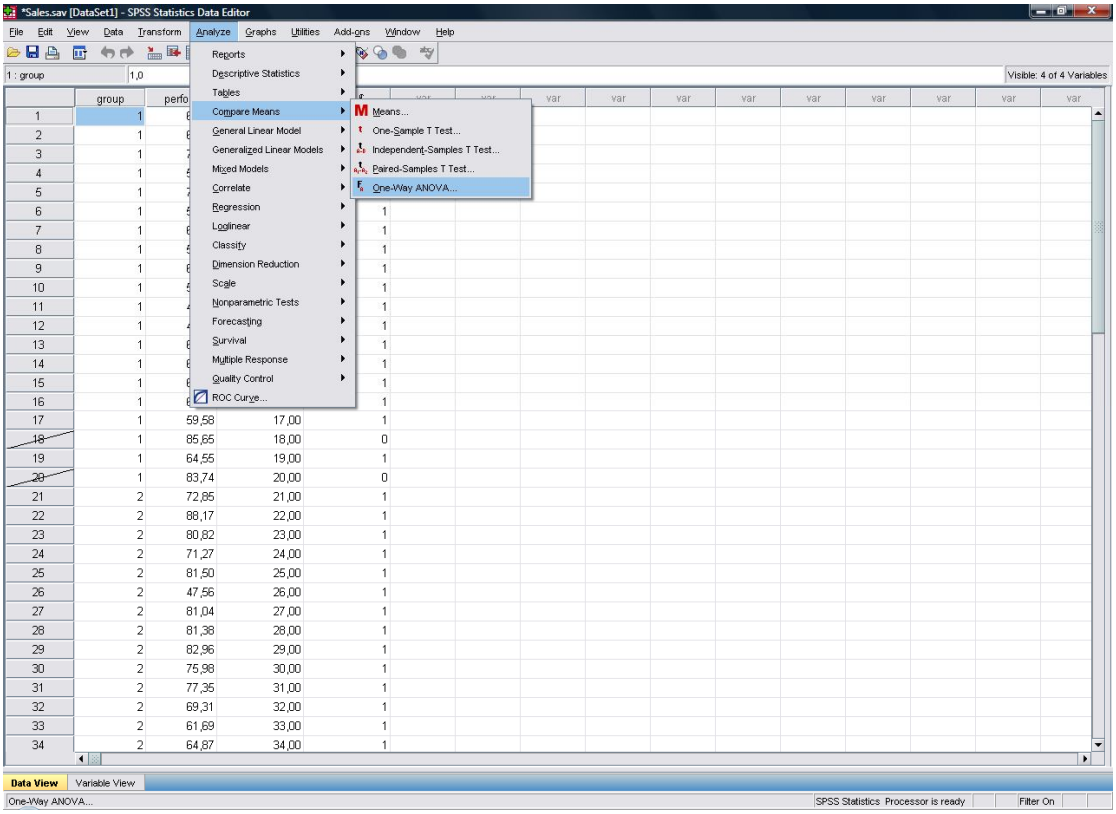
	Sales training group	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Score on training exam	1	,111	18	,200*	,964	18	,676
	2	,134	20	,200*	,948	20	,344
	3	,153	20	,200*	,962	20	,582

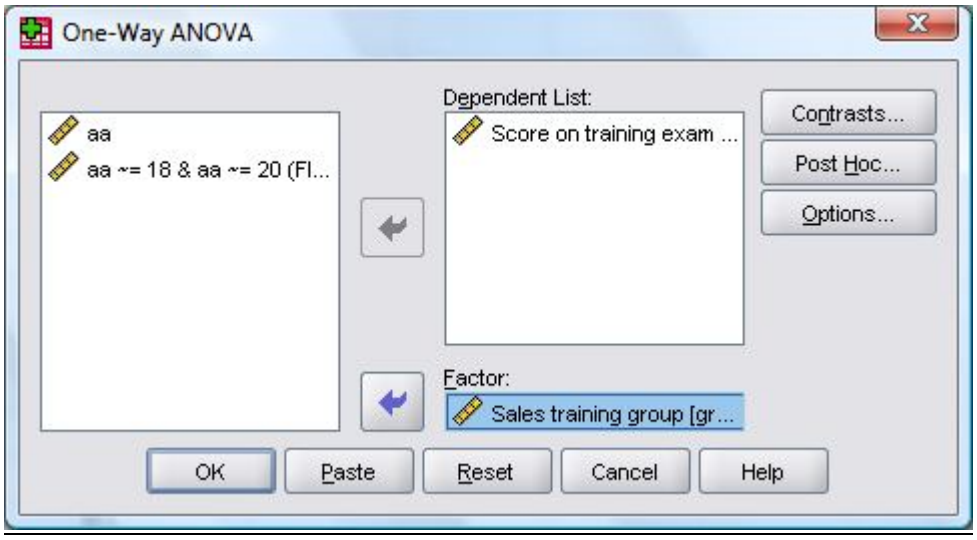
a. Lilliefors Significance Correction

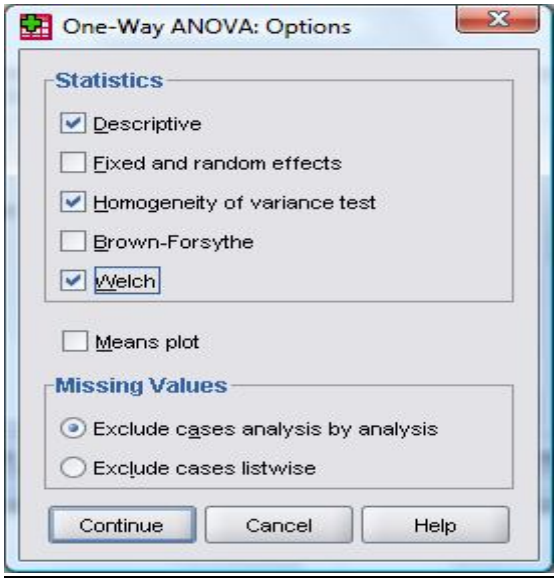
*. This is a lower bound of the true significance.

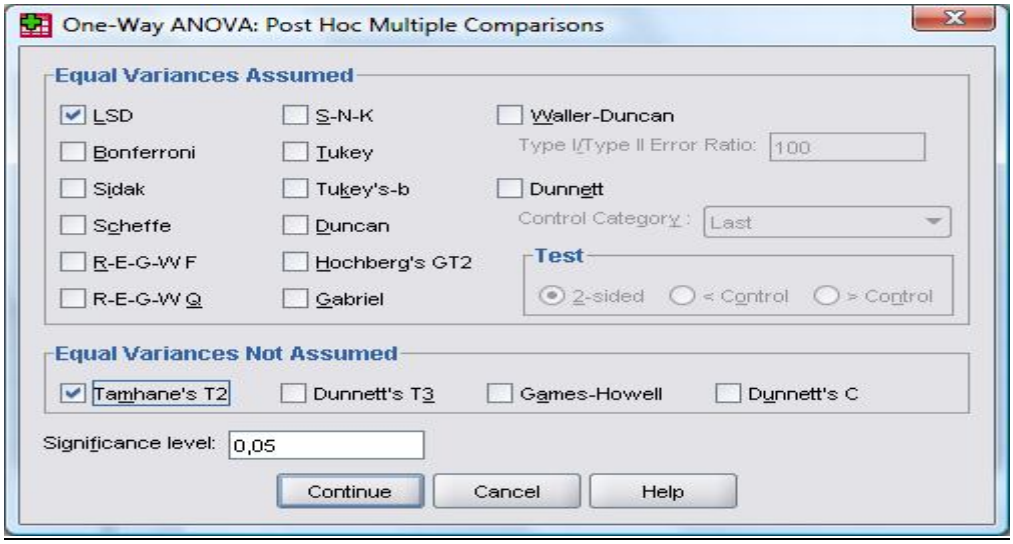
Η υπόθεση της κανονικότητας δεν απορρίπτεται για κάποιον από τους τρεις πληθυσμούς και θα εξετάσουμε την προς έλεγχο υπόθεση παραμετρικά.

Προσοχή: Παρότι δηλώνουμε τις παρακάτω επιλογές δεν χρειάζονται όλες αυτές οι επιλογές, όπως αναλυτικά θα εξηγηθεί.


*Sales.sav [DataSet1] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze 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
Quality Control
ROC Curve...
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
Data View
Variable View
One-Way ANOVA...
SPSS Statistics Processor is ready
Filter On


One-Way ANOVA
aa
aa ~= 18 & aa ~= 20 (Fl...
Dependent List:
Score on training exam ...
Factor:
Sales training group [gr...
Contrasts...
Post Hoc...
Options...
OK
Paste
Reset
Cancel
Help


One-Way ANOVA: Options
Statistics
Descriptive
Fixed and random effects
Homogeneity of variance test
Brown-Forsythe
Welch
Means plot
Missing Values
Exclude cases analysis by analysis
Exclude cases listwise
Continue
Cancel
Help


One-Way ANOVA: Post Hoc Multiple Comparisons
Equal Variances Assumed
LSD
S-N-K
Waller-Duncan
Type I/Type II Error Ratio: 100
Bonferroni
Tukey
Sidak
Tukey's-b
Dunnett
Control Category : Last
Scheffe
Duncan
R-E-G-W F
Hochberg's GT2
Test
2-sided
< Control
> Control
R-E-G-W Q
Gabriel
Equal Variances Not Assumed
Tamhane's T2
Dunnett's T3
Games-Howell
Dunnett's C
Significance level: 0,05
Continue
Cancel
Help

Από τη θεωρία γνωρίζουμε ότι αν χρησιμοποιηθεί το F-test του πίνακα Ανάδια ή το τεστ του Welch καθορίζεται από την ικανοποίηση ή όχι της ισότητας των πληθυσμιακών διακυμάνσεων. Η υπόθεση αυτή ελέγχεται από το στατιστικό τεστ του Levene. Όταν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων απορρίπτεται χρησιμοποιούμε το στατιστικό του Welch. Στο συγκεκριμένο παράδειγμα αν πούμε ότι θέλουμε να εργαστούμε με επίπεδο σημαντικότητας 5% τότε απορρίπτεται η ισότητα των πληθυσμιακών διακυμάνσεων. Και επομένως θα χρησιμοποιηθεί το τεστ του Welch.

Test of Homogeneity of Variances

Score on training exam

Levene Statistic	df1	df2	Sig.
4,511	2	55	,015

Προκύπτει ότι υπάρχουν στατιστικά σημαντικές διαφορές στη μέση επίδοση των 3 πληθυσμών (τεστ Welch p-τιμή <0.001).

Robust Tests of Equality of Means

Score on training exam

	Statistic[a]	df1	df2	Sig.
Welch	28,831	2	30,976	,000

a. Asymptotically F distributed.

Προβαίνουμε σε πολλαπλές συγκρίσεις και καθώς έχει απορριφθεί η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων θα στηριχθούμε στα αποτελέσματα της μεθόδου Tamhanne.

Multiple Comparisons

Dependent Variable:Score on training exam

	(I) Sales trainig group	(J) Sales training group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	-11,77844*	2,70979	,000	-17,2090	-6,3479
		3	-17,49002*	2,70979	,000	-22,9206	-12,0595
	2	1	11,77844*	2,70979	,000	6,3479	17,2090
		3	-5,71158*	2,63752	,035	-10,9973	-,4259
	3	1	17,49002*	2,70979	,000	12,0595	22,9206
		2	5,71158*	2,63752	,035	,4259	10,9973
Tamhane	1	2	-11,77844*	3,15221	,002	-19,6737	-3,8832
		3	-17,49002*	2,29786	,000	-23,3790	-11,6010
	2	1	11,77844*	3,15221	,002	3,8832	19,6737
		3	-5,71158	2,56883	,102	-12,2771	,8539
	3	1	17,49002*	2,29786	,000	11,6010	23,3790
		2	5,71158	2,56883	,102	-,8539	12,2771

*. The mean difference is significant at the 0.05 level.

Από τον πίνακα Multiple Comparisons προκύπτει ότι η μέση επίδοση της ομάδας 1 διαφέρει στατιστικά σημαντικά από αυτές των 2 και 3 όντας χειρότερη.

Αναφορά: Θέλουμε να εξετάσουμε αν υπάρχει στατιστικά σημαντική διαφορά στη μέση επίδοση ως προς τις 3 διαφορετικές ομάδες πωλητών.

Για να μπορούμε να αποφανθούμε χρησιμοποιώντας μεθόδους της παραμετρικής στατιστικής θα πρέπει να πληρούνται οι ακόλουθες υποθέσεις:

1. Τα δείγματά μας να είναι τυχαία επιλεγμένα
2. Να μην υπάρχουν ακραίες τιμές στα δειγματικά δεδομένα κάθε πληθυσμού που να ξεπερνούν σε ποσοστό το 10%.

3. Κάθε πληθυσμός να περιγράφεται ικανοποιητικά από την κανονική κατανομή.

Η πρώτη από τις προϋποθέσεις σχετίζεται με τον τρόπο που επιλέξαμε τα δείγματά μας και ικανοποιείται.

Έλεγχος ακραίων τιμών

Ο έλεγχος των ακραίων τιμών στο δείγμα των τιμών που περιγράφουν την επίδοση της πρώτης ομάδας πωλητών έδειξε ότι οι παρατηρήσεις 18 και 20 με τιμές στην επίδοση 85.65 και 83.74 αντίστοιχα είναι ακραίες (20 διαθέσιμες, άρα ποσοστό ίσο του 10%) και αποκλείονται από την περαιτέρω ανάλυση. Ο έλεγχος των ακραίων τιμών στα δείγματα των τιμών που καταγράφεται η επίδοση της πρώτης καθώς και της δεύτερης ομάδας πωλητών έδειξε ότι δεν υπάρχουν ακραίες τιμές.

Έλεγχος κανονικότητας

Ο έλεγχος της υπόθεσης ότι τα δεδομένα που καταγράφεται η επίδοση της πρώτης ομάδας πωλητών ακολουθούν κανονική κατανομή δεν απορρίπτεται (τεστ Shapiro-Wilk p-τιμή=0.676).

Ο έλεγχος της υπόθεσης ότι τα δεδομένα που καταγράφεται η επίδοση της δεύτερης ομάδας πωλητών ακολουθούν κανονική κατανομή δεν απορρίπτεται (τεστ Shapiro-Wilk p-τιμή=0.344).

Ο έλεγχος της υπόθεσης ότι τα δεδομένα που καταγράφεται η επίδοση της τρίτης ομάδας πωλητών ακολουθούν κανονική κατανομή δεν απορρίπτεται (τεστ Shapiro-Wilk p-τιμή=0.582).

Έλεγχος ίσων διακυμάνσεων

Η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων της επίδοσης των 3 διαθέσιμων ομάδων πωλητών απορρίπτεται σε επίπεδο σημαντικότητας 5% (test Levene p-τιμή=0.015<0.05).

Έλεγχος ισότητας πληθυσμιακών μέσων τιμών-Πολλαπλές συγκρίσεις

Επομένως για να ελέγξουμε αν υπάρχει στατιστικά σημαντική διαφορά στη μέση πληθυσμιακή επίδοση των 3 ομάδων πωλητών θα χρησιμοποιήσουμε το τεστ του Welch με επίπεδο σημαντικότητας 5%.

Υπάρχει στατιστικά σημαντική διαφορά στη μέση πληθυσμιακή επίδοση των 3 ομάδων πωλητών (τεστ Welch p-τιμή<0,001). Θέλοντας να εντοπίσουμε που υπάρχουν οι στατιστικά σημαντικές διαφορές ως προς τη μέση πληθυσμιακή επίδοση των 3 ομάδων

πωλητών θα χρησιμοποιήσουμε τη μέθοδο των πολλαπλών συγκρίσεων του Tamhane με επίπεδο σημαντικότητας 5%.

Η μέση επίδοση της πρώτης ομάδας πωλητών διαφέρει στατιστικά σημαντικά από τις αντίστοιχες της δεύτερης και τρίτης ομάδας (βλέπε πίνακα Multiple Comparisons., Tamhane, $p=0.002$ και $p<0.001$ αντίστοιχα) και μάλιστα είναι χειρότερη όπως προκύπτει από τον πίνακα που ακολουθεί ή από τον πίνακα των πολλαπλών συγκρίσεων.

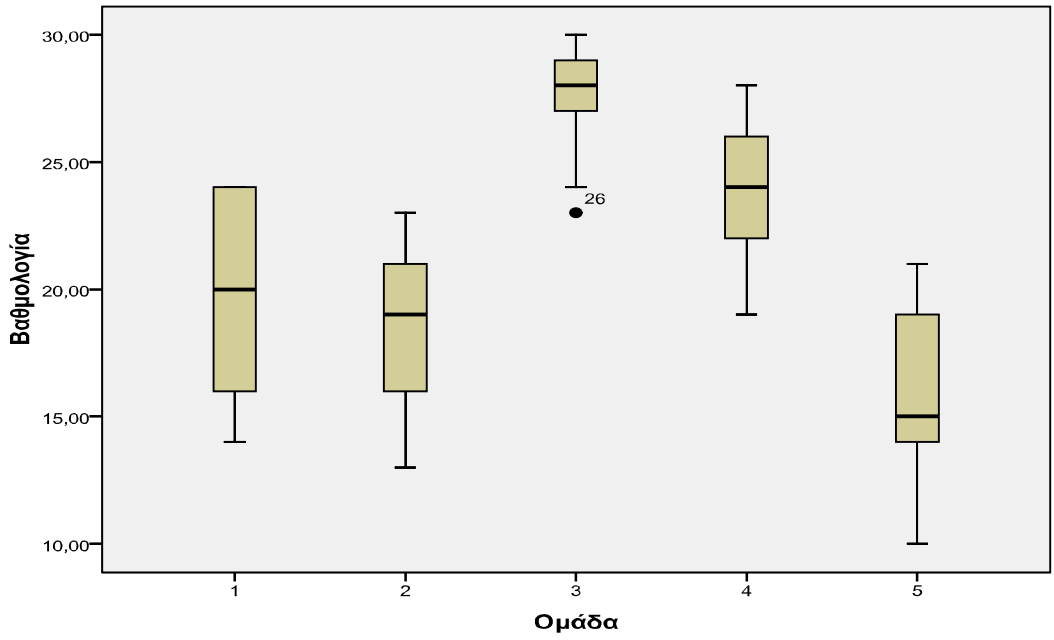
	N	Mean
1	18	61,7892
2	20	73,5677
3	20	79,2792
Total	58	71,8818

<u>Παράδειγμα 2ο</u> Αρχείο School.sav*

Έχουμε 5 ομάδες μαθητών: οι δύο πρώτες διδάχθηκαν με την ίδια μέθοδο διδασκαλίας ενώ οι τρεις επόμενες με διαφορετική μέθοδο. **Θέλουμε να ελέγξουμε αν είναι εφικτό αν υπάρχει στατιστικά σημαντική διαφορά στη μέση βαθμολογία στις 5 ομάδες.**

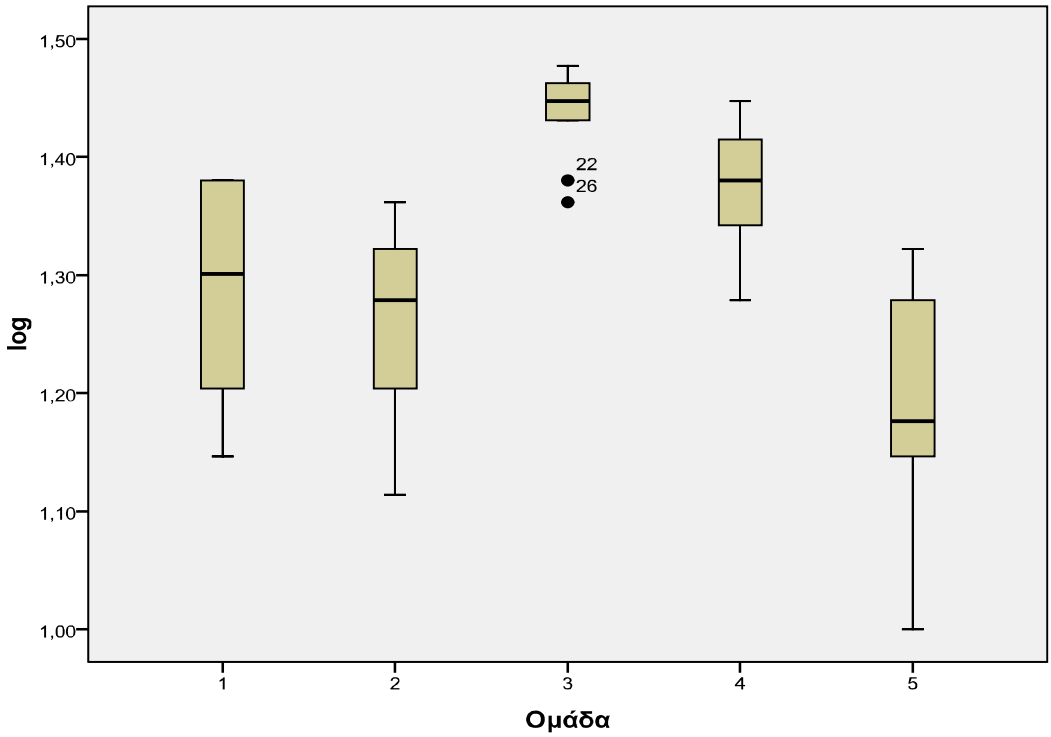
Αποτελέσματα:

Σχήμα 1 Θηκογράμματα 1-5



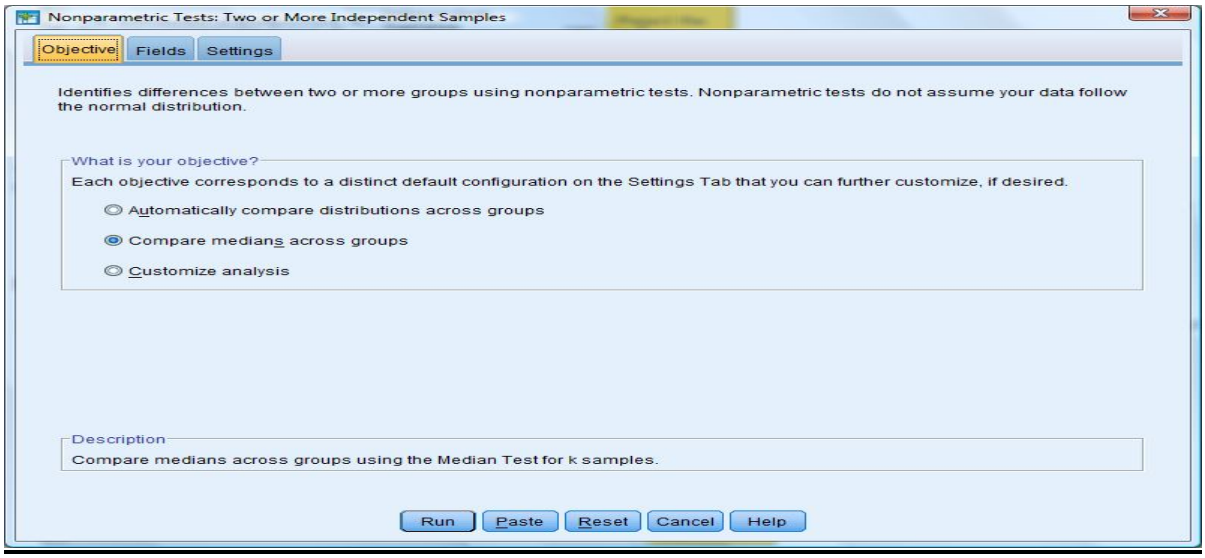
Τουλάχιστον 1 ακραία στις 9 διαθέσιμες δειγματικές της 3ης ομάδας. Επομένως υπάρχει πρόβλημα ακραίων τιμών. Ο μετασχηματισμός του λογαρίθμου δε διορθώνει το πρόβλημα.

Σχήμα 2 Θηκογράμματα 6-10



Άρα πρέπει να πάμε μη παραμετρικά και να ελέγξουμε αν υπάρχει διαφορά στη διάμεσο της επίδοσης των 5 πληθυσμών.

Αρχικά δηλώνεται η μεταβλητή Ομάδα (group) ότι είναι ποιοτική (μέσω του πλαισίου Variable View) και έπειτα χρησιμοποιείται ένας εναλλακτικός τρόπος όπως φαίνεται παρακάτω μέσω της διαδικασίας Analyze Nonparametrics Independent Samples



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Βαθμολογία is the same across categories of Ομάδα.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
2	The medians of Βαθμολογία are the same across categories of Ομάδα.	Independent-Samples Median Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Ποια η καλύτερη ομάδα? Τα αποτελέσματα γενικεύονται?

Η αναφορά των αποτελεσμάτων αφήνεται ως άσκηση.

Άσκηση

Αρχείο Cartoon*

Θέλουμε να ελέγξουμε αν η μέση βαθμολογία στα cartoon διαφοροποιείται ανάλογα με το επίπεδο μόρφωσης.

ΚΕΦΑΛΑΙΟ ΟΓΔΟΟ

Γραμμική παλινδρόμηση

Σε προηγούμενο κεφάλαιο είδαμε ότι η γραφική παράσταση δύο μεταβλητών είναι ένα πρώτο βήμα για τη διαπίστωση της ύπαρξης μίας σχέσης μεταξύ δύο μεταβλητών. Στην παλινδρόμηση το ενδιαφέρον επικεντρώνεται στην εύρεση του καλύτερου γραμμικού μοντέλου που μας δείχνει τον τρόπο με τον οποίο p το πλήθος ανεξάρτητες μεταβλητές επιδρούν σε μία ποσοτική μεταβλητή. Αναζητούμε, επομένως, το μαθηματικό μοντέλο που περιγράφει με τον καλύτερο δυνατό τρόπο τις τιμές της εξαρτημένης μεταβλητής συναρτήσει των τιμών των ανεξάρτητων μεταβλητών. Η εύρεση ενός τέτοιου μοντέλου μας δίνει τη δυνατότητα τόσο να μοντελοποιήσουμε ένα φυσικό-τυχαίο φαινόμενο όσο και να κάνουμε προβλέψεις για τις τιμές της εξαρτημένης μεταβλητής, όταν οι ανεξάρτητες θεωρούνται δεδομένες.

Όταν έχουμε μόνο μία ανεξάρτητη μεταβλητή λέμε ότι έχουμε το μοντέλο της απλής γραμμικής παλινδρόμησης. Το μοντέλο αυτό χρησιμοποιείται για την πρόβλεψη των τιμών μίας εξαρτημένης μεταβλητής από τις τιμές μίας ανεξάρτητης μεταβλητής, όταν αυτές είναι συσχετισμένες. Η ανεξάρτητη μεταβλητή μπορεί να είναι είτε κατηγορική είτε συνεχής, ενώ η εξαρτημένη είναι συνεχής. Γενίκευση του μοντέλου της απλής γραμμικής παλινδρόμησης για p το πλήθος ανεξάρτητες μεταβλητές αποτελεί η πολλαπλή παλινδρόμηση.

Σχόλιο: Μία ανεξάρτητη κατηγορική μεταβλητή με k κατηγορίες-τιμές υπεισέρχεται στο μοντέλο της γραμμικής παλινδρόμησης με τη χρήση k-1 δείκτριων μεταβλητών, ενώ όταν η εξαρτημένη μεταβλητή είναι κατηγορική τότε χρησιμοποιούνται μεθοδολογίες της Λογιστικής Παλινδρόμησης. Οι μεθοδολογίες αυτές ξεφεύγουν από το σκοπό αυτών των σημειώσεων.

8.1 Προσαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης

Στην ενότητα αυτή θα περιγράψουμε τη μεθοδολογία που ακολουθείται για την προσαρμογή ενός μοντέλου απλής γραμμικής παλινδρόμησης.

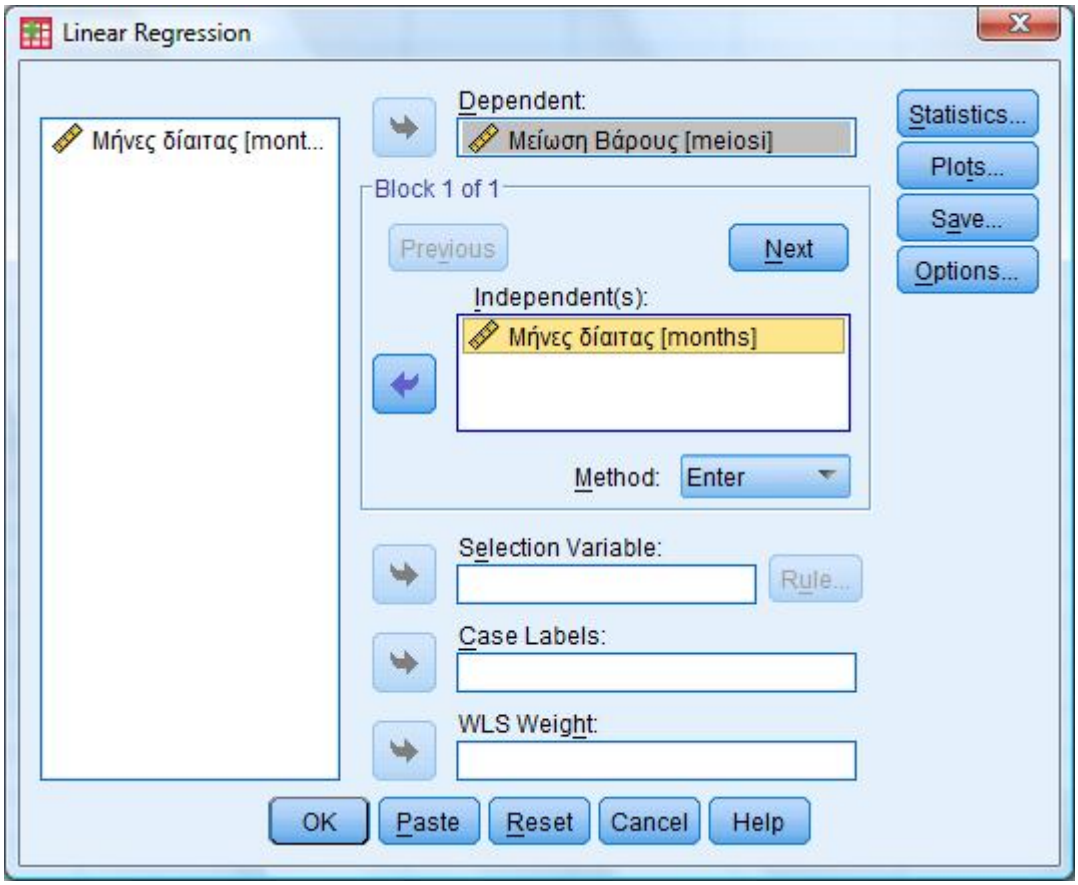
Υλοποίηση στο S.P.S.S. (βλέπε Καρακώστας, 2002, σελ. 22)

Οι παρακάτω τιμές είναι το βάρος (σε λίβρες) που έχασαν 10 άτομα αφού ακολούθησαν κάποια δίαιτα για ορισμένους μήνες. Είναι δυνατή η πρόβλεψη της απώλειας βάρους από τους μήνες δίαιτας.

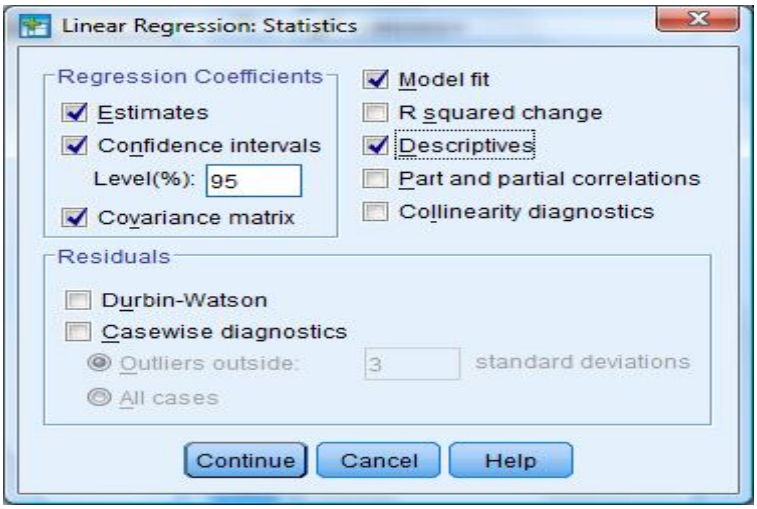
Μήνες Δίαιτας	Μείωση Βάρους
4	17
17	64
14	53
1	1
10	45
22	71
9	38
12	40
4	11
7	24

1. Το πρώτο βήμα για την ανάλυση του παραπάνω προβλήματος είναι να ορίσουμε ποια είναι η εξαρτημένη και ποια η ανεξάρτητη μεταβλητή. Προφανώς το ρόλο της ανεξάρτητης παίζει η μεταβλητή Μήνες δίαιτας, ενώ το ρόλο της εξαρτημένης η Μείωση Βάρους.
2. Θέλοντας να διαπιστώσουμε αν η προσαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης αιτιολογείται προβαίνουμε στη γραφική παράσταση των δεδομένων της εξαρτημένης ως προς την ανεξάρτητη (βλέπε παράγραφο για Διάγραμμα Διασποράς). Αν η γραφική αυτή παράσταση μας υποδεικνύει ότι η σχέση των δύο μεταβλητών δεν είναι γραμμική, τότε η υιοθέτηση του μοντέλου της απλής γραμμικής παλινδρόμησης είναι λανθασμένη. Τρόποι αντιμετώπισης αυτού του προβλήματος αναφέρονται στην επόμενη παράγραφο και στην ενότητα «Ορθότητα μοντέλου».
3. Προχωρούμε έπειτα στην προσαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης επιλέγοντας από το αρχικό παράθυρο του στατιστικού πακέτου S.P.S.S.: **Analyze→Regression→Linear**. Στο νέο παράθυρο διαλόγου που προκύπτει τοποθετείται η Μείωση Βάρους ως εξαρτημένη μεταβλητή (Dependent) και οι Μήνες δίαιτας ως

ανεξάρτητη μεταβλητή (Independent), αντίστοιχα, ενώ στο πεδίο Method επιβεβαιώνουμε ότι η επιλογή Enter έχει καθοριστεί.

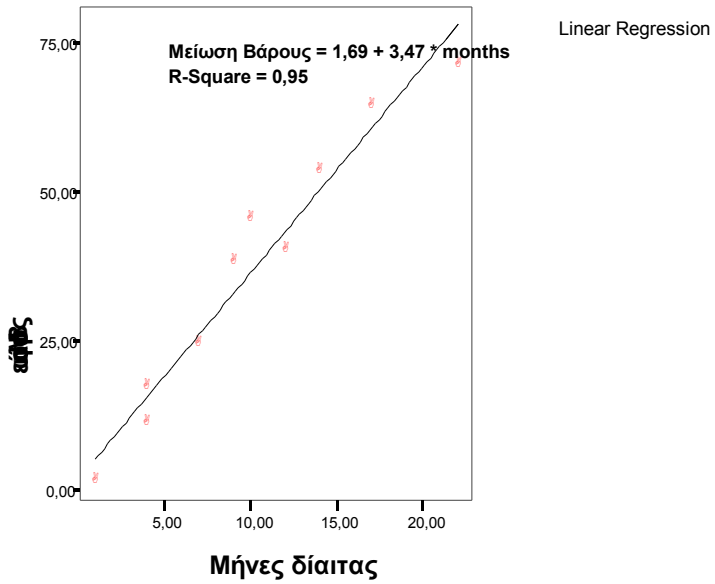


4. Από την επιλογή Statistics επιλέγουμε, προς το παρόν, τα ακόλουθα, τα αποτελέσματα των οποίων θα δούμε μέσω της ερμηνείας των αποτελεσμάτων, πατάμε Continue και OK:



Ερμηνεία αποτελεσμάτων

Η γραφική παράσταση που προκύπτει, αν ζητήσουμε να προσαρμοστεί και η ευθεία της γραμμικής παλινδρόμησης (διπλό κλικ στο γράφημα και έπειτα δεξί κλικ και επιλογή Add Fit Line at Total) είναι η ακόλουθη:



Παρατηρούμε ότι η γραφική αυτή παράσταση μας δείχνει ότι η σχέση των δύο μεταβλητών είναι γραμμική σε αρκετά ικανοποιητικό βαθμό και επομένως είναι λογικό να προσαρμόσουμε το μοντέλο της απλής γραμμικής παλινδρόμησης. Στο ίδιο συμπέρασμα καταλήγουμε ερμηνεύοντας και το αποτέλεσμα για το συντελεστή συσχέτισης του Pearson (βλέπε πίνακα Correlations, r=0.976, p-τιμή<0,001), παρότι θα πρέπει να είμαστε επιφυλακτικοί καθώς (όπως έχει ήδη αναφερθεί στο 3^{ο} Κεφάλαιο) αυτός επηρεάζεται από την ύπαρξη ακραίων τιμών, ενώ ο στατιστικός έλεγχος αν υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ της μείωσης βάρους και του αριθμού των μηνών που διεξήχθη η δίαιτα υποθέτει την ύπαρξη διδιάστατης κανονικότητας.

Correlations

		Μείωση Βάρους	Μήνες δίαιτας
Pearson Correlation	Μείωση Βάρους	1,000	,976
	Μήνες δίαιτας	,976	1,000
Sig. (1-tailed)	Μείωση Βάρους	.	,000
	Μήνες δίαιτας	,000	.
N	Μείωση Βάρους	10	10
	Μήνες δίαιτας	10	10

Θέλοντας να κατασκευάσουμε ένα μοντέλο πρόβλεψης της μείωσης του βάρους από τους μήνες δίαιτας προσαρμόζουμε το μοντέλο:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,...,10,$$

όπου Y_i η Μείωση Βάρους του i-οστού ατόμου (η μέση απώλεια βάρους είναι 36,4 λίβρες και η τυπική απόκλιση 22,97922 λίβρες) και X_i οι Μήνες Δίαιτας του i-οστού ατόμου, αντίστοιχα (η μέση διάρκεια δίαιτας είναι 10 μήνες και η τυπική απόκλιση 6,46357 μήνες, βλέπε πίνακα Descriptive Statistics).

Descriptive Statistics

	Mean	Std. Deviation	N
Μείωση Βάρους	36,4000	22,97922	10
Μήνες δίαιτας	10,0000	6,46357	10

Ο έλεγχος της υπόθεσης ότι δεν υπάρχει παλινδρόμηση έδειξε ότι η υπόθεση αυτή απορρίπτεται (βλέπε Πίνακα ΑΝΑΔΙΑ, $F = \frac{MS_{reg}}{MS_{res}} = 162,430, p - \text{τιμή} < 0.001$).

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4529,322	1	4529,322	162,430	,000(a)
	Residual	223,078	8	27,885		
	Total	4752,400	9			

a Predictors: (Constant), Μήνες δίαιτας
b Dependent Variable: Μείωση Βάρους

<u>Σχόλιο</u>: Από τον πίνακα ANOVA έχουμε όλες τις πληροφορίες που περιέχονται σε ένα ΠΙΝΑΚΑ ΑΝΑΔΙΑ: Άθροισμα Τετραγώνων (Sum of Squares) της Παλινδρόμησης (Regression), των Υπολοίπων (Residual), καθώς και συνολικό άθροισμα τετραγώνων (Total), βαθμοί ελευθερίας (df), μέσα τετράγωνα (Mean Square) της παλινδρόμησης και των υπολοίπων, τιμή του F-στατιστικού τεστ για τον έλεγχο της υπόθεσης $\beta_1 = 0$ και αντίστοιχη p-τιμή).

Με τη μέθοδο των ελαχίστων τετραγώνων προκύπτουν, οι ακόλουθοι εκτιμητές (οι λεγόμενοι εκτιμητές ελαχίστων τετραγώνων των παραμέτρων του μοντέλου, στήλη Unstandardized Coefficients B)

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 1.693$$

και

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = 3.471 .$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	1,693	3,194		,530	,611	-5,674	9,059
	Μήνες δίαιτας	3,471	,272	,976	12,7	,000	2,843	4,099

a. Dependent Variable: Μείωση Βάρους

Το γεγονός αυτό σημαίνει ότι υπό την προϋπόθεση ότι το εκτιμώμενο μοντέλο είναι σωστό ισχύει ότι:

$$\hat{Y} = 1.693 + 3.471X,$$

δηλαδή μπορούμε να πούμε ότι $\hat{\beta}_0 = 1.693$ κιλά είναι η απώλεια βάρους όταν κάποιος δεν κάνει δίαιτα (άρα γίνεται αντιληπτό ότι το μοντέλο με σταθερό όρο δεν είναι λογικό) και $\hat{\beta}_1 = 3.471$ κιλά είναι η απώλεια βάρους που θα έχει κάποιος αν κάνει <u>ένα</u> μήνα περισσότερο δίαιτα (γενικά ισχύει ότι αν $\hat{\beta}_1 > 0$ αύξηση της τιμής της ανεξάρτητης μεταβλητής κατά μία μονάδα επιφέρει αύξηση των τιμών της εξαρτημένης κατά $\hat{\beta}_1$ μονάδες, ενώ όταν $\hat{\beta}_1 < 0$ αύξηση της τιμής της ανεξάρτητης κατά μία μονάδα επιφέρει ελάττωση των τιμών της εξαρτημένης κατά $\hat{\beta}_1$ μονάδες, και θα πρέπει να ελέγχουμε αν τα αποτελέσματα αυτά συμφωνούν με τη φύση του προβλήματος).

<u>Παρατήρηση</u>: Η εκτιμώμενη εξίσωση $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ δεν θα πρέπει να χρησιμοποιείται για την πρόβλεψη των τιμών της εξαρτημένης μεταβλητής για τιμές της ανεξάρτητης πέρα του

πεδίου τιμών αυτής με το οποίο δημιουργήθηκε το μαθηματικό μας μοντέλο. Δηλαδή για το συγκεκριμένο παράδειγμα δε μπορεί να προβλεφθεί η απώλεια βάρους για π.χ. 35 μήνες δίαιτας.

Επιπρόσθετα προκύπτει ότι οι μήνες δίαιτας εξηγούν το 95.3% της μεταβλητότητας της μείωσης βάρους (βλέπε πίνακα Model Summary, $R^2 = SS_{reg} / SS_{tot} = 0.953$). Το αποτέλεσμα αυτό είναι αρκετά ικανοποιητικό.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,976(a)	,953	,947	5,28060

a Predictors: (Constant), Μήνες δίαιτας

<u>Σχόλιο</u>: Για να είναι εφικτή η σύγκριση μοντέλων που έχουν την ίδια εξαρτημένη μεταβλητή και διαφορετικό αριθμό ανεξάρτητων μεταβλητών έχει υιοθετηθεί ο προσαρμοσμένος συντελεστής R^2. Αυτός υπολογίζεται από τη σχέση Adjusted $R^2 = R^2 - \frac{(1-R^2)(k-1)}{(n-k)}$, όπου k το πλήθος των ανεξάρτητων μεταβλητών συμπεριλαμβανομένου του σταθερού όρου και n το μέγεθος του δείγματος.

Επιπλέον, υπό την προϋπόθεση ότι τα σφάλματα ε_i, $i = 1,...,10,$ ακολουθούν κανονική κατανομή με μέση τιμή 0, σταθερή διακύμανση σ^2 και είναι ασυσχέτιστα μεταξύ τους ανά δύο, δηλαδή $Cov(\varepsilon_i, \varepsilon_j) = 0$ για $i, j = 1,...,10,$ με $i \neq j$, ισχύουν τα ακόλουθα:

α) το 95% Διάστημα Εμπιστοσύνης για τους συντελεστές του μοντέλου της παλινδρόμησης είναι (-5.674, 9.0591) και (2.843,4.099), αντίστοιχα (βλέπε στήλη 95% Confidence Interval for B).

β) Επιπλέον, συμπεραίνουμε ότι στο μοντέλο δεν πρέπει να συμπεριληφθεί σταθερός όρος, καθώς δεν απορρίπτεται η υπόθεση $H_0 : \beta_0 = 0$, ενώ δικαιολογείται το μοντέλο της παλινδρόμησης καθώς απορρίπτεται η μηδενική υπόθεση και αυτό διότι:

$$t = \frac{\hat{\beta}_0}{\sqrt{\hat{V}ar(\hat{\beta}_0)}} = 0.530, \text{ p-τιμή=0.611>0.05,}$$

και

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{V}ar(\hat{\beta}_1)}} = 12.7, \text{ p-τιμή}<0.001.$$

Σχόλιο: Από τον πίνακα Coefficient Correlations υπολογίζονται οι εκτιμητές των $Cov(\hat{\beta}_i, \hat{\beta}_j)$, $i, j = 0,1$.

Προσοχή: Στην επόμενη ενότητα θα δούμε τρόπους ελέγχου των υποθέσεων που διατυπώθηκαν πριν το προηγούμενο σχόλιο.

8.2 Έλεγχος των υποθέσεων της απλής γραμμικής παλινδρόμησης

Το μοντέλο της γραμμικής παλινδρόμησης στηρίζεται στις ακόλουθες υποθέσεις για τα σφάλματα ε_i, $i = 1,...,n$:

α) ακολουθούν κανονική κατανομή (με μέση τιμή 0),

β) έχουν σταθερή διακύμανση σ^2 και

γ) είναι ασυσχέτιστα μεταξύ τους ανά δύο.

Επιπλέον, υποθέτουμε ότι

δ) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,...,n$, η ανεξάρτητη μεταβλητή συνδέεται με τις εξαρτημένες μέσω της γραμμικής σχέσης

Στις προηγούμενες υποθέσεις θα πρέπει να προσθέσουμε τις υποθέσεις:
ε) της μη ύπαρξης ακραίων τιμών και
στ) της μη ύπαρξης επηρεάζουσων παρατηρήσεων.

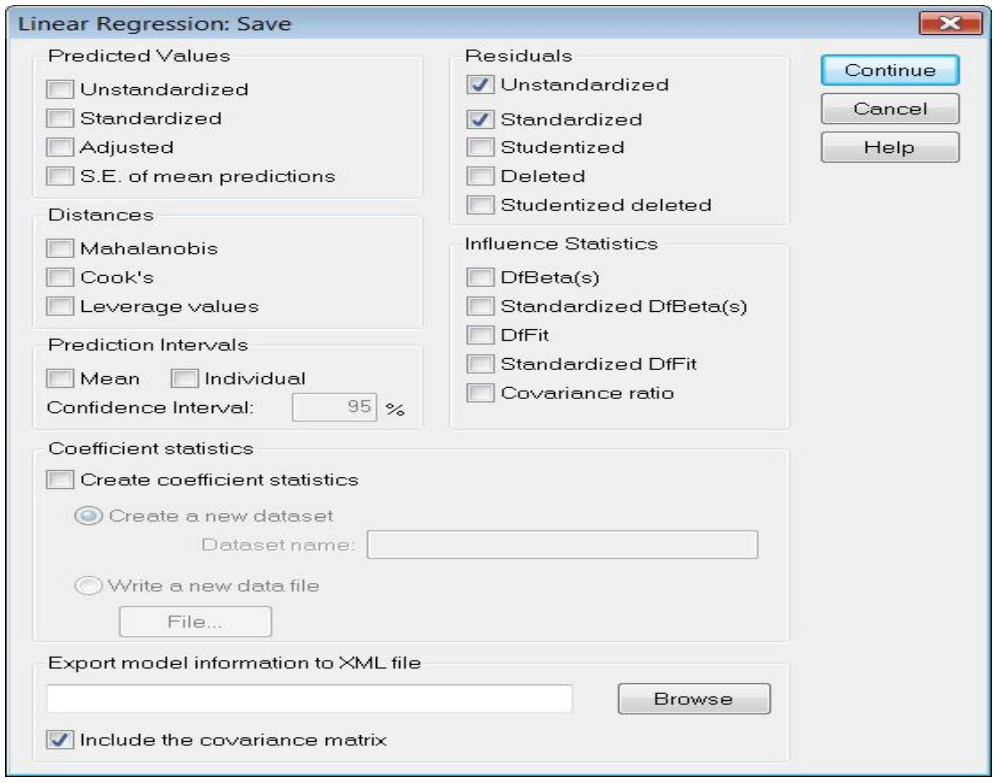
Σε αυτήν την παράγραφο θα υποδείξουμε τρόπους ελέγχου των παραπάνω υποθέσεων. Επιπλέον, θα σχολιάσουμε εν συντομία τα προβλήματα που δημιουργούνται όταν δεν ικανοποιούνται και τέλος θα προτείνουμε ή θα παραπέμψουμε τον αναγνώστη σε λύσεις για την αντιμετώπισή τους.

8.2.1 Έλεγχος κανονικότητας των σφαλμάτων

Ο έλεγχος της κανονικής κατανομής των σφαλμάτων γίνεται με την βοήθεια είτε των υπολοίπων $e_i = Y_i - \hat{Y}_i,\ i = 1,...,n,$ είτε των τυποποιημένων υπολοίπων $e_{si} = \frac{e_i}{\sqrt{MSres}}$. Με τη βοήθεια του S.P.S.S. μπορούμε να προχωρήσουμε τόσο σε γραφικό όσο και σε στατιστικό έλεγχο, χρησιμοποιώντας τις πιο πάνω ποσότητες και τη διαδικασία Explore.

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση είτε των Unstandardized Residuals είτε των Standardized Residuals (μη τυποποιημένα και τυποποιημένα υπόλοιπα αντίστοιχα).



Έπειτα, ελέγχουμε με γραφικούς τρόπους (βλέπε Q-Q plot και Detrended Q-Q plot) και με το στατιστικό τεστ των Shapiro-Wilk (βλέπε διαδικασία Explore, 2^{ο} Κεφάλαιο).

Για το παράδειγμα της προηγούμενης παραγράφου, προκύπτει ότι η υπόθεση της κανονικότητας των υπολοίπων δεν απορρίπτεται (τεστ Shapiro-Wilk, p-τιμή=0.784>0,05).

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,155	10	,200(*)	,960	10	,784

* This is a lower bound of the true significance.

a Lilliefors Significance Correction

Τρόποι διόρθωσης του προβλήματος

Εάν η υπόθεση της κανονικότητας των σφαλμάτων του μοντέλου μας δεν μπορεί να γίνει δεκτή τότε καταφεύγουμε σε ένα μετασχηματισμό των τιμών της εξαρτημένης μεταβλητής έτσι ώστε να επιτευχθεί η κανονικότητα. Η μορφή του ιστογράμματος των υπολοίπων ίσως μας υποδεικνύει ποιος μετασχηματισμός είναι κατάλληλος. Ενδεικτικά αν έχουμε ιστόγραμμα με ουρά για μεγάλες τιμές, τότε είναι κατάλληλος ο μετασχηματισμός του λογαρίθμου, ενώ αν η ουρά παρατηρείται για τις μικρές τιμές, θεωρούμε το μετασχηματισμό της ρίζας. Εναλλακτικά είναι διαθέσιμος ο μετασχηματισμός των Box and Cox (1964).

Αν το μέγεθος του δείγματος είναι μεγάλο (λόγω του Κεντρικού Οριακού Θεωρήματος) χρησιμοποιούμε την κανονικότητα των σφαλμάτων προσεγγιστικά, με τη διαφοροποίηση ότι οι κρίσιμες πιθανότητες (οι p-τιμές δηλαδή) είναι προσεγγιστικές και όχι ακριβείς.

Συνέπειες της μη κανονικότητας των σφαλμάτων

Η μη κανονικότητα των σφαλμάτων έχει τις ακόλουθες συνέπειες:

α) Λάθος διαστήματα εμπιστοσύνης και μη σωστοί έλεγχοι υποθέσεων για τις παραμέτρους του μοντέλου.

β) Οι εκτιμητές ελαχίστων τετραγώνων δεν είναι Α.Ο.Ε.Δ.

Παρατήρηση: Διάφορα άλλα προβλήματα παραβίασης των υποθέσεων του μοντέλου μπορούν να έχουν ως συνέπεια τη μη κανονικότητα των υπολοίπων. Μέλημά μας είναι η διόρθωση των υπόλοιπων προβλημάτων και όχι της μη κανονικότητας, ειδικά αν έχουμε μεγάλο σε μέγεθος δείγμα.

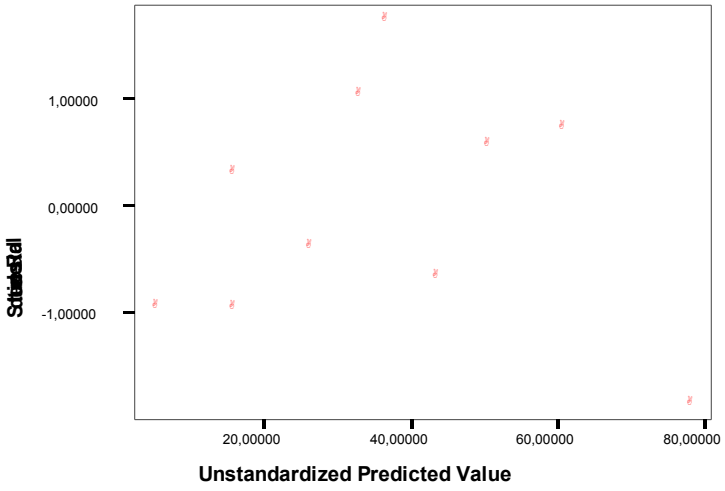
8.2.2 Έλεγχος σταθερής διακύμανσης των σφαλμάτων

Ο έλεγχος της σταθερής διακύμανσης των σφαλμάτων γίνεται (βλέπε μεταξύ άλλων Seber, 1977, σελ. 165) με τη γραφική παράσταση είτε των τυποποιημένων υπολοίπων (Standardized Residuals) είτε των μαθητικοποιημένων υπολοίπων (Studentized Residuals) ως προς τις εκτιμώμενες τιμές (Unstandardized Predicted Values). Αν η διακύμανση είναι σταθερή στο γράφημα που προκύπτει παρατηρούμε ότι τα υπόλοιπα κατανέμονται τυχαία γύρω από μία οριζόντια γραμμή που περνά από το 0.

Αντίθετα αν διαπιστώσουμε για παράδειγμα είτε αύξηση είτε ελάττωση της διακύμανσης με τις εκτιμώμενες τιμές, υπάρχει πρόβλημα σταθερής διακύμανσης. Κάτι τέτοιο δεν πρέπει να θεωρείται ασυνήθιστο. Είναι αναμενόμενο να συμβεί, για παράδειγμα, αν έχουμε ως εξαρτημένη μεταβλητή τις Αποδοχές ενός υπαλλήλου και ως ανεξάρτητη τα χρόνια των σπουδών (βλέπε και πρόβλημα συσχετισμένων σφαλμάτων).

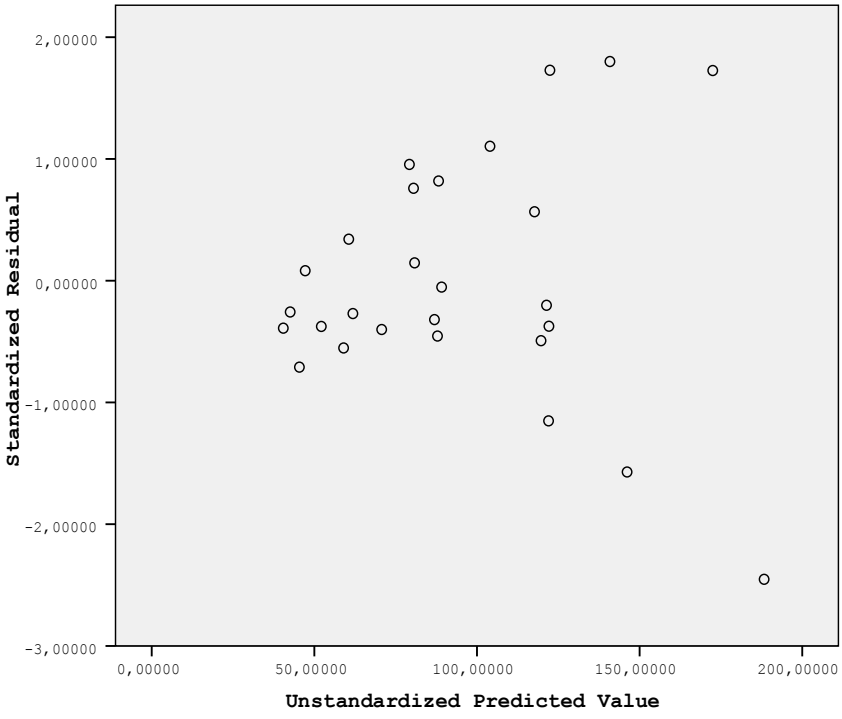
Υλοποίηση στο S.P.S.S.

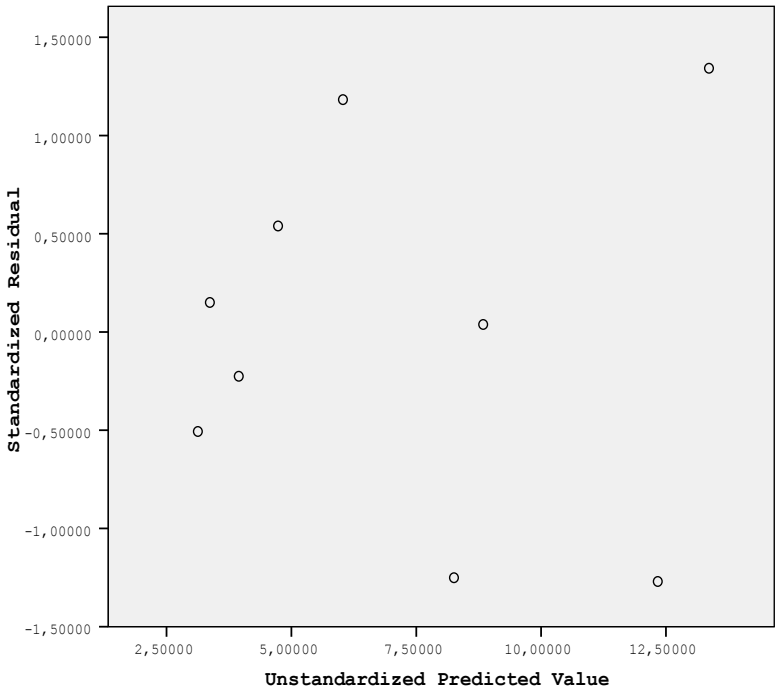
Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των Unstandardized Predicted Values και των Studentized Residuals (μη τυποποιημένες εκτιμώμενες τιμές και μαθητικοποιημένα υπόλοιπα, αντίστοιχα). Έπειτα, κάνουμε τη γραφική παράσταση αυτών π.χ. μέσω της διαδικασίας Graphs →Interactive →Scatter plot.



Από το γράφημα αυτό παρατηρούμε ότι τα υπόλοιπα κατανέμονται τυχαία γύρω από το μηδέν, αλλά υπάρχει και μία παρατήρηση η οποία είναι κάπως ασυνήθιστη (βλέπε δεξιά κάτω γωνία).

Στη συνέχεια παρατίθενται κάποιες ενδεικτικές γραφικές παραστάσεις βασισμένες σε παραδείγματα των Chatterjee, S. and Price, B. (1977), από τις οποίες συμπεραίνουμε ότι η υπόθεση της σταθερής διακύμανσης απορρίπτεται.





Τρόποι διόρθωσης του προβλήματος

Η χρήση των γενικευμένων εκτιμητών ελαχίστων τετραγώνων ή η προσθήκη κάποιου όρου στο μοντέλο ή ένας κατάλληλος μετασχηματισμός των τιμών της εξαρτημένης μεταβλητής συνιστούν τρόπους διόρθωσης του προβλήματος της μη σταθερής διακύμανσης. Ενδεικτικά αναφέρουμε (βλέπε για περισσότερες λεπτομέρειες Rawlings (1988) και Καρακώστας (2002)) ότι οι συνηθέστεροι μετασχηματισμοί είναι: η τετραγωνική ρίζα (όταν η εξαρτημένη μεταβλητή περιγράφει τον αριθμό των γεγονότων σε κάποιο χρονικό διάστημα δηλ. ακολουθεί Poisson κατανομή), ο λογάριθμος (το εύρος των τιμών της εξαρτημένης είναι μεγάλο και λαμβάνει θετικές τιμές) και ο αντίστροφος μετασχηματισμός (η πλειοψηφία των τιμών κοντά στο μηδέν αλλά υπάρχουν και κάποιες αρκετά μεγάλες τιμές).

Συνέπειες της μη σταθερής διακύμανσης των σφαλμάτων

Η μη σταθερή διακύμανση των σφαλμάτων έχει τις ακόλουθες συνέπειες:

α) Λάθος εκτίμηση της διακύμανσης των εκτιμητών των παραμέτρων του μοντέλου.

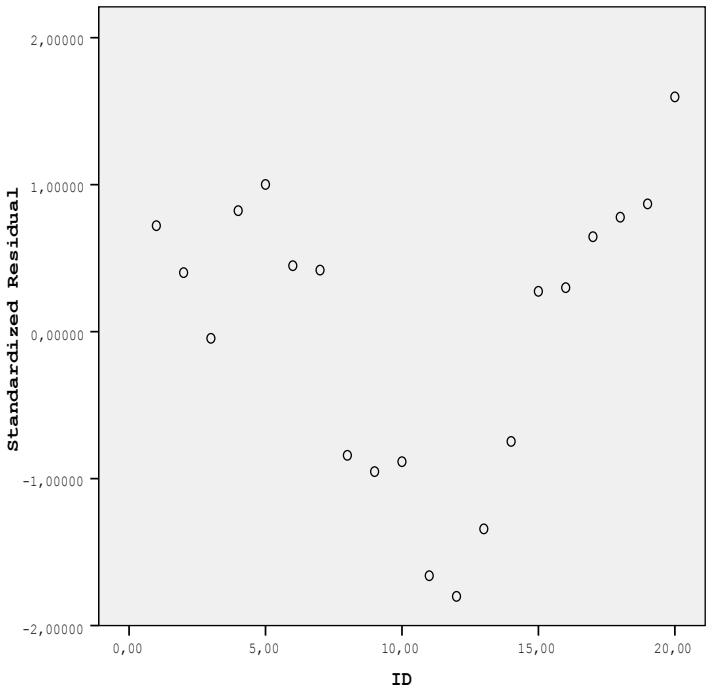
β) Μη αξιόπιστα διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου,

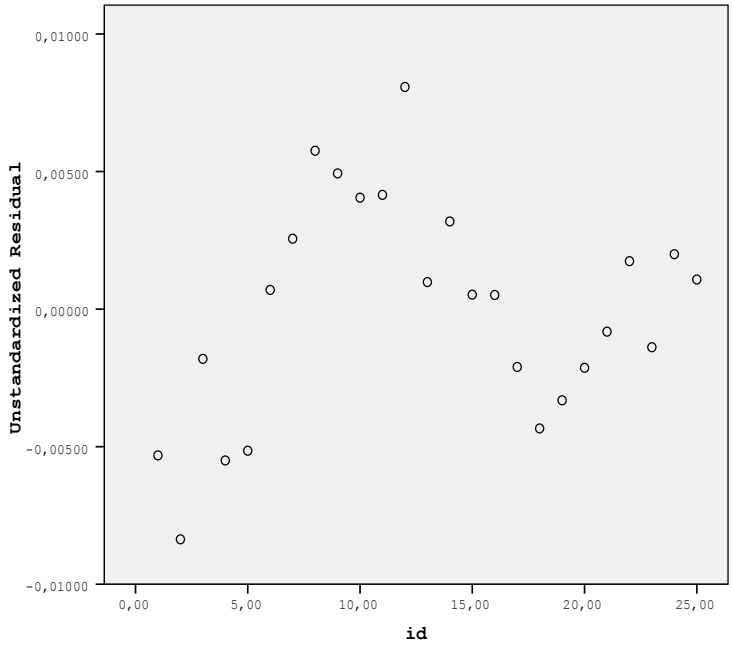
γ) Μη αξιόπιστοι έλεγχοι υποθέσεων για τις παραμέτρους του μοντέλου.

8.2.3 Έλεγχος ασυσχέτιστου των σφαλμάτων

Η ύπαρξη συσχετισμένων σφαλμάτων μπορεί να οφείλεται σε πολλούς λόγους. Είναι σύνηθες φαινόμενο στην περίπτωση που τα δεδομένα έχουν καταγραφεί σε χρονολογική σειρά. Στη συνέχεια παραθέτουμε κάποιους από τους τρόπους ελέγχου της ύπαρξης ή μη συσχετισμένων σφαλμάτων, υλοποιώντας τους στο S.P.S.S. για το παράδειγμα της προηγούμενης παραγράφου.

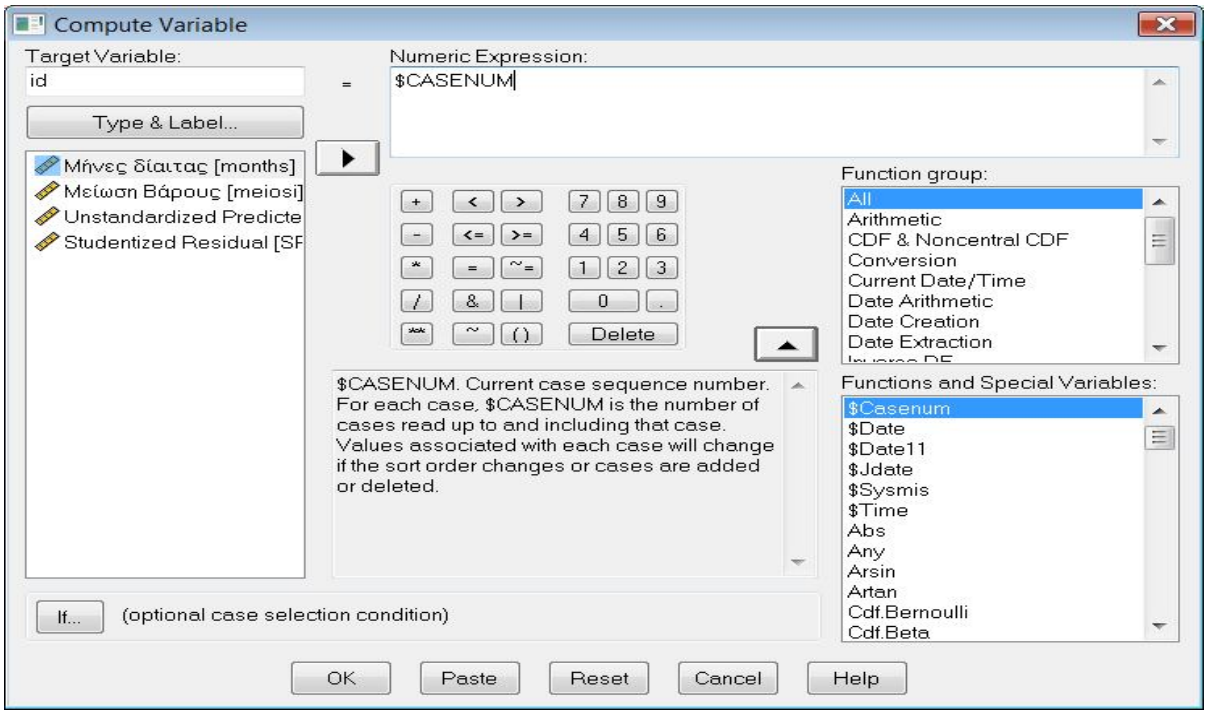
1. Γραφικά ο έλεγχος της ύπαρξης ή μη αυτοσυσχέτιστων σφαλμάτων γίνεται με την γραφική παράσταση των υπολοίπων (ή των μαθητικοποιημένων υπολοίπων) ως προς την χρονολογική σειρά των παρατηρήσεων. Αν η εικόνα μιας τέτοιας γραφικής παράστασης έχει κυματοειδή μορφή ή οι τιμές των υπολοίπων σχετίζονται με τη χρονολογική σειρά (π.χ. στην αρχή μεγάλες τιμές έπειτα μικρές κ.ο.κ.) τότε οδηγούμαστε στο συμπέρασμα ότι υπάρχει αυτοσυσχέτιση μεταξύ των σφαλμάτων του μοντέλου μας. Ενδεικτικές τέτοιες γραφικές παραστάσεις είναι οι ακόλουθες που είναι βασισμένες σε παραδείγματα των Chatterjee, S. and Price, B. (1977)





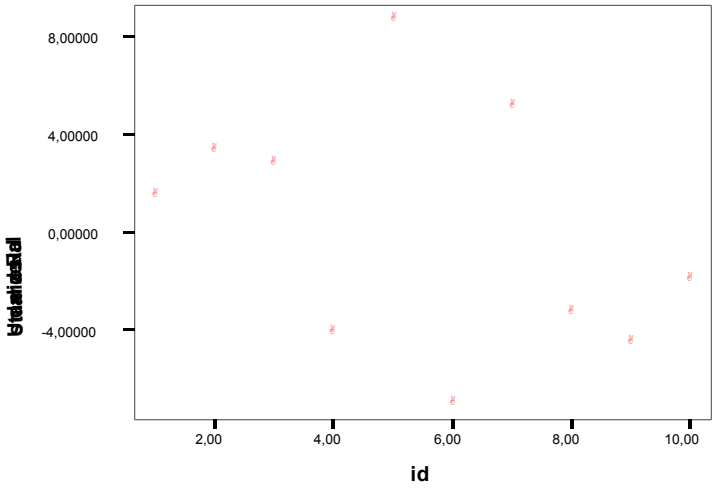
Υλοποίηση στο S.P.S.S.

Δημιουργούμε μία νέα στήλη-μεταβλητή με την ονομασία π.χ. ID. Στη στήλη αυτή καταγράφεται ο αύξων αριθμός της παρατήρησης (συνάρτηση $CASENUM) και έπειτα μέσω π.χ. της διαδικασίας Graphs →Interactive →Scatter plot αποκτούμε το γράφημα που επιθυμούμε.



Ερμηνεία αποτελεσμάτων

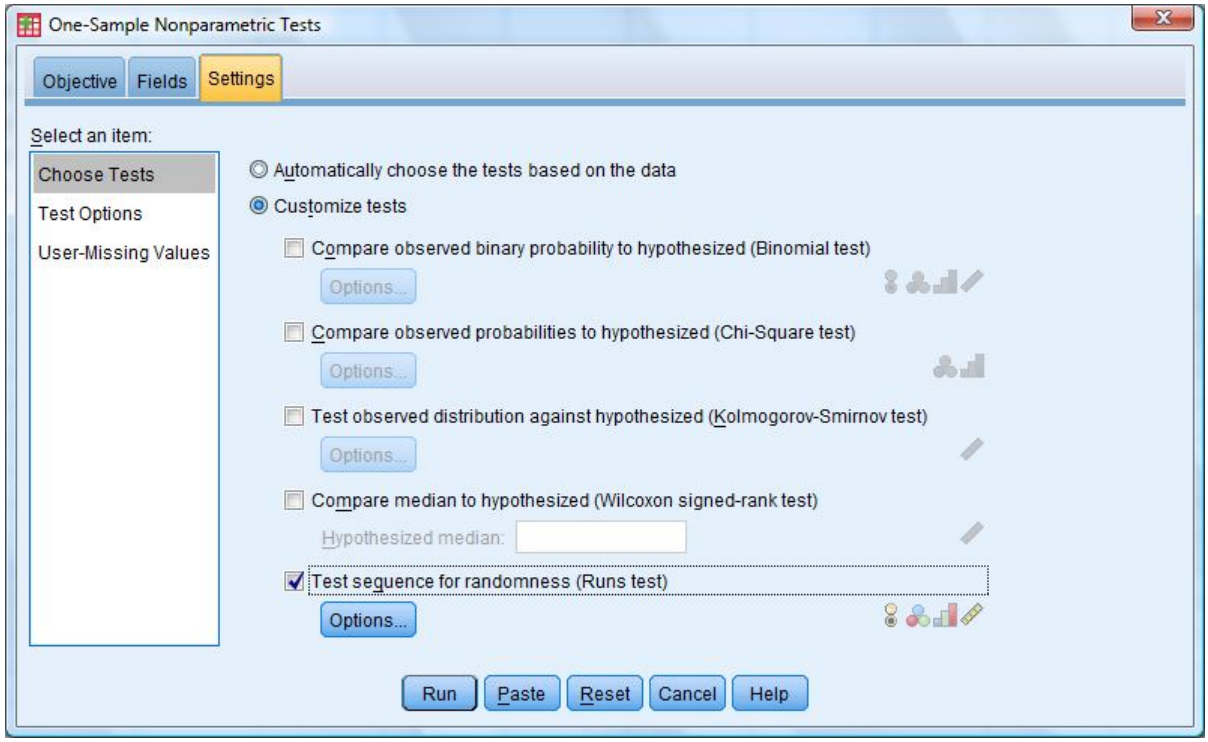
Η εικόνα της γραφικής παράστασης των υπολοίπων ως προς την χρονολογική σειρά δεν έχει κάποια ιδιαίτερη κυματοειδή μορφή, επομένως φαίνεται να μην απορρίπτεται η υπόθεση των ασυσχέτιστων σφαλμάτων.



2. Υπάρχουν και στατιστικά τεστ που ελέγχουν αν τα σφάλματα είναι συσχετισμένα ή όχι. Ένα από τα γνωστά τεστ είναι το τεστ των ροών που στηρίζεται στην ακολουθία-διάταξη των προσήμων των υπολοίπων (είναι διαταγμένα σε χρονολογική σειρά).

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των Standardized Residuals (τυποποιημένα υπόλοιπα). Έπειτα, από το κεντρικό παράθυρο διαλόγου επιλέγουμε: Analyze→NonParametric Tests→One Sample και από το πλαίσιο Settings να επιλέξουμε το Runs test



Από την p-τιμή του ελέγχου αποφασίζουμε αν υπάρχει αυτοσυσχέτιση ή όχι (αν p-τιμή>0.05 δεν υπάρχει αυτοσυσχέτιση).

Ερμηνεία αποτελεσμάτων

Η υπόθεση της τυχαιότητας των σφαλμάτων δεν απορρίπτεται με το τεστ των ροών (p-τιμή>0.05)

3. Ένας άλλος στατιστικός τρόπος εξέτασης της αυτοσυσχέτισης πρώτου βαθμού επιτυγχάνεται με το στατιστικό των Durbin-Watson (Linear Regression Statistics). Το στατιστικό αυτό ελέγχει την μηδενική υπόθεση της μη ύπαρξης αυτοσυσχέτισης έναντι της εναλλακτικής ότι υπάρχει θετική αυτοσυσχέτιση πρώτου βαθμού (γραμμική). Η τιμή d αυτού του στατιστικού συγκρίνεται με τις τιμές d_l και d_u που δίνονται από κατάλληλους πίνακες. Ισχύει ότι αν $d < d_l$ τότε απορρίπτεται η υπόθεση των ασυσχέτιστων σφαλμάτων. Αν $d > d_u$ η υπόθεση δεν μπορεί να απορριφθεί, ενώ αν $d_l < d < d_u$ δεν μπορούμε να πάρουμε απόφαση.

Παρατήρηση Ο έλεγχος της μηδενικής υπόθεσης της μη ύπαρξης αυτοσυσχέτισης έναντι της εναλλακτικής ότι υπάρχει αρνητική αυτοσυσχέτιση πρώτου βαθμού γίνεται ανάλογα χρησιμοποιώντας την τιμή $d^* = 4 - d$.

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Statistics, επιλέγουμε το πλαίσιο Durbin Watson.

Ερμηνεία αποτελεσμάτων

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,976(a)	,953	,947	5,28060	3,070

a Predictors: (Constant), Μήνες δίαιτας
b Dependent Variable: Μείωση Βάρους

Η τιμή του στατιστικού των Durbin-Watson είναι ίση με 3.070 και καθώς $d_l = 0.87913$ και $d_u = 1.31971$ (τιμές που προκύπτουν από ειδικούς πίνακες, βλέπε Καρακώστας 2002) η υπόθεση των ασυσχέτιστων σφαλμάτων δεν απορρίπτεται.

4. Ένας εναλλακτικός γραφικός τρόπος ελέγχου της ύπαρξης αυτοσυσχέτισης k βαθμού αποτελεί η γραφική παράσταση των υπολοίπων $e_1, e_2, ..., e_n$ ως προς τις τιμές $(-, -, ..., e_k, ..., e_{n-1})$. Αν από αυτό το γράφημα προκύπτει μία γραμμική τάση τότε έχουμε αυτοσυσχέτιση k βαθμού.

Υλοποίηση στο S.P.S.S.

Είναι απαραίτητος ο σχηματισμός, η δημιουργία μίας νέας στήλης όπου θα δίνονται οι τιμές των υπολοίπων $(-, -, ..., e_k, ..., e_{n-1})$. Επιτυγχάνεται με χρήση της συνάρτησης LAG(Variable,k), όπου στο πλαίσιο Variable εισάγουμε τη μεταβλητή των υπολοίπων και στο πλαίσιο k το βαθμό της αυτοσυσχέτισης που θέλουμε να ελέγξουμε. Έπειτα μέσω π.χ. της διαδικασίας Graphs →Interactive →Scatter plot αποκτούμε το γράφημα που επιθυμούμε.

Τρόποι διόρθωσης του προβλήματος

Η άρση της αυτοσυσχέτισης επιτυγχάνεται μεταξύ άλλων είτε με κατάλληλο μετασχηματισμό των μεταβλητών είτε με εισαγωγή νέων μεταβλητών. Για λεπτομέρειες σχετικά με αυτούς τους τρόπους παραπέμπουμε τον αναγνώστη στο σύγγραμμα των Chaterjee and Price (1977). Ένας άλλος τρόπος είναι με χρήση γενικευμένων εκτιμητών ελαχίστων τετραγώνων (βλέπε Rawlings (1988)). Στο πλαίσιο αυτού του προπτυχιακού μαθήματος απλά θα επισημαίνουμε την ύπαρξη αυτοσυσχέτισης και τις συνέπειες αυτής και θα προβαίνουμε σε διόρθωση του προβλήματος στην ειδική περίπτωση της ύπαρξης αυτοσυσχέτισης πρώτου βαθμού (βλέπε Άσκηση 1 ενότητα 8.3).

Συνέπειες

Η ύπαρξη αυτοσυσχέτισης μεταξύ των σφαλμάτων του μοντέλου έχει τις ακόλουθες συνέπειες (Chaterjee and Price (1977)):

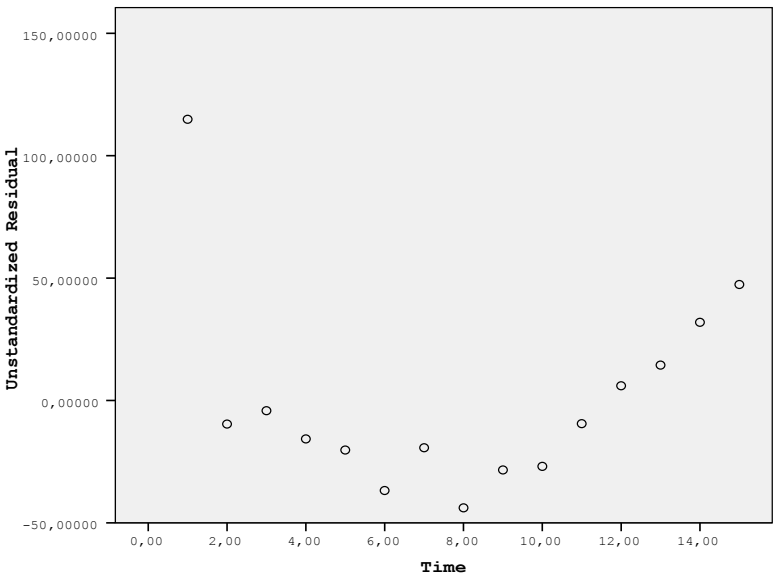
α) Οι εκτιμητές ελαχίστων τετραγώνων είναι αμερόληπτοι, αλλά όχι ΑΟΕΔ.

β) Ο εκτιμητής του σ και τα τυπικά σφάλματα των συντελεστών της παλινδρόμησης μπορεί να υποεκτιμούνται. Αυτό οδηγεί σε μη αξιόπιστα αποτελέσματα για τα διαστήματα εμπιστοσύνης και για τους ελέγχους υποθέσεων για τις παραμέτρους του μοντέλου.

8.2.4 Έλεγχος ορθότητας μοντέλου

Γραφικά, ο έλεγχος της ορθότητας του μοντέλου γίνεται (βλέπε μεταξύ άλλων Norusis (2002)) με την γραφική παράσταση των υπολοίπων ως προς την ανεξάρτητη μεταβλητή. Αν δεν παρατηρηθεί κάποια ιδιαίτερη μορφή και τα σημεία βρίσκονται τυχαία γύρω από το μηδέν το μοντέλο μπορεί να θεωρηθεί ορθό. Αν δούμε κάποια ιδιαίτερη γραφική παράσταση τότε η εξαρτημένη και η ανεξάρτητη μεταβλητή μπορεί να μην συνδέονται με μία γραμμική σχέση.

Μία ενδεικτική γραφική παράσταση που υποδεικνύει πρόβλημα ορθότητας μοντέλου (εισαγωγή δευτεροβάθμιου όρου) βασισμένη σε παράδειγμα των Chatterjee, S. and Price, B. (1977)) είναι η ακόλουθη (time είναι η ανεξάρτητη μεταβλητή):



Παρατήρηση: Από τη γραφική παράσταση των υπολοίπων ως προς τις τιμές μίας ανεξάρτητης μεταβλητής που δεν είναι στο μοντέλο μπορούμε να αποφασίσουμε αν πρέπει η συγκεκριμένη μεταβλητή να συμπεριληφθεί στο μοντέλο ή όχι. Έτσι αν παρατηρηθεί κάποια σχέση τότε ίσως πρέπει να συμπεριληφθεί αυτή η ανεξάρτητη μεταβλητή στο μοντέλο.

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των Unstandardized Residuals (μη τυποποιημένα υπόλοιπα). Έπειτα, κάνουμε τη γραφική παράσταση αυτών π.χ. μέσω της διαδικασίας Graphs →Interactive →Scatter plot.

Τρόποι διόρθωσης του προβλήματος

Κάποιες μη γραμμικές σχέσεις είναι δυνατό να αναχθούν σε γραμμικές με κατάλληλους μετασχηματισμούς. Στον παρακάτω πίνακα αναφέρονται κάποιες από αυτές καθώς και ο μετασχηματισμός που οδηγεί σε γραμμικά μοντέλα (βλέπε Rawlings (1988, σελ. 306-308), Chaterjee and Price (1977, σελ. 29)). Φυσικά όλες οι μη γραμμικές σχέσεις δεν είναι δυνατό να μετατραπούν σε γραμμικές. Η μελέτη και προσαρμογή μη γραμμικών μοντέλων ξεφεύγει από τους σκοπούς αυτών των σημειώσεων.

Μη γραμμική σχέση	Κατάλληλος Μετασχηματισμός
$Y_i = aX_i^{\beta}\varepsilon_i$	Λογάριθμος
$Y_i = a\exp(\beta X_i)\varepsilon_i$	Λογάριθμος
$Y_i = \frac{X_i}{a + \beta X_i + \varepsilon_i}$	Αντίστροφος
$Y_i = \frac{a}{1 + \gamma \exp(-\beta X_i)\varepsilon_i}$	$Y^* = \ln\left(\frac{a}{Y} - 1\right)$

Συνέπειες του μη ορθού μοντέλου

Οι συνέπειες ενός μη ορθού μοντέλου είναι:

α) λάθος ερμηνεία των παραμέτρων του μοντέλου,

β) λάθος προβλέψεις,

γ) λάθος εκτίμηση της κοινής διακύμανσης των σφαλμάτων.

Από το τελευταίο προκύπτει ως επακόλουθη συνέπεια

δ) η μη εγκυρότητα των όποιων διαστημάτων εμπιστοσύνης και ελέγχων υποθέσεων για τις παραμέτρους του μοντέλου.

8.2.5 Έλεγχος ακραίων τιμών

Σε μερικές περιπτώσεις το μοντέλο φαίνεται να είναι ορθό για την πλειοψηφία των δεδομένων, αλλά υπάρχει ένα υπόλοιπο που η απόλυτη τιμή του είναι πολύ μεγαλύτερη από τα άλλα υπόλοιπα. Κάτι τέτοιο μπορεί να οφείλεται σε λάθος καταγραφή των δεδομένων αλλά και όχι. Στη δεύτερη περίπτωση η μελέτη της ακραίας τιμής είναι εξίσου σημαντική όσο και η μελέτη του υπόλοιπου συνόλου δεδομένων καθώς μπορεί να μας δώσει σημαντικές πληροφορίες. Έτσι, η αυτόματη απομάκρυνση των ακραίων τιμών δεν συνίσταται.

Ένας τρόπος ελέγχου της ύπαρξης ή μη ακραίων παρατηρήσεων στα δεδομένα μας γίνεται με τη βοήθεια των τυποποιημένων ή μαθητικοποιημένων υπολοίπων. Τότε (βλέπε

μεταξύ άλλων Field, 2005, σελ. 164) παρατηρήσεις των οποίων η απόλυτη τιμή των υπολοίπων αυτών είναι μεγαλύτερη του τρία (για να είμαστε περισσότερο ακριβείς του 3.29) θεωρούνται ακραίες και συνηθέστερα αποκλείονται από την περαιτέρω ανάλυση. Αν περισσότερο από 1% των τυποποιημένων υπολοίπων έχουν απόλυτες τιμές μεγαλύτερες του 2.5 (για την ακρίβεια του 2.58) υποδεικνύεται ότι το μοντέλο έχει κακή προσαρμογή. Στο ίδιο συμπέρασμα καταλήγουμε αν 5% των διαθέσιμων παρατηρήσεων έχουν απόλυτες τιμές των τυποποιημένων υπολοίπων μεγαλύτερες του 2 (του 1.96 για την ακρίβεια όταν το επίπεδο σημαντικότητας είναι 5%). Τέλος, παρατηρήσεις με απόλυτες τιμές των τυποποιημένων υπολοίπων μεταξύ 2 και 3 (1.96 και 3.29 αν θέλουμε να είμαστε πιο ακριβείς) θεωρούνται ως πιθανές ακραίες. Η τελική απόφαση για το αν είναι ακραίες ή όχι γίνεται με τη βοήθεια ενός στατιστικού ελέγχου.

Ο στατιστικός έλεγχος για την ύπαρξη ακραίων παρατηρήσεων γίνεται με την βοήθεια των μαθητικοποιημένων διαγραφόμενων υπολοίπων (studentized deleted residuals). Αν τα σφάλματα του μοντέλου ακολουθούν κανονική κατανομή, τότε η κατανομή των studentized deleted residuals είναι t-κατανομή με n-p-1 βαθμούς ελευθερίας. Απόλυτες τιμές των μαθητικοποιημένων διαγραφόμενων υπολοίπων για μία παρατήρηση μεγαλύτερες του $t_{n-p-1,a/2}$ = IDF.T(1-a/2,n-p-1) υποδεικνύουν τη συγκεκριμένη παρατήρηση ως ακραία.

Τρόποι διόρθωσης του προβλήματος

Το πρώτο μέλημα μας όταν έχουμε αρκετές ακραίες τιμές είναι η εύρεση ενός μετασχηματισμού που θα διορθώσει το πρόβλημα. Αν η εύρεση ενός τέτοιου μετασχηματισμού είναι αδύνατη τότε είτε θα απορρίψουμε τις ακραίες τιμές και θα προχωρήσουμε στην ανάλυση των υπόλοιπων δεδομένων (τακτική που οδηγεί πολλές φορές σε απώλεια σημαντικής πληροφορίας) είτε θα χρησιμοποιήσουμε μεθόδους ανθεκτικές στην ύπαρξη ακραίων τιμών (βλέπε Huber (1973)). Υπάρχει βέβαια και η επιλογή της διεξαγωγής της έρευνας τόσο με τις ακραίες όσο και χωρίς τις ακραίες τιμές και την επισήμανση των όποιων διαφορετικών αποτελεσμάτων.

Συνέπειες της ύπαρξης ακραίων τιμών

Η παρουσία ακραίων τιμών στο δείγμα μας έχει σαν συνέπεια οι εκτιμητές των παραμέτρων καθώς και οι διακυμάνσεις αυτών να μην έχουν τις γνωστές ιδιότητες των

εκτιμητών ελαχίστων τετραγώνων. Άμεση συνέπεια αυτού είναι η μη εγκυρότητα των όποιων διαστημάτων εμπιστοσύνης ή ελέγχου υποθέσεων για τις παραμέτρους του μοντέλου.

8.2.6 Επηρεάζουσες παρατηρήσεις

Είναι πιθανό δύο ή περισσότερες πειραματικές μονάδες να επιδρούν σημαντικά στο μοντέλο παλινδρόμησης. Τέτοιες παρατηρήσεις ονομάζονται επηρεάζουσες.

Έτσι για παράδειγμα οι συντελεστές των παραμέτρων του μοντέλου αλλάζουν αρκετά όταν οι τιμές των συγκεκριμένων πειραματικών μονάδων εξαιρούνται από τον υπολογισμό τους. Μία τέτοια κατάσταση είναι ανεπιθύμητη καθώς θέλουμε ένα μοντέλο παλινδρόμησης που να μην εξαρτάται από τις τιμές ενός μικρού αριθμού πειραματικών μονάδων, αλλά όλες οι πειραματικές μονάδες να συνεισφέρουν όσο γίνεται το ίδιο στον υπολογισμό των συντελεστών αυτών. Θα πρέπει να δοθεί ξεχωριστή σημασία στις συγκεκριμένες πειραματικές μονάδες που είναι επηρεάζουσες παρατηρήσεις και ίσως πρέπει να παρουσιαστούν τα αποτελέσματα των αναλύσεων με και χωρίς αυτές.

Τρόποι ελέγχου

Από το πλαίσιο Influence Statistics του Linear Regression Save μπορούμε να ζητήσουμε την αποθήκευση διάφορων ποσοτήτων για την εξέταση αυτού του προβλήματος:

DfBeta(s): Η διαφορά στις τιμές των συντελεστών της παλινδρόμησης αν δεν ληφθεί υπόψη η συγκεκριμένη πειραματική μονάδα. Υπολογίζεται και για τον σταθερό όρο. Οι τυποποιημένες τιμές παρατίθενται στη στήλη Standardized DfBeta. Απόλυτες τιμές αυτών μεγαλύτερες από $2/\sqrt{n}$ μας υποδεικνύουν παρατήρηση που επιδρά στην εκτίμηση των συντελεστών της παλινδρόμησης (βλέπε Belsley *et al.* (1980)).

DfFit: Μετρά τη διαφορά στην προσαρμογή, δηλαδή στην εκτιμώμενη τιμή, αν δεν συμπεριληφθεί η συγκεκριμένη παρατήρηση στους υπολογισμούς. Δίνονται και οι αντίστοιχες τυποποιημένες τιμές Standardized DfFit. Απόλυτες τιμές αυτών μεγαλύτερες του $2\sqrt{\frac{p+1}{n}}$ υποδεικνύουν επηρεάζουσες παρατηρήσεις (βλέπε Belsley *et al.* (1980)).

Σχόλιο: Κάποιοι συγγραφείς διαφοροποιούν τα παραπάνω κριτήρια στην περίπτωση που το μέγεθος του δείγματος είναι μικρότερο του 30. Υποστηρίζουν ότι σε μία τέτοια περίπτωση

μία παρατήρηση είναι επηρεάζουσα για τιμές των παραπάνω τυποποιημένων δεικτών μεγαλύτερες της μονάδας.

Covariance ratio: Το πηλίκο της ορίζουσας του πίνακα διακυμάνσεων συνδιακυμάνσεων χωρίς η συγκεκριμένη παρατήρηση να λαμβάνεται υπόψη στους υπολογισμούς προς την αντίστοιχη ορίζουσα όταν αυτή η παρατήρηση λαμβάνεται υπόψη. Τιμές μεγαλύτερες (μικρότερες αντίστοιχα) του $1+3\frac{p+1}{n}$ (του $1-3\frac{p+1}{n}$ αντίστοιχα) υποδεικνύουν επηρεάζουσα παρατήρηση (βλέπε Belsley et al. (1980)).

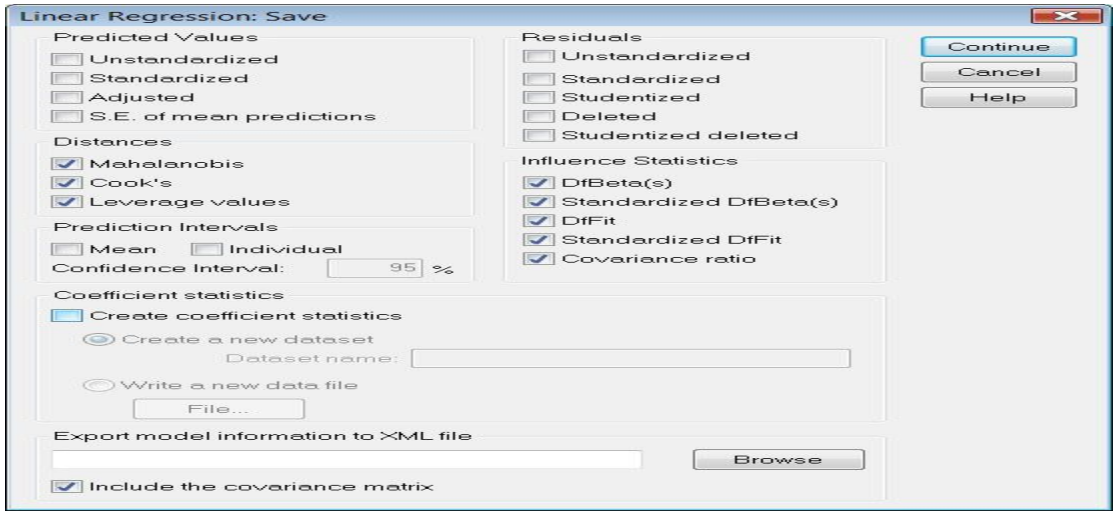
Απόσταση του Cook: καθορίζει πόσο οι τιμές των υπολοίπων όλων των περιπτώσεων θα μεταβληθούν, αν η συγκεκριμένη τιμή δε ληφθεί υπόψη στους υπολογισμούς των συντελεστών του μοντέλου. Αποδεικνύεται ότι το στατιστικό αυτό ακολουθεί μία F κατανομή με 2 και n-2 βαθμούς ελευθερίας. Τιμές της απόστασης του Cook για μία παρατήρηση μεγαλύτερες του $F_{2,n-2,a} = IDF.F(1-a,2,n-2)$ υποδεικνύουν τη συγκεκριμένη παρατήρηση ως επηρεάζουσα.

Απόσταση Mahalanobis: καθορίζει την απόσταση των πειραματικών μονάδων από τη μέση τιμή της προβλεπόμενης τιμής. Το ενδιαφέρον επικεντρώνεται στις πειραματικές μονάδες με μεγάλες τιμές σε αυτή, αλλά δυστυχώς δεν υπάρχει ένα γενικό cut-off point (βλέπε Barnett and Lewis, 1978).

Leverage values: Μετρούν την επίδραση μίας πειραματικής μονάδας στην προσαρμογή του μοντέλου της παλινδρόμησης. Οι κεντρικές Leverage values λαμβάνουν τιμές από 0 (όχι ενδείξεις επίδρασης) έως $(n-1)/n$.

Υλοποίηση στο S.P.S.S.

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Linear Regression και από την επιλογή Save, ζητούμε την αποθήκευση των



Για τον εντοπισμό πιθανής επηρεάζουσας παρατήρησης υπολογίζουμε, αρχικά, το λεγόμενο cut-off point (δηλαδή τις ποσότητες $2/\sqrt{n}$, $2\sqrt{\frac{p+1}{n}}$, $1\pm 3\frac{p+1}{n}$ και $F_{2,n-2,a}$). Έπειτα, δημιουργούμε μία νέα στήλη-μεταβλητή με την ονομασία π.χ. ID. Στη στήλη αυτή καταγράφεται ο αύξων αριθμός της παρατήρησης (συνάρτηση $CASENUM). Έπειτα επιλέγουμε Graphs→ Legacy Dialog→ Scatter/Dot και Simple Scatter. Στο νέο παράθυρο διαλόγου που προκύπτει τοποθετούμε στον άξονα των Y π.χ. την απόσταση Cook, ενώ στον άξονα των X την νέα μεταβλητή ID. Κάνοντας διπλό κλικ στο γράφημα που προκύπτει και έπειτα δεξί κλικ ζητούμε την προσθήκη γραμμών αναφοράς στον άξονα Y, Add→ Y Axis Reference Line και στο πλαίσιο Position δηλώνουμε την κατάλληλη τιμή του cut-off point. Αν υπάρχουν σημεία που παραβιάζουν την προς έλεγχο σχέση με το cut-off point με δεξί κλικ και επιλογή του Show Data Labels μας υποδεικνύεται ο αύξων αριθμός της πειραματικής μονάδας (εναλλακτικά αφού επιλέξουμε το σημείο, δεξί κλικ και επιλογή του Go to Case)

8.3 Ασκήσεις

1. Στο αρχείο autocorrelation1.sav καταγράφονται τα τετραμηνιαία δεδομένα από το 1952 έως το 1956 που αφορούν τις δαπάνες και τις αποταμιεύσεις μετρούμενες σε δισ. δολάρια. Οι οικονομολόγοι ενδιαφέρονται για την μεταβολή στις δαπάνες που προκαλούνται από τη μεταβολή στις αποταμιεύσεις (Chatterjee and Price (1977, σελ. 124)).

2. Μία εταιρεία θέλει να κατανοήσει τη σχέση μεταξύ οικοδομικών αδειών (housing starts) και της ανάπτυξης του πληθυσμού. Στο αρχείο autocorrelation2.sav δίνονται τα δεδομένα

αυτά για 25 χρόνια. Επιπλέον, σε μία τρίτη στήλη δίνεται η τιμή ενός δείκτη που μετρά την οικονομική δυνατότητα (mortgage money) (Chatterjee and Price (1977, σελ. 133)).

3. Μία εταιρεία της Αμερικής παράγει και πουλάει εξαρτήματα σκι. Θέλει να προβλέψει τις πωλήσεις της με βάση ένα δείκτη (PDI) που μετρά το εισόδημα. Δίνονται στο αρχείο autocorrelation3.sav τα δεδομένα που αφορούν 40 τρίμηνα από το 1964-1973. (Chatterjee and Price (1977, σελ. 138))

4. Στο αρχείο chatterjeep.44.sav καταγράφονται ο αριθμός των προϊστάμενων και υφιστάμενων 27 εταιρειών. Μπορεί να δημιουργηθεί ένα μοντέλο πρόβλεψης του αριθμού των προϊστάμενων από τον αριθμό των υφιστάμενων; (Chatterjee and Price (1977, σελ. 44))

5. Στο αρχείο chatterjeep.40.sav καταγράφονται το ποσοστό των πτήσεων και ο αριθμός των ατυχημάτων 9 αεροπορικών εταιρειών. Μπορεί να δημιουργηθεί ένα μοντέλο πρόβλεψης του αριθμού των ατυχημάτων από το ποσοστό των πτήσεων; (Chatterjee and Price (1977, σελ. 40))

6. Στο αρχείο δεδομένων chatterjee21.sav καταγράφονται 30 παρατηρήσεις και 2 μεταβλητές που αφορούν την ακροαματικότητα πριν το δελτίο ειδήσεων (lead in) και την ακροαματικότητα του δελτίου ειδήσεων (newsrate). Θέλουμε να εξετάσουμε αν το πρόγραμμα πριν τις ειδήσεις επηρεάζει την ακροαματικότητα των ειδήσεων. (Chatterjee and Price (1977, σελ. 21))

7*. Με σκοπό να μελετηθεί αν τα γενικά έξοδα, σε ετήσια βάση, μιας οικογένειας μπορούν να προβλέψουν τα έξοδα που γίνονται για την εκπαίδευση των παιδιών της οικογένειας συγκεντρώθηκαν τα δεδομένα του αρχείου EducSpend.sav. Η μελέτη αυτή έγινε πιλοτικά και οι οικογένειες επιλέχθηκαν τυχαία από μια γεωγραφική περιοχή μιας πόλης. Η μεταβλητή Pay εκφράζει το ποσό των γενικών εξόδων, ενώ η μεταβλητή Spend εκφράζει τα ειδικά έξοδα για την εκπαίδευση, σε χιλιάδες δολάρια. Με βάση τα δεδομένα του συγκεκριμένου αρχείου να δοθεί μια απάντηση στο αρχικό ερώτημα.

8*. Μια ομάδα γιατρών θέλησε να εξετάσει αν ο ρυθμός θνησιμότητας των γυναικών που πάσχουν από καρκίνο του στήθους επηρεάζεται από την θερμοκρασία. Αν ναι να βρουν ένα μοντέλο με το οποίο θα μπορούν να κάνουν αξιόπιστες προβλέψεις για τον μέσο ρυθμό θνησιμότητας (σε ετήσια βάση) από την μέση ετήσια θερμοκρασία. Για τον σκοπό αυτό

κατέγραψαν τα δεδομένα στο αρχείο BreastCancer.sav, όπου mortality είναι ο ρυθμός θνησιμότητας και Temperature η μέση θερμοκρασία (σε βαθμούς Fahrenheit), για την συγκεκριμένη χρονιά. Ζητείται με βάση τα δεδομένα αυτά να διατυπώσουμε τα συμπεράσματά μας σχετικά με το αρχικό ερώτημα των γιατρών.

9*. Το Οικονομικό Επιμελητήριο μιας χώρας θέλησε να εξετάσει αν οι μεταβολές του πληθωρισμού επηρεάζουν και πώς τα τραπεζικά επιτόκια, σε μηνιαία βάση. Για τον σκοπό αυτό συγκέντρωσε στοιχεία, για τον πληθωρισμό και τα επιτόκια, για τους τελευταίους 191 μήνες. Τα στοιχεία αυτά περιέχονται στο αρχείο BankInflatRate.sav. Στο αρχείο αυτό BankRate είναι το μέσο τραπεζικό επιτόκιο, (σε μηνιαία βάση) και InflatRate ο μέσος μηνιαίος πληθωρισμός.

10*. Τα δεδομένα τα οποία παρουσιάζονται στο αρχείο HeihgtWeight12.sav είναι ένα τυχαίο δείγμα 63 παιδιών ηλικίας 12 ετών από ένα σχολικό συγκρότημα. Σκοπός μας είναι να ελέγξουμε αν και σε ποιο βαθμό το ύψος (σε ίντσες) ενός παιδιού μπορεί να προσδιορίσει το βάρος του (σε λίβρες). Να αναλυθούν τα δεδομένα και να διατυπωθούν τα όποια συμπεράσματα.

11*. Στο αρχείο BirthRatio.sav έχουν συγκεντρωθεί τα αποτελέσματα μιας έρευνας με σκοπό να εξετασθεί κατά πόσον είναι δυνατόν να προβλέψουμε το λόγο του βάρους προς το ύψος (μεταβλητή Ratio στο αρχείο) σε νεογέννητα παιδιά ηλικίας μερικών μηνών (μεταβλητή Age στο αρχείο). (Ο πληθυσμός στον οποίο αναφέρεται το συγκεκριμένο δείγμα είναι αυτός των γυναικών που γέννησαν σε ένα συγκεκριμένο Νοσοκομείο).

Οι ασκήσεις που επισημαίνονται με * καθώς και τα αντίστοιχα σύνολα δεδομένα προέρχονται από το υλικό διδασκαλίας του κ. Κ. Καρακώστα (βλέπε Καρακώστας (2004)).

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΗ

Γναρδέλλης, Χ. (2009). Ανάλυση δεδομένων με το PASW Statistics 17.0. Εκδόσεις Παπαζήση.

Ζωγράφος, Κ. (2003). *Μαθήματα Πιθανοτήτων και Στατιστικής*. Πανεπιστήμιο Ιωαννίνων.

Καρακώστας Κ. (2002). *Γραμμικά Μοντέλα: Παλινδρόμηση, Ανάλυση Διακύμανσης*. Πανεπιστήμιο Ιωαννίνων.

Καρακώστας Κ. (2004). Στατιστική εφαρμοσμένη στις κοινωνικές επιστήμες. Διδακτικές Σημειώσεις. ΠΜΣ Τμήματος Φιλοσοφίας-Παιδαγωγικής-Ψυχολογίας Πανεπιστημίου Ιωαννίνων Επιμέλεια Σούρλου Ευλαμπία.

Μπούτσικας, Μ. (2004). Σημειώσεις Μαθήματος «Στατιστικά Προγράμματα». Τμήμα Στατ. & Ασφ. Επιστήμης, Πανεπιστήμιο Πειραιώς.

Παπαϊωάννου, Τ. και Λουκάς, Σ. (2002). *Εισαγωγή στη Στατιστική*, Δεύτερη Έκδοση, Εκδόσεις Σταμούλη.

Παπαϊωάννου, Τ. και Φερεντίνος, Κ. (2000). *Ιατρική Στατιστική και Στοιχεία Βιομαθηματικών*, Βελτιωμένη Έκδοση, Εκδόσεις Σταμούλη.

Τσάντας, Ν., Μωυσιάδης, Χ., Μπαγιάτης, Ν. και Χατζηπαντελής, Θ. (1999). *Ανάλυση δεδομένων με τη βοήθεια στατιστικών πακέτων S.P.S.S., Excel, S-Plus*, Εκδόσεις Ζήτη.

Χατζηνικολάου, Δ. (2002). *Στατιστική για οικονομολόγους*, Β Έκδοση.

ΞΕΝΟΓΛΩΣΣΗ

Barnett, V. and Lewis, T. (1978). *Outliers in statistical data*. Wiley, New York.

Belsley, David A, Kuh, Edwin and Welsch, Roy E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley Series in Probability and Mathematical Statistics. *John Wiley & Sons.*

Box, G. E. P. and Cox, D. R. (1964). *An Analysis of Transformations*, Journal of the Royal Statistical Society, 211-243, discussion 244-252.

Chatterjee, S. and Price, B. (1977). *Regression analysis by examples*. John Wiley & Sons, Inc.

Coakes, S. and Steed, L (1999). *S.P.S.S. Analysis without Anguish*. Wiley.

Field, A. P. (2005). *Discovering statistics using S.P.S.S. (Second Edition)*. London: Sage.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics***, 1,** 799--821.

Landau, S. and Everitt (2004). *A Handbook of Statistical Analyses using S.P.S.S.*. Chapman and Hall.

Neter, J., Kutner, M., Nachtsheim, C. and Wassserman, W. (1996). *Applied linear statistical models*. 4th Edition, Irwin, Inc.

Norousis, M. (2002). *S.P.S.S. 11.0 Guide to Data Analysis*. Prentice Hall, Inc.

Rawlings, J. O. (1988). *Applied regression analysis: a research tool*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Rencher, A. C. (2000). *Linear Models in Statistics*. Wiley.

Searle, S. R. (1971). *Linear models*. John Wiley & Sons, Inc.

Seber, G. A. F. (1977). Linear regression analysis. John Wiley & Sons, Inc.