ΣΤΑΤΙΣΤΙΚΗ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ ΠΑΚΕΤΟΥ IBM SPSS 22

STATISTICAL PACKAGE for the SOCIAL SCIENCES

ΤΣΑΓΡΗΣ ΜΙΧΑΗΛ

Email: mtsagris@yahoo.gr

ΑΘΗΝΑ και Nottingham
Μάρτιος 2014

Περιεχόμενα

Ένας μικρός πρόλογος 5
1.1 Σύντομη εισαγωγή στη Στατιστική 7
1.2 Σύντομη αναφορά στις δειγματοληπτικές τεχνικές 8
1.3 Σύντομη αναφορά στα είδη των ερευνών 9
2.1 Ανοίγοντας το IBM SPSS 22 10
2.2 Τα παράθυρα του SPSS και οι επιλογές τους 11
2.3 Η εντολή Select Cases 14
2.4 Η επιλογή Transform 16
2.5 Το μενού της επιλογής Analyze 21
3. Το Bootstrap στο IBM SPSS 22 25
3.1 Μία σύντομη εισαγωγή 25
3.2 Σύντομη περιγραφή του αλγόριθμου 25
4.1 Περιγραφικά μέτρα για συνεχείς μεταβλητές 27
4.2 Περιγραφικά μέτρα για κατηγορικές μεταβλητές 34
4.3 Ιστογράμματα 35
4.4 Κυκλικά διαγράμματα 41
4.5 Ραβδογράμματα 45
5.1 Έλεγχος κανονικότητας 49
5.2 Διαστήματα εμπιστοσύνης 52
5.3 Συντελεστές γραμμικής συσχέτισης 53
5.4 X^2 Έλεγχος ανεξαρτησίας για κατηγορικές μεταβλητές 57
5.5 Relative Risk και Odds ratio 61
5.6 Έλεγχος αλλαγής κατάστασης μιας δίτιμης κατηγορικής μεταβλητής 64
5.7 Αξιοπιστία ή βαθμός ταύτισης δύο κατηγορικών μεταβλητών (κάππα του Cohen) 66
5.8 Αξιοπιστία ενός ερωτηματολογίου 67
5.9 Καμπύλη χαρακτηριστικού λειτουργικού δέκτη (ROC curve) 69
5.10 Έλεγχοι υποθέσεων για το μέσο και τη διάμεσο ενός δείγματος (έλεγχος t και έλεγχος του Wilcoxon) 72
5.11 Έλεγχοι υποθέσεων για τη διαφορά των μέσων δύο ανεξάρτητων δειγμάτων (έλεγχος t και έλεγχος των Mann-Whitney-Wilcoxon) 77
5.12 Έλεγχοι υποθέσεων για τη διαφορά των μέσων δύο εξαρτημένων δειγμάτων (έλεγχος t και έλεγχος Wilcoxon για δείγμα ζευγών παρατηρήσεων) 80
6.1 Μια μικρή εισαγωγή στη γραμμική παλινδρόμηση 84
6.2 Διαγράμματα διασποράς 84
6.3 Απλή γραμμική παλινδρόμηση 87
6.4 Πολλαπλή γραμμική παλινδρόμηση 93
6.5 Παραβίαση των υποθέσεων στη γραμμική παλινδρόμηση 96
6.7 Μέθοδοι πολλαπλής παλινδρόμησης 98
6.8 Πολλαπλή γραμμική παλινδρόμηση με κατηγορική(ές) μεταβλητή(ές) 99
7.1 Ανάλυση διακύμανσης κατά ένα παράγοντα (One-way ANOVA) 109
7.2 Ανάλυση διακύμανσης με τη μέθοδο του Welch και των Brown-Forsythe .. 115
7.3 Μη παραμετρική ανάλυση διακύμανσης (έλεγχος των Kruskal-Wallis) 117
7.4 Ανάλυση διακύμανσης κατά δύο παράγοντες (Two-way ANOVA) 119
7.6 Ανάλυση διακύμανσης για εξαρτημένα δείγματα 122
7.7 Μη παραμετρική ανάλυση διακύμανσης για εξαρτημένα δείγματα 125
8.1 Προχωρημένη απλή παλινδρόμηση (μία ανεξάρτητη μεταβλητή) 127
8.2 Λογιστική παλινδρόμηση για δίτιμη εξαρτημένη μεταβλητή 129

8.3 Διαχωριστική ανάλυση 137
Βιβλιογραφία 144
Λεξικό στατιστικών όρων 147

Ένας μικρός πρόλογος

Οι σημειώσεις αυτές γράφτηκαν κυρίως για τους μη στατιστικούς που ασχολούνται με τη στατιστική και ειδικότερα την εφαρμοσμένη στατιστική. Το επίκεντρο ήταν η χρήση του SPSS για την υλοποίηση στατιστικών αναλύσεων. Η θεωρία που κρύβεται από πίσω είναι το υπόκεντρο. Πιστεύω όμως ότι ο αναγνώστης θα είναι σε θέση να εκτελέσει μία βασική στατιστική ανάλυση την οποία στη συνέχεια μπορεί να επεκτείνει με λίγο περισσότερο ψάξιμο.

Δεν θέλω να πω πολλά εδώ, εκτός από το ότι αυτή η έκδοση είναι μία αναθεωρημένη έκδοση της αρχικής του Απριλίου 2008 (για το SPSS 15) στην οποία έχουν διορθωθεί πολλά λάθη. Θα ήμουν υπόχρεος για οποιαδήποτε υπόδειξη σφάλματος στο παρόν κείμενο ώστε να βελτιωθεί ακόμα πιο πολύ.

Η δεύτερη έκδοση ήταν τον Ιανουάριο του 2010 και αναφερόταν πάλι στο SPSS 15 στην οποία όμως είχαμε προσθέσει και το αγγλοελληνικό λεξικό στατιστικών όρων. Η τρίτη έκδοση (πάλι για το SPSS 15) ήταν το Νοέμβριο του 2010. Στην τέταρτη έκδοση (Ιούλιος 2011, για το IBM SPSS 19) των σημειώσεων αναφερθήκαμε και στο bootstrap, μία σχετικά νέα τεχνική στη στατιστική που έχει φανεί πολύ χρήσιμη. Στις κοινωνικές επιστήμες πιστεύω θα αργήσει να διδαχτεί εκτός και αν σε κάποια τμήματα διδάσκεται το SPSS μόνο.

Το IBM SPSS 19 εκτός από το bootstrap έχει αλλάξει και το περιβάλλον στα μη παραμετρικά. Εμείς εδώ θα δούμε τις κλασσικές επιλογές για τη μη παραμετρική στατιστική, αλλά ο αναγνώστης αν θέλει μπορεί να δει και την άλλη επιλογή για τις ίδιες αναλύσεις. Οδηγούν στο ίδιο τέρμα, απλά πάνε από άλλο δρόμο και ίσως πιο χρονοβόρο.

Στην πέμπτη έκδοση των σημειώσεων (για το IBM SPSS 19, Ιανουάριος 2013) είχαμε προσθέσει τον έλεγχο του McNemar για την αλλαγή κατάστασης μιας δίτιμης μεταβλητής καθώς και την καμπύλη λειτουργικού χαρακτηριστικού δέκτη (ROC curve).

Η IBM έβγαλε καινούριες εκδόσεις και φτάσαμε πια στην 22η. Έχουν αλλάξει κάποια γραφικά (έχουμε και Windows 7 πια, οπότε παίζουν και αυτό το ρόλο τους στα γραφικά) και κάποια μενού μέσα, αλλά όλα θα τα δούμε στην πορεία και όπως θα δούμε είναι μικρές οι αλλαγές. Προσθέσαμε ένα κεφάλαιο ακόμα και κάποιες παραγράφους στα ήδη υπάρχοντα κεφάλαια. Αλλάξαμε επίσης και τη γραμματική μας σε κάποια σημεία. Ελπίζουμε λοιπόν η έκτη έκδοση των σημειώσεων να έχει αμελητέα λάθη.

Για περισσότερες σημειώσεις στατιστικής (και για το SPSS) ο αναγνώστης καλείται να δει το statlink.tripod.com .

Τσαγρής Μιχαήλ

Μάρτιος 2014

1.1 Σύντομη εισαγωγή στη Στατιστική

Υπάρχουν δύο ενδεχόμενα για την προέλευση του όρου "στατιστική". Εικάζεται ότι προέρχεται από την αρχαία ελληνική λέξη «στατίζω» που σημαίνει τοποθετώ, ταξινομώ, συμπεραίνω. Το άλλο ενδεχόμενο είναι ότι προέρχεται από τη λατινική λέξη «status» που σημαίνει πολιτεία, κράτος. Η λέξη αυτή αρχικά χρησιμοποιήθηκε για δεδομένα που αφορούν στον πληθυσμό μίας χώρας. Έρευνες έχουν δείξει ότι η αρχαιότερη συλλογή στοιχείων έγινε στην Κίνα το 2238 π.Χ. Πρόκειται για μία απογραφή του πληθυσμού η οποία διεξήχθη υπό την αυτοκρατορία του Yao. Υπάρχουν ενδείξεις ότι απογραφές είχαν διενεργήσει και άλλοι λαοί κατά τη αρχαιότητα, όπως οι Αιγύπτιοι και οι Πέρσες. Γνωρίζουμε επίσης ότι απογραφή διεξήχθη και κατά την περίοδο γέννησης του Χριστού από τον καίσαρα Αύγουστο. Ο Σωκράτης αναφέρει τον όρο "στατιστική" στο έργο του «Ξενοφώντος απομνημονεύματα» και ο Σωκράτης στο έργο του «Πολιτεία». Η στατιστική με την πάροδο των χρόνων αναπτύχθηκε δειλά στην αρχή και ραγδαία κατά τα τέλη του 19^{ου} αιώνα και μετά για να φτάσει στη στο σήμερα.

Η στατιστική είναι η επιστήμη που ασχολείται με τη συλλογή δεδομένων, την περιγραφή τους και την εξαγωγή τεκμηριωμένων αποτελεσμάτων με τη χρήση επιστημονικά αποδεκτών τεχνικών. Αν θέλαμε να δώσουμε έναν άλλο ορισμό στον όρο "στατιστική" θα επιλέγαμε αυτόν που έδωσε ο πατέρας της σύγχρονης στατιστικής Ronald Fisher (1890-1962):

Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- Το σχεδιασμό της διαδικασίας συλλογής δεδομένων
- Τη συνοπτική και αποτελεσματική παρουσίαση τους
- Την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων.

Οι βασικές μορφές της στατιστικής είναι αυτές της περιγραφικής στατιστικής και της επαγωγικής στατιστικής. Η μεν πρώτη ασχολείται με την περιγραφή των δεδομένων του δείγματος και η δεύτερη με τη εξαγωγή χρήσιμων συμπερασμάτων για τον πληθυσμό.

Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε έναν πληθυσμό καλούνται μεταβλητές. Οι δυνατές τιμές που μπορεί να πάρει μία μεταβλητή ονομάζονται τιμές της μεταβλητής. Οι μεταβλητές με τη σειρά τους διακρίνονται σε ποιοτικές ή κατηγορικές και ποσοτικές. Οι ποιοτικές μεταβλητές μπορεί να είναι είτε ονομαστικού τύπου στις οποίες οι τιμές αναφέρονται μόνο σε κατηγορίες, π.χ. ομάδα αίματος είτε διατακτικού τύπου στις οποίες οι συγκρίσεις της μορφής «μεγαλύτερη», «μικρότερη» ή «ίση» έχουν νόημα, π.χ. απάντηση σε ερωτηματολόγιο ικανοποίησης. Οι ποσοτικές μεταβλητές μπορεί να είναι είτε συνεχείς είτε διακριτές. Συνεχείς είναι οι μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή σε ένα διάστημα τιμών. Διακριτές είναι οι μεταβλητές που μπορούν να πάρουν διακριτές (μεμονωμένες) τιμές. Επιπλέον οι ποσοτικές μεταβλητές μπορούν να διαχωριστούν σε άλλες δύο κατηγορίες ανάλογα με τον τρόπο μέτρησης τους. Όταν εκτός από την προφανή διάταξη των τιμών τους έχει νόημα και η μεταξύ τους απόσταση, όπως π.χ. στα έτη ζωής, τότε μιλάμε για ποσοτικές μεταβλητές που μετριόνται σε κλίμακα διαστήματος. Αν εκτός από την διάταξη και το μέγεθος του διαστήματος μεταξύ των τιμών έχει έννοια και ο λόγος των τιμών τότε μιλάμε για ποσοτικές μεταβλητές που μετριόνται σε κλίμακα λόγου. Οι τιμές για παράδειγμα στα προϊόντα μετριόνται σε κλίμακα λόγου. Έχει νόημα να πούμε ότι η τιμή ενός προϊόντος σε μία λαϊκή αγορά είναι κατά

ένα ποσοστό μεγαλύτερη ή μικρότερη από την τιμή του ίδιου προϊόντος σε μία άλλη λαϊκή αγορά.

1.2 Σύντομη αναφορά στις δειγματοληπτικές τεχνικές

Θα αναφερθούμε τώρα στη δειγματοληψία, τη διαδικασία δηλαδή συλλογής δεδομένων από έναν πληθυσμό. Το σύνολο των δεδομένων που θα συλλεχθούν, το οποίο θα είναι προφανώς ένα υποσύνολο του πληθυσμού, ονομάζεται δείγμα. Το δείγμα θα καλείται αντιπροσωπευτικό όταν όλα τα στοιχεία του πληθυσμού έχουν την ίδια πιθανότητα επιλογής. Υπάρχουν διάφορες τεχνικές με τις οποίες μπορούμε να αντλήσουμε δεδομένα από έναν πληθυσμό. Οι κύριες δειγματοληπτικές τεχνικές είναι τέσσερις.

- Η απλή τυχαία δειγματοληψία είναι η πιο απλή περίπτωση. Επιλέγουμε τυχαία, στοιχεία (ή μονάδες) από το σύνολο του πληθυσμού.
- Η στρωματοποιημένη δειγματοληψία είναι η περίπτωση κατά την οποία χωρίζουμε τον πληθυσμό σε στρώματα και μετά επιλέγουμε τυχαία τα στοιχεία από κάθε στρώμα.
- Η δειγματοληψία κατά ομάδες είναι μία τεχνική δειγματοληψίας στην οποία χωρίζουμε τον πληθυσμό σε πολλές ομάδες (όχι στρώματα), όπου η κάθε ομάδα περιέχει ένα πλήθος στοιχείων. Επιλέγουμε τυχαία ομάδες από το σύνολο των ομάδων και συμπεριλαμβάνουμε στο δείγμα όλα τις μονάδες των επιλεγμένων ομάδων.
- Η τέταρτη περίπτωση είναι η συστηματική δειγματοληψία. Ας υποθέσουμε ότι έχουμε στα χέρια μας ένα μακρύ κατάλογο με τα στοιχεία του πληθυσμού αριθμημένα. Τότε μπορούμε να εφαρμόσουμε το εξής: διαλέγουμε ένα στοιχείο στην αρχή του καταλόγου, έστω το στοιχείο που βρίσκεται στην $4^{η}$ γραμμή του καταλόγου. Τα επόμενα θα επιλεγούν με ένα βήμα όπως για παράδειγμα το 10. Δηλαδή, θα επιλέξουμε το $14^{ο}$ στοιχείο, το $24^{ο}$ στοιχείο και ούτω καθεξής.

Όπως είναι φυσικό οι διάφορες τεχνικές μπορούν να συνδυαστούν και να προκύψουν πιο σύνθετα δειγματοληπτικά σχήματα. Οι περιπτώσεις που αναφέραμε εδώ είναι οι πιο απλές και συνάμα δημοφιλείς τεχνικές δειγματοληψίας. Σε αυτό το σημείο πρέπει να κάνουμε αναφορά στα είδη των πληθυσμών προς αποφυγή λανθασμένων συμπερασμάτων, τον αντικειμενικό πληθυσμό (target population), τον υπό μελέτη πληθυσμό (study population) και το δειγματοληπτικό πλαίσιο (sampling frame) και στην έννοια του δειγματοληπτικού σφάλματος.

Ο αντικειμενικός πληθυσμός είναι το σύνολο των ατόμων ή στοιχείων των οποίων ένα ή περισσότερα χαρακτηριστικά θέλουμε να εξετάσουμε. Ο υπό μελέτη πληθυσμός είναι υποσύνολο συνήθως του αντικειμενικού πληθυσμού (μπορεί και να ταυτίζεται). Για παράδειγμα ο αντικειμενικός πληθυσμός μίας μελέτης θα μπορούσε να είναι το σύνολο των ελλήνων μαθητών του δημοτικού σχολείου, ενώ ο υπό μελέτη πληθυσμός να αφορά μόνο στους μαθητές της Αττικής. Το δειγματοληπτικό πλαίσιο είναι το σύνολο των ατόμων (ή στοιχείων) που έχουν πραγματικά δυνατότητα επιλογής στο δείγμα (η πηγή του δείγματος). Στο ίδιο παράδειγμα, λόγω κόστους οι μαθητές κάποιων απομακρυσμένων περιοχών της Αττικής δεν έχουν τη δυνατότητα να συμπεριληφθούν στο δείγμα. Το πλαίσιο δηλαδή στην ουσία αποτελεί μία υποδιαίρεση του υπό μελέτη πληθυσμού, μπορεί όμως και να ταυτίζεται με αυτόν.

Το δειγματοληπτικό σφάλμα είναι η διαφορά ανάμεσα στα αποτελέσματα μίας δειγματοληψίας και μίας απογραφής (100% δείγμα). Όλοι έχουμε ακούσει αυτόν τον ορισμό κατά την περίοδο των εκλογών. Είναι το περιθώριο σφάλματος που

αναφέρουν οι εταιρείες δημοσκοπήσεων όταν κάνουν πρόβλεψη για το ποσοστό που θα "πάρει" ένα κόμμα, το "συν πλην 3 ποσοστιαίες μονάδες".

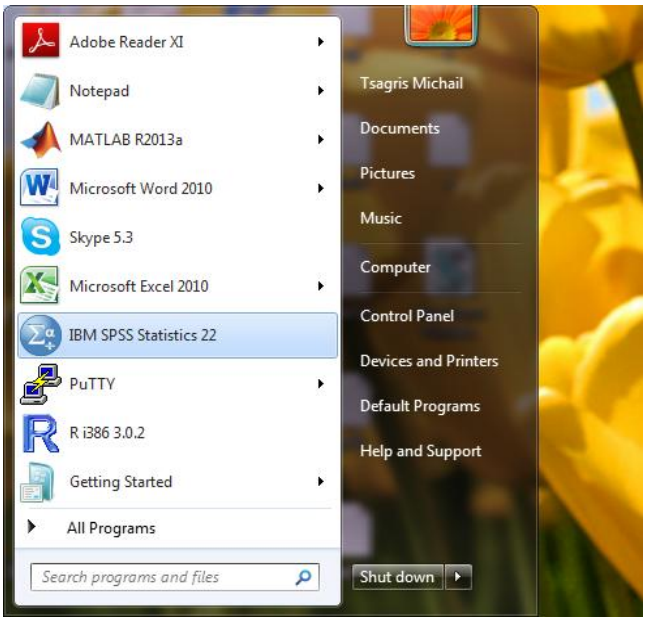
1.3 Σύντομη αναφορά στα είδη των ερευνών

Οι ιατρικές μελέτες χωρίζονται σε δύο κατηγορίες, η μία, είναι οι πειραματικές μελέτες στις οποίες ο ερευνητής παρεμβαίνει ενεργητικά στον τρόπο σχεδιασμού του πειράματος, του καθορισμού των ομάδων κ.ά. Χαρακτηριστικό παράδειγμα τέτοιων μελετών είναι οι κλινικές δοκιμές. Η άλλη κατηγορία είναι οι μελέτες παρατήρησης ή μη πειραματικές μελέτες στις οποίες ο ερευνητής δεν παρεμβαίνει αλλά απλά παρατηρεί και καταγράφει. Σε αυτήν την κατηγορία περιλαμβάνονται οι διατμηματικές μελέτες, οι προοπτικές μελέτες (διαμήκεις-διαχρονικές, μελέτες κοορτής, μελέτες παρακολούθησης ή follow-up studies είναι εναλλακτικές ονομασίες αυτών των ερευνών) και οι αναδρομικές μελέτες ή μελέτες μαρτύρων-ασθενών (retrospective or case-control studies). Στις διατμηματικές μελέτες τα στοιχεία του δείγματος κατηγοριοποιούνται με βάση την έκθεση τους σε μία νόσο, χωρίς να λαμβάνουμε υπόψη τον παράγοντα χρόνο. Οι προοπτικές μελέτες λαμβάνουν υπόψη τους τον παράγοντα χρόνο. Το βασικό χαρακτηριστικό αυτών των μελετών είναι ότι επιλέγουμε μία ομάδα ανθρώπων της οποίας την εξέλιξη παρατηρούμε στο πέρασμα του χρόνου. Υπάρχουν έρευνες που διήρκησαν πολλά χρόνια. Χαρακτηριστικό παράδειγμα η έρευνα θνησιμότητας των βρετανών ιατρών των Doll & Hill στην οποία συμμετείχαν 40637 ιατροί και είχε διάρκεια 10 ετών (δημοσιεύτηκε στο British Medical Journal το 1964). Τέλος οι αναδρομικές μελέτες είναι αυτές που κάνουν την αντίθετη κατά κάποιο τρόπο διαδικασία από τις προοπτικές μελέτες. Σε αυτές τις μελέτες κοιτάζουμε πίσω χρονικά και κατηγοριοποιούμε τα στοιχεία ανάλογα με κάποιο χαρακτηριστικό

2.1 Ανοίγοντας το IBM SPSS 22

Έχοντας εγκατεστημένο το στατιστικό πακέτο SPSS, μπορούμε να προχωρήσουμε στη χρήση του για τη διεξαγωγή στατιστικών αναλύσεων ή και μόνο συνοπτική παρουσίαση στατιστικών στοιχείων που μας αφορούν. Για να ανοίξουμε το SPSS κάνουμε διπλό κλικ πάνω στο εικονίδιο που βρίσκεται στην επιφάνεια εργασίας (αν υπάρχει) ειδάλλως πηγαίνουμε στο μενού εργασιών των windows.

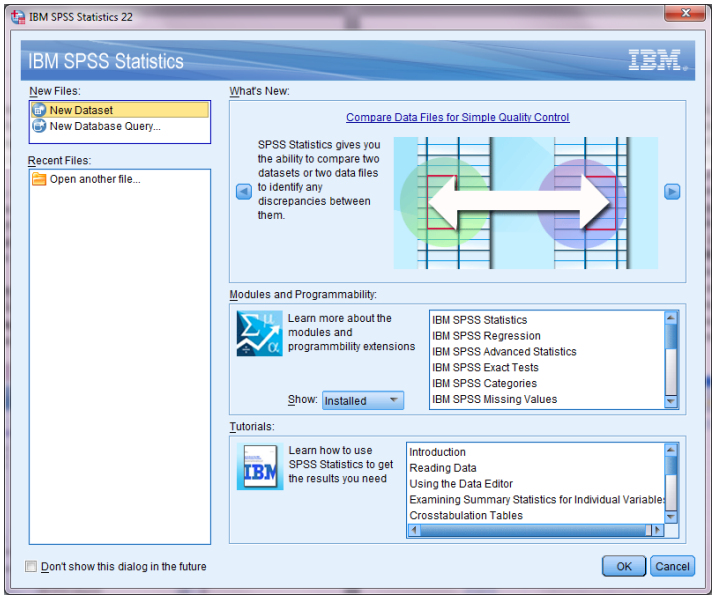
Επιλέγουμε **all programs→SPSS(PAW)→IBM Statistics SPSS 22** (μπορεί να είναι διαφορετική σε άλλους η σειρά ή και η ονομασία) και μετά αριστερό κλικ, όπως φαίνεται και από την παρακάτω εικόνα. Στην εικόνα 1 βέβαια έχει μπει ήδη στη λίστα συχνής εκκίνησης.



Εικόνα 1

Μόλις το πακέτο φορτώσει θα μας εμφανιστεί μία οθόνη γεμάτη κελιά, ένα κενό φύλο εργασίας (**IBM SPSS Data Editor**), όπως στην περίπτωση του MS Excel, αλλά και το πλαίσιο διαλόγου της εικόνας 2. Στο παράθυρο που εμφανίζεται υπάρχει ένα ερώτημα σχετικά με το τι θέλουμε το SPSS να κάνει.

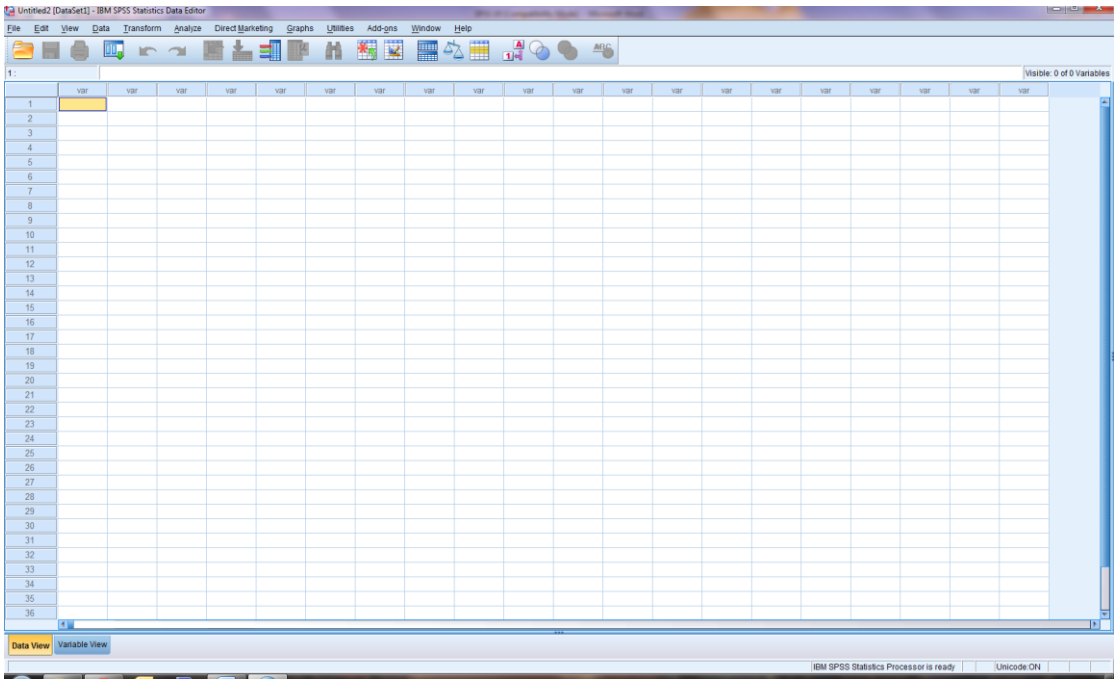
- Εμείς θα επιλέξουμε **New dataset** επάνω δεξιά **και μετά** OK.
- Αν επιλέξουμε **Don't show this dialogue in the future** τότε απενεργοποιούμε αυτό το πλαίσιο διαλόγου κατά τις επόμενες εκκινήσεις του SPSS.
- Τι κάνουν οι άλλες επιλογές για να πω την αλήθεια δεν ξέρω ακριβώς (ίσως ξέρω λίγο), οπότε ο αναγνώστης καλείται να παίξει λίγο αν θέλει.



Εικόνα 2

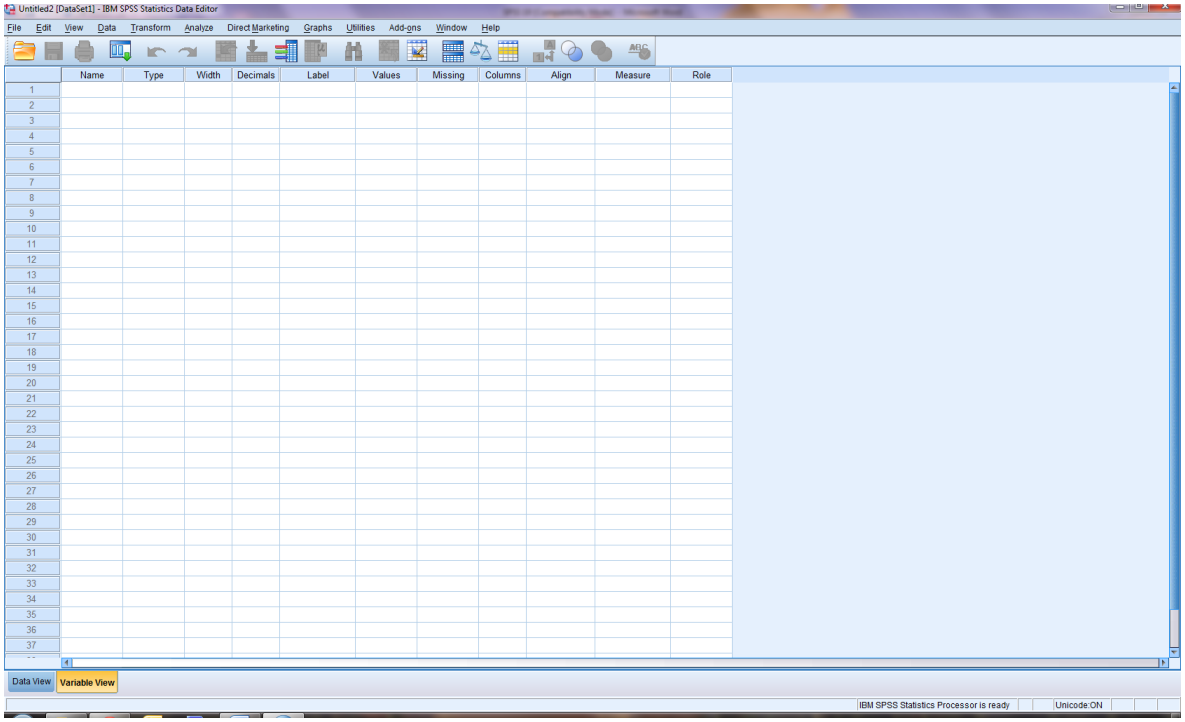
2.2 Τα παράθυρα του SPSS και οι επιλογές τους

Το παράθυρο IBM SPSS Data Editor είναι αυτό που φαίνεται στην παρακάτω εικόνα. Η γραμμή τίτλου είναι η μπλε γραμμή που φαίνεται στο πάνω μέρος του παραθύρου. Το μενού επιλογών είναι παρόμοιο με αυτό που συναντάται στο MS Office. Είναι η σειρά που φαίνεται κάτω από τη γραμμή τίτλου και περιλαμβάνει τις εξής επιλογές του παραθύρου: **File, Edit, View, Data, Transform, Analyze, Direct Marketing, Graphs, Utilities, Add-ons**, **Window** και **Help**. Η γραμμή εργαλείων βρίσκεται κάτω από το μενού επιλογών και αποτελείται από εικονίδια χρήσιμα για λειτουργίες που χρησιμοποιούνται συχνά, όπως αποθήκευση, εκτύπωση, άνοιγμα κάποιου αρχείου. Οι γραμμές κύλισης βρίσκονται στα δεξιά και στο κάτω μέρος του παραθύρου και μας βοηθάνε να μετακινηθούμε πάνω-κάτω και δεξιά-αριστερά. Στο κάτω μέρος του παραθύρου (δεξιά) εμφανίζεται ένα μήνυμα που λέει **IBM SPSS Statistics Processor is ready**. Η γραμμή αυτή στην οποία εμφανίζεται αυτό το μήνυμα είναι η γραμμή κατάστασης. Όταν το SPSS διεξάγει κάποιον υπολογισμό, έχει μία διεργασία σε εξέλιξη, ή τερματίσει μία οποιαδήποτε διεργασία θα εμφανίζεται το αντίστοιχο μήνυμα. Στο πάνω μέρος των κελιών εμφανίζεται το όνομα των κελιών. Από τη στιγμή που δεν έχουμε δώσει ονόματα στα κελιά απλά εμφανίζετε η λέξη **VAR00001**, **VAR00002** και ούτω καθεξής. Επίσης στο αριστερό κάτω μέρος του παραθύρου υπάρχουν δύο επιλογές. Η μία είναι η **Data View** και η άλλη είναι η **Variable View**. Αυτή τη στιγμή είναι επιλεγμένη η πρώτη επιλογή. Δηλαδή το παράθυρο που μπορούμε να περάσουμε δεδομένα είναι αυτό που φαίνεται στην οθόνη. Τα δεδομένα τα περνάμε κατά τον ίδιο τρόπο με το Excel, δηλαδή κάθετα αν πρόκειται για τιμές της ίδιας μεταβλητής. Πληκτρολογούμε τον αριθμό και πατάμε **Enter**.



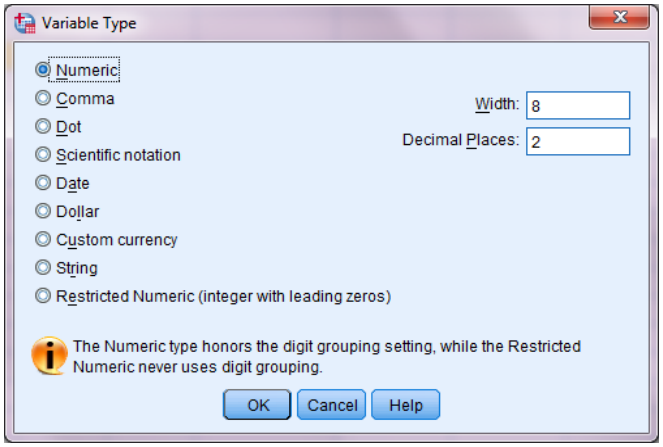
Εικόνα 3

Αν επιλέξουμε τη δεύτερη επιλογή (**Variable View**) θα εμφανιστεί το παράθυρο της εικόνας 4. Η πρώτη στήλη έχει τίτλο **Name**. Στα κελιά της πρώτης στήλης δίνουμε τα ονόματα των στηλών των δεδομένων που βρίσκονται στο Data Editor. Έτσι, στο πρώτο κελί αντιστοιχεί το όνομα της πρώτης στήλης των δεδομένων, στο δεύτερο αντιστοιχεί το όνομα της δεύτερης στήλης και ούτω καθεξής.



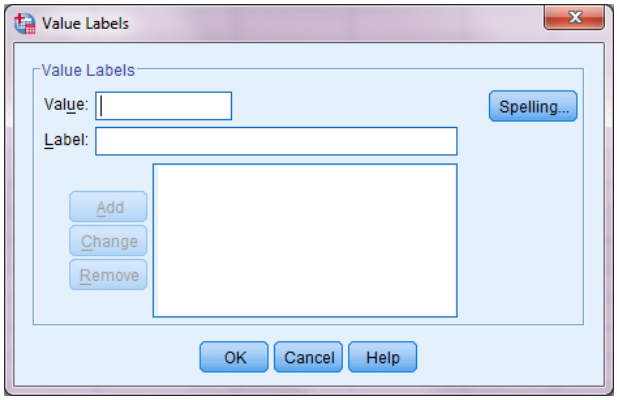
Εικόνα 4

Αν πάμε σε ένα κελί της δεύτερη στήλης (**Type**) και πατήσουμε στο δεξί μέρος του κελιού (το εικονιδιάκι με τις τρεις τελείες) το τότε θα εμφανιστεί το παράθυρο της εικόνας 5. Αυτό μας δίνει τη δυνατότητα να επιλέξουμε τον τύπο των δεδομένων για κάθε στήλη του Data Editor. Ξανά το πρώτο κελί αντιστοιχεί στα δεδομένα της πρώτης στήλης του Data Editor, το δεύτερο κελί στα δεδομένα της δεύτερης στήλης και ούτω καθεξής. Η επιλογή **Numeric** είναι προεπιλεγμένη από το πακέτο διότι τα δεδομένα μας είναι τις πιο πολλές φορές αριθμητικά. Αν επιλέξουμε **String** τότε τα δεδομένα μας θα είναι σε μορφή χαρακτήρα ή απλά θα είναι γράμματα. Με το **Width** ορίζουμε το μέγιστο πλήθος των ψηφίων που θα έχουν τα αριθμητικά δεδομένα και με το **Decimal Places** ορίζουμε το πλήθος των ψηφίων που θα βρίσκονται δεξιά της υποδιαστολής. Το τρίτο και το τέταρτο κελί του παραθύρου της εικόνας 4 αναφέρονται στο πλήθος των ψηφίων αριστερά και δεξιά της υποδιαστολής, όπως ήδη αναφέραμε. Το κελί με τίτλο **Label** είναι η ετικέτα των δεδομένων. Αν δηλώσουμε ονόματα στα δεδομένα μας στους πίνακες που θα εμφανίζονται με τα αποτελέσματα από το SPSS θα εμφανίζονται οι στήλες των δεδομένων με τα ονόματα τους. Αν όμως δηλώσουμε και ετικέτες ή μόνο ετικέτες στις στήλες των δεδομένων θα εμφανίζονται τα αποτελέσματα όπου κάθε στήλη δεδομένων θα έχει για όνομα αυτό που έχουμε ορίσει στις ετικέτες των στηλών. Το τι θα εμφανίζεται μπορεί να επιλεγεί από το χρήστη από την υπό-επιλογή **Options** που βρίσκεται μέσα στην επιλογή **Edit** την οποία θα δούμε παρακάτω.



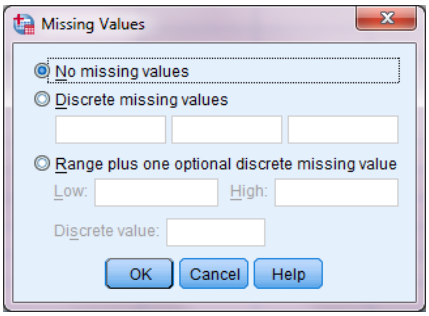
Εικόνα 5

Το κελί **Values** (πάλι κλικ πάνω στο εικονιδιάκι με τις τρεις τελείες δεξιά ενός κελιού αυτής της στήλης) οδηγεί στο παράθυρο της εικόνας 6. Έστω ότι έχουμε ποιοτικά δεδομένα και θέλουμε να τα περάσουμε στο SPSS. Για να γίνει αυτό κωδικοποιούμε εξαρχής τα δεδομένα δίνοντας τους τιμές για κάθε μία κατηγορία. Για παράδειγμα θέλουμε να καταχωρήσουμε σε μία στήλη του SPSS Data Editor το φύλο κάποιων ανθρώπων για τους οποίος συλλέξαμε κάποια στοιχεία. Η πιο συνήθης κωδικοποίηση είναι της μορφής 0 για τους άντρες και 1 για τις γυναίκες. Για να θυμόμαστε λοιπόν που αντιστοιχεί ο κάθε αριθμός θα εκχωρήσουμε και το φύλο στο SPSS. Γράφουμε 0 στο πρώτο λευκό τετραγωνάκι που αντιστοιχεί στο **Value.** Στο **Value Label** θα πληκτρολογήσουμε το φύλο δηλαδή *man.* Στη συνέχεια πατάμε **Add** ώστε να καταχωρηθεί το φύλο στον αριθμό. Το ίδιο θα κάνουμε και για τις γυναίκες. Η διαδικασία θα επαναληφθεί όσες φορές χρειαστεί. Για παράδειγμα σε ένα ερωτηματολόγιο που υπάρχουν περισσότερες από δύο απαντήσεις θα επαναλάβουμε τη διαδικασία τόσες φορές όσες είναι και οι απαντήσεις.



Εικόνα 6

Στο κελί **Missing** (πάλι κλικ πάνω στο εικονιδιάκι με τις τρεις τελείες δεξιά ενός κελιού αυτής της στήλης) θα ορίσουμε τις χαμένες παρατηρήσεις. Για παράδειγμα κάποιοι εκ των ερωτηθέντων σε ένα ερωτηματολόγιο δεν έχουν απαντήσει σε όλες τις ερωτήσεις. Το παράθυρο που θα εμφανιστεί είναι αυτό που βρίσκεται στην εικόνα 7. Πρέπει να προσέξουμε πως θα ορίσουμε τις χαμένες τιμές. Για παράδειγμα σε ένα ερωτηματολόγιο που οι απαντήσεις είναι σε κλίμακα Likert από 1 έως 5 τις στις χαμένες τιμές θα δώσουμε έναν αριθμό που δε βρίσκεται μεταξύ 1 και 5. Θα πρέπει να είναι δηλαδή ένας αριθμός που δε συναντάται στα δεδομένα της κάθε στήλης ξεχωριστά. Το κελιά **Columns** και **Align** αναφέρονται στο μέγεθος της στήλης και στη στοίχιση των δεδομένων στην κάθε στήλη. Τέλος, η τελευταία στήλη (**Measure**) αναφέρεται στον τύπο των δεδομένων. Αν τα δεδομένα αφορούν σε ποσοτικές μετρήσεις (**Scale**), διατεταγμένες (**Ordinal**) ή ονομαστικές (**Nominal**).



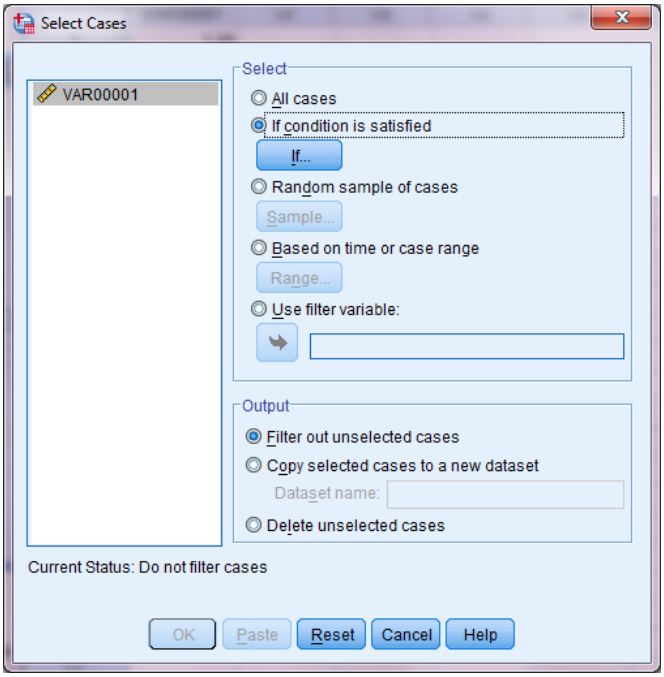
Εικόνα 7

2.3 Η εντολή Select Cases

Επιλέγοντας **Data→Select Cases…** θα εμφανιστεί το παράθυρο της εικόνας 8. Με αυτήν την επιλογή δίνεται η δυνατότητα στο χρήστη να επιλέξει ένα μέρος των δεδομένων, το οποίο θα χρησιμοποιηθεί στις αναλύσεις. Η επιλογή **All cases** είναι προεπιλεγμένη από το πακέτο. Αν επιλέξουμε τη δεύτερη επιλογή (**If condition is satisfied**) και μετά **If…** θα οδηγηθούμε στο παράθυρο της εικόνας 9.

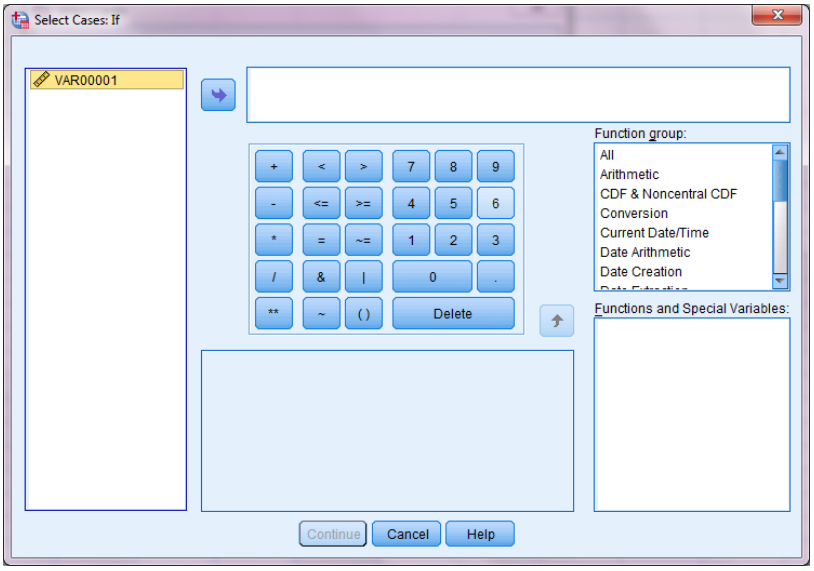
Αυτό που επιδιώκουμε να κάνουμε με αυτήν την επιλογή είναι να επιλέξουμε από μία ή περισσότερες στήλες δεδομένων κάποια δεδομένα που ικανοποιούν κάποια συνθήκη. Περνώντας την/τις μεταβλητή/τές από το αριστερό κουτάκι στο δεξιό κουτάκι του παραθύρου της εικόνας 9 δίνουμε στο SPSS να καταλάβει τις στήλες

δεδομένων, των οποίων τα δεδομένα θέλουμε να ικανοποιούν κάποια συνθήκη. Υπάρχει μία λίστα από συναρτήσεις λογικές και μη, τις οποίες μπορούμε να χρησιμοποιήσουμε. Για παράδειγμα οι τελεστές > , < , >= , <= δηλώνουν ανισοϊσότητες. Μπορούμε δηλαδή να επιλέξουμε τα δεδομένα μίας στήλης που να είναι μικρότερα ή μεγαλύτερα από μία τιμή. Με βάση τις μαθηματικές συναρτήσεις μπορούμε να ζητήσουμε από το SPSS να επιλέξει τα δεδομένα εκείνα για τα οποία η απόλυτη τους τιμή είναι μικρότερη από μία καθορισμένη τιμή.



Εικόνα 8

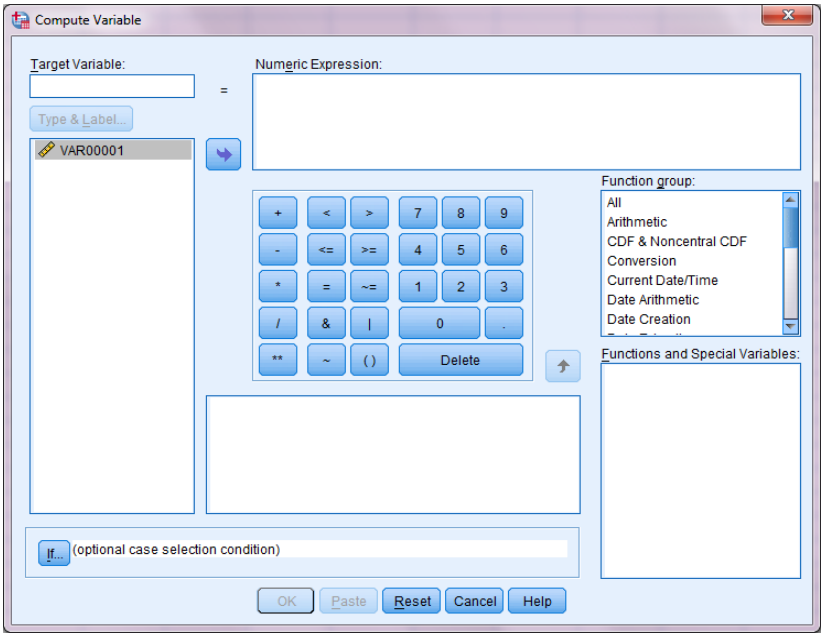
Η επιλογή **random sample of cases** που υπάρχει στο παράθυρο της εικόνας 8 μας δίνει τη δυνατότητα να επιλέξουμε τυχαία είτε ένα ποσοστό των δεδομένων, είτε ένα δείγμα από τα δεδομένα καθορίζοντας φυσικά το μέγεθος του δείγματος. Η επιλογή **Based on time or case range** μας δίνει τη δυνατότητα να επιλέξουμε δεδομένα τα οποία βρίσκονται μέσα σε κάποια περιοχή ή κάποια όρια. Ιδιαίτερη προσοχή θα πρέπει να δοθεί όσον αφορά τα δεδομένα τα οποία δεν επιλέγονται με βάση κάποια από τις προηγούμενες επιλογές. Η επιλογή **Filter out unselected cases** είναι προεπιλεγμένη από το πακέτο. Σε αυτήν την περίπτωση τα δεδομένα απλά δε θα υπολογίζονται στις επόμενες αναλύσεις. Θα εμφανιστεί μία νέα στήλη που θα περιέχει τιμές 0 και 1 ανάλογα με το αν τα δεδομένα ικανοποιούν ή όχι τη συνθήκη. Επίσης θα δούμε στη στήλη που περιέχει την αρίθμηση των γραμμών μία διαγώνιο γραμμή να έχει "διαγράψει" κατά κάποιο τρόπο τις γραμμές των δεδομένων που δεν ικανοποιούν τη συνθήκη. Αν επιλέξουμε **Delete unselected cases**, τότε τα δεδομένα θα διαγραφούν από το αρχείο. Αν όμως επιλέξουμε **Copy selected cases to a new dataset** τότε τα δεδομένα που ικανοποιούν τη συνθήκη θα αποθηκευτούν σε ένα νέο αρχείο δεδομένων του SPSS στο οποίο θα πρέπει να δώσουμε ένα όνομα πληκτρολογώντας το στο λευκό κουτάκι που θα ενεργοποιηθεί (**Dataset name**).



Εικόνα 9

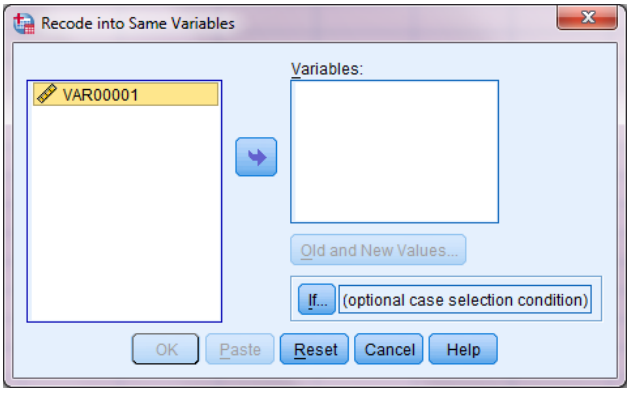
2.4 Η επιλογή Transform

Μία πολύ χρήσιμη επιλογή από το μενού επιλογών είναι αυτή της μετατροπής των δεδομένων (**Transform**). Η πρώτη εντολή που εμπεριέχεται στην επιλογή **Transform** είναι η **Compute Variable**. Επιλέγοντας αυτήν την εντολή θα εμφανιστεί το παράθυρο της εικόνας 10. Το λευκό κουτάκι που λέγεται **Target Variable** πρέπει να συμπληρωθεί με ένα όνομα. Εκεί θα αποθηκευτεί η μετασχηματισμένη στήλη δεδομένων. Η μετασχηματισμένη στήλη μπορεί είτε να αποθηκευτεί στην ίδια στήλη είτε σε διαφορετική. Περνώντας τις στήλες από το αριστερό κουτάκι στο κουτάκι που λέγεται **Numeric Expression** ορίζουμε τις στήλες οι οποίες θα μετασχηματιστούν. Το κουτάκι **Function group** περιέχει διάφορα είδη συναρτήσεων όπως μαθηματικές, στατιστικές, μετατροπής και άλλες. Για κάθε είδος συναρτήσεων που επιλέγουμε, στο κουτάκι που βρίσκεται ακριβώς από κάτω βλέπουμε τις διαθέσιμες συναρτήσεις. Αυτές είναι συναρτήσεις που μας βοηθάνε στο μετασχηματισμό των δεδομένων. Βέβαια, μπούμε να γράψουμε μία δική μας συνάρτηση μετατροπής η οποία δε βρίσκεται στη λίστα με τις ήδη υπάρχουσες συναρτήσεις.

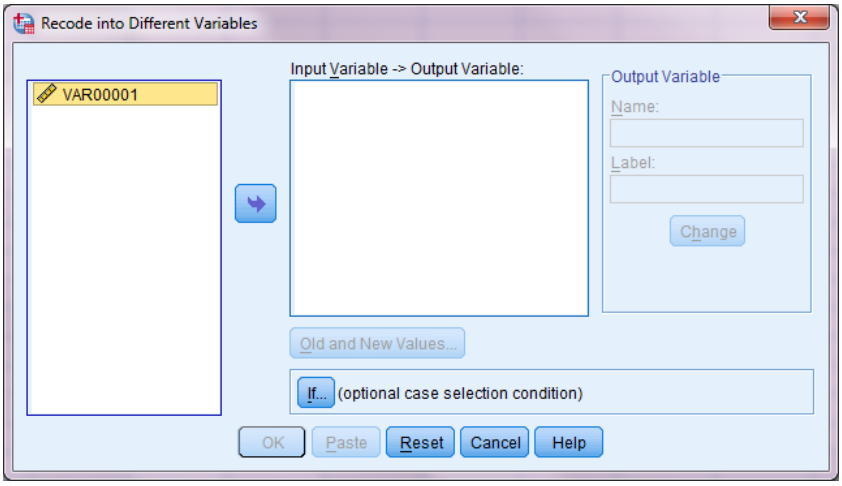


Εικόνα 10

Η εντολή επανακωδικοποίησης είναι εντολή κωδικοποίησης των ήδη υπαρχουσών στηλών δεδομένων. Ένα παράδειγμα χρησιμοποίησης αυτής της εντολής είναι το εξής: έστω ότι έχουμε συλλέξει ηλικίες ατόμων από 20 έως 90+ έτη. Αντί να δουλεύουμε με τις ηλικίες αυτές καθεαυτές θέλουμε να τις κατηγοριοποιήσουμε σε ομάδες ηλικιών, έστω 7 τον αριθμό. Μπορούμε να επιλέξουμε είτε να σώσουμε τις ομάδες ηλικιών στη στήλη των ήδη υπαρχουσών ηλικιών (οπότε θα χαθούν οι ηλικίες), είτε σε μία άλλη στήλη. Θα επιλέξουμε δηλαδή είτε **Recode into Same Variables**, είτε **Recode into Different Variables** αντίστοιχα. Αν επιλέξουμε να σώσουμε τη νέα στήλη των ομάδων ηλικιών στην ίδια στήλη των ηλικιών, διαγράφοντας ουσιαστικά τις ηλικίες θα εμφανιστεί το παράθυρο της εικόνας 11. Αν επιλέξουμε να αποθηκεύσουμε τη στήλη των ηλικιακών ομάδων σε άλλη στήλη θα εμφανιστεί το παράθυρο της εικόνας 12.



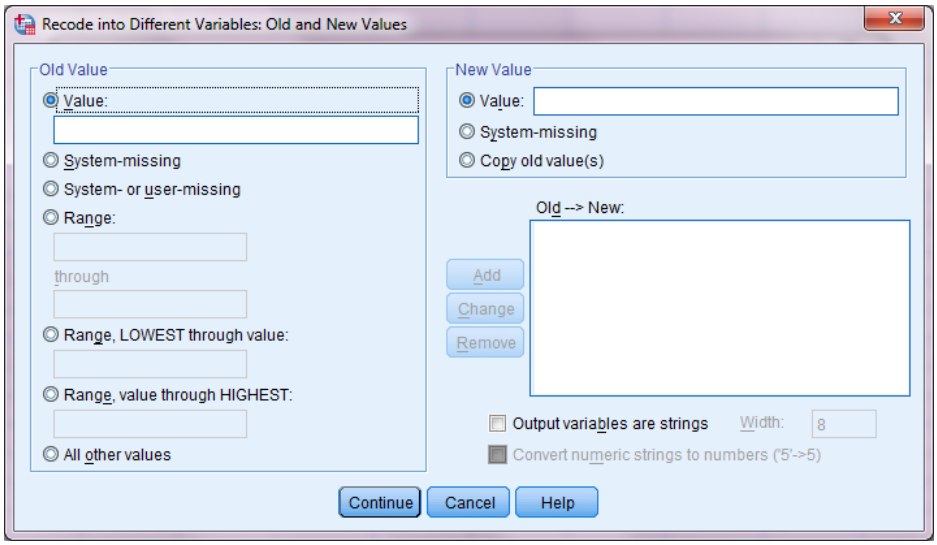
Εικόνα 11



Εικόνα 12

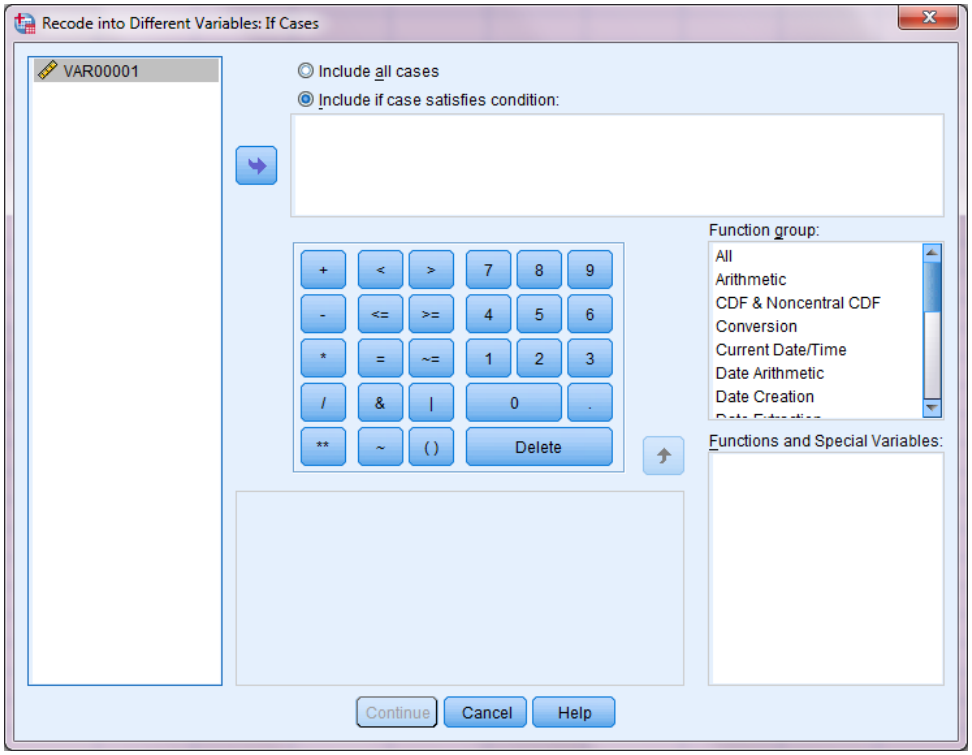
Και στις δύο περιπτώσεις θα πρέπει να περάσουμε τη στήλη των δεδομένων που θέλουμε να μετασχηματίσουμε από το αριστερό στο δεξιό λευκό κουτάκι. Μόλις το κάνουμε αυτό θα ενεργοποιηθεί η επιλογή **Old and New Values** που βρίσκεται κάτω από το δεξιό κουτάκι. Επιλέγοντας αυτήν την επιλογή εμφανίζεται το παράθυρο της εικόνας 13. Στην περιοχή **Old Value** που βρίσκεται στο αριστερό μέρος του παραθύρου θα επιλέξουμε να πληκτρολογήσουμε τις τιμές των δεδομένων που θα μετασχηματίσουμε. Στο παράδειγμα με τις ηλικιακές ομάδες θα επιλέξουμε το **Range** και θα πληκτρολογήσουμε στα δύο λευκά κουτάκια που θα ενεργοποιηθούν το εύρος των τιμών. Για παράδειγμα η πρώτη ηλικιακή ομάδα είναι οι ηλικίες από 20 έως 30 έτη. Άρα θα πληκτρολογήσουμε το 20 στο πρώτο κουτάκι και το 30 στο δεύτερο κουτάκι. Με αυτόν τον τρόπο δηλώνουμε το εύρος των τιμών που θέλουμε να μετασχηματίσουμε. Υπάρχουν άλλες δύο επιλογές που μπορούμε να ορίσουμε εύρη ή διαστήματα τιμών. Είτε από τη χαμηλότερη τιμή έως κάποια τιμή, είτε από κάποια τιμή έως την υψηλότερη. Στη συνέχεια πηγαίνουμε στο δεξιό μέρος του παραθύρου στην περιοχή **New Value.** Στο λευκό κουτάκι που βρίσκεται δεξιά της προεπιλογής **Value** θα πληκτρολογήσουμε τη νέα τιμή για τη συγκεκριμένη ηλικιακή ομάδα. Αφού ξεκινήσαμε με την πρώτη ηλικιακή ομάδα θα βάλουμε τον αριθμό 1. Εν συνεχεία θα πατήσουμε το κουτάκι **Add** για να καταχωρηθεί η αλλαγή στο SPSS.
Μόλις το κάνουμε αυτό θα εμφανιστεί στο μεγάλο λευκό κουτάκι η καταχωρημένη αλλαγή. Συνεχίζουμε κατά τον ίδιο τρόπο και για τις υπόλοιπες ηλικιακές ομάδες. Αν έχουμε επιλέξει οι νέες κωδικοποιημένες τιμές να αποθηκευτούν στη στήλη με τις ήδη υπάρχουσες τιμές, δε χρειάζεται να κάνουμε τίποτα άλλο. Αν όμως έχουμε επιλέξει να αποθηκεύσουμε τις νέες τιμές σε διαφορετική στήλη τότε πατώντας **Continue** θα επιστρέψουμε στο παράθυρο της εικόνας 12 το οποίο είναι διαφορετικό από αυτό της εικόνας 11. Κάτω από το **Output Variable** θα πρέπει να δώσουμε στη νέα στήλη ένα όνομα και μία ετικέτα (ενεργοποιούνται στην αρχή όταν περάσουμε τη στήλη των δεδομένων από το αριστερό στο δεξιό κουτάκι). Στη συνέχεια θα πατήσουμε **Change** ώστε να γίνει η αλλαγή στο όνομα. Αφού τελειώσουμε θα εμφανιστεί στο SPSS Data Editor μία νέα στήλη που θα περιέχει τις κωδικοποιημένες τιμές της αρχικής στήλης. Και στις δύο περιπτώσεις το SPSS αντιστοιχεί στις ήδη υπάρχουσες τιμές τις νέες τιμές ανάλογα με το διάστημα στο οποίο βρίσκονται. Για τις ηλικίες δηλαδή από 20 έως και 29 θα αντιστοιχήσει την τιμή 1 (η ηλικία 30 θα συμπεριληφθεί στη δεύτερη ομάδα, όπως και κάθε άνω άκρο των κλάσεων ή ομάδων). Για τις ηλικίες από 30 έως 39 θα αντιστοιχήσει την τιμή 2. Οι κλάσεις που

θα δημιουργηθούν θα είναι τύπου [,). Καλό θα είναι σε αυτό το σημείο να επιλέξουμε στο Data Editor να εμφανίσει το παράθυρο Variable View, για να καθορίσουμε την κάθε κωδικοποιημένη τιμή που ανήσυχη. Στο παράδειγμα με τις ηλικίες επιλέγοντας **Values** και με τη διαδικασία η οποία έχει ήδη περιγραφεί να ορίσουμε την κάθε τιμή σε ποια ηλικιακή ομάδα αντιστοιχεί. Για παράδειγμα για την τιμή 1 μπορούμε να πληκτρολογήσουμε στο παράθυρο που θα εμφανιστεί (εικόνα 6) «**ages between 20 and 30**». Αυτό βοηθάει ώστε στις αναλύσεις να μην εμφανίζονται νούμερα π.χ. 1, 2, 3, ... και να πρέπει να εξηγούμε ότι η τιμή 1 αντιστοιχεί στην πρώτη ηλικιακή ομάδα και ούτω καθεξής. Με αυτόν τον τρόπο για κάθε τιμή θα εμφανίζεται το μήνυμα που έχουμε πληκτρολογήσει στο παράθυρο (εικόνα 6).



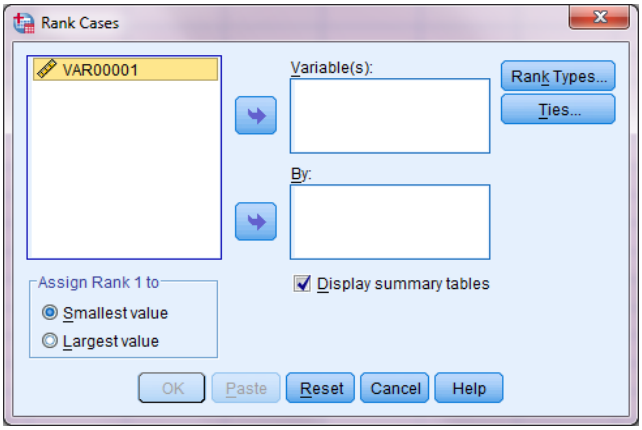
Εικόνα 13

Ανεξάρτητα από την επιλογή αποθήκευσης της μετασχηματισμένης στήλης δεδομένων, αν πατήσουμε την επιλογή **If** θα εμφανιστεί το παράθυρο της εικόνας 14. Με αυτήν την επιλογή κωδικοποιούμε μόνο τις τιμές που ικανοποιούν κάποια λογική συνθήκη. Εδώ, όπως και στο παράθυρο της εικόνας 10, πρώτα επιλέγουμε τη συνάρτηση από τη λίστα των συναρτήσεων, την ανεβάζουμε πάνω με το βελάκι και μετά περνάμε τη στήλη με την οποία θα δουλέψουμε δεξιά.



Εικόνα 14

Με την εντολή **Rank Cases** αναθέτουμε τάξεις μεγέθους στα δεδομένα των στηλών που επιλέγουμε. Το παράθυρο που εμφανίζεται σε αυτήν την περίπτωση είναι αυτό της εικόνας 15. Περνάμε τη στήλη στα δεδομένα της οποίας θέλουμε να τοποθετήσουμε τάξεις μεγέθους στο πάνω δεξιό κουτάκι. Κάτω αριστερά μας δίνεται η δυνατότητα να επιλέξουμε από που θα αρχίζουν οι τάξεις μεγέθους. Η προεπιλεγμένη επιλογή του SPSS είναι αυτή που δίνει στη μικρότερη τιμή την τιμή 1 στην αμέσως επόμενη την τιμή 2 και ούτω καθεξής. Μπορούμε να επιλέξουμε αν θέλουμε απλά τάξεις μεγέθους για τα δεδομένα ή ποσοστά από το **Rank Types.** Πατώντας το **Ties** το SPSS μας ρωτάει τι τάξη μεγέθους να αναθέσει στην περίπτωση ισοβαθμούντων τάξεων μεγέθους. Η προεπιλογή είναι ο μέσος όρος των τάξεων. Για παράδειγμα η πέμπτη και η έκτη τιμή είναι ίσες. Οι τάξεις μεγέθους που θα αντιστοιχούσαν στις δύο αυτές τιμές αν ήταν διαφορετικές θα ήταν η 5 και η 6 αντίστοιχα. Σε αυτήν την περίπτωση που έχουμε δύο ίσες τιμές το SPSS θα αναθέσει και στις δύο τιμές το μέσο όρο των δύο τάξεων μεγέθους των τιμών, δηλαδή το 5.5 και για τις δύο τιμές.



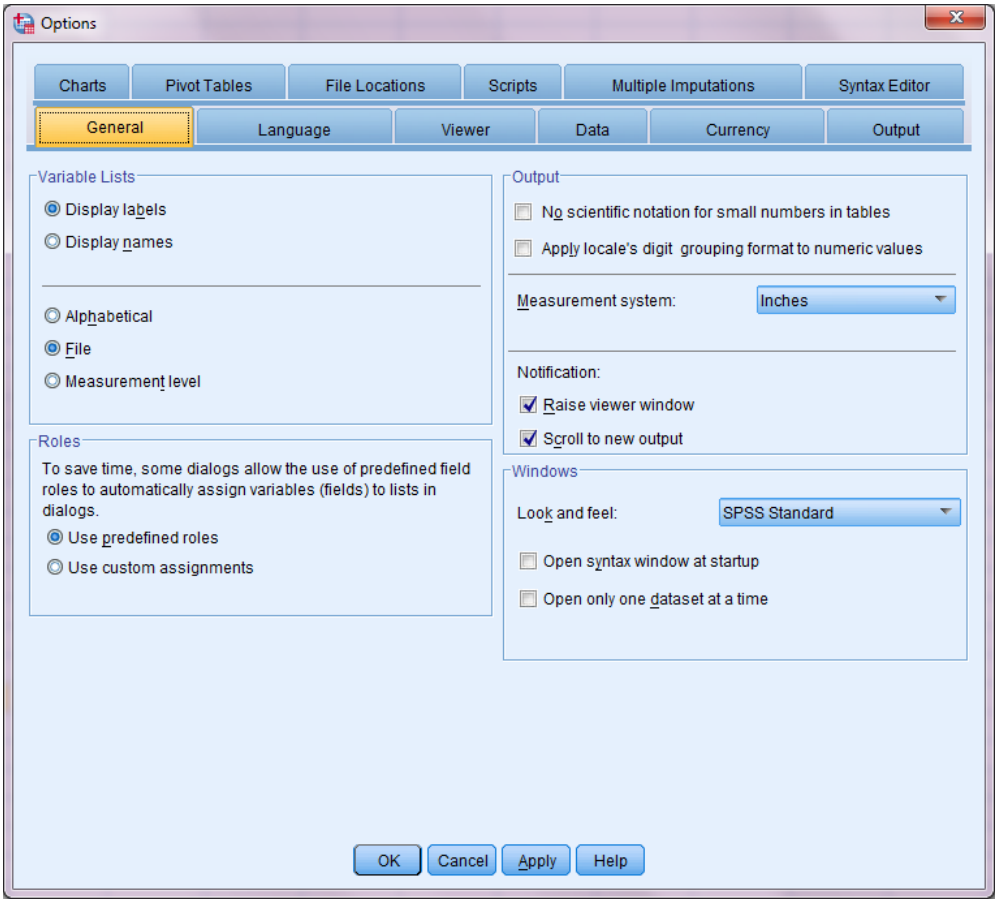
Εικόνα 15

Η εντολές **Create Time Series** και **Replace Missing values** είναι για λίγο πιο προχωρημένους, οπότε τις παραλείπουμε προς το παρόν.

2.5 Το μενού της επιλογής Analyze

Πριν δούμε την επιλογή **Analyze** από το μενού επιλογών του **Data Editor** ας δούμε τι σημαίνουν οι επιλογές που αποτελούν το μενού.

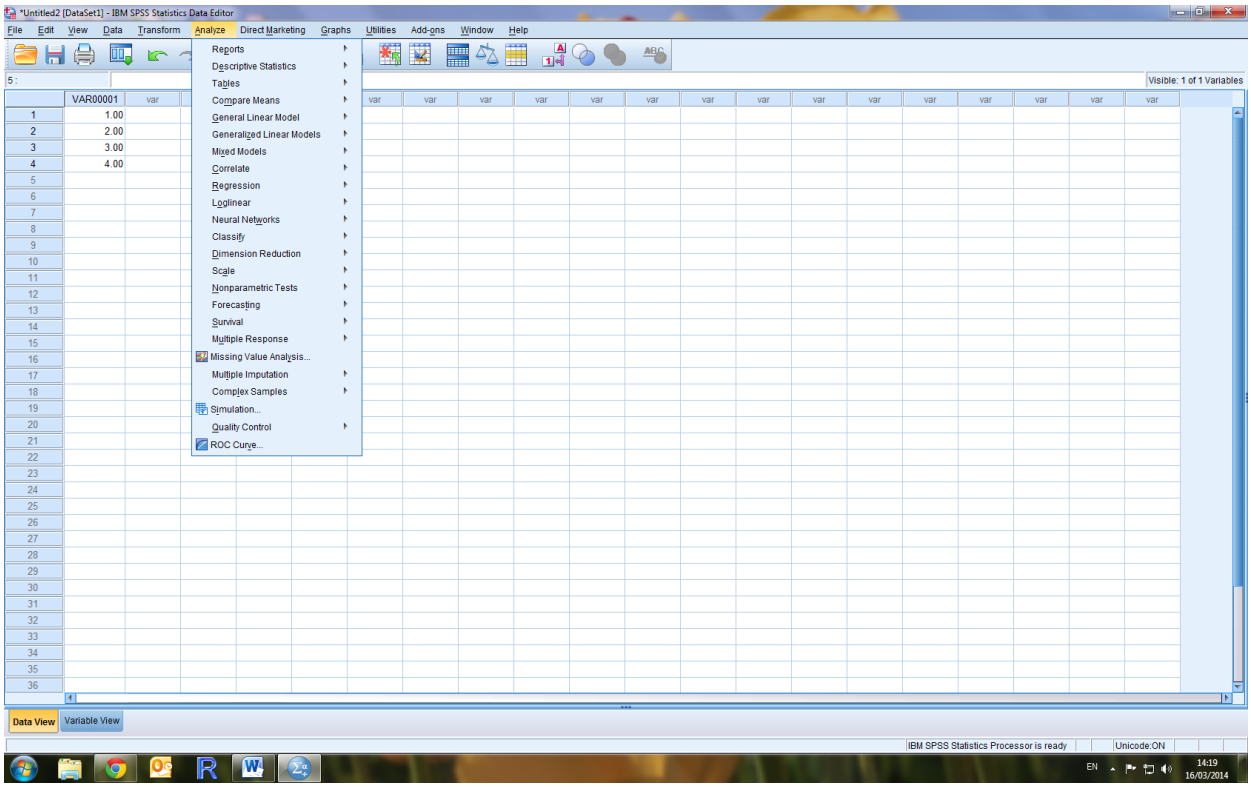
- Η επιλογή **File** χρησιμοποιείται για να δημιουργήσουμε ή να ανοίξουμε ένα νέο αρχείο δεδομένων, ή να αποθηκεύσουμε/εκτυπώσουμε το υπάρχον αρχείο δεδομένων.
- Η επιλογή **Edit** χρησιμοποιείται για την επεξεργασία δεδομένων, όπως αντιγραφή επικόλληση κ.ά. Επίσης υπάρχει η επιλογή **Options** η οποία εμφανίζει το παράθυρο της εικόνας 16a. Εδώ μας παρέχονται γενικά επιλογές του SPSS όπως η εμφάνιση των αποτελεσμάτων και των πινάκων. Αν πατήσουμε το **Viewer** θα μας πάει στο παράθυρο της εικόνας 16.



Εικόνα 16

- Η επιλογή **View** παρέχει λίγες πληροφορίες, μία από αυτές είναι να βλέπουμε τα δεδομένα χωρίς κελιά.
- Η επιλογή **Data** μας παρέχει δυνατότητες όπως επιλογή στηλών δεδομένων με τις οποίες θέλουμε να δουλέψουμε (την είδαμε προηγουμένως), διάταξης των δεδομένων, ορισμό νέων μεταβλητών κ.ά.
- Η επιλογή **Transform** παρέχει δυνατότητες όπως αυτέ που είδαμε προηγουμένως.
- Η επιλογή **Analyze** είναι η καρδιά των επιλογών του SPSS αφού περιέχει σχεδόν όλες τις εντολές ανάλυσης των δεδομένων και την οποία θα δούμε σε λίγο.
- Η επιλογή **Direct Marketing** περιέχει εργαλεία χρήσιμα σε όσους κάνουν ανάλυση σε τεχνικές Marketing.
- Η επιλογή **Graphs** περιέχει όλα τα διαγράμματα που μπορούμε να δημιουργήσουμε και την οποία θα δούμε αργότερα.
- Η επιλογή **Utilities** μας επιτρέπει να δημιουργήσουμε σετ στηλών, να δούμε πληροφορίες για κάθε στήλη ξεχωριστά κ.ά.
- Η επιλογή **Add-ons** έχει διάφορες επιλογές οι οποίες παραπέμπουν στην ηλεκτρονική διεύθυνση του SPSS.
- Η επιλογή **Window** μας επιτρέπει να κάνουμε split του φύλου δεδομένων ή να ελαχιστοποιήσουμε το παράθυρο εργασίας.
- Η τελευταία επιλογή του μενού επιλογών αλλά πολύ χρήσιμη είναι αυτή της βοήθειας (**Help**).

Επιλέγοντας **Analyze** από το μενού θα εμφανιστεί το υπομενού αυτής της επιλογής που περιέχει σχεδόν όλες τις δυνατές στατιστικές τεχνικές που παρέχει το πακέτο.



Εικόνα 17

Πιο συγκεκριμένα οι εντολές που περιέχονται στην επιλογή **Analyze** είναι οι εξής:

- **Reports**: περιέχει δυνατότητες παρουσίασης κάποιων στοιχείων για τα δεδομένα.
- **Descriptive Statistics**: περιέχει δυνατότητες εμφάνισης περιγραφικών μέτρων των δεδομένων, γραφημάτων, πινάκων δεδομένων κ.ά.
- **Tables**: παρέχονται δυνατότητες δημιουργίας πολύπλοκων πινάκων.
- **Compare Means**: περιλαμβάνονται οι εντολές ελέγχων υποθέσεων για τους μέσους. Θα τους δούμε όμως πιο αναλυτικά παρακάτω.
- **General Linear Model**: υπάρχουν οι δυνατότητες χρησιμοποίησης μοντέλων ανάλυσης διακύμανσης. Θα δούμε κάποια από αυτά παρακάτω.
- **Generalized Linear Models**: περιέχει μία πληθώρα δυνατοτήτων χρησιμοποίησης γενικευμένων γραμμικών μοντέλων.
- **Mixed Models**: η εντολή αφορά σε μικτά γραμμικά μοντέλα.
- **Correlate**: περιέχει συντελεστές συσχέτισης, μερικής συσχέτισης και υπολογισμού αποστάσεων
- **Regression**: περιέχει δυνατότητες χρησιμοποίησης απλής και πολλαπλής γραμμικής και μη γραμμικής παλινδρόμησης, λογιστικής παλινδρόμησης κ.ά.
- **Loglinear**: παρέχει δυνατότητες χρησιμοποίησης λογαριθμικών μοντέλων.
- **Neural Networks**: περιέχει εργαλεία για νευρωνικά δίκτυα.

- **Classify**: εμπεριέχει πολλές πολυμεταβλητές, στατιστικές και μη, τεχνικές ομαδοποίησης δεδομένων ή μεταβλητών.
- **Dimension Reduction**: περιέχει πολυμεταβλητές τεχνικές μείωσης μεταβλητών, όπως παραγοντική ανάλυση, ανάλυση αντιστοιχιών.
- **Scale**: περιέχει τεχνικές πολυδιάστατης κλιμακοποίησης και ανάλυσης αξιοπιστίας η οποία χρησιμοποιείται κατά κόρον σε ψυχομετρικούς ελέγχους, ελέγχους προσωπικότητας, ικανοτήτων.
- **Nonparametric Tests**: υπάρχει λίστα με μη παραμετρικές στατιστικές τεχνικές. Θα δούμε τη χρησιμότητα τους αργότερα.
- **Forecasting**: η επιλογή αυτή περιέχει διάφορες τεχνικές ανάλυσης χρονολογικών σειρών.
- **Survival**: υπάρχουν τεχνικές ανάλυσης χρόνων ζωής από ιατρικές μελέτες.
- **Multiple Response**: παρέχεται η δυνατότητα δημιουργίας διχοτομικών (0 και 1 δεδομένων) μεταβλητών ή ψευδομεταβλητών όπως αλλιώς ονομάζονται από μεταβλητές με πολλές κατηγορίες.
- **Missing Value Analysis:** η εντολή αφορά στην ανάλυση εκλιπουσών τιμών.
- **Multiple Imputation**: η εντολή αφορά στην ανάλυση εκλιπουσών τιμών.
- **Complex Samples**: περιέχει μία σειρά από διαδικασίες δειγματοληψίας.
- **Simulation**: η εντολή αναφέρεται σε διαδικασίες προσομοίωσης.
- **Quality Control**: αφορά σε διαδικασίες στατιστικού ελέγχου ποιότητας.
- **ROC Curve:** η εντολή αφορά σε καμπύλες χαρακτηριστικού λειτουργικού δέκτη

3. Το Bootstrap στο IBM SPSS 22

Η SPSS ενδυνάμωσε πολύ το συγκεκριμένο στατιστικό πακέτο της με την προσθήκη της επιλογής bootstrap στην 19η (ίσως και από την 17η, δε θυμάμαι). Σε αυτό το σημείο λοιπόν πιστεύω θα ήταν καλό να προσπαθήσουμε να καταλάβουμε τι είναι αυτή η τεχνική ή αλγόριθμος, καλύτερα.

3.1 Μία σύντομη εισαγωγή

Ήταν το1979 όταν ο Bradley Efron δημοσίευσε για πρώτη φορά την ιδέα του. Πρόκειται για τη θέα της στατιστικής από μία άλλη άποψη, πιο υπολογιστική. Θα προσπαθήσουμε να εξηγήσουμε τον αλγόριθμο χωρίς να γίνουμε πολύ τεχνικοί. Να αναφέρουμε για την ιστορία ότι το όνομα προήλθε από το παραμύθι "οι περιπέτειες του Βαρόνου Μινχάουζεν". Σε μία περιπέτεια του ο Βαρόνος βυθιζόταν στη θάλασσα. Για να σωθεί θα έπρεπε να τραβήξει ένα σκοινί. Δεν είχε όμως σκοινί, οπότε άρχισε να τραβάει τα κορδόνια της μπότας του και έτσι "τράβηξε τον εαυτό του" επάνω στην επιφάνεια της θάλασσας.

Στη στατιστική λέμε όταν υπολογίζουμε π.χ. το μέσο (όρο) ενός δείγματος ότι εκτιμούμε την πραγματική τιμή του μέσου του πληθυσμού, από τον οποίο προήλθε το δείγμα. Εν συνεχεία λέμε ότι ο μέσος ακολουθεί ασυμπτωτικά (δηλαδή καθώς το μέγεθος του δείγματος μεγαλώνει και τείνει προς το άπειρο) την κανονική κατανομή. Όταν κατασκευάζουμε ένα 95% διάστημα εμπιστοσύνης για την πραγματική τιμή του μέσου η ερμηνεία βασίζεται σε ασυμπτωτικά αποτελέσματα.

Η ερμηνεία του 95% διαστήματος εμπιστοσύνης για το μέσο είναι η εξής: Αν είχαμε τη δυνατότητα να επαναλάβουμε τη δειγματοληψία ν φορές και κάθε φορά εκτιμούσαμε το μέσο και κατασκευάζαμε 95% διαστήματα εμπιστοσύνης, θα αναμέναμε το 95% αυτών να έχουν συμπεριλάβει την πραγματική τιμή του μέσου.

Το bootstrap μας δίνει τη δυνατότητα αυτή, να προσομοιάσουμε δηλαδή αυτή τη δυνατότητα επαναδειγματοληψίας βασισμένοι στο αρχικό μας δείγμα. Όπως ο Βαρόνος έσωσε τον εαυτό του τραβώντας τον ίδιο του τον εαυτό έτσι και το bootstrap θα μας δώσει μία εκτίμηση για την κατανομή του δειγματικού μέσου βασισμένο στο ίδιο το δείγμα.

3.2 Σύντομη περιγραφή του αλγόριθμου

Θα περιγράψουμε τη διαδικασία με τη βοήθεια της λοταρίας. Βάζουμε κάποια μπαλάκια με αριθμούς σε ένα βάζο, έστω n. Βάζουμε το χέρι μας μέσα και τραβάμε ένα μπαλάκι. Σημειώνουμε τον αριθμό του σε ένα χαρτί και το ρίχνουμε πάλι μέσα. Ξαναβάζουμε το χέρι μας μέσα και τραβάμε ένα μπαλάκι και πάλι σημειώνουμε τον αριθμό του και το ρίχνουμε μέσα στο βάζο (αυτή η διαδικασία λέγεται δειγματοληψία με επανατοποθέτηση). Αυτή τη διαδικασία θα την επαναλάβουμε n φορές, διότι τόσα μπαλάκια έχουμε δηλαδή στη διάθεση μας. Προφανώς κάποια μπαλάκια μπορεί να έχουν επιλεγεί περισσότερες από μία φορές, αλλά αυτό δεν αποτελεί πρόβλημα.

Αφού λοιπόν έχουμε επιλέξει n μπαλάκια, έχουμε τελειώσει την πρώτη δειγματοληψία bootstrap και έχουμε το πρώτο δείγμα bootstrap. Παίρνουμε τα νούμερα από τα μπαλάκια και υπολογίζουμε π.χ. το μέσο όρο τους. Αυτός θα είναι ο πρώτος μέσος bootstrap. Θα επαναλάβουμε τη διαδικασία που μόλις περιγράψαμε πολλές φορές, έστω 1000. Στο τέλος της ημέρας θα έχουμε υπολογίσει 1000 μέσους (bootstrap). Με αυτό τον τρόπο θα έχουμε "κατασκευάσει" την κατανομή του δειγματικού μέσου. Αυτή η κατανομή θα "μιμείται" την πραγματική κατανομή του

μέσου που είναι η κανονική. Για εμάς προφανώς τα μπαλάκια είναι οι παρατηρήσεις που έχουμε, το δείγμα μεγέθους n.

Έχοντας λοιπόν την (ψευδό-) κατανομή του μέσου μπορούμε να εκτιμήσουμε το τυπικό σφάλμα του (υπολογίζοντας την τυπική απόκλιση των 1000 αυτών τιμών bootstrap) και να φτιάξουμε και διαστήματα εμπιστοσύνης για την πραγματική τιμή του μέσου. Αφού έχουμε 1000 τιμές για το μέσο μπορούμε να τις διατάξουμε από τη μικρότερη στη μεγαλύτερη και να πάρουμε το 95% των κεντρικών τιμών. Δηλαδή θα αφήσουμε τις 25 πιο χαμηλές και υψηλές τιμές έξω. Η $26^{η}$ και η $975^{η}$ τιμή θα αποτελούν τα κάτω και άνω άκρα του 95% διαστήματος εμπιστοσύνης για το μέσο.

Επίσης μας δίνεται η δυνατότητα να εκτιμήσουμε τη μεροληψία του δειγματικού μέσου. Την απόσταση δηλαδή του εκτιμώμενου μέσου από τον πραγματικό μέσο. Ο μέσος του δείγματος θα παίξει το ρόλο του πραγματικού μέσου. Ο μέσος που θα προκύψει από το bootstrap θα παίξει το ρόλο του εκτιμώμενου μέσου. Η διαφορά τους είναι μία εκτίμηση της πραγματικής μεροληψίας.

Θα χρησιμοποιήσουμε το bootstrap μόνο όταν θέλουμε διαστήματα εμπιστοσύνης, διότι ο έλεγχος υποθέσεων προϋποθέτει μία τροποποίηση των δεδομένων πριν τη διεξαγωγή του αλγορίθμου.

4.1 Περιγραφικά μέτρα για συνεχείς μεταβλητές

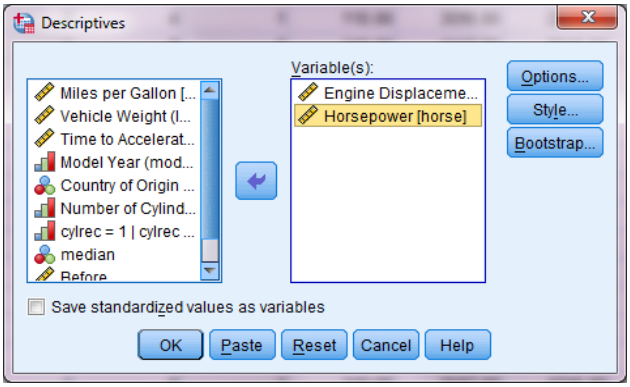
Πριν μιλήσουμε για τον τρόπο εξαγωγής των περιγραφικών μέτρων στο SPSS, καλό θα ήταν να αλλάξουμε τον όρο στήλη των δεδομένων σε μεταβλητή, αφού κάθε στήλη αναπαριστά μία μεταβλητή στην οποία είναι εκχωρημένες οι τιμές της. Τα δεδομένα τα οποία θα χρησιμοποιήσουμε βρίσκονται στο φάκελο SPSS 15 και είναι δεδομένα που αφορούν αυτοκίνητα (**Cars.sav**)[*]. Τα περιγραφικά μέτρα χωρίζονται σε μέτρα κεντρικής τάσης ή θέσης, μέτρα διασποράς και μέτρα ασυμμετρίας και κύρτωσης. Τα μέτρα θέσης δίνουν πληροφορίες για τις κεντρικές τιμές του δείγματος. Αυτά είναι ο μέσος, η διάμεσος, η επικρατούσα τιμή και τα εκατοστημόρια. Τα εκατοστημόρια είναι τιμές του δείγματος οι οποίες "κόβουν" το δείγμα σε συγκεκριμένα συνήθως σημεία. Για παράδειγμα το πρώτο τεταρτημόριο είναι η τιμή του δείγματος η οποία έχει την εξής ιδιότητα: το πολύ 25% των παρατηρήσεων βρίσκεται κάτω από αυτήν την τιμή. Το δεύτερο τεταρτημόριο είναι η τιμή που αφήνει το πολύ το 50% των παρατηρήσεων κάτω από αυτή. Το τρίτο τεταρτημόριο είναι η τιμή για την οποία ισχύει ότι το πολύ το 25% των παρατηρήσεων βρίσκεται πάνω από αυτή. Η διάμεσος είναι η τιμή που "κόβει" τις παρατηρήσεις του δείγματος στη μέση (ταυτίζεται με το δεύτερο τεταρτημόριο) και η κορυφή είναι η παρατήρηση με τη μεγαλύτερη συχνότητα εμφάνισης.

Τα μέτρα διασποράς δίνουν πληροφορίες για το πως εκτείνονται οι παρατηρήσεις γύρω από το "κέντρο" τους. Αυτά είναι το εύρος, η τυπική απόκλιση, η διακύμανση, ο συντελεστής μεταβλητότητας και το ενδοτεταρτημοριακό εύρος. Ο συντελεστής μεταβλητότητας ορίζεται ως το πηλίκο της τυπικής απόκλισης με το μέσο πολλαπλασιασμένο επί %. Είναι ένα μέτρο ομοιογένειας του δείγματος και δη σχετικής διασποράς, όχι απόλυτης διασποράς. Χρησιμοποιείται επίσης και για τη σύγκριση μεταβλητών εκφρασμένων σε διαφορετικά μεγέθη. Δεχόμαστε ότι ένα δείγμα είναι ομοιογενές όταν η τιμή του συντελεστή δεν ξεπερνά το 10%. Τα μέτρα ασυμμετρίας και κύρτωσης είναι ο συντελεστής ασυμμετρίας και ο συντελεστής κύρτωσης αντίστοιχα. Είναι μέτρα που αφορούν στη μορφή της κατανομής των δεδομένων και θα συζητηθούν παρακάτω.

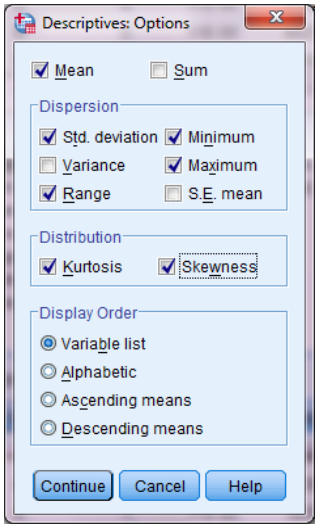
Πατώντας **Analyze→Descriptive Statistics→Descriptives** θα εμφανιστεί το παράθυρο της εικόνας 18. Περνάμε δεξιά τις μεταβλητές των οποίων τα περιγραφικά μέτρα θέλουμε να εμφανιστούν. Εμείς επιλέξαμε τις μεταβλητές που αφορούν στη ιπποδύναμη και στον κυβισμό των αυτοκινήτων. Πατώντας **Options** θα εμφανιστεί το παράθυρο της εικόνας 19 στο οποίο μας δίνεται η δυνατότητα να επιλέξουμε εμείς ποια περιγραφικά μέτρα θέλουμε να εμφανιστούν.

Ο μέσος (mean), η τυπική απόκλιση (Std. deviation), η ελάχιστη (Minimum) και η μέγιστη (Maximum) τιμή είναι προεπιλεγμένα από το SPSS. Με την επιλογή **Display Order** επιλέγουμε με ποια σειρά να εμφανιστούν τα αποτελέσματα.

[*] Αν δεν έχετε στην διάθεση σας τα δεδομένα στείλτε μου ένα e-mail να σας τα στείλω.



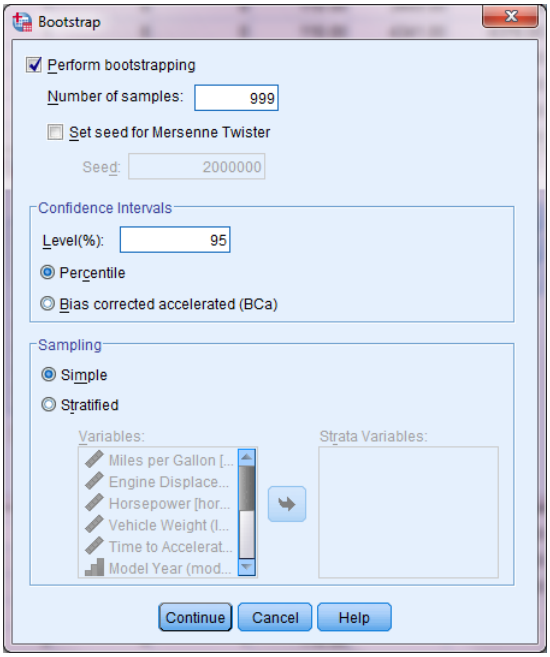
Εικόνα 18



Εικόνα 19

Εμείς θα επιλέξουμε κάποια μέτρα και μετά θα πατήσουμε **Continue.** Έτσι θα γυρίσουμε στο αρχικό παράθυρο της εικόνας 18. Σε αυτό το παράθυρο υπάρχει στο κάτω μέρος μία επιλογή (**Save standardized values as variables**). Με αυτήν την επιλογή το SPSS δημιουργεί μία νέα στήλη για κάθε μεταβλητή που έχουμε επιλέξει η οποία περιέχει τις τυποποιημένες τιμές της μεταβλητής. Οι τυποποιημένες τιμές μίας μεταβλητής είναι οι ίδιες τιμές μετασχηματισμένες όμως έτσι ώστε να έχουν μέση τιμή ίση με το μηδέν και διακύμανση ίση με τη μονάδα. Ο τύπος μετασχηματισμού είναι ο εξής: $(x_i-\mu)/\sigma$, όπου x_i μία τιμή της μεταβλητής X, μ ο μέσος της μεταβλητής και σ η τυπική απόκλιση της μεταβλητής X.

Στο παράθυρο της εικόνας 18 υπάρχει και η επιλογή **Bootstrap**. Αν την επιλέξουμε θα μας οδηγήσει στο παράθυρο της εικόνας 20. Εκεί θα "τικάρουμε" την επιλογή **Perform bootstrapping.** στο κουτάκι από κάτω θα βάλουμε αντί για 1000, 999 δείγματα bootstrap. Ο αριθμός είναι συνήθως ίσος με 999, αλλά μπορείτε να βάλετε και 99 και 9999 ή άλλο νούμερο. Ο αριθμός πρέπει να είναι κατά 1 λιγότερο από έναν αριθμό που είναι πολλαπλάσιος του 100 και ο λόγος θα γίνει κατανοητός λίγο αργότερα. Πιο κάτω στο **Confidence Intervals** θα αφήσουμε το 95% (εκτός και αν θέλουμε άλλο βαθμό εμπιστοσύνης, π.χ. 90%). Ακριβώς από κάτω θα επιλέξουμε **Percentile** και μετά **Continue** για να επιστρέψουμε στο παράθυρο της εικόνας 18, όπου θα πατήσουμε **OK**. Το αποτέλεσμα φαίνεται στο σχήμα 1 παρακάτω.



Εικόνα 20

Descriptive Statistics

				Bootstrap[a]			
						95% Confidence Interval	
		Statistic	Std. Error	Bias	Std. Error	Lower	Upper
Engine Displacement (cu. inches)	N	400		0	0	400	400
	Range	451					
	Minimum	4					
	Maximum	455					
	Mean	195.02		-.26	5.39	183.96	205.07
	Std. Deviation	105.579		-.181	2.911	99.164	110.698
	Skewness	.674	.122	.003	.092	.500	.871
	Kurtosis	-.821	.243	.013	.165	-1.096	-.438
Horsepower	N	400		0	0	400	400
	Range	184					
	Minimum	46					
	Maximum	230					
	Mean	104.83		-.07	1.98	100.94	108.68
	Std. Deviation	38.522		-.047	1.549	35.443	41.521
	Skewness	1.044	.122	.001	.106	.836	1.259
	Kurtosis	.591	.243	.009	.332	-.026	1.299
Valid N (listwise)	N	400		0	0	400	400

a. Unless otherwise noted, bootstrap results are based on 999 bootstrap samples

Σχήμα 1: Περιγραφικά μέτρα για τις δύο μεταβλητές.

Τα περιγραφικά μέτρα που παρουσιάζονται εδώ είναι με τη σειρά τα εξής: Το πλήθος των στοιχείων (N), το εύρος (Range) το οποίο υπολογίζεται ως η διαφορά της μικρότερης (Minimum) τιμής από τη μεγαλύτερη (Maximum). Ο μέσος (Mean) των μεταβλητών και η τυπικό διακύμανση (Std. Deviation). Η διακύμανση είναι ο μέσος όρος των τετραγωνικών αποκλίσεων των τιμών από τη μέση τιμή. Η τυπική απόκλιση προκύπτει από την τετραγωνική ρίζα της διακύμανσης.

Ο συντελεστής ασυμμετρίας (skewness) δίνει πληροφορίες για την ασυμμετρία της κατανομής των δεδομένων. Τιμές κοντά στο μηδέν παρέχουν ενδείξεις ότι η κατανομή των παρατηρήσεων είναι συμμετρική. Αρνητικές τιμές του συντελεστή ασυμμετρίας είναι ένδειξη ότι η κατανομή παρουσιάζει αρνητική ή αριστερή ασυμμετρία. Τέλος η κατανομή είναι θετικά ή δεξιά ασύμμετρη όταν έχουμε θετικές τιμές. Όταν η κατανομή είναι θετικά ασύμμετρη ο μέσος των παρατηρήσεων είναι μεγαλύτερος από τη διάμεσο η οποία είναι μεγαλύτερη με τη σειρά της από την κορυφή. Το ακριβώς αντίθετο ισχύει για την περίπτωση της αρνητικής ασυμμετρίας. Δηλαδή ο μέσος είναι μικρότερος από τη διάμεσο η οποία είναι μικρότερη από την κορυφή. Για την περίπτωση της συμμετρικής κατανομής αυτά τα τρία μέτρα ταυτίζονται.

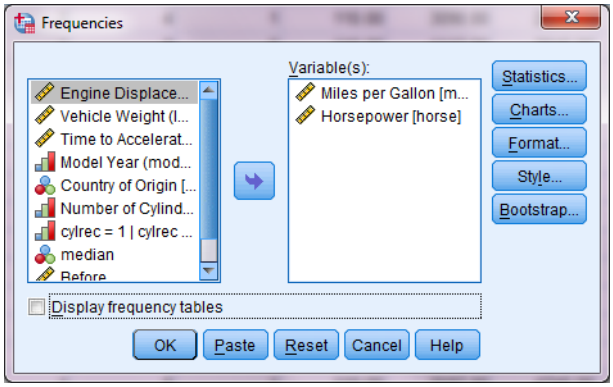
Ο συντελεστής κύρτωσης αναφέρεται στην κυρτότητα της κατανομής των δεδομένων. Αρνητικές τιμές σημαίνουν ότι η κατανομή είναι πλατύκυρτη ενώ θετικές τιμές ότι είναι λεπτόκυρτη. Τιμές κοντά στο μηδέν είναι ένδειξη ότι η κατανομή είναι μεσόκυρτη. Όταν αναφερόμαστε στην κυρτότητα μιας κατανομής αναφερόμαστε στα άκρα της κατανομής ή "ουρές" της κατανομής όπως αλλιώς λέγονται. Οι "παχιές" ουρές είναι ένδειξη πλτατύκυρτης κατανομής. Αντίθετα οι "λεπτές" ουρές αποτελούν ένδειξη πως η κατανομή είναι λεπτύκυρτη.

Όπως μπορούμε να δούμε από το σχήμα 1, οι διαφορές μεταξύ των τιμών που υπολογίστηκαν από το δείγμα και των τιμών που υπολογίστηκαν από το bootstrap δε διαφέρουν και πολύ (εκτός από τη διακύμανση). Η στήλη **Bias** που περιέχει τις διαφορές αυτές αλλά όχι τις τιμές που υπολογίστηκαν από το bootstrap. Οι διαφορές είναι υπολογισμένες ως εξής: τιμές bootstrap - δειγματικές τιμές. Άρα για να υπολογίσουμε τις εκτιμήσεις από το bootstrap απλά προσθέτουμε στις εκτιμήσεις που έχουμε, τις τιμές της στήλης **Bias**. Θα τονίσουμε σε αυτό το σημείο ότι το bootstrap για τη διακύμανση ή την τυπική απόκλιση καλό είναι να αποφεύγεται. Αυτό το λένε οι Casella and Burger στο βιβλίο τους και δεν έχουμε λόγο να τους αμφισβητήσουμε.

Ο λόγος που οι διαφορές είναι πολύ μικρές είναι διότι το μέγεθος του δείγματος ήταν μεγάλο και για τις δύο μεταβλητές (400 παρατηρήσεις). Για αυτό το λόγο στις επόμενες αναλύσεις θα χρησιμοποιούμε μικρότερο δείγμα, μεγέθους 20. Θα επιλέξουμε λοιπόν κάποιες παρατηρήσεις στην τύχη. Πήραμε για παράδειγμα τις 20 πρώτες τιμές από τον κυβισμό και την ιπποδύναμη.

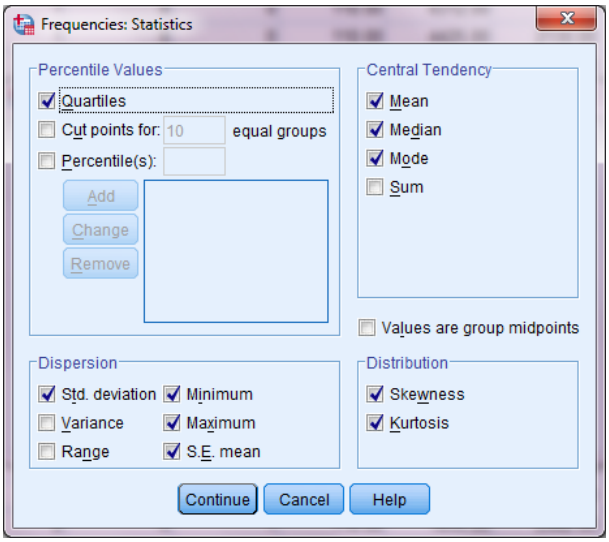
Ας δούμε τώρα μία άλλη επιλογή από το μενού επιλογών η οποία παρέχει περισσότερα περιγραφικά μέτρα. Αυτή τη φορά όμως θα επιλέξουμε ένα δείγμα 20 παρατηρήσεων από αυτές τις δύο μεταβλητές. Οι νέες μεταβλητές έχουν σχεδόν το ίδιο όνομα με πριν. Ο λόγος που πήραμε ένα μικρό δείγμα είναι για να δούμε πόσο διαφορετικές είναι οι εκτιμήσεις από το bootstrap.

Πατάμε **Analyze→Descriptive Statistics→Frequencies** και θα εμφανιστεί στην οθόνη το παράθυρο της εικόνας 21. Κάτω δεξιά έχει μία επιλογή να εμφανίσει πίνακες συχνοτήτων για τις επιλεγμένες μεταβλητές. Εμείς δεν το χρειαζόμαστε στην παρούσα φάση άρα το "ξε-τικάρουμε". Αν πατήσουμε **Charts**, θα μας εμφανίσει ένα παράθυρο με επιλογές γραφημάτων τα οποία όμως θα δούμε παρακάτω. Πατώντας **Statistics** θα εμφανιστεί το παράθυρο της εικόνας 22.



Εικόνα 21

Υπάρχουν επιλογές εμφάνισης όλων των περιγραφικών μέτρων και είναι διαχωρισμένα ανάλογα με το είδος τους (κεντρικής τάσης, διασποράς, κατανομής και ποσοστιαία σημεία). Πατώντας **Continue** επιστρέφουμε στο παράθυρο της εικόνας 21. Εκεί επιλέγουμε **Bootstrap** και θα εμφανιστεί το παράθυρο της εικόνας 20 πάλι. Επιλέγουμε αριθμό δειγμάτων bootstrap 999 και μετά **Continue** και μετά **OK** για να εμφανιστεί το σχήμα 2 στο Output του SPSS. Αν επιλέξουμε **Quartiles** θα μας εμφανίσει επίσης το πρώτο (25%), το δεύτερο (50% διάμεσος) και το τρίτο (75%) τεταρτημόριο. Το τυπικό σφάλμα ή τυπική απόκλιση του μέσου ορίζεται ως η τυπική απόκλιση του δείγματος διαιρεμένη με την τετραγωνική ρίζα του μεγέθους του δείγματος (N). Η διακύμανση είναι ο μέσος όρος των τετραγωνικών αποκλίσεων των τιμών από τη μέση τιμή. Η τυπική απόκλιση προκύπτει από την τετραγωνική ρίζα της διακύμανσης.



Εικόνα 22

Στο σημείο αυτό που είδαμε αυτούς τους δύο τρόπους εξαγωγής κάποιων περιγραφικών μέτρων για συνεχείς μεταβλητές και αφού είδαμε και το bootstrap μπορούμε να πούμε μερικά λόγια για τα αποτελέσματα του bootstrap.

Καταρχήν το bootstrap δεν έδωσε εκτιμήσεις για όλα τα στατιστικά. Τα 95% διαστήματα εμπιστοσύνης που υπολογίζονται δεν βασίζονται στου κλασικό τύπο που θα δούμε αργότερα, ο εκτιμητής +/- 1.96 φορές το τυπικό σφάλμα του. Υπολογίζεται

με βάση έναν άλλο τύπο που αναφέραμε όταν εξηγούσαμε τον υπολογιστικό αυτό αλγόριθμο. Έστω ότι η διαδικασία έχει επαναληφθεί Β φορές. Τότε έχουμε Β τιμές τις οποίες τις διατάσσουμε από τη μικρότερη στη μεγαλύτερη. Εν συνεχεία βρίσκουμε κάποια συγκεκριμένες τιμές από αυτές τις διατεταγμένες τιμές. Έστω ότι θέλουμε βαθμό εμπιστοσύνης ίσο με α. Θα πάρουμε την (Β+1)*(α/2) μικρότερη και μεγαλύτερη τιμή. Θα κόψουμε δηλαδή από την κατανομή των Β τιμών αυτών το α% των τιμών, α/2% από κάτω και α/2% από πάνω. Είπαμε όμως προηγουμένως να προτιμούνται τιμές που δεν είναι πολλαπλάσια του 100 αλλά πολλαπλάσια του 99. Έτσι, αν π.χ. Β=999, τότε Β+1=1000, άρα θα πρέπει να βρούμε την $25^{η}$ και $976^{η}$ τιμή κατά αύξουσα σειρά. Αυτός είναι ο τρόπος υπολογισμού που χρησιμοποιεί και το IBM SPSS 22.

Statistics

				Bootstrap[b]			
						95% Confidence Interval	
			Statistic	Bias	Std. Error	Lower	Upper
N	Valid	Miles per Gallon	392	0	0	392	392
		Horsepower	392	0	0	392	392
	Missing	Miles per Gallon	0	0	0	0	0
		Horsepower	0	0	0	0	0
Mean		Miles per Gallon	23.45	.01	.41	22.64	24.26
		Horsepower	104.21	.00	2.00	100.21	108.33
Std. Error of Mean		Miles per Gallon	.394				
		Horsepower	1.931				
Median		Miles per Gallon	22.75	-.12	.77	21.00	24.00
		Horsepower	93.00	.11	2.11	90.00	97.00
Mode		Miles per Gallon	13				
		Horsepower	150				
Std. Deviation		Miles per Gallon	7.805	-.005	.237	7.334	8.271
		Horsepower	38.233	-.087	1.605	35.005	41.200
Skewness		Miles per Gallon	.457	-.004	.089	.288	.624
		Horsepower	1.098	-.005	.108	.895	1.311
Std. Error of Skewness		Miles per Gallon	.123				
		Horsepower	.123				
Kurtosis		Miles per Gallon	-.516	-.001	.170	-.818	-.168
		Horsepower	.753	-.004	.349	.138	1.466
Std. Error of Kurtosis		Miles per Gallon	.246				
		Horsepower	.246				
Minimum		Miles per Gallon	9				
		Horsepower	46				
Maximum		Miles per Gallon	47				
		Horsepower	230				
Percentiles	25	Miles per Gallon	17.00	.14	.60	16.00	18.00
		Horsepower	75.00	.97	1.67	74.00	80.00
	50	Miles per Gallon	22.75	-.12	.77	21.00	24.00
		Horsepower	93.00	.11	2.11	90.00	97.00
	75	Miles per Gallon	29.00	.12	.75	27.98	30.50
		Horsepower	125.00	.23	7.92	110.00	140.00

b. Unless otherwise noted, bootstrap results are based on 999 bootstrap samples

Σχήμα 2: Περιγραφικά μέτρα για συνεχείς μεταβλητές.

4.2 Περιγραφικά μέτρα για κατηγορικές μεταβλητές

Είδαμε τι κάνουμε όταν οι μεταβλητές μας είναι συνεχείς, αλλά τι γίνεται όταν είναι κατηγορικές. Κατηγορικές όπως για παράδειγμα η ομάδα αίματος, το φύλο, ο βαθμός ικανοποίησης ή αρεσκείας, οι χώρα που μένουν άνθρωποι, η ομάδα που υποστηρίζουν μερικοί σε κάποιο ομαδικό άθλημα, ή η χώρα προέλευσης των αυτοκινήτων στο παράδειγμα μας. Προφανώς, αυτές οι μεταβλητές δεν παίρνουν αριθμητικές τιμές, π.χ. 1, 2, ...

Θα τονίσουμε επίσης ότι αν υπάρχει κάποια διάταξη στις μεταβλητές αυτές τότε θα λέμε ότι είναι διατακτικές μεταβλητές ή κατηγορικές διατακτικής κλίμακας. Για παράδειγμα, ο βαθμός ικανοποίησης κάποιου ατόμου από τη δουλειά του: «καθόλου», «*μέτρια*», «λίγο», «πολύ». Αυτές οι περιπτώσεις φανερώνουν διάταξη. Αν όμως δεν υπάρχει διάταξη, π.χ. το φύλο: άντρας, γυναίκα ή η ομάδα αίματος, τότες θα λέμε ότι αυτές οι μεταβλητές αυτές είναι ονομαστικέ μεταβλητές ή κατηγορικές ονομαστικής κλίμακας.

Στην περίπτωση λοιπόν των κατηγορικών μεταβλητών θα πάμε πάλι στο παράθυρο της εικόνας 21. Την επιλογή κάτω αριστερά **Display frequency tables** την είχαμε "ξε-τικάρει" προηγουμένως. Τώρα όμως δεν θα την "ξε-τικάρουμε" διότι αυτό είναι που θέλουμε να εμφανιστεί. Στην επιλογή όμως **Statistics** και **Bootstrap** (δεν υπάρχει λόγος για bootstrap) δεν θα επιλέξουμε τίποτα. Πατάμε **OK** και το αποτέλεσμα φαίνεται στο σχήμα 3.

Country of Origin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	American	253	62.3	62.5	62.5
	European	73	18.0	18.0	80.5
	Japanese	79	19.5	19.5	100.0
	Total	405	99.8	100.0	
Missing	System	1	.2		
Total		406	100.0		

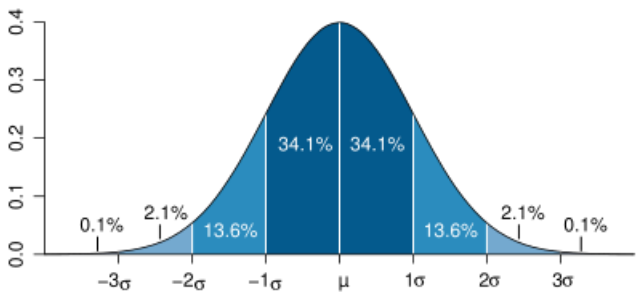
Σχήμα 3: Περιγραφικά μέτρα για κατηγορική μεταβλητή.

Έχουμε για κάθε χώρα από τις 3 τον αριθμό των αυτοκινήτων που προέρχονται από αυτές, σε απόλυτους αριθμούς και σε σχετικούς. Στο SPSS έχουμε 406 γραμμές αλλά για ένα αυτοκίνητο δε γνωρίζουμε τη χώρα προέλευσης (Missing value 1). Βλέπουμε ότι η χώρα με τα περισσότερα αυτοκίνητα είναι η Αμερική. Η δεύτερη στήλη περιέχει τα ποσοστά των τιμών αυτών, συμπεριλαμβανομένης και της εκλιπούσας τιμής. Τα ποσοστά αυτά έχουν υπολογιστεί διαιρώντας κάθε αριθμό με το 406 και μετά πολλαπλασιάζοντας με το 100 για να έρθουν σε μορφή ποσοστού. Η Τρίτη στήλη (**Valid Percent**) όμως περιέχει τα αποτελέσματα της διαίρεσης με το 405 και όχι με το 406. Το SPSS δηλαδή έχει υπολογίσει τα ποσοστά για τις χώρες που έχουμε παρατηρήσεις, αφού για ένα αυτοκίνητο δε γνωρίζουμε τη χώρα προέλευσης, μπορούμε να το αφήσουμε εκτός και να δουλέψουμε με τα υπόλοιπα. Η τέταρτη στήλη είναι τα αθροιστικά ποσοστά. Το ποσοστό της Αμερικής είναι 62.5%, της Ευρώπης 18%, άρα μαζί και τα δύο μας δίνουν το 80.5%. Αν προσθέσουμε και της Ιαπωνίας τότε φτάνουμε το 100%.

4.3 Ιστογράμματα

Πριν μιλήσουμε για τον τρόπο εμφάνισης των ιστογραμμάτων στο SPSS, καλό θα είναι να αναφέρουμε κάποια πράγματα για τα ιστογράμματα. Έστω ότι έχουμε τιμές από μία ποσοτική μεταβλητή. Αν το πλήθος των τιμών είναι πολύ μεγάλο μπορούμε να τις απεικονίσουμε διαγραμματικά με το ιστόγραμμα συχνοτήτων. Στον οριζόντιο άξονα τοποθετούνται οι κλάσεις των τιμών (ή αλλιώς οι ομάδες των τιμών, τις οποίες έχουμε κατηγοριοποιήσει). Στον κάθετο άξονα τοποθετούνται οι συχνότητες εμφάνισης των τιμών, που είναι ομαδοποιημένες. Με αυτόν το τον τρόπο σχηματίζουμε ορθογώνια, το μήκος των οποίων είναι ίσο με το εύρος των τιμών που έχουν συμπεριληφθείς στο κάθε ιστόγραμμα. Τα ορθογώνια είναι "κολλημένα" το ένα στο άλλο. Ενώνοντας τώρα το μέσο της πάνω πλευράς όλων των ορθογωνίων με μία γραμμή καταλήγουμε στο πολύγωνο συχνοτήτων. Καθώς τώρα ο αριθμός των κλάσεων τείνει στο άπειρο, η πολυγωνική γραμμή γίνεται με τη σειρά της ομαλή καταλήγοντας στη γραμμή που ονομάζεται καμπύλη συχνοτήτων.

Η πιο γνωστή κατανομή αλλά και η πιο βολική (και χρήσιμη) είναι η κανονική κατανομή. Στην περίπτωση αυτή η καμπύλη συχνοτήτων των δεδομένων σχηματίζει μία "καμπάνα". Η κατανομή αυτή εξετάστηκε πάρα πολύ από το Γερμανό μαθηματικό Carl Friedrich Gauss, γι' αυτό και μερικές φορές συναντάται με το όνομα κατανομή Gauss ή Γκαουσιανή κατανομή. Η κατανομή έχει τη μορφή που παρουσιάζεται παρακάτω στην εικόνα 23.

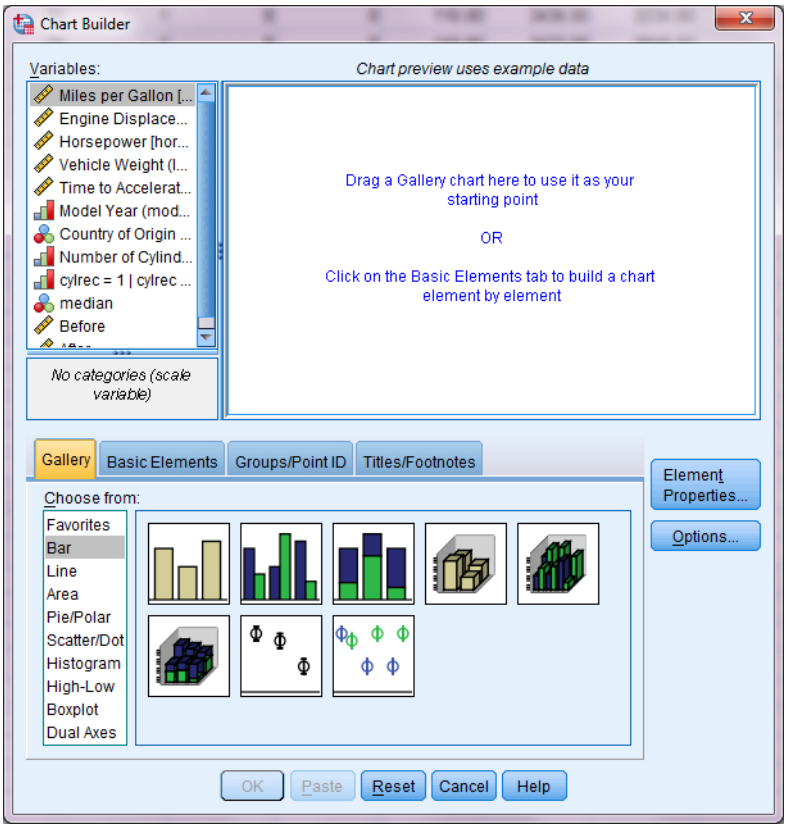


Εικόνα 23: Η κανονική κατανομή.

Η κανονική κατανομή είναι συμμετρική και μεσόκυρτη κατανομή, άρα ισχύει ότι η διάμεσος, η επικρατούσα τιμή και η μέση της τιμή ταυτίζονται. Επίσης μία άλλη χρήσιμη ιδιότητα της κανονικής κατανομής η οποία ισχύει και για άλλες μη κανονικές συμμετρικές κατανομές είναι η εξής: το 68% περίπου των παρατηρήσεων βρίσκεται στο διάστημα (μ-σ , μ+σ), το 95% περίπου των παρατηρήσεων βρίσκεται στο διάστημα (μ-2σ , μ+2σ) και το 99.7% περίπου των παρατηρήσεων βρίσκεται στο διάστημα (μ-3σ , μ+3σ). Με μ συμβολίζουμε το μέσο και σ την τυπική απόκλιση της κατανομής.

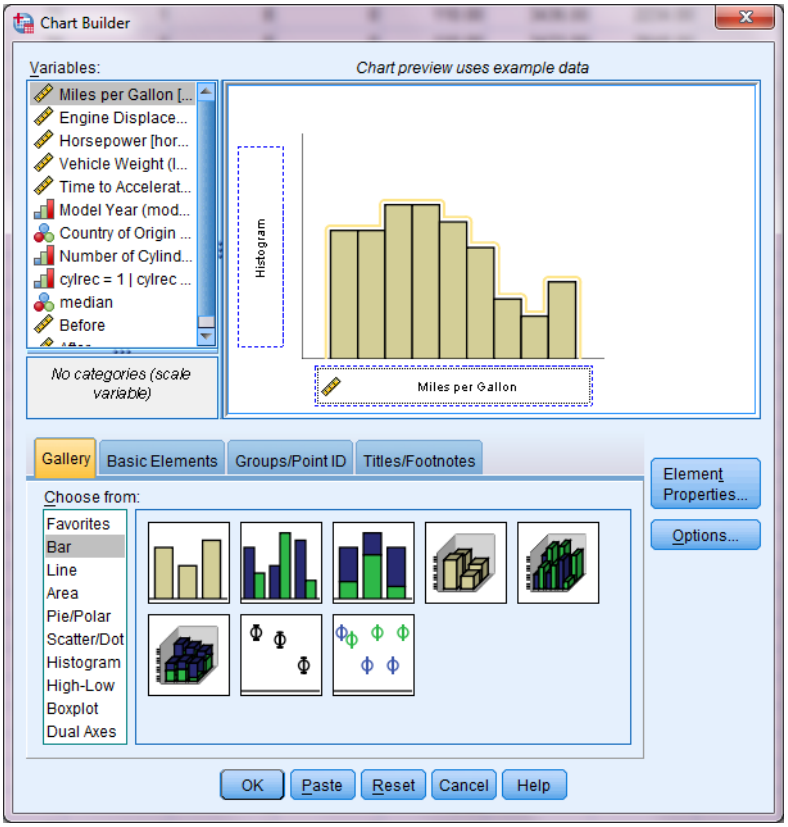
Αν τώρα υποθέσουμε ότι έχουμε μία ποιοτική μεταβλητή ή μία ποσοτική μεταβλητή με μικρό εύρος διακριτών τιμών ή με λίγες κλάσεις ομαδοποιημένων τιμών τότε μπορούμε να χρησιμοποιήσουμε το κυκλικό διάγραμμα (ή διάγραμμα πίτας). Το κυκλικό διάγραμμα χρησιμοποιείται όταν έχουμε ποιοτικές μεταβλητές για να απεικονίσουμε τις συχνότητες εμφάνισης των κατηγοριών ή το ποσοστό εμφάνισης που αντιστοιχεί σε κάθε κατηγορία μίας ποιοτικής μεταβλητής. Η κατασκευή του είναι απλή, διαιρούμε τη συχνότητα εμφάνισης μίας κατηγορίας της ποιοτικής μεταβλητής με το άθροισμα των συχνοτήτων όλων των κατηγοριών της

μεταβλητής και πολλαπλασιάζουμε το 360^o. Με αυτόν τον τρόπο καθορίζουμε τις μοίρες της κάθε "φέτας" στο διάγραμμα που αντιστοιχεί σε κάθε κατηγορία. Αν για παράδειγμα μία κατηγορία μίας ποιοτικής μεταβλητής εμφανίζεται σε ένα ποσοστό 50%, το κομμάτι της "πίτας" που "ανήκει" σε αυτήν την κατηγορία είναι ίσο με $50\%X360^o=180^o$. Ας δούμε όμως τώρα πως κατασκευάζουμε ιστογράμματα στο SPSS. Η επιλογή **Graphs** έχει δύο υποεπιλογές μέσω των οποίων μπορούμε να κατασκευάσουμε ένα ιστόγραμμα συχνοτήτων. Ας δούμε την πρώτη επιλογή. Πατάμε **Graphs→Chart Builder** και θα εμφανιστεί το παράθυρο της εικόνας 24.



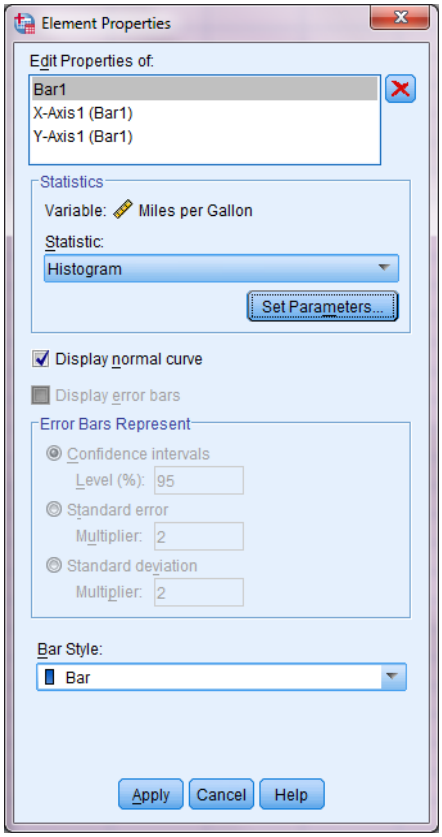
Εικόνα 24

Οι μεταβλητές που φαίνονται στο δεξιό κουτάκι είναι από τα δεδομένα που αφορούν σε μετρήσεις αυτοκινήτων και είναι διαθέσιμα από το αρχείο του SPSS. Πρώτα πρέπει να επιλέξουμε τον τύπο γραφήματος που θέλουμε από την επιλογή **Gallery**. Εκεί εμείς θα επιλέξουμε **Histogram** και από αυτά την πρώτη επιλογή αριστερά. Έπειτα θα επιλέξουμε τη μεταβλητή που θέλουμε και θα τη σύρουμε με το ποντίκι στον **X-axis?** (οριζόντιος άξονας). Το παράθυρο της εικόνας 24β θα ανοίξει δίπλα από το παράθυρο της εικόνας 24α. Το παράθυρο αυτό είναι στην ουσία το παράθυρο των ιδιοτήτων του διαγράμματος (επιλογή **Element Properties** που εμφανίζεται δεξιά του παραθύρου της εικόνας 24α). Στο παράθυρο αυτό θα επιλέξουμε να εμφανιστεί η καμπύλη της κανονικής κατανομής. Αυτό θα βοηθήσει στο να δούμε γραφικά πόσο μακριά είμαστε από την κανονική κατανομή για αυτή τη μεταβλητή. Επιλέγουμε **Display normal curve** στο παράθυρο αυτό και μετά **Apply**.

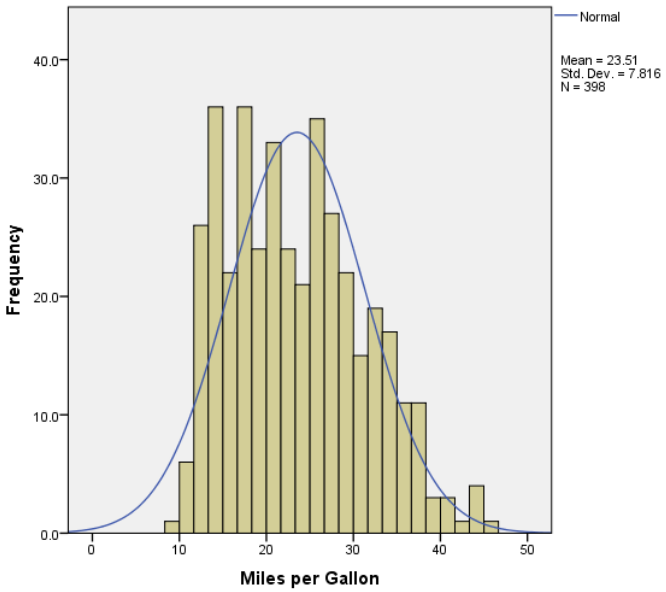


Εικόνα 24α

Στην επιλογή **Titles/Footnotes** μπορούμε να δώσουμε τίτλο/ υπότιτλο/ υποσημείωση στο διάγραμμα μας. Πατώντας **OK** θα εμφανιστεί το γράφημα του σχήματος 4.



Εικόνα 24β

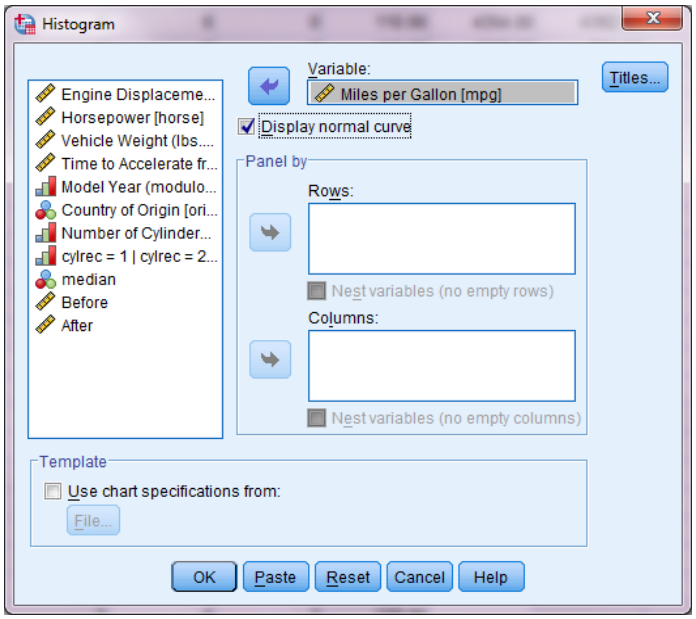


Σχήμα 4: Ιστόγραμμα συχνοτήτων.

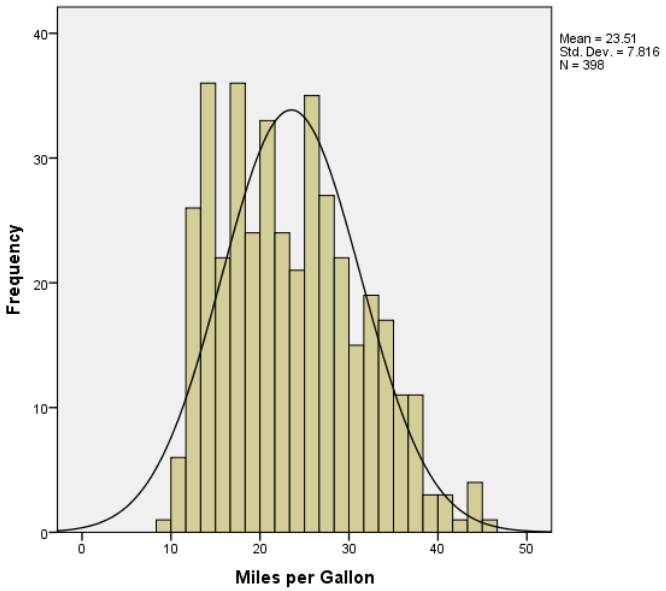
Ένας άλλος τρόπος κατασκευής του ιστογράμματος συχνοτήτων είναι πάλι από την επιλογή **Graphs**. Τώρα όμως θα επιλέξουμε **Legacy Dialogs** και μετά **Histogram** για να εμφανιστεί το παράθυρο της εικόνας 25. Περνάμε τη μεταβλητή της οποίας το ιστόγραμμα θέλουμε να κατασκευάσουμε στο δεξιό ορθογώνιο

κουτάκι. Ακριβώς από κάτω μπορούμε να επιλέξουμε αν θέλουμε να εμφανιστεί η γραμμή της κανονικής κατανομής με την επιλογή **Display normal curve.**

Και σε αυτήν τη περίπτωση όμως δεν μπορούμε να κατασκευάσουμε δύο ιστογράμματα με μία επιλογή μόνο. Πατώντας **OK** θα εμφανιστεί το ιστόγραμμα του σχήματος 5. Το διάγραμμα είναι ίδιο με αυτό του σχήματος 4, εκτός από μία διαφορά. Η γραμμή της κανονικής κατανομής είναι μαύρη, όχι μπλε.

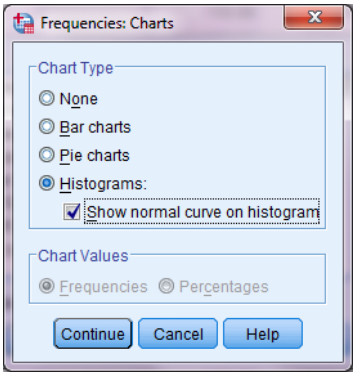


Εικόνα 25



Σχήμα 5: Ιστόγραμμα συχνοτήτων.

Ο τρίτος και ίσως ο πιο βολικός τρόπος κατασκευής ιστογραμμάτων συχνοτήτων παρέχεται από την επιλογή **Analyze**. Πατάμε **Analyze→Descriptive Statistics→Frequencies** και θα εμφανιστεί το παράθυρο της εικόνας 21. Εάν σε αυτό το παράθυρο επιλέξουμε την επιλογή **Charts** θα εμφανιστεί ένα άλλο παράθυρο, αυτό της εικόνας 26.

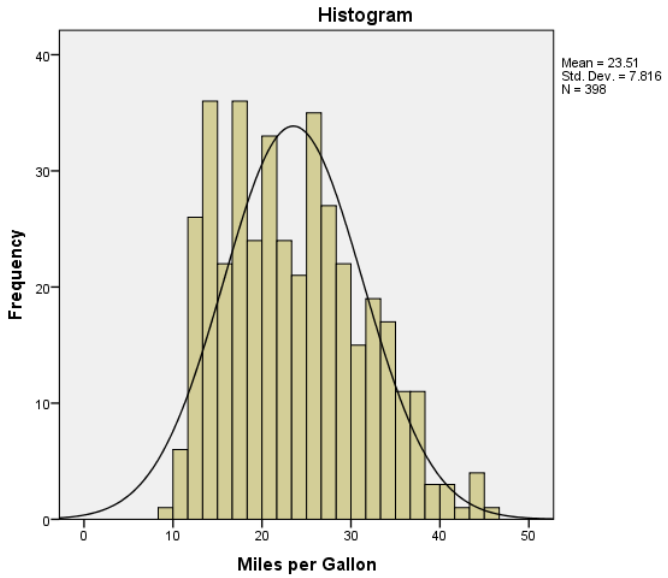


Εικόνα 26

Στο παράθυρο της εικόνας 21 επιλέγαμε για ποιες μεταβλητές θέλουμε να εμφανιστούν τα περιγραφικά μέτρα. Επομένως πρέπει να έχουμε περάσει τουλάχιστον μία μεταβλητή στο δεξιό κουτάκι για να εμφανιστεί το ιστόγραμμα συχνοτήτων της. Απλά επιλέγουμε την επιλογή **Histograms** και "κλικάρουμε" την επιλογή **With normal curve** αν επιθυμούμε την εμφάνιση της καμπύλης γραμμής της κανονικής κατανομής. Μπορούμε να μην επιλέξουμε κανένα περιγραφικό μέτρο να εμφανιστεί. Σε αυτήν την περίπτωση θα εμφανιστεί μόνο το ιστόγραμμα συχνοτήτων.

Ένα πλεονέκτημα της κατασκευής ιστογραμμάτων συχνοτήτων από αυτήν την επιλογή είναι ότι μπορούμε να "ζητήσουμε" την εμφάνιση ιστογραμμάτων συχνοτήτων για περισσότερες από μία μεταβλητές. Πατώντας **Continue** λοιπόν γυρίζουμε στο παράθυρο της εικόνας 21 και μετά **OK** για να εμφανιστεί το σχήμα 6. Έχοντας από-επιλέξει την επιλογή εμφάνισης του πίνακα συχνοτήτων (όπως και προηγουμένως) αφού έχουμε συνεχή μεταβλητή.

Παρατηρούμε ότι σε αυτό το διάγραμμα έχουν εμφανιστεί κάτω δεξιά η μέση τιμή των τιμών της μεταβλητής, η τυπική απόκλιση και το πλήθος των τιμών. Το ιστόγραμμα με άλλα λόγια είναι ακριβώς το ίδιο, σε αντίθεση με αυτό που παράγεται με τον πρώτο τρόπο και με το οποίο συναντάμε κάποιες μικρές διαφορές ως προς την εμφάνιση. Πάνω από το ιστόγραμμα εμφανίστηκε και ένας πίνακας που μας δίνει κάποιες πληροφορίες για το πλήθος των τιμών. Οι τιμές που χρησιμοποιήθηκαν στην κατασκευή του παραπάνω ιστογράμματος είναι ίσες με 398. Έχουμε και 8 εκλιπούσες τιμές (το μέγεθος του δείγματος είναι ίσο με 406). Επίσης να αναφέρουμε ότι το ιστόγραμμα που κατασκευάστηκε με τον πρώτο τρόπο έχει τον κατακόρυφο άξονα των συχνοτήτων "κολλημένο" στο ιστόγραμμα. Επίσης το πλαίσιο στη δεξιά πλευρά "ακουμπάει" το ιστόγραμμα. Στο δεύτερο και στο τρίτο ιστόγραμμα βλέπουμε ότι κάτι τέτοιο δεν ισχύει. Η καμπύλη γραμμή της κανονικότητας στο πρώτο ιστόγραμμα δεν ξεκινάει από το μηδέν και ούτε καταλήγει στον οριζόντιο άξονα. Στο τρίτο ιστόγραμμα όμως βλέπουμε ότι το πλαίσιο είναι μεγαλύτερο από το ιστόγραμμα επιτρέποντας στην καμπύλη γραμμή της κανονικής κατανομής να ξεκινήσει από τον οριζόντιο άξονα και να καταλήξει σε αυτόν.



Σχήμα 6: Ιστόγραμμα συχνοτήτων.

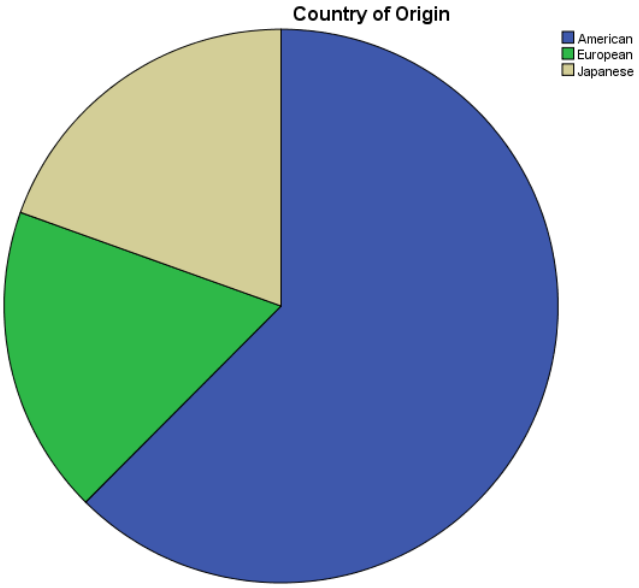
4.4 Κυκλικά διαγράμματα

Για να κατασκευάσουμε ένα κυκλικό διάγραμμα (για κατηγορικές μεταβλητές ως επί των πλείστων) έχουμε πάλι πολλές επιλογές από το μενού. Εμείς όμως θα εξηγήσουμε μία από αυτές, την πιο χρήσιμη. Επιλέγοντας να εμφανιστεί το παράθυρο της εικόνας 26 θα "κλικάρουμε" την επιλογή **Pie charts**. Στο κάτω μέρος μπορούμε να επιλέξουμε αν θέλουμε να εμφανιστούν οι συχνότητες ή τα ποσοστά των χωρών. Εμείς επιλέξαμε τα ποσοστά διότι είναι προτιμότερο να έχουμε την ποσοστιαία κατανομή από την απόλυτη. Φανταστείτε στις εκλογές να έδιναν τα αποτελέσματα σε αριθμό ψήφων αντί για ποσοστά!

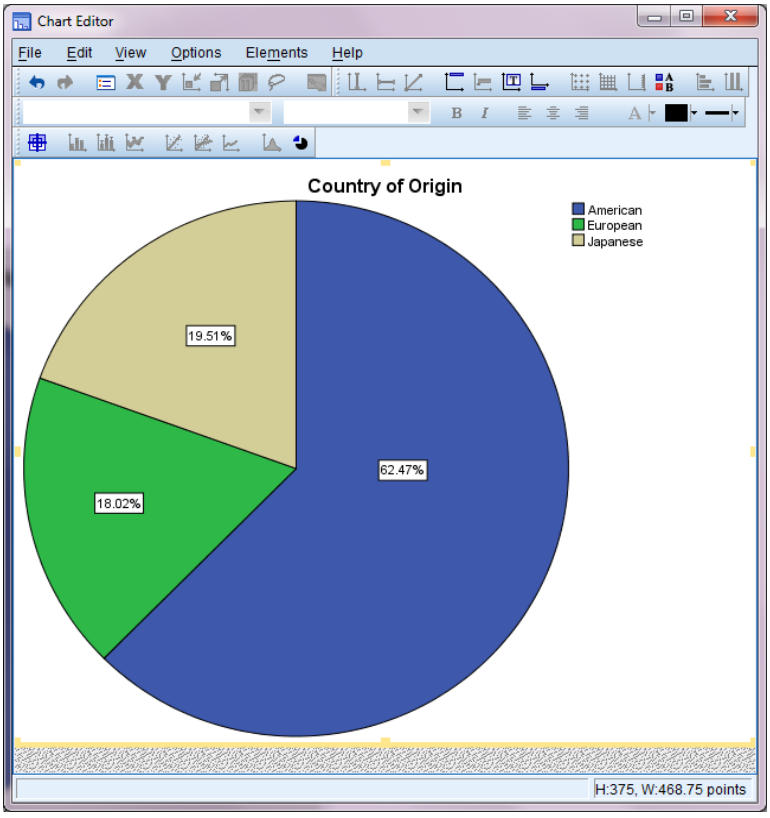
Αν στο παράθυρο της εικόνας 21 επιλέξουμε **Display frequency tables** μαζί με το διάγραμμα θα εμφανιστεί και ο πίνακας συχνοτήτων και σχετικών συχνοτήτων για κάθε κατηγορία της ποιοτικής μεταβλητής που έχουμε επιλέξει. Εμείς επιλέξαμε τη μεταβλητή έτος κατασκευής για τα δεδομένα που αφορούν σε αυτοκίνητα. Το κυκλικό διάγραμμα που εμφανίστηκε ήταν αυτό του σχήματος 6. Παρατηρούμε ότι όπως και στο ιστόγραμμα που κατασκευάστηκε από αυτήν την επιλογή, εμφανίστηκε ένας πίνακας πάνω από το διάγραμμα που μας ενημερώνει για το πλήθος των τιμών αυτής της μεταβλητής που συμμετέχουν στην κατασκευή του διαγράμματος και το πλήθος των τιμών που έχουν "χαθεί". Στην περίπτωση του έτους κατασκευής έχει "χαθεί" το έτος κατασκευής για ένα αυτοκίνητο. Με αυτήν την επιλογή μπορούμε όπως και προηγουμένως να επιλέξουμε την κατασκευή κυκλικών διαγραμμάτων για πολλές κατηγορικές μεταβλητές μαζί. Στο υπόμνημα που εμφανίζεται δεξιά του διαγράμματος και εξηγεί το κάθε χρώμα σε ποιο έτος κατασκευής αντιστοιχεί έχει συμπεριληφθεί στο διάγραμμα μία εκλιπούσα τιμή (Missing).

Παρατηρούμε όμως ότι στο διάγραμμα δεν εμφανίστηκαν τα ποσοστά που ζητήσαμε για κάθε χώρα. Για να εμφανιστούν υπάρχουν δύο επιλογές. Η μία είναι να κάνουμε διπλό "κλικ" με τον κέρσορα πάνω στο διάγραμμα. Ο δεύτερος είναι να κάνουμε αριστερό "κλικ" με το ποντίκι πάνω στο διάγραμμα και να επιλέξουμε **Edit Content→In Separate Window** και θα εμφανιστούν τα παράθυρα των εικόνων 27α και 27β. Στο παράθυρο της εικόνας 27α θα επιλέξουμε **Elements→Show Data**

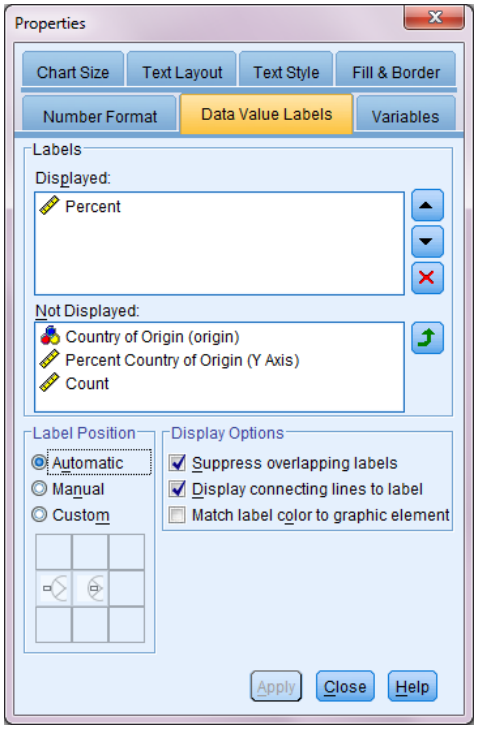
Labels και μετά κλείνουμε το παράθυρο θα εμφανιστεί το διάγραμμα του σχήματος 7. Στην ουσία στο ίδιο γράφημα εμφανίζονται τώρα τα ποσοστά.



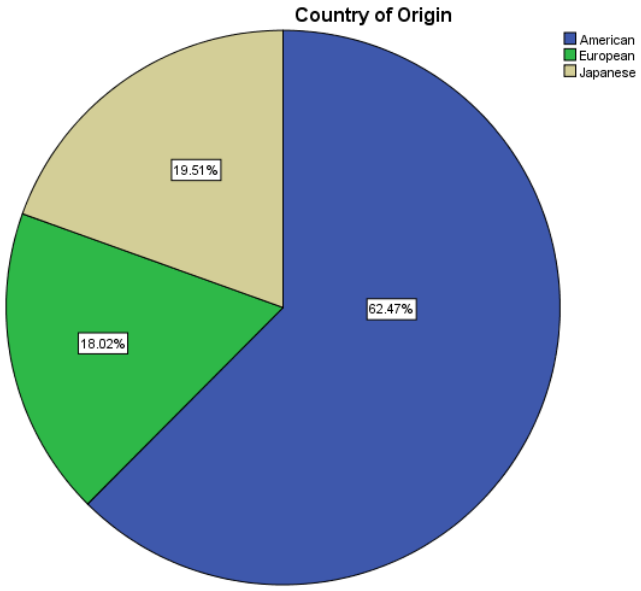
Σχήμα 7: Κυκλικό διάγραμμα.



Εικόνα 27α

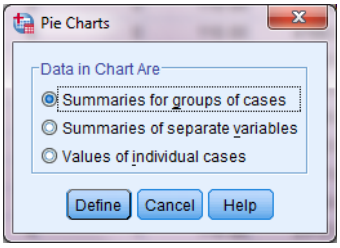


Εικόνα 27β



Σχήμα 8: Κυκλικό διάγραμμα με ποσοστά.

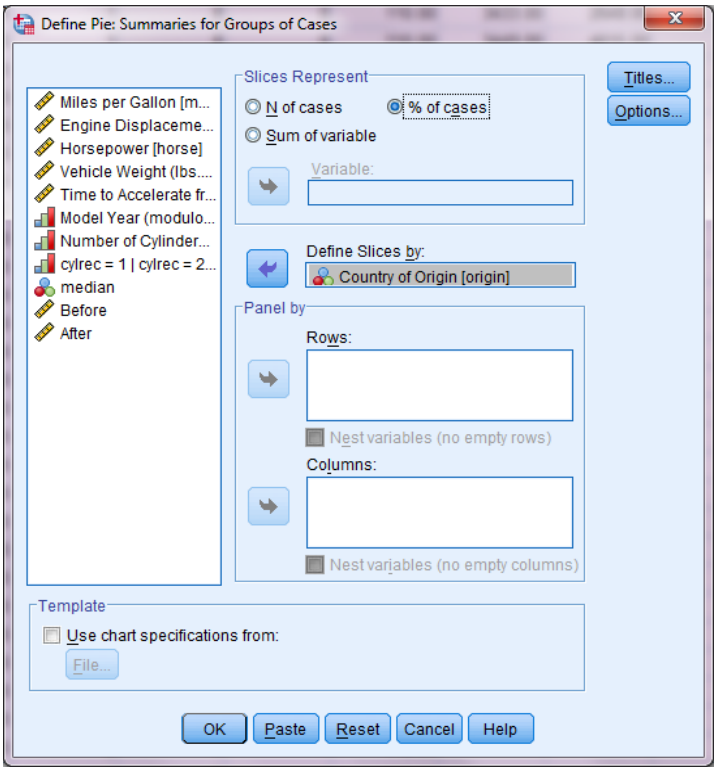
Από ότι μπορούμε να δούμε οι εκλιπούσες τιμές έχουν αγνοηθεί. Αν θέλουμε να κατασκευάσουμε ένα κυκλικό διάγραμμα με έναν άλλο τρόπο τότε μπορούμε να πατήσουμε **Graphs→Legacy Dialogs→Pie** και στο παράθυρο της εικόνας 26 που θα εμφανιστεί θα επιλέξουμε **Summaries of group of cases** και μετά **Define.**



Εικόνα 28

Στο παράθυρο που θα εμφανιστεί (εικόνα 29) θα περάσουμε την ποιοτική μεταβλητή για την οποία θέλουμε να κατασκευάσουμε το κυκλικό διάγραμμα στο λευκό κουτάκι **Define Slices by**. Πάνω στην επιλογή **Slices Represent** επιλέξαμε **% of cases** για να εμφανιστούν πάλι τα ποσοστά στο διάγραμμα.

Κάτω δεξιά στο παράθυρο μας δίνεται η επιλογή να τοποθετήσουμε τίτλους στο διάγραμμα (επιλογή **Titles**). Ακριβώς από κάτω υπάρχει η επιλογή **Options**. Πατώντας την επιλογή αυτή μπορούμε να επιλέξουμε αν θέλουμε οι εκλιπούσες τιμές να συμπεριληφθούν στο διάγραμμα ή όχι. Είναι ήδη προεπιλεγμένη (από το πακέτο) η επιλογή να μη συμπεριλαμβάνονται στο διάγραμμα. Οπότε πατώντας **OK** θα εμφανιστεί το διάγραμμα του σχήματος 7. Αν θέλουμε τα ποσοστά να εμφανίζονται πρέπει να ξανακάνουμε την ίδια διαδικασία που περιγράψαμε προηγουμένως.



Εικόνα 29

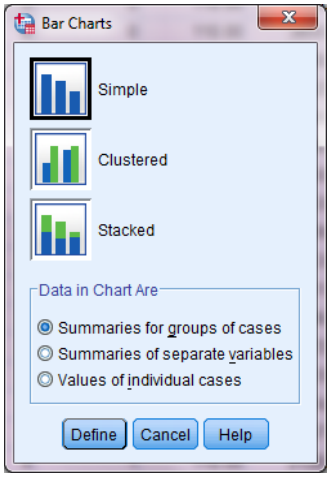
Είδαμε έναν τρόπο κατασκευής κυκλικών διαγραμμάτων ή διαγραμμάτων πίτας όπως αλλιώς λέγονται. Το μειονέκτημα τους όμως είναι ότι δεν εμφανίζουν τα ποσοστά των τιμών των μεταβλητών και πρέπει να τα τοποθετήσουμε μετά.

Επιλέγοντας **Graphs→Chart Builder** εμφανίζονται τα παράθυρα των εικόνων 24α και 24β. Στην επιλογή **Gallery** θα επιλέξουμε **Pie/Polar** και θα το

σύρουμε πάνω στο λευκό κουτάκι. Θα επιλέξουμε την κατηγορικά μεταβλητή της οποίας τις τιμές θέλουμε να εμφανιστούν στο κυκλικό διάγραμμα και θα τη "σύρουμε" με το ποντίκι στο λευκό κουτάκι **Slice by**. Στο παράθυρο της εικόνας 24β στην επιλογή Statistic θα επιλέξουμε **Percentage(?)** και μετά **Apply**. Έπειτα στο παράθυρο της εικόνας 24α θα πατήσουμε **OK** και το αποτέλεσμα θα είναι το ίδιο με αυτό του σχήματος 7. Ο τρόπος για να εμφανίσουμε τα ποσοστά είναι ήδη γνωστός.

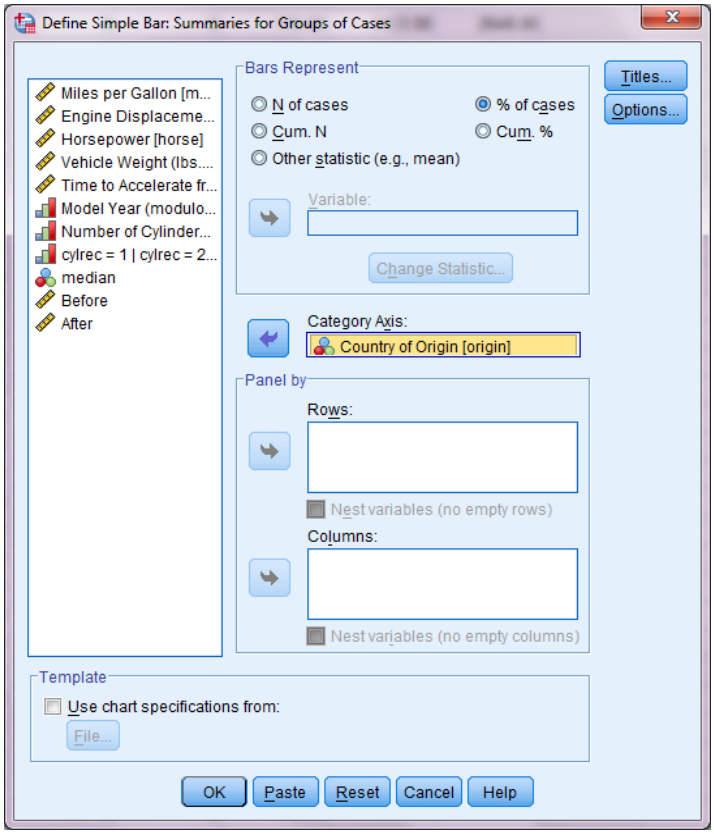
4.5 Ραβδογράμματα

Ας υποθέσουμε πάλι ότι έχουμε κατηγορικές μεταβλητές και θέλουμε να δούμε το κυκλικό διάγραμμα τι μορφή θα είχε αν ήταν σε στήλες. Θυμίζει λίγο το ιστόγραμμα ως προς τα ορθογώνια (ράβδους), με μία σημαντική διαφορά ότι τα ορθογώνια δεν είναι "κολλημένα" μεταξύ τους. Υπάρχουν πολλές επιλογές για να κατασκευάσουμε ραβδογράμματα. Εμείς θα παρουσιάσουμε μία από αυτές (ένας άλλος τρόπος είναι από το παράθυρο της εικόνας 26). Πατώντας **Graphs→Legacy Dialogs→Bar** θα εμφανιστεί το παράθυρο της εικόνας 30. Εκεί θα επιλέξουμε το πρώτο εικονίδιο (**Simple**) και μετά **Define** για να μεταβούμε στο παράθυρο της εικόνας 31 (σχεδόν ίδιο με αυτό της εικόνας 29). Στο παράθυρο της εικόνας 31 θα περάσουμε την ποιοτική μεταβλητή για την οποία θα κατασκευάσουμε το ραβδόγραμμα στο λευκό κουτάκι **Category Axis:**. Από το παράθυρο 31 μας δίνεται η δυνατότητα εμφάνισης των ποσοστών αντί για των συχνοτήτων. Το διάγραμμα είναι το ίδιο, απλά στον κατακόρυφο άξονα αντί για τις συχνότητες θα βρίσκονται τα ποσοστά των τιμών (ή επιπέδων) της κατηγορικής μεταβλητής.

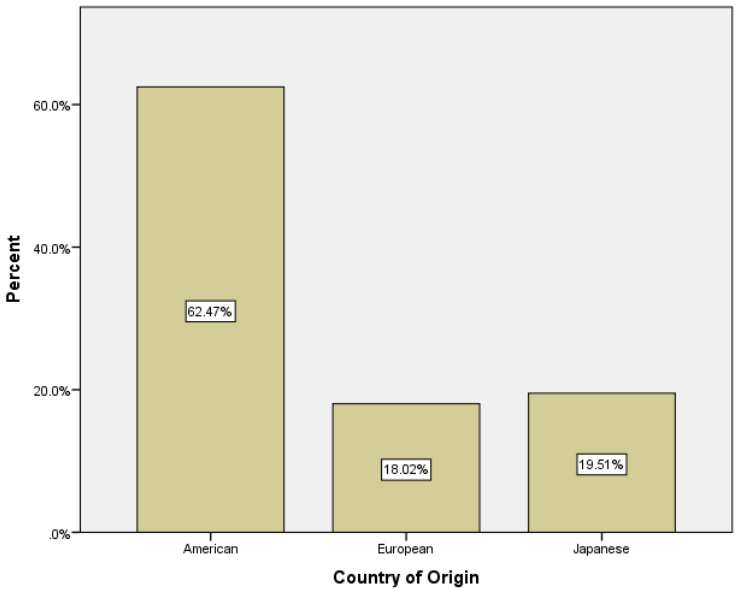


Εικόνα 30

Η επιλογή **Titles** μας δίνει τη δυνατότητα να τοποθετήσουμε τίτλους στο διάγραμμα και η επιλογή **Options** μας επιτρέπει να εμφανίσουμε αν θέλουμε και ένα ακόμα ραβδόγραμμα που θα περιέχει το πλήθος των χαμένων τιμών. Εμείς επιλέξαμε τη μεταβλητή που εκφράζει τη χώρα προέλευσης των αυτοκινήτων. Πατώντας λοιπόν **OK**, το αποτέλεσμα φαίνεται στο σχήμα 8. Το διάγραμμα στο σχήμα 8 εμφανίζει και τα ποσοστά των τιμών. Η διαδικασία για να γίνει αυτό έχει περιγραφεί προηγουμένως.



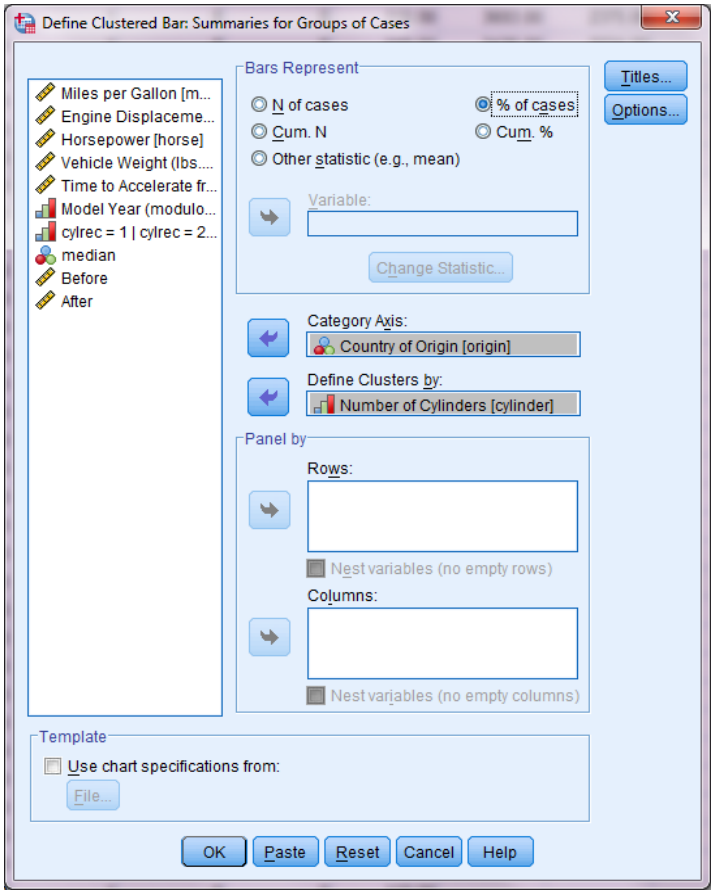
Εικόνα 31



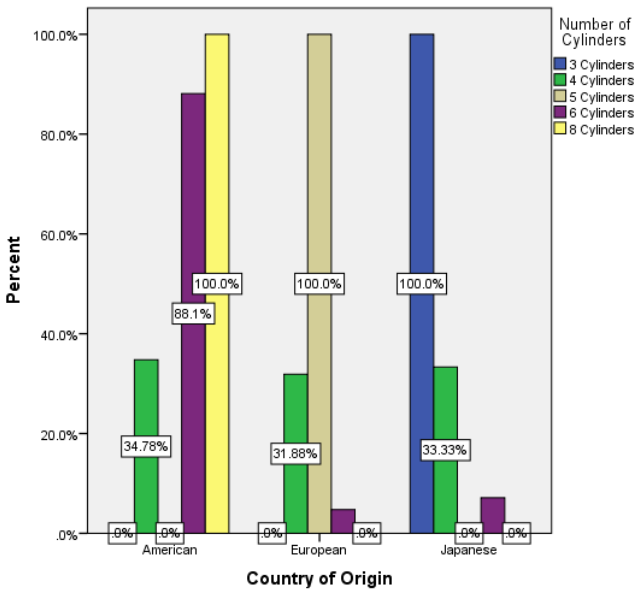
Σχήμα 9: Ραβδόγραμμα συχνοτήτων.

Στην περίπτωση που θέλουμε να δούμε ένα ραβδόγραμμα για δύο ποιοτικές μεταβλητές μαζί θα επιλέξουμε το δεύτερο εικονίδιο στην εικόνα 30 (**Clustered**) και μετά **Continue** για να οδηγηθούμε στο παράθυρο της εικόνας 32, το οποίο είναι ελαφρώς διαφορετικό από αυτό της εικόνας 31. Οι ποιοτικές μεταβλητές θα περαστούν στα λευκά κουτάκια **Category Axis:** και **Define Clusters by:**. Εμείς

περάσαμε στο πρώτο κουτάκι που αφορά στη χώρα προέλευσης των αυτοκινήτων και στο δεύτερο κουτάκι τη μεταβλητής που δηλώνει τον αριθμό των κυλίνδρων στα αυτοκίνητα. Το σχήμα αναπαριστά για κάθε χώρα προέλευσης το πλήθος των αυτοκινήτων ανάλογα με τον αριθμό των κυλίνδρων τους. Αν θέλαμε για τον κάθε αριθμό κυλίνδρων να εμφανίζεται το πλήθος των αυτοκινήτων ανάλογα με τη χώρα προέλευσης τους, απλά θα αλλάζαμε τη σειρά με την οποία περάσαμε τις μεταβλητές στα λευκά κουτάκια στο παράθυρο της εικόνας 32.



Εικόνα 32



Σχήμα 10: Ραβδόγραμμα δύο ποιοτικών μεταβλητών.

5.1 Έλεγχος κανονικότητας

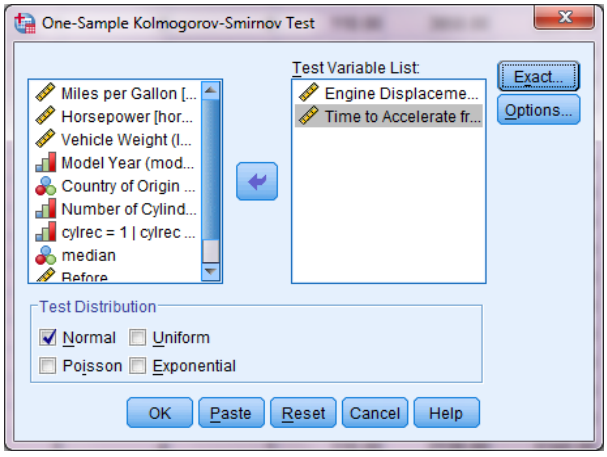
Αναφέραμε προηγουμένως πως όταν το ιστόγραμμα συχνοτήτων των ποσοτικών μεταβλητών έχει το σχήμα "καμπάνας", τότε λέμε ότι τα δεδομένα ακολουθούν την κανονική κατανομή ή κατανέμονται κανονικά. Το ιστόγραμμα όμως δεν είναι "ικανό" να μας απαντήσει στη ερώτηση αν είναι κανονικά τα δεδομένα ή αν προέρχονται από μία κανονική κατανομή με ένα μέσο και μία διακύμανση.

Μπορούμε να κατασκευάσουμε δύο γραφήματα με το SPSS, το **P-P Plot** και το **Q-Q Plot** (Επιλέγοντας **Analyze→Descriptive Statistics→P-P Plots** ή **Q-Q Plots**). Με αυτά τα γραφήματα ελέγχουμε οπτικά την ύπαρξη κανονικότητας στα δεδομένα. Όσο πιο κοντά στην ευθεία είναι τα σημεία του σχήματος τόσο πιο πολλές είναι οι ενδείξεις ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Το μάτι όμως πάλι μπορεί να "πέσει έξω" και να ξεγελαστούμε. Για αυτό το λόγο καταφεύγουμε σε έλεγχο κανονικότητας για να απαντήσουμε στην προηγούμενη ερώτηση.

Ο έλεγχος κανονικότητας υπάγεται σε μία ευρύτερη οικογένεια ελέγχων, τη λεγόμενη «έλεγχοι υποθέσεων». Όταν ακούμε για ελέγχους υποθέσεων μας έρχονται πολλά πράγματα στο μυαλό. Κάποια από αυτά είναι η μηδενική υπόθεση (**Null Hypothesis** ή **Ho**), η εναλλακτική υπόθεση (**Alternative Hypothesis** ή $\mathbf{H_1}$), το επίπεδο στατιστικής σημαντικότητας (**α**) και το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας (**p-value** ή **Significance**). Οι υποθέσεις είναι της ακόλουθης μορφής:

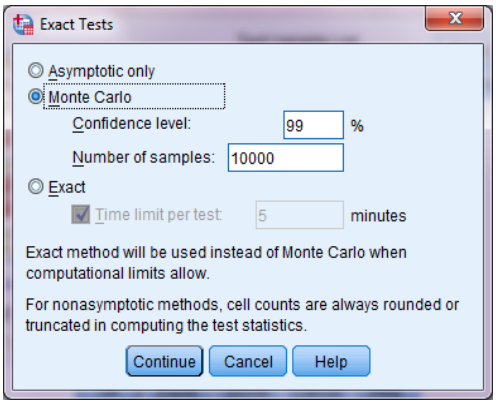
Ho: Η κατανομή των δεδομένων δε διαφέρει από την κανονική κατανομή
H_1: Η κατανομή των δεδομένων διαφέρει από την κανονική κατανομή

Για τη διεξαγωγή των ελέγχων υποθέσεων χρησιμοποιούνται κάποιοι μαθηματικοί τύποι, που καλούνται ελεγχοσυναρτήσεις. Με βάση το αποτέλεσμα τους οδηγούμαστε στο συμπέρασμα ότι η μηδενική υπόθεση απορρίπτεται ή όχι. Στη συγκεκριμένη περίπτωση η μηδενική υπόθεση την οποία θέλουμε να ελέγξουμε είναι ότι τα δεδομένα ακολουθούν την κανονική ή ότι προέρχονται από ένα πληθυσμό που ακολουθεί την κανονική κατανομή. Η εναλλακτική είναι ότι τα δεδομένα δεν ακολουθούν την κανονική κατανομή. Το επίπεδο στατιστικής σημαντικότητας ορίζεται συνήθως ίσο με 0.05 ή 5%. Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας (p-value) ορίζεται ως η πιθανότητα η τιμή του ελέγχου (ελεγχοσυνάρτησης) να πάρει μία τιμή τόσο ακραία ή περισσότερο ακραία από αυτή που πήρε στο συγκεκριμένο δείγμα κάτω από τη μηδενική υπόθεση. Αν η p-value είναι μικρότερη του 0.05, τότε λέμε ότι η μηδενική υπόθεση απορρίπτεται. Αν η p-value είναι μεγαλύτερη ή ίση του 0.05, τότε λέμε ότι η μηδενική υπόθεση δεν απορρίπτεται. Το SPSS εμφανίζει τις τιμές των παρατηρηθέντων επιπέδων στατιστικής σημαντικότητας και τις ονομάζει (**Asymptotic**) **Significances.** Ο λόγος που χρειαζόμαστε την κανονικότητα των δεδομένων, είναι για να έχουν ισχύ κάποιες στατιστικές τεχνικές που θα χρησιμοποιήσουμε όπως οι έλεγχοι υποθέσεων για τους μέσους, η γραμμική παλινδρόμηση, η ανάλυση διακύμανσης κ.ά. Ας δούμε τώρα στο SPSS πως θα διεξάγουμε ελέγχους κανονικότητας. Πατάμε **Analyze→Non Parametric Tests→Legacy Dialogs→1-Sample K-S** και εμφανίζεται το παράθυρο της εικόνας 33.



Εικόνα 33

Όπως βλέπετε περάσαμε δύο μεταβλητές στο δεξιό κουτάκι, για τις οποίες θα ελέγξουμε αν οι τιμές τους ακολουθούν την κανονική κατανομή. Παρατηρούμε ότι η επιλογή για τον έλεγχο κανονικότητας είναι ήδη προεπιλεγμένος από το SPSS (**Normal**). Επιλέγοντας **Options** εμφανίζεται ένα άλλο παράθυρο στο οποίο μπορούμε να επιλέξουμε και την εμφάνιση ενός πίνακα με κάποια περιγραφικά μέτρα που αφορούν αυτές τις μεταβλητές. Πατώντας **Exact** θα εμφανιστεί το παράθυρο της εικόνας 34. Το SPSS έχει ως προεπιλογή το **Asymptotic only**. Αυτό σημαίνει ότι θα διεξάγει το έλεγχο κανονικότητας των **Kolmogorov-Smirnov** όπως επιλέξαμε άλλωστε. Αν επιλέξουμε την επιλογή που βρίσκεται ακριβώς από κάτω, δηλαδή το **Monte Carlo**, θα ενεργοποιηθούν και τα επόμενα δύο λευκά κουτάκια, το **Confidence level** και το **Number of samples**. Με την επιλογή **Monte Carlo** "ζητάμε" από το SPSS να χρησιμοποιήσει και την τεχνική της προσομοίωσης για να κάνει τον έλεγχο της κανονικότητας. Δε θα επεκταθούμε περισσότερο στην τεχνική της προσομοίωσης, παρά μόνο θα πούμε ότι διεξάγει 10000 (προεπιλογή) ελέγχους κανονικότητας και για κάθε ένα υπολογίζει την p-value. Στο τέλος εμφανίζει το μέσο όρο αυτών των 10000 p-value και ένα 99% διάστημα εμπιστοσύνης για τον μέσο όρο αυτών των p-value βασισμένο προφανώς στις 10000 p-value. Το bootstrap δε χρειάζεται εδώ. Θα πούμε όμως πιο πολλά για τα διαστήματα εμπιστοσύνης παρακάτω.



Εικόνα 34

One-Sample Kolmogorov-Smirnov Test

			Engine Displacement (cu. inches)	Time to Accelerate from 0 to 60 mph (sec)
N			406	406
Normal Parameters[a,b]	Mean		194.04	15.50
	Std. Deviation		105.207	2.821
Most Extreme Differences	Absolute		.183	.047
	Positive		.183	.047
	Negative		-.114	-.032
Test Statistic			.183	.047
Asymp. Sig. (2-tailed)			**.000**[c]	**.030**[c]
Monte Carlo Sig. (2-tailed)	**Sig.**		**.000**[d]	**.310**[d]
	99% Confidence Interval	Lower Bound	.000	.298
		Upper Bound	.000	.322

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. Based on 10000 sampled tables with starting seed 624387341.

Σχήμα 11: Έλεγχος κανονικότητας.

Πατώντας **Continue** επιστρέφουμε στο αρχικό παράθυρο της εικόνας 33 και μετά **OK** για να εμφανιστεί το σχήμα 11 στο Output του SPSS.

Η υπόθεση την οποία θέλουμε να ελέγξουμε είναι ότι οι μεταβλητές ακολουθούν την κανονική κατανομή. Και για τις δύο μεταβλητές εμφανίζεται το μέγεθος του δείγματος (406). Από ότι φαίνεται για αυτές τις μεταβλητές δεν έχουμε εκλιπούσες τιμές. Εμφανίζονται επίσης ο μέσος και η τυπική απόκλιση για κάθε μεταβλητή. Για τον έλεγχο της κανονικότητας μας ενδιαφέρουν δύο τιμές, η **Asymp. Sig. (2-tailed)** και η **Monte Carlo Sig**. Πρόκειται για τις p-value που υπολογίζονται για κάθε μέθοδο ξεχωριστά. Ο έλεγχος των **Kolmogorov-Smirnov** είναι ένα απλά στη μία περίπτωση η p-value υπολογίζεται με βάση τη "συμβατική" μέθοδο, ενώ στην άλλη βασίζεται στην τεχνική **Monte Carlo**.

Παρατηρούμε για την πρώτη τιμή ότι οι p-value που υπολογίστηκαν και με τις δύο μεθόδους είναι ίσες με το μηδέν. Όπως αναφέραμε προηγουμένως αν η p-value είναι μικρότερη από το 0.05, τότε απορρίπτουμε την υπόθεση της κανονικότητας των δεδομένων. Άρα η υπόθεση ότι οι μετρήσεις που αφορούν τον κυβισμό των αυτοκινήτων, κατανέμονται κανονικά απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας α=0.05 ή α=5%. Ειδάλλως, μπορούμε να πούμε ότι υπάρχουν ενδείξεις ότι αυτές οι μετρήσεις δεν ακολουθούν την κανονική κατανομή.

Για τη δεύτερη μεταβλητή όμως, που είναι η επιτάχυνση των αυτοκινήτων, βλέπουμε ότι η p-value με το συμβατικό τρόπο (**Asymp. Sig. (2-tailed)**) είναι ίση με 0.326, ενώ η p-value που υπολογίστηκε με βάση την τεχνική της προσομοίωσης είναι ίση με 0.318. Και στις δύο περιπτώσεις δηλαδή τα παρατηρηθέντα επίπεδα στατιστικής σημαντικότητας (p-value) είναι μεγαλύτερα του 0.05. Αυτό σημαίνει ότι

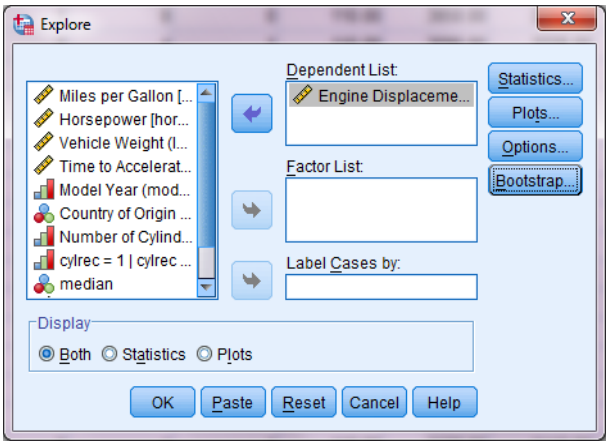
η υπόθεση της κανονικότητας για τις επιταχύνσεις των αυτοκινήτων δεν απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας α=0.05 ή α=5%. Με άλλα λόγια υπάρχουν ενδείξεις ότι οι μετρήσεις που αφορούν τις επιταχύνσεις των αυτοκινήτων ακολουθούν την κανονική κατανομή.

5.2 Διαστήματα εμπιστοσύνης

Πριν αναφερθούμε στη χρησιμότητα των διαστημάτων εμπιστοσύνης και πως κατασκευάζονται στο SPSS, ας δώσουμε το σωστό ορισμό τους. Στην προσπάθεια να εκτιμήσουμε την πραγματική τιμή του μέσου, χρησιμοποιούμε το μέσο ενός δείγματος. Στη συνέχεια κατασκευάζουμε ένα 95% διάστημα εμπιστοσύνης με βάση ένα μαθηματικό τύπο, ο οποίος είναι στην ουσία 2 τυπικά σφάλματα αριστερά και δεξιά της τιμής του μέσου που βρήκαμε για το δείγμα. Αν επαναλάβουμε τη δειγματοληψία ν φορές θα εκτιμήσουμε ν διαφορετικούς μέσους και προφανώς ν διαφορετικά (πολλά θα είναι αλληλοεπικαλυπτόμενα) διαστήματα εμπιστοσύνης.

Αυτό που ευελπιστούμε είναι ότι στο 95% των ν περιπτώσεων τα διαστήματα εμπιστοσύνης που υπολογίσαμε θα έχουν περικλείσει, ή "πιάσει", ή "χτυπήσει" την τιμή του πραγματικού μέσου. Επομένως, αν για παράδειγμα "παίρναμε" κάθε φορά 100 δείγματα από έναν πληθυσμό και κατασκευάζαμε 100 διαστήματα εμπιστοσύνης για τη μέση τιμή μίας μεταβλητής και επαναλαμβάναμε τη διαδικασία άπειρες τιμές, κατά μέσο όρο στο 95% των περιπτώσεων θα είχαμε φτιάξει διαστήματα εμπιστοσύνης που θα είχαν "πιάσει" τον πραγματικό μέσο του πληθυσμού. Το 95% θα το ονομάζουμε βαθμό ή επίπεδο εμπιστοσύνης. Το υπόλοιπο 5% είναι αυτό που έχουμε ήδη ορίσει επίπεδο στατιστικής σημαντικότητας.

Για να κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για τη μέση τιμή μίας μεταβλητής εργαζόμαστε ως εξής: πατάμε **Analyze→Descriptive Statistics→Explore** και θα εμφανιστεί το παράθυρο της εικόνας 35. Από την επιλογή **Plots** μπορούμε να επιλέξουμε αν θέλουμε να εμφανιστεί ένα ιστόγραμμα των ή της μεταβλητής που θα περάσουμε στο άνω λευκό κουτάκι (**Dependent List:**). Κάτω αριστερά (**Display**) θα επιλέξουμε **Statistics** διότι δε θέλουμε την εμφάνιση του ιστογράμματος. Πατώντας **OK** θα προκύψει ένας πίνακας που δίνει πληροφορίες για το δείγμα και ο πίνακας του σχήματος 12.



Εικόνα 35

Descriptives

			Statistic	Std. Error
Engine Displacement (cu. inches)	Mean		194.04	5.221
	95% Confidence Interval for Mean	Lower Bound	**183.78**	
		Upper Bound	**204.30**	
	5% Trimmed Mean		188.28	
	Median		148.50	
	Variance		11068.589	
	Std. Deviation		105.207	
	Minimum		4	
	Maximum		455	
	Range		451	
	Interquartile Range		199	
	Skewness		.692	.121
	Kurtosis		-.791	.242

Σχήμα 12: Περιγραφικά μέτρα με 95% διάστημα εμπιστοσύνης.

Στο παράθυρο της εικόνας 35 υπάρχει δεξιά η επιλογή bootstrap. Δεν την επιλέξαμε αυτή τη φορά διότι το παράθυρο είναι ίδιο με αυτό της εικόνας 20. Ο αναγνώστης όμως καλείται να το επιλέξει για εξοικειωθεί με τη χρήση του bootstrap.

Ο παραπάνω πίνακας περιέχει τα περιγραφικά μέτρα στα οποία έχουμε ήδη αναφερθεί μαζί με λίγα ακόμα για τα οποία δεν έχουμε μιλήσει. Η μεταβλητή επιλέχθηκε τυχαία. Η πρώτη γραμμή του πίνακα περιέχει τη μέση τιμή για τις τιμές αυτής της μεταβλητής. Οι επόμενες δύο τιμές είναι το κάτω και το άνω άκρο του 95% διαστήματος εμπιστοσύνης για τον πραγματικό μέσο του πληθυσμού. Η επόμενη γραμμή είναι ο μέσος των τιμών της μεταβλητής από την οποία έχουμε αφαιρέσει το 5% των μεγαλύτερων και μικρότερων τιμών. Το Interquartile range (ενδοτεταρτημοριακό εύρος) είναι η διαφορά μεταξύ τρίτου και πρώτου τεταρτημόριου. Σε αυτό το εύρος βρίσκεται το 50% των κεντρικών παρατηρήσεων της μεταβλητής.

5.3 Συντελεστές γραμμικής συσχέτισης

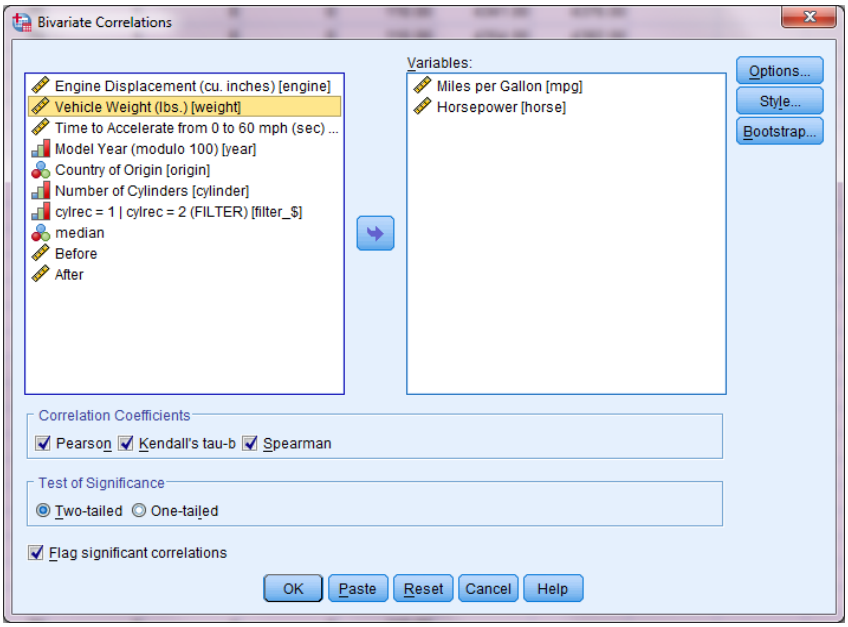
Πριν περάσουμε στη διεξαγωγή ελέγχων υποθέσεων για τους μέσους των μεταβλητών θα αναφερθούμε στη γραμμική συσχέτιση μεταξύ δύο ποσοτικών μεταβλητών. Οι συντελεστές που θα παρουσιαστούν παρακάτω, αναφέρονται στη γραμμικής φύσεως σχέση που μπορεί να συνδέει τις δύο μεταβλητές. Οι τιμές που μπορεί να πάρει ένας συντελεστής συσχέτισης είναι από -1 έως +1. Αρνητικές τιμές του συντελεστή γραμμικής συσχέτισης δύο μεταβλητών σημαίνει ότι έχουμε την ύπαρξη αρνητικής γραμμικής συσχέτισης. Δηλαδή, οι μεγαλύτερες τιμές της μίας μεταβλητής τείνουν να αντιστοιχούν στις μικρότερες τιμές της άλλης μεταβλητής. Θετικές τιμές του συντελεστή γραμμικής συσχέτισης είναι ένδειξη θετικής γραμμικής συσχέτισης μεταξύ των δύο μεταβλητών. Δηλαδή, οι μεγαλύτερες τιμές της μίας μεταβλητής τείνουν να αντιστοιχούν στις μεγαλύτερες τιμές της άλλη μεταβλητής. Τιμές κοντά στο μηδέν αποτελούν ένδειξη ότι δεν υπάρχει στατιστικά σημαντική

γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Όσο πιο μεγάλες είναι οι τιμές του συντελεστή, ή όσο πιο κοντά βρίσκονται στη μονάδα (σε απόλυτη τιμή πάντα), τόσο πιο ισχυρή είναι η γραμμική συσχέτιση μεταξύ τους. Οι πιο γνωστοί συντελεστές γραμμικής συσχέτισης είναι οι συντελεστές του **Pearson**, του **Spearman** και του **Kendall**. Η μηδενική και η εναλλακτική υπόθεση εδώ είναι οι εξής:

Ho: ρ=0 ή δεν υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών
H_1: ρ≠0 ή υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών

Ο συντελεστής συσχέτισης του Pearson υποθέτει κανονικότητα των δεδομένων, σε αντίθεση με τους άλλους δύο που δεν υποθέτουν κανονικότητα των δεδομένων. Βέβαια, για μεγάλα δείγματα, μεγέθους 30 παρατηρήσεων και πάνω και όσο το μέγεθος του δείγματος μεγαλώνει η θεωρία μας λέει ότι οι τιμές των συντελεστών "πλησιάζουν" η μία την άλλη. Δηλαδή, για τα δεδομένα που αφορούν στα αυτοκίνητα, δεν έχουμε κάποιο πρόβλημα να χρησιμοποιήσουμε οποιονδήποτε συντελεστή ανεξαρτήτως κατανομής των δεδομένων. Η κύρια διαφορά των συντελεστών είναι ότι ο συντελεστής του Pearson υπολογίζεται με βάση τα δεδομένα, ενώ οι άλλοι δύο υπολογίζονται με βάση τις τάξεις μεγέθους των δεδομένων. Ειδικότερα, ο συντελεστής του Spearman είναι ο συντελεστής του Pearson στην ουσία υπολογισμένος για τις τάξεις μεγέθους των δεδομένων. Το γεγονός λοιπόν ότι οι συντελεστές του Spearman και του Kendall υπολογίζονται με βάση τις τάξεις μεγέθους των δεδομένων είναι που επιτρέπει την ελευθερία ως προς τη μη ικανοποίηση της κανονικότητας των μεταβλητών. Βέβαια, εγώ προτιμάω του Pearson άσχετα από κανονικότητα και αυτόν συνιστώ.

Για να υπολογίσουμε τους τρεις αυτούς συντελεστές συσχέτισης στο SPSS επιλέγουμε τα εξής: **Analyze→Correlate→Bivariate** και εμφανίζεται το παράθυρο της εικόνας 36.



Εικόνα 36

Στο δεξιό κουτάκι πρέπει να περάσουμε τουλάχιστον δύο μεταβλητές, διότι οι συντελεστές συσχέτιση υπολογίζονται για ζεύγη μεταβλητών. Οπότε αν περάσουμε

περισσότερες από δύο μεταβλητές, θα υπολογιστούν οι συντελεστές γραμμικής συσχέτισης για όλα τα ζεύγη των μεταβλητών. Βλέπουμε από την εικόνα 36 ότι μόνο ο συντελεστής του Pearson είναι επιλεγμένος. Αν θέλουμε να εμφανιστούν και οι άλλοι δύο συντελεστές απλά τους επιλέγουμε. Παρατηρήστε ότι στο κάτω αριστερό μέρος του παραθύρου είναι επιλεγμένη μία επιλογή (**Flag significant correlations**). Η επιλογή **Options** μας δίνει τη δυνατότητα εμφάνισης των μέσων, των τυπικών αποκλίσεων και των πληθών των τιμών για κάθε μεταβλητή. Πατώντας **OK** θα εμφανιστούν τα σχήματα 13 και 14.

Αυτή τη φορά επιλέξαμε την εφαρμογή του bootstrap και μας έδωσε μία εκτίμηση της μεροληψίας. Υποθέτοντας ότι η τιμή του δείγματος είναι η πραγματική τιμή τότε ο μέση τιμή των 999 bootstrap τιμών για το συντελεστή συσχέτισης αποτελεί μία εκτίμηση για τη δειγματική τιμή του συντελεστή. Η διαφορά τους είναι μία εκτίμηση της μεροληψίας. Βέβαια όπως αναφέραμε και προηγουμένως έχουμε μεγάλα δείγματα εδώ, οπότε το bootstrap δεν θα δώσει κάτι διαφορετικό.

Correlations

				Miles per Gallon	Horsepower
Miles per Gallon	Pearson Correlation			1	**-.771****
	Sig. (2-tailed)				**.000**
	N			392	392
	Bootstrap[b]	Bias		0	-.001
		Std. Error		0	.016
		95% Confidence Interval	Lower	1	-.803
			Upper	1	-.741
Horsepower	Pearson Correlation			-.771**	1
	Sig. (2-tailed)			.000	
	N			392	392
	Bootstrap[b]	Bias		-.001	0
		Std. Error		.016	0
		95% Confidence Interval	Lower	-.803	1
			Upper	-.741	1

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

b. Unless otherwise noted, bootstrap results are based on 999 bootstrap samples

Σχήμα 13: Συντελεστής συσχέτισης του Pearson.

Βλέπουμε ότι για όλες τις τιμές των συντελεστών γραμμικής συσχέτισης υπάρχουν δύο αστεράκια. Αυτό γίνεται μέσω της επιλογής **Flag significant correlations**. Οι συντελεστές συσχέτισης που υπολογίστηκαν για αυτά τα ζεύγη μεταβλητών ανίχνευσαν κάποιες στατιστικά σημαντικές συσχετίσεις μεταξύ όλων των ζευγών μεταβλητών. Κάτω από κάθε τιμή του συντελεστή συσχέτισης εμφανίζεται μία p-value (**Sig. (2-tailed)**). Η p-value που έχει υπολογιστεί για κάθε συντελεστή ξεχωριστά αναφέρεται στον έλεγχο της υπόθεσης ότι στο συγκεκριμένο ζεύγος μεταβλητών δεν υπάρχει γραμμική συσχέτιση (δηλαδή ότι ο συντελεστής συσχέτισης για το ζεύγος είναι ίσος με το μηδέν).

Correlations

					Miles per Gallon	Horsepower
Kendall's tau_b	Miles per Gallon	Correlation Coefficient			1.000	**-.674****
		Sig. (2-tailed)			.	**.000**
		N			392	392
		Bootstrap[b]	Bias		.000	-.001
			Std. Error		.000	.018
			95% Confidence Interval	Lower	1.000	-.708
				Upper	1.000	-.636
	Horsepower	Correlation Coefficient			-.674**	1.000
		Sig. (2-tailed)			.000	.
		N			392	392
		Bootstrap[b]	Bias		-.001	.000
			Std. Error		.018	.000
			95% Confidence Interval	Lower	-.708	1.000
				Upper	-.636	1.000
Spearman's rho	Miles per Gallon	Correlation Coefficient			1.000	**-.849****
		Sig. (2-tailed)			.	**.000**
		N			392	392
		Bootstrap[b]	Bias		.000	.001
			Std. Error		.000	.017
			95% Confidence Interval	Lower	1.000	-.878
				Upper	1.000	-.811
	Horsepower	Correlation Coefficient			-.849**	1.000
		Sig. (2-tailed)			.000	.
		N			392	392
		Bootstrap[b]	Bias		.001	.000
			Std. Error		.017	.000
			95% Confidence Interval	Lower	-.878	1.000
				Upper	-.811	1.000

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

b. Unless otherwise noted, bootstrap results are based on 999 bootstrap samples

Σχήμα 14: Συντελεστές Kendall και Spearman (μη παραμετρικοί).

Αφού το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι μικρότερο του 0.05, συμπεραίνουμε ότι αυτή η υπόθεση απορρίπτεται σε α=0.05. Άρα υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ του ζεύγους. Στην περίπτωση που η p-value είναι μικρότερη του 0.01, τότε ο συντελεστής συσχέτισης εμφανίζεται με δύο αστεράκια αντί για μόνο ένα. Προσέξτε και το μήνυμα που υπάρχει κάτω από

κάθε πίνακα που εξηγεί τι σημαίνουν τα δύο αστεράκια. Στο παράδειγμα μας όλες οι p-value είναι μικρότερες και το 0.001.

Σε αυτό το σημείο καλό θα ήταν να αναφέρουμε ότι ο συντελεστής του Kendall μπορεί να χρησιμοποιηθεί και στην περίπτωση που έχουμε κατηγορικές μεταβλητές οι οποίες όμως είναι υποχρεωτικά σε κλίμακα διάταξης. Είναι δηλαδή διατακτικές κατηγορικές μεταβλητές. Ακόμα να αναφέρουμε ότι με το συντελεστή γραμμικής συσχέτισης ελέγχουμε αν σε ένα ζεύγος μεταβλητών υπάρχει γραμμική συσχέτιση μόνο. Δηλαδή μπορεί να υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών, αλλά όχι γραμμικής φύσεως. Σε αυτήν την περίπτωση αυτή η σχέση που συνδέει τις δύο μεταβλητές δεν μπορεί να ανιχνευτεί με το συντελεστή γραμμικής συσχέτισης. Οπότε προσοχή στην ερμηνεία που δίνουμε στο συντελεστή συσχέτισης. Να υπενθυμίσουμε επίσης ότι η λογική με την οποία απορρίπτουμε ή όχι μία υπόθεση είναι πάντα η ίδια. Αν το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι μικρότερο του 0.05 η υπόθεση απορρίπτεται. Στην αντίθετη περίπτωση δεν απορρίπτεται.

5.4 X^2 Έλεγχος ανεξαρτησίας για κατηγορικές μεταβλητές

Στην προηγούμενη παράγραφο είδαμε πως υπολογίζουμε το συντελεστή γραμμικής συσχέτισης για την περίπτωση ποσοτικών μεταβλητών. Τι γίνεται όμως στην περίπτωση που έχουμε κατηγορικές μεταβλητές; Σε αυτήν την περίπτωση χρησιμοποιούμε τον X^2 έλεγχο ανεξαρτησίας. Η απαιτούμενη κλίμακα μέτρησης των μεταβλητών είναι η ονομαστική, παρόλο που και μεταβλητές με διατακτική κλίμακα μπορούν να χρησιμοποιηθούν. Ο X^2 έλεγχος ανεξαρτησίας χρησιμοποιείται για τον έλεγχο της υπόθεσης ότι δύο κατηγορικές μεταβλητές είναι ανεξάρτητες μεταξύ τους. Οι κατηγορικές μεταβλητές μπορούν να έχουν οσαδήποτε επίπεδα (ή κατηγορίες), αρκεί βέβαια η κάθε μία να έχει τουλάχιστον δύο επίπεδα. Όπως θα δούμε παρακάτω όταν διεξάγουμε αυτόν τον έλεγχο ανεξαρτησίας με το SPSS, θα εμφανίζεται και ένας πίνακας. Αυτός ο πίνακας θα περιέχει τις συχνότητες εμφάνισης όλων των δυνατών συνδυασμών ζευγών των επιπέδων των κατηγορικών μεταβλητών. Οι υποθέσεις σε αυτήν την περίπτωση είναι οι εξής:

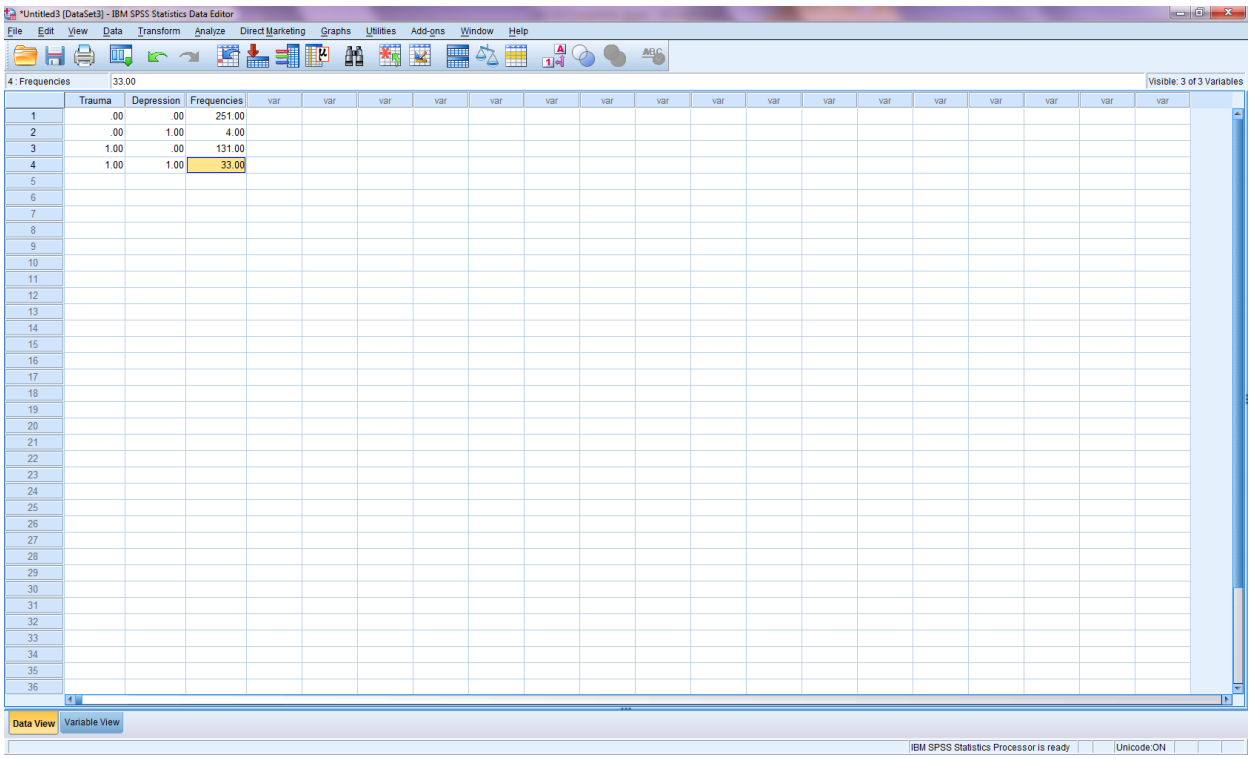
Ho: υπάρχει ανεξαρτησία μεταξύ των δύο μεταβλητών
H_1: δεν υπάρχει ανεξαρτησία μεταξύ των δύο μεταβλητών

Ας υποθέσουμε για παράδειγμα ότι έχουμε έναν 2X2 πίνακα. Στον πίνακα 1 έχουμε ταξινομήσει ένα δείγμα 419 γυναικών ανάλογα με το αν πάσχουν από κατάθλιψη και αν είχαν κάποια τραυματική εμπειρία στη ζωή τους. Το ερώτημα είναι αν υπάρχει εξάρτηση μεταξύ της τραυματικής εμπειρίας και της κατάθλιψης. Η μηδενική υπόθεση είναι πάντα αυτή που δεν υποθέτει εξάρτηση (υποθέτει ανεξαρτησία μεταξύ των μεταβλητών). Η προϋπόθεση που απαιτείται από τον X^2 έλεγχο ανεξαρτησίας είναι οι συχνότητες των κελιών να είναι τουλάχιστον ίσες με 5. Το SPSS χρησιμοποιεί το άλλο είδος υπόθεσης που θέλει τις αναμενόμενες συχνότητες των κελιών να είναι τουλάχιστον ίσες με 5. Ένα αποδεκτό ποσοστό κελιών που θα έχουν συχνότητες μικρότερες του 5 είναι το 25%, δηλαδή το πολύ ένα στα τέσσερα κελιά να έχει μία τιμή μικρότερη του 5 χωρίς να μειώνεται σημαντικά η αποτελεσματικότητα του ελέγχου. Αυτό ισχύει βέβαια και για πίνακες που έχουν περισσότερα κελιά. Αν αυτή η υπόθεση δεν ικανοποιείται, τότε κοιτάζουμε την p-value που υπολογίζεται με βάση τον ακριβή έλεγχο του Fisher (**Fisher's exact test**) ή το Monte Carlo.

	Κατάθλιψη		
Τραυματική Εμπειρία	Όχι (0)	Ναι (1)	Σύνολο
Όχι (0)	251	4	255
Ναι (1)	131	33	164
Σύνολο	382	37	419

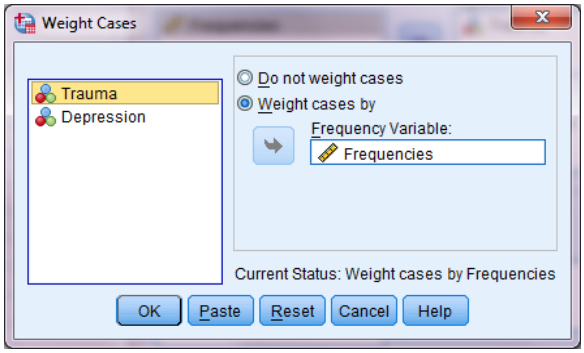
Πίνακας 1: Αριθμός ατόμων με τραυματική εμπειρία και κατάθλιψη.

Ο ακριβής έλεγχος του Fisher θα διεξαχθεί μόνο στην περίπτωση που έχουμε 2X2 πίνακες όπως στο παράδειγμα. Σε αυτήν την περίπτωση το Monte Carlo δεν υπολογίζεται. Στην περίπτωση λοιπόν που έχουμε τον πίνακα 2X2 λόγου χάριν όπως εδώ έτοιμο, τότε πρέπει να πληκτρολογήσουμε τα δεδομένα στο SPSS Data Editor. Μέσα στις παρενθέσεις έχουμε τοποθετήσει κάποιους αριθμούς (0 και 1) για να μας διευκολύνουν στο να περάσουμε τα δεδομένα στο SPSS. Πρέπει να προσέξουμε ώστε ο κάθε συνδυασμός γραμμής και στήλης να περιέχει τον αριθμό του κελιού που πρέπει. Περνώντας τα δεδομένα στο SPSS θα έχουν την εξής μορφή (σημασία έχει ο κάθε συνδυασμός γραμμής και στήλης να περιέχει το σωστό αριθμό):



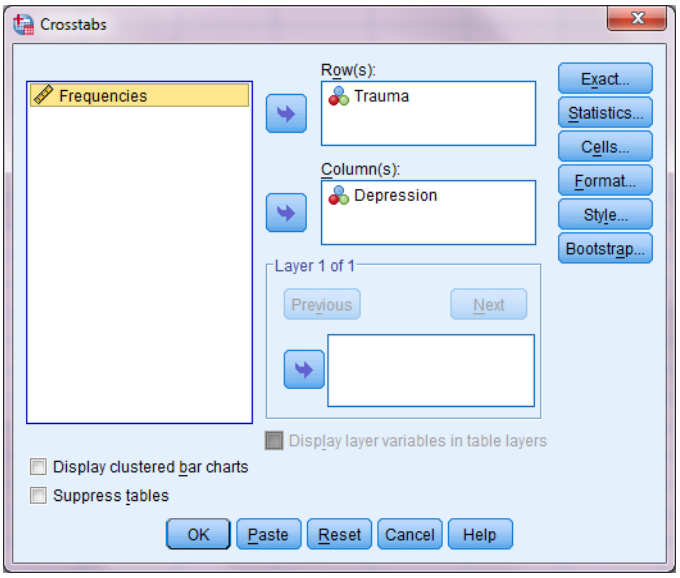
Εικόνα 37

Εμείς εδώ περάσαμε πρώτα τις γραμμές και μετά τις στήλες του πίνακα. Η σειρά δε μετράει, αυτό που μετράει είναι ο κάθε συνδυασμός γραμμής και στήλης να περιέχει το σωστό αριθμό. Για να "δώσουμε" στο SPSS να "καταλάβει" ότι η τρίτη στήλη περιέχει τις συχνότητες των κελιών θα επιλέξουμε τα εξής: **Data→Weight Cases** και θα εμφανιστεί το παράθυρο της εικόνας 38.



Εικόνα 38

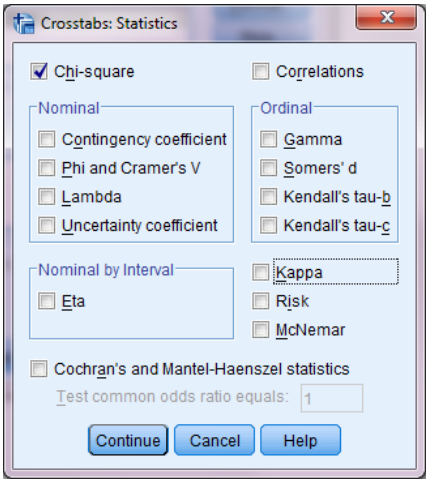
Εμείς θα επιλέξουμε **Weight cases by** και στο λευκό κουτάκι από κάτω θα περάσουμε την τρίτη στήλη (Frequencies για την περίπτωση μας) ή τη στήλη που περιέχει τις συχνότητες των κελιών. Μετά πατάμε **OK** και θα δείτε ότι το παράθυρο της εικόνας 38 θα κλείσει. Το σημαδάκι δίπλα από το όνομα της μεταβλητής δηλώνει ότι είναι κατηγορική ονομαστικής κλίμακας (δείτε στην αρχή του αυτού εγχειρίδιου για το κελί **Measure**, στο **Variable View** που αναφέρουμε για τους τύπους των μεταβλητών). Μετά επιλέγουμε **Analyze→Descriptive Statistics→Crosstabs** και θα εμφανιστεί το παράθυρο της εικόνας 39.



Εικόνα 39

Επιλέγοντας **Display clustered bar charts** θα εμφανιστεί ένα ραβδόγραμμα ίδιας μορφής με αυτό που εμφανίστηκε στη δεύτερη περίπτωση της παραγράφου 3.4 Επιλέγοντας **Exact** θα εμφανιστεί το παράθυρο της εικόνας 34 από όπου και θα επιλέξουμε τη διεξαγωγή του ελέγχου Monte Carlo. Εδώ το bootstrap δε χρειάζεται. Πατώντας **Statistics** θα εμφανιστεί το παράθυρο της εικόνας 40, στο οποίο θα επιλέξουμε **Chi-square**. Πατάμε **Continue** για να γυρίσουμε στο αρχικό παράθυρο, της εικόνας 39. Πατώντας μετά **OK**, θα εμφανιστούν διάφοροι πίνακες και ένα ραβδόγραμμα, αυτό το σχήματος 16. Ο πρώτος πίνακας περιέχει πληροφορίες για το μέγεθος του δείγματος, ο δεύτερος πίνακας είναι ο αρχικός πίνακας του οποίου τα

δεδομένα χρησιμοποιήσαμε (πίνακας 1). Και οι δύο πίνακες παραλείπονται για ευνόητους λόγους. Ο τρίτος πίνακας είναι αυτός του σχήματος 15.



Εικόνα 40

Chi-Square Tests[c]

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	42.675[a]	1	**.000**	**.000**	.000	
Continuity Correction[b]	40.402	1	.000			
Likelihood Ratio	44.365	1	**.000**	**.000**	.000	
Fisher's Exact Test				**.000**	.000	
Linear-by-Linear Association	42.573[d]	1	.000	.000	.000	.000
N of Valid Cases	419					

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 14.48.

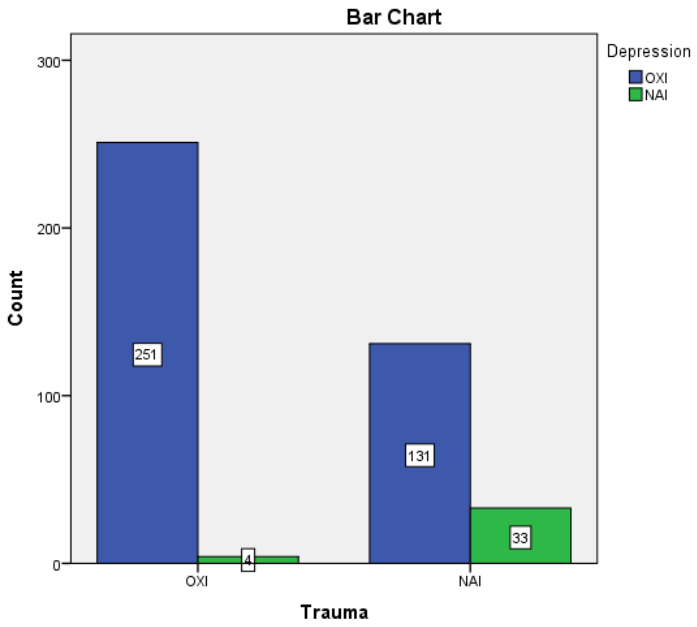
b. Computed only for a 2x2 table

c. For 2x2 crosstabulation, exact results are provided instead of Monte Carlo results.

d. The standardized statistic is 6.525.

Σχήμα 15: Αποτελέσματα X^2 ελέγχου ανεξαρτησίας.

Παρατηρούμε ότι το SPSS έχει εμφανίσει διάφορα μηνύματα κάτω από το σχήμα 15. Το τελευταίο μήνυμα αναφέρει ότι για την περίπτωση των 2X2 πινάκων ο έλεγχος του Fisher υπολογίζεται αντί του Monte Carlo. Το δεύτερο μήνυμα μας πληροφορεί για το αν ικανοποιείται η προϋπόθεση ισχύος του X^2 ελέγχου. Θέλουμε το πολύ το 25% των κελιών να έχουν τιμές μικρότερες από 5. Αν δεν ισχύει αυτό τότε δεν εμπιστευόμαστε τα αποτελέσματα του X^2 ελέγχου, παρά μόνο του Fisher για την περίπτωση δισδιάστατων πινάκων και του Monte Carlo για την περίπτωση πινάκων με περισσότερες από δύο γραμμές ή/και στήλες. Στη βιβλιογραφία αναφέρεται και μία πιο αυστηρή προϋπόθεση όσον αφορά τα κελιά με αριθμούς μικρότερους του 5 και η οποία θέλει όλα τα κελιά να έχουν τιμές μεγαλύτερες του 5.



Σχήμα 16: Ραβδόγραμμα ποιοτικών μεταβλητών

Αυτό που κοιτάζουμε από τον πίνακα του σχήματος 15 είναι οι p-value για κάθε έλεγχο. Κοιτάζουμε το **Asymp. Sig. (2 sided)** και το **Exact Sig. (2-sided)** για τα **Pearson Chi-Square** και **Fisher's Exact Test**. Μπορούμε φυσικά να κοιτάξουμε και τις p-value για το **Likelihood Ratio**.

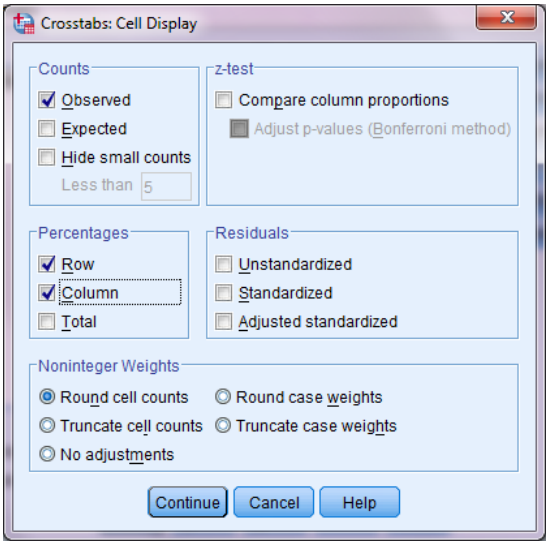
Ας δούμε όμως τώρα την περίπτωση στην οποία δεν έχουμε έναν πίνακα με τιμές αλλά τις μεταβλητές σε στήλες όπως στο παράδειγμα με τα αυτοκίνητα. Οι επιλογές και τα παράθυρα είναι ακριβώς τα ίδια. Σε αυτήν την περίπτωση δε χρειάζεται να ανοίξουμε το παράθυρο της εικόνας 38. Ανοίγουμε το παράθυρο της εικόνας 39 και απλά περνάμε τις δύο κατηγορικές μεταβλητές στα δύο λευκά κουτάκια. Οι επιλογές είναι ίδιες με πριν. Το SPSS στο Output θα εμφανίσει τώρα τους ίδιους πίνακες αποτελεσμάτων με προηγουμένως. Στην προηγούμενη περίπτωση εμφάνιζε τον πίνακα με τους συνδυασμούς των τιμών των κατηγορικών μεταβλητών, παρόλο που τον γνωρίζαμε εξαρχής, εδώ θα τον υπολογίσει για να τον εμφανίσει.

Στο σχήμα 16 βλέπουμε τις συχνότητες για κάθε περίπτωση πάνω στις ράβδους. Όταν εμφανίζεται το σχήμα δεν έχει αυτά τα νούμερα. Για το πως μπαίνουν δείτε τις σχετικές οδηγίες των εικόνων 27α και 27β. Όσο για τις τιμές των μεταβλητών (NAI και OXI εδώ) δείτε στην αρχή που αναφέρουμε το κελί ***Values*** για να βάζετε ονόματα στις τιμές των μεταβλητών (εικόνα 6).

5.5 Relative Risk και Odds ratio

Έχουμε ήδη δει πως διεξάγουμε τον X^2 έλεγχο για τον έλεγχο της ύπαρξης ανεξαρτησίας μεταξύ δύο κατηγορικών μεταβλητών. Ο λόγος για τον οποίο περνάμε τα δεδομένα ως 1 και 2 είναι για να δηλώσουμε τις γραμμές και τις στήλες. Δηλαδή, στην πρώτη στήλη του SPSS το 1 φανερώνει την πρώτη γραμμή και το 2 τη δεύτερη γραμμή. Στη δεύτερη στήλη το 1 φανερώνει την πρώτη στήλη και το 2 τη δεύτερη. Τώρα θα δούμε πως υπολογίζουμε το Odds Ratio (OR), τα Relative Risk (RR) και τι σημαίνουν αυτά.

Θα περάσουμε τις στήλες και τις γραμμές του πίνακα κατά τον ίδιο τρόπο. Δηλαδή, η πρώτη στήλη στο SPSS θα αντιστοιχεί στις γραμμές και η δεύτερη στις στήλες. Η τρίτη προφανώς θα περιέχει τις συχνότητες των κελιών. Το Odds Ratio ορίζεται ανεξαρτήτως του πως θα ορίσουμε τις στήλες ή τις γραμμές, αλλά το Relative Risk απαιτεί αυτή τη σειρά στο SPSS. Ο λόγος είναι απλός, το Relative Risk προϋποθέτει την ύπαρξη μίας ανεξάρτητης μεταβλητής η οποία επηρεάζει μία άλλη, την εξαρτημένη. Τον τρόπο με τον οποίο περνάμε τέτοιους πίνακες στο SPSS, τον έχουμε ήδη δει. Στο παράθυρο της εικόνας 40 θα επιλέξουμε το **Risk**. Θα πατήσουμε **Cells** για να εμφανιστεί το παράθυρο της εικόνας 41 στο οποίο θα επιλέξουμε να εμφανιστούν τα ποσοστά των γραμμών και των στηλών (**Row**, **Column**). Πατώντας **Continue** και **OK** θα εμφανιστεί ο παραπάνω πίνακας μαζί με τα ποσοστά των γραμμών και των στηλών και ο πίνακας του σχήματος 17.



Εικόνα 41

Ζητήσαμε από το SPSS να εμφανίσει τα ποσοστά γραμμών και στηλών για να γίνει κατανοητή η έννοια του Odds ratio και των σχετικών κινδύνων (Relative Risk). Οι γραμμές του πίνακα φανερώνουν το αν οι γυναίκες έχουν τραυματική εμπειρία (1) ή όχι (2) και οι στήλες φανερώνουν το αν οι γυναίκες έχουν κατάθλιψη (1) ή όχι (2). Το 20.1% των γυναικών που είχαν τραυματική εμπειρία έχουν πάθει κατάθλιψη. Το 1.6% των γυναικών που δεν είχαν τραυματική εμπειρία έχει πάθει κατάθλιψη. Είναι προφανές ότι για κάθε κατηγορία γυναικών που έπαθαν/δεν έπαθαν κατάθλιψη τα υπόλοιπα ποσοστά αναφέρονται στις γυναίκες που δεν πάσχουν από κατάθλιψη (79.9% και 98.4% αντίστοιχα).

Trauma * Depression Crosstabulation

			Depression		
			OXI	NAI	Total
Trauma	OXI	Count	251	4	255
		% within Trauma	98.4%	1.6%	100.0%
		% within Depression	65.7%	10.8%	60.9%
	NAI	Count	131	33	164
		% within Trauma	79.9%	20.1%	100.0%
		% within Depression	34.3%	89.2%	39.1%
Total		Count	382	37	419
		% within Trauma	91.2%	8.8%	100.0%
		% within Depression	100.0%	100.0%	100.0%

Σχήμα 17: Πίνακας με τις συχνότητες και τα ποσοστά.

Ας δούμε τώρα πως προκύπτουν οι σχετικοί κίνδυνοι και ποια η σημασία τους. Ο κίνδυνος μία γυναίκα που είχε τραυματική εμπειρία στο παρελθόν σε σχέση με μία γυναίκα που δεν είχε, να πάθει κατάθλιψη είναι ίσος με το λόγο των δύο πρώτων ποσοστών: 20.1%/1.6%=12.828. Δηλαδή ανάμεσα στις γυναίκες που πάσχουν από κατάθλιψη το ποσοστό των γυναικών που είχαν τραυματική εμπειρία είναι ίσο με 12.828 φορές το ποσοστό των γυναικών που δεν είχαν τραυματική εμπειρία. Άρα ο κίνδυνος μία γυναίκα να πάθει κατάθλιψη είναι κατά 11.828 φορές μεγαλύτερος για μία γυναίκα που έχει τραυματική εμπειρία σε σχέση με μία γυναίκα που δεν έχει. Η πιθανότητα δηλαδή μία γυναίκα να πάθει κατάθλιψη και είχε τραυματική εμπειρία στο παρελθόν είναι αυξημένη κατά 1182.8% σε σχέση με μία γυναίκα που δεν είχε τραυματική εμπειρία στο παρελθόν. Ο αντίστοιχος κίνδυνος για μία γυναίκα που είχε τραυματική εμπειρία στο παρελθόν σε σχέση με μία γυναίκα που δεν είχε, να μην πάθει κατάθλιψη είναι ίσος με 9.9%/98.4%=0.812. Δηλαδή η πιθανότητα μία γυναίκα που είχε τραυματική εμπειρία στο παρελθόν, να μην πάθει κατάθλιψη είναι ίση με 0.812 φορές την αντίστοιχη πιθανότητα για μία γυναίκα που δεν είχε τραυματική εμπειρία στο παρελθόν.

Ο λόγος των δύο σχετικών κινδύνων ισούται με το Odds ratio. Δηλαδή ο κίνδυνος μία γυναίκα πάθει κατάθλιψη έναντι του να μην πάθει είναι 15.807 μεγαλύτερος για τις γυναίκες που είχαν τραυματική εμπειρία σε σχέση με τις γυναίκες που δεν είχαν τραυματική εμπειρία. Ας δούμε όμως προκύπτει πραγματικά ο λόγος των Odds και τι είναι τα Odds. Το ποσοστό των καταθλιπτικών γυναικών που είχαν τραυματική εμπειρία στο παρελθόν είναι ίσο με 89.2%, ενώ το αντίστοιχο ποσοστό των γυναικών που δεν είχαν τραυματική εμπειρία στο παρελθόν είναι ίσο με 100%-89.2%=8.8%. Δηλαδή το 8.8% είναι το συμπληρωματικό ποσοστό του 89.2%. Συμπληρωματικό με την έννοια ότι είναι το ποσοστό που χρειάζεται το πρώτο ποσοστό για να δώσει άθροισμα 100%. Προκύπτει ότι το πηλίκο ανάμεσα στα δύο ποσοστά είναι ίσο με 89.2%/10.2%=8.745098. Επομένως ανάμεσα στις καταθλιπτικές γυναίκες το ποσοστό αυτών που είχαν τραυματική εμπειρία είναι 8.745 φορές το ποσοστό αυτών που δεν είχαν τραυματική εμπειρία. Μπορούμε να πούμε δηλαδή ότι στις καταθλιπτικές γυναίκες, 100 γυναίκες που δεν είχαν τραυματική εμπειρία στο παρελθόν αντιστοιχούν σε 874.5 γυναίκες που είχαν τραυματική εμπειρία. Αυτό το

πηλίκο λέγεται Odds. Πιο συγκεκριμένα τα Odds μίας καταθλιπτικής γυναίκας να έχει τραυματική εμπειρία στο παρελθόν σε σχέση με το να μην έχει είναι 8.745 (προς 1). Για τις γυναίκες όμως που δεν έχουν κατάθλιψη, το πηλίκο του πλήθους των γυναικών που είχαν τραυματική προς το πλήθος των γυναικών που δεν είχαν τραυματική εμπειρία είναι ίσο με 34.3%/65.7%=0.52207. Στις μη καταθλιπτικές γυναίκες, για κάθε 100 που είχαν τραυματική εμπειρία αντιστοιχούν 52 που δεν είχαν τραυματική εμπειρία στο παρελθόν. Επομένως για μία καταθλιπτική γυναίκα τα Odds του να έχει τραυματική εμπειρία στο παρελθόν σε σχέση με το να μην έχει είναι 0.522 (προς 1). Το πηλίκο των δύο Odds λέγεται Odds ratio. Άρα τα odds να είναι μία γυναίκα καταθλιπτική είναι 15.807 φορές μεγαλύτερα για τις γυναίκες που έχουν τραυματική εμπειρία σε σχέση με τις γυναίκες που δεν έχουν τραυματική εμπειρία.

Ο λόγος των Odds είναι συνώνυμο του ελέγχου X^2, ελέγχει δηλαδή την ανεξαρτησία μεταξύ δύο κατηγορικών μεταβλητών (μόνο όμως για 2X2 πίνακες, δηλαδή η κάθε κατηγορική μεταβλητή έχει δύο τιμές). Στον παρακάτω πίνακα βλέπουμε επίσης και τα 95% διαστήματα εμπιστοσύνης για τους σχετικούς κινδύνους και το λόγο των Odds. Αν το 95% διάστημα εμπιστοσύνης για το Odds ratio συμπεριλαμβάνει τη μονάδα, συμπεραίνουμε ότι ο λόγος αυτός δεν είναι στατιστικά σημαντικός, άρα και η εξάρτηση μεταξύ των δύο κατηγορικών μεταβλητών δεν είναι στατιστικά σημαντική. Το ίδιο ισχύει και για τους σχετικούς κινδύνους. Στο παράδειγμα όμως βλέπουμε ότι κανένα διάστημα εμπιστοσύνης δεν περιλαμβάνει τη μονάδα. Αφού λοιπόν το 95% διάστημα εμπιστοσύνης για το λόγο των Odds δεν συμπεριλαμβάνει τη μονάδα, συμπεραίνουμε ότι υπάρχουν ενδείξεις ότι η υπόθεση της ανεξαρτησίας μεταξύ της τραυματικής εμπειρίας και της κατάθλιψης πρέπει να απορριφθεί. Δηλαδή η εξάρτηση μεταξύ τους είναι στατιστικά σημαντική.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Trauma (OXI / NAI)	**15.807**	5.482	45.578
For cohort Depression = OXI	1.232	1.139	1.333
For cohort Depression = NAI	.078	.028	.216
N of Valid Cases	419		

Σχήμα 18: Odds Ratio και Relative Risk.

5.6 Έλεγχος αλλαγής κατάστασης μιας δίτιμης κατηγορικής μεταβλητής

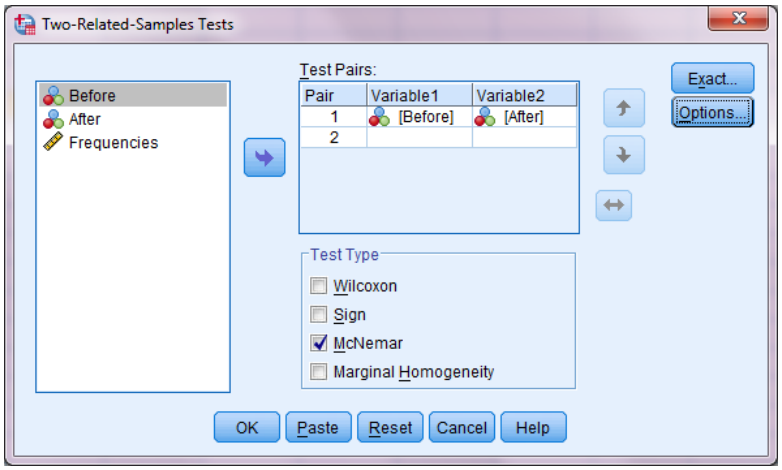
Ας υποθέσουμε ότι έχουμε τα ίδια δεδομένα με πριν αλλά το πρόβλημα είναι άλλο. Έχουμε 419 άτομα όπως και πριν από τα οποία τα 164 είναι καταθλιπτικά. Αποφασίζουμε λοιπόν να εφαρμόσουμε μία θεραπεία σε αυτά τα άτομα. Από ότι βλέπουμε στον πίνακα τα 131 μετά τη θεραπεία δεν είναι καταθλιπτικά.

Το ερώτημα που τίθεται είναι αν κατά πόσο υπήρξε αλλαγή της κατάστασης των ατόμων μετά σε σχέση με πριν. Μας ενδιαφέρουν δηλαδή ζεύγη (0,1) και (1,0). Ο έλεγχος σε αυτήν την περίπτωση είναι ο έλεγχος του McNemar. Ο τρόπος για να περάσουμε τα δεδομένα στο SPSS είναι ο ίδιος με προηγουμένως (βλέπε παράθυρα εικόνων 37-38).

Κατάθλιψη	Μετά		
Πριν	**Όχι (0)**	**Ναι (1)**	**Σύνολο**
Όχι (0)	251	4	255
Ναι (1)	131	33	164
Σύνολο	382	37	419

Πίνακας 2: Καταθλιπτικά άτομα πριν και μετά τη θεραπεία.

Για να διεξάγουμε αυτόν τον έλεγχο στο SPSS πάμε ως εξής: **Analyze→Non Parametric Tests→Legacy Dialogs→2 Related Samples** και θα εμφανιστεί το παράθυρο της εικόνας 42. Έχουμε περάσει τις μεταβλητές πριν και μετά στην πρώτη γραμμή στα δεξιά. Έχουμε επιπλέον επιλέξει τον έλεγχο McNemar. Επιλέγοντας **Exact** θα εμφανιστεί το παράθυρο της εικόνας 34 από όπου και θα επιλέξουμε τη διεξαγωγή του Monte Carlo ελέγχου.



Εικόνα 42

Πατώντας **OK** στο παράθυρο της εικόνας 42 θα εμφανιστεί το σχήμα 19.

Test Statistics[a]

	Before & After
N	419
Chi-Square[b]	117.600
Asymp. Sig.	**.000**
Exact Sig. (2-tailed)	**.000[c]**
Exact Sig. (1-tailed)	.000[c]
Point Probability	.000[c]

a. McNemar Test

b. Continuity Corrected

c. Exact results are provided instead of Monte Carlo for this test.

Σχήμα 19: Έλεγχος McNemar.

Η p-value βασισμένη στη X^2 κατανομή με 1 βαθμό ελευθερίας είναι μικρότερη του 0.005 όπως βλέπουμε στον πίνακα του σχήματος 20. Η p-value υπολογισμένη με βάση την προσομοίωση Monte Carlo είναι επίσης πολύ χαμηλή (p-value<0.001). Το SPSS χρησιμοποιεί και τη διόρθωση για τον υπολογισμό της p-value. Άρα έχουμε αλλαγή της κατάστασης μετά σε σχέση με πριν. Και κρίνοντας από τον πίνακα 2 βλέπουμε ότι επήλθε μείωση των καταθλιπτικών ατόμων μετά τη θεραπεία.

Να τονίσουμε σε αυτό το σημείο ότι αν οι μεταβλητή μας δεν ήταν δίτιμη αλλά είχε περισσότερες κατηγορίες και ήταν είτε ονομαστικής (π.χ. πρόθεση ψήφου) είτε διατακτικής κλίμακας (βαθμός ικανοποίησης) τότε θα έπρεπε να διεξάγουμε τον έλεγχο περιθώριας ομοιογένειας των Stuart-Maxwell.

Το SPSS παρέχει αυτόν τον έλεγχο (παράθυρο της εικόνας 42, **Marginal Homogeneity**) αλλά μόνο για διατακτικές μεταβλητές. Δηλαδή μπορείτε να χρησιμοποιήσετε τον έλεγχο της περιθώριας ομοιογένειας αλλά μόνο για μεταβλητές τύπου βαθμός ικανοποίησης. Η διαδικασία είναι ίδια όπως αυτή που περιγράψαμε μόλις τώρα. Για περισσότερες λεπτομέρειες σχετικά με αυτόν τον έλεγχο δείτε το άρθρο των Kuritz et al. (1988).

5.7 Αξιοπιστία ή βαθμός ταύτισης δύο κατηγορικών μεταβλητών (κάππα του Cohen)

Ας υποθέσουμε ότι έχουμε στη διάθεση μας τα αποτελέσματα δύο κριτών οι οποίοι έχουν κατατάξει κάποιους αθλητές με βάση τις επιδόσεις τους σε 5 κατηγορίες. Δύο κριτές μίας σχολικής εξέτασης όπου το αποτέλεσμα είναι είτε θετικό είτε αρνητικό. Ή αν θέλετε έχουμε δύο μεθόδους που ελέγχουν αν ένας άνθρωπος έχει μία αρρώστια ή όχι. Θέλουμε να δούμε το βαθμό συμφωνίας των δύο κριτών ή των δύο μεθόδων. Για να πάρουμε μία απάντηση θα υπολογίσουμε το συντελεστή κάππα του Cohen[†].

Στο παράδειγμα μας έχουμε τις απαντήσεις δύο κριτών για κάποιο θέμα. Οι απαντήσεις που έδωσαν οι 2 κριτές είναι A, B ή C. Θα κάνουμε σχεδόν την ίδια διαδικασία που κάναμε για τον X^2 έλεγχο ανεξαρτησίας (**Analyze→Descriptive Statistics→Crosstabs** και θα εμφανιστεί το παράθυρο της εικόνας 39). Εκεί θα περάσουμε στα δεξιά κουτάκια τις απαντήσεις των δύο κριτών, δύο μεταβλητές μία θα τη βάλουμε στο κουτάκι για τις στήλες και μία στο κουτάκι για τις γραμμές. Μετά θα επιλέξουμε το **Statistics** για να πάμε στο παράθυρο της εικόνας 40. Εκεί θα επιλέξουμε το **Kappa** προς τα κάτω και δεξιά. Πατάμε **Continue** και **OK** και το αποτέλεσμα θα είναι ένας πίνακας σαν του σχήματος 20 και ένας πίνακας σαν του σχήματος 21. Στον πίνακα του σχήματος 20 βλέπουμε ότι οι δύο κριτές συμφωνούν στις 67+12+7=86 από τις 103 απαντήσεις τους.

Ο συντελεστής τους Cohen παίρνει τιμές από 0 (απόλυτη ασυμφωνία) μέχρι 1 (απόλυτη συμφωνία). Η εδώ τιμή του είναι ίση με 0.647, και η p-value του ελέγχου ότι η τιμή του είναι ίση με το μηδέν είναι μικρότερη το 0.001. Αυτό σημαίνει ότι ο βαθμός συμφωνίας μεταξύ των δύο κριτών είναι στατιστικά σημαντικά υψηλός. Δεν είναι όσο ψηλός όσο ίσως να θέλαμε, αλλά είναι στατιστικά σημαντικός.

[†] Δείτε τη σελίδα της Wikipedia για τον τρόπου υπολογισμού αυτού του συντελεστή. http://en.wikipedia.org/wiki/Cohen's_kappa

Rater 1 * Rater 2 Crosstabulation

Count

		Rater 2			
		A	B	C	Total
Rater 1	A	**67**	8	2	77
	B	0	**12**	3	15
	C	0	4	**7**	11
Total		67	24	12	103

Σχήμα 20: Πίνακας διπλής εισόδου με τις απαντήσεις των δύο κριτών.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	**.647**	.070	8.823	**.000**
N of Valid Cases		103			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

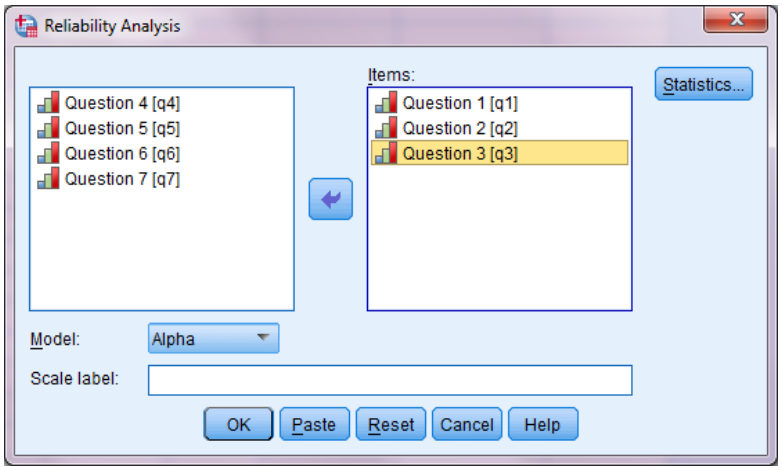
Σχήμα 21: Συντελεστής κάππα του Cohen.

5.8 Αξιοπιστία ενός ερωτηματολογίου

Για την ανάλυση των ερωτηματολογίων είναι απαραίτητη η εκτίμηση της αξιοπιστίας τους. Για τη μέτρηση της αξιοπιστίας των ερωτηματολογίων χρησιμοποιούμε το άλφα του Cronbach (Cronbach's alpha). Οι τιμές των συντελεστών αξιοπιστίας κυμαίνονται από 0 έως 1. Οι συντελεστές άλφα όπως και ο συντελεστής διχοτομικής αξιοπιστίας (split-half coefficient) αλλά και πολλοί άλλοι μετρούν στην ουσία την εσωτερική συνέπεια και όχι την αξιοπιστία ενός ερωτηματολογίου.

Το άλφα του Cronbach προτιμάται αντί του συντελεστή διχοτομικής αξιοπιστίας. Ο συντελεστή του ημίκλαστου όπως αλλιώς λέγεται ο συντελεστής διχοτομικής αξιοπιστίας κόβει τυχαία τις ερωτήσεις του ερωτηματολογίου στα δύο και υπολογίζει το συντελεστή συσχέτισης ανάμεσα στα δύο σκορ που προκύπτουν από τα δύο "μισά" του ερωτηματολογίου. Επίσης πρέπει να χρησιμοποιηθεί η φόρμουλα διόρθωσης των Spearman-Brown. Το πρόβλημα με το διαχωρισμό των ερωτήσεων του ερωτηματολογίου στα δύο είναι ότι αυτός ο συντελεστής βασίζεται σε ένα μόνο διαμερισμό του ερωτηματολογίου. Αρκεί να σκεφτεί κανείς για ένα ερωτηματολόγιο 20 ερωτήσεων υπάρχουν 184756 διαφορετικοί τρόποι να "κόψουμε" τις ερωτήσεις στα δύο και μπορούμε να υπολογίσουμε δηλαδή τόσους διαφορετικούς συντελεστές διχοτομικής αξιοπιστίας. Ο συντελεστής του Cronbach έρχεται να λύσει αυτό το πρόβλημα. Μαθηματικά η τιμή του άλφα ισούται με το μέσο όρο όλων αυτών των διαφορετικών συντελεστών διχοτομικής αξιοπιστίας. Επιπλέον, οι συντελεστές άλφα δε χρειάζονται διόρθωση μέσω της φόρμουλας των Spearman-Brown.

Ένα άρθρο από τον Charter, Richard A. που δημοσιεύτηκε το 2003[‡] έρχεται να επιβεβαιώσει ότι η συχνότητα χρησιμοποίησης του διχοτομικού συντελεστή αξιοπιστίας τείνει να μειώνεται με την πάροδο των ετών, ενώ ο συντελεστής άλφα κερδίζει έδαφος. Επίσης να αναφερθεί ότι για τιμές των συντελεστών αξιοπιστίας μεγαλύτερες του 0.70, σημαίνει ότι το ερωτηματολόγιο είναι αξιόπιστο. Για να βρούμε την αξιοπιστία ενός ερωτηματολογίου με το SPSS εκτελούμε τα εξής: **Analyze→Scale→Reliability Analysis** και θα εμφανιστεί το παράθυρο της εικόνας 44.



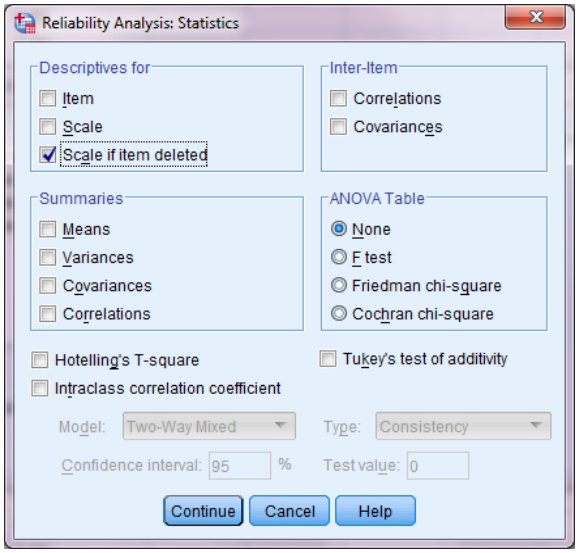
Εικόνα 44

Στο συγκεκριμένο παράδειγμα, οι 7 μεταβλητές αποτελούν τις απαντήσεις σε ένα ερωτηματολόγιο 7 ερωτήσεων. Οι δυνατές απαντήσεις ήταν σε διατακτική κλίμακα, από 1 έως 5. Όχι ακριβώς αριθμητικές μεταβλητές, αλλά το προσπερνάμε. Αυτές θα τις περάσουμε στο δεξιά κουτάκι. Στο σημείο που λέει **Model** είναι προεπιλεγμένη η επιλογή **Alpha**. Το SPSS θα υπολογίσει δηλαδή το συντελεστή αξιοπιστίας άλφα του Cronbach. Πατώντας **Statistics** θα εμφανιστεί το παράθυρο της εικόνας 45. Επιλέξαμε όπως βλέπετε την επιλογή **Scale if item deleted**. Το SPSS θα υπολογίσει με αυτόν τον τρόπο την αξιοπιστία του ερωτηματολογίου αν εξαιρούσαμε μία ερώτηση κάθε φορά. Αυτό είναι μία ένδειξη για το αν θα έπρεπε μία ερώτηση να απαλειφθεί από το ερωτηματολόγιο. Αν για παράδειγμα η αξιοπιστία του ερωτηματολογίου ανεβαίνει χωρίς αυτήν την ερώτηση είναι μία ένδειξη για περαιτέρω μελέτη αυτής της ερώτησης. Το αντίστροφο βέβαια ισχύει. Αν η αξιοπιστία του ερωτηματολογίου πέφτει χωρίς αυτήν την ερώτηση είναι μία ένδειξη καταλληλότητας αυτής της ερώτησης. Πατώντας σε αυτό το παράθυρο **Continue** και μετά **OK** το αποτέλεσμα φαίνεται παρακάτω.

Βλέπουμε στο πινακάκι του σχήματος 22 ότι η τιμή του άλφα είναι ίση με 0.756. Είπαμε προηγουμένως ότι για να θεωρηθεί αξιόπιστος ένα ερωτηματολόγιο πρέπει να έχει αξιοπιστία μεγαλύτερη από 0.70. Σε κάποιες επιστήμες όμως όπως στην ιατρική η επιθυμητή αξιοπιστία πρέπει να είναι μεγαλύτερη του 0.90 ή και 0.95. Στο επόμενο σχήμα βλέπουμε του συντελεστές άλφα δίπλα από κάθε ερώτηση όταν έχει αφαιρεθεί αυτή η ερώτηση. Παρατηρούμε ότι όταν αφαιρεθεί η τρίτη ερώτηση, η αξιοπιστία πέφτει πιο πολύ, ενώ οποιαδήποτε ερώτηση και αν αφαιρεθεί δεν

[‡] The journal of General Psychology: A breakdown of reliability coefficients by best type and reliability method, and the clinical implications of low reliability.

ανεβαίνει η αξιοπιστία του ερωτηματολογίου. Αν όμως αφαιρέσουμε την ερώτηση 6, δε χάνουμε και πολλά.



Εικόνα 45

Reliability Statistics

Cronbach's Alpha	N of Items
.756	7

Σχήμα 22: Συντελεστής άλφα του Cronbach.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Question 1	20.7010	22.420	.508	.720
Question 2	20.6907	21.653	.637	.695
Question 3	20.8247	21.875	.409	.744
Question 4	21.3299	23.286	.416	.738
Question 5	21.1856	20.424	.629	.690
Question 6	21.0825	23.243	.349	**.754**
Question 7	21.0309	22.759	.413	.739

Σχήμα 23: Αξιοπιστία του ερωτηματολογίου αν αφαιρούνται ερωτήσεις.

5.9 Καμπύλη χαρακτηριστικού λειτουργικού δέκτη (ROC curve)

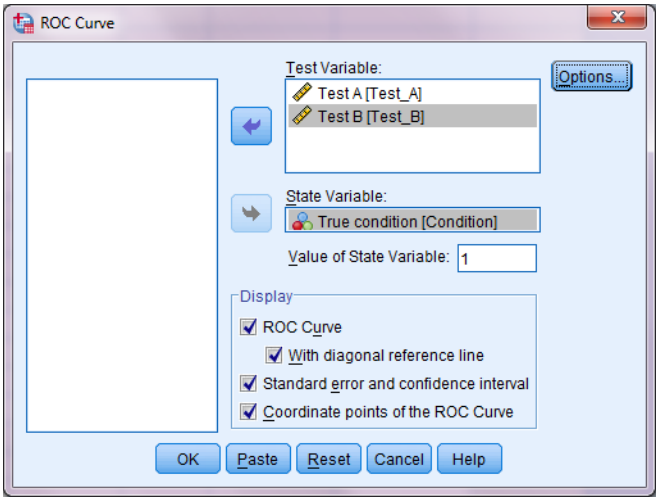
Ας υποθέσουμε για λίγο ότι είμαστε ιατροί και έχουμε στη διάθεση μας δεδομένα σχετικά με μία ασθένεια. Γνωρίζουμε αν ο ασθενής έχει την ασθένεια ή όχι και έχουμε κατασκευάσει δύο νέους διαγνωστικούς ελέγχους για να ανιχνεύουμε αν ο ασθενής έχει την ασθένεια ή όχι. Για να αξιολογήσουμε πόσο καλοί είναι αυτοί οι

διαγνωστικοί έλεγχοι αυτοί θα χρησιμοποιήσουμε την καμπύλη χαρακτηριστικού λειτουργικού δέκτη (ROC Curve).

Τα δεδομένα παρουσιάζονται στον πίνακα 3 και η τιμή 1 σημαίνει ότι ο ασθενής έχει την ασθένεια ενώ η τιμή 0 σημαίνει ότι δεν έχει την ασθένεια. Για να πάρουμε την ROC καμπύλη πάμε ως εξής: **Analyze→ROC Curve** και θα εμφανιστεί το παράθυρο της εικόνας 43. Σε αυτό το παράθυρο έχουμε περάσει τις στήλες με τα αποτελέσματα των δύο ελέγχων στο δεξιά πάνω κουτάκι (**Test Variable**) και τη στήλη με τις πραγματικές τιμές της ασθένειας στο κάτω δεξιά κουτάκι (**Test Variable**). Δηλώσαμε στο κουτάκι **Value of State Variable:** ότι η τιμή 1 δηλώνει την ύπαρξη της ασθένειας και επιλέξαμε όλες τις επιλογές εκτός από την τελευταία.

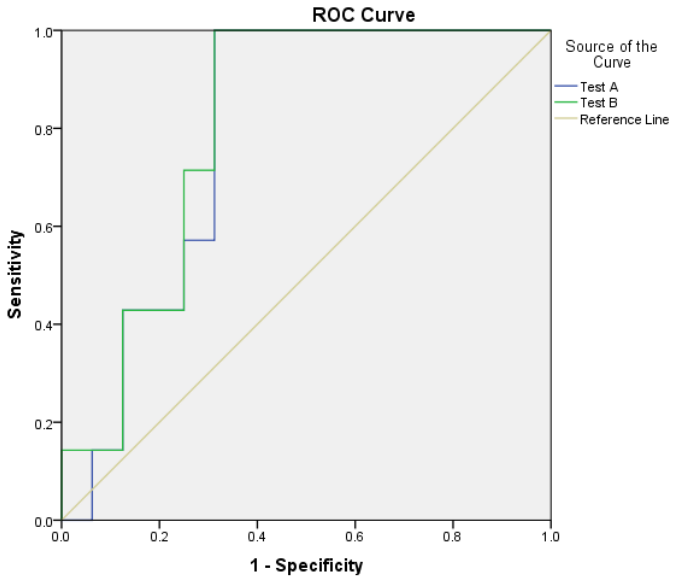
Ασθένεια (1 σημαίνει ΝΑΙ)	**Τιμή διαγνωστικού ελέγχου Α**	**Τιμή διαγνωστικού ελέγχου Β**
1	53	52.19
1	57	62.56
1	58	65.52
1	63	67.94
0	36	33.61
0	37	33.31
0	37	39.39
0	67	69.53
0	28	30.39
0	29	42.66
0	30	35.95
0	80	79.79
1	70	73.35
1	70	74.70
0	72	76.49
0	66	68.74
0	35	38.61
1	75	81.67
0	36	35.74
0	46	49.50
0	38	38.40
0	39	39.00
0	61	64.16

Πίνακας 3: Πραγματική τιμή της ασθένειας και προβλεπόμενη με βάση τους δύο διαγνωστικούς ελέγχους.



Εικόνα 43

Καλό θα ήταν για την τελευταία επιλογή να αποθηκεύονταν αυτά τα σημεία στις στήλες του SPSS δίπλα από τις στήλες των δεδομένων μας, κάτι που δεν γίνεται σε αυτήν την έκδοση του SPSS. Πατώντας **OK** θα εμφανιστούν τα αποτελέσματα των σχημάτων 24 και 25.



Σχήμα 24: Καμπύλη χαρακτηριστικού λειτουργικού δέκτη.

Οι δύο ROC καμπύλες παρουσιάζονται στο σχήμα 20. Μπορούμε να έχουμε και παραπάνω καμπύλες στο ίδιο σχήμα[§]. Η πράσινη γραμμή αντιστοιχεί στο διαγνωστικό έλεγχο Α και η μπλε γραμμή στο διαγνωστικό έλεγχο Β. Η διαγώνιος γραμμή (Reference line) είναι η γραμμή για την οποία ισχύει ότι τα αποτελέσματα (οι προβλέψεις) βασίζονται στην τύχη. Όταν η καμπύλη του ελέγχου βρίσκεται πάνω από τη διαγώνιο σημαίνει ότι η πρόβλεψη (ή η ταξινόμηση σε ασθενείς και μη, ή

[§] Το παράδειγμα δεν είναι και πολύ καλό, διότι όπως βλέπετε χάνονται οιμ γραμμές, αλλά τουλάχιστον παίρνετε την ιδέα.

οποιαδήποτε άλλη ταξινόμηση) είναι καλή. Το αντίθετο ισχύει αν η καμπύλη βρίσκεται κάτω από τη διαγώνιο γραμμή.

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Test A	**.786**	.095	**.033**	.600	.971
Test B	**.804**	.091	**.023**	.625	.982

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Σχήμα 25: Πληροφορίες για την καμπύλη ROC.

Το εμβαδόν της καμπύλης (Area Under Curve, AUC) είναι και το «ζουμί» στην ουσία της καμπύλης αυτής. Θα δανειστούμε τα λόγια της Ειρήνη Αδαμίδη (Αδαμίδη, 2012) από τη διπλωματική της εργασία για να εξηγήσουμε την έννοια του εμβαδού της καμπύλης. Συμπερασματικά το εμβαδόν της περιοχής κάτω από την καμπύλη ROC εκφράζει την πιθανότητα του ελέγχου να ταξινομήσει η δοκιμασία ορθά ένα τυχαίο ζεύγος ενός πάσχοντος και ενός μη πάσχοντος από το υπό εξέταση νόσημα και λαμβάνει τιμές από 0 έως 1. Όσο μεγαλύτερη είναι η τιμή του εμβαδού κάτω από την καμπύλη, τόσο μεγαλύτερη είναι η ακρίβεια του διαγνωστικού ελέγχου προς ανίχνευση ασθενών, ή μη, ατόμων.

Η 4η στήλη του πίνακα στο σχήμα 21 περιέχει τις p-value για το έλεγχο της μη καλή ορθής ταξινόμησης του διαγνωστικού ελέγχου. Αν το εμβαδόν κάτω από την καμπύλη είναι 0.5 τότε ο διαγνωστικός έλεγχος δεν είναι καλός στο να ανιχνεύει την ασθένεια. Η p-value και για τους δύο διαγνωστικούς ελέγχους είναι μικρότερη του 0.05. Άρα μπορούμε να ισχυριστούμε σε επίπεδο 5% ότι οι δύο διαγνωστικοί έλεγχοι δίνουν καλά αποτελέσματα. Η 5η και η 6η στήλη περιέχουν τα άνω και κάτω άκρα του 95% διαστήματος εμπιστοσύνης για την πραγματική τιμή του εμβαδόν κάτω από την καμπύλη. Όπως μπορούμε να δούμε, κανένα από τα δύο διαστήματα εμπιστοσύνης δεν περιέχει την τιμή 0.5.

Το IBM SPSS 22 δεν περιέχει έλεγχο μεταξύ δύο ή περισσότερων ROC καμπύλων. Για περισσότερες πληροφορίες σχετικά με την ROC καμπύλη, μπορείτε να δείτε τη διπλωματική εργασία της Ειρήνης Αδαμίδη. Επίσης στην παράγραφο 8.2 έχουμε κάποιες επιπλέον πληροφορίες με ένα άλλο παράδειγμα.

5.10 Έλεγχοι υποθέσεων για το μέσο και τη διάμεσο ενός δείγματος (έλεγχος t και έλεγχος του Wilcoxon)

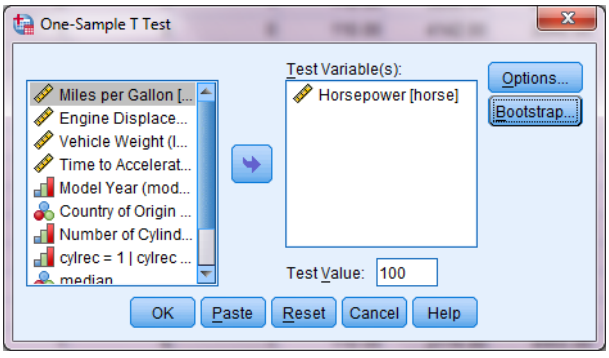
Όταν μιλήσαμε για τους συντελεστές συσχέτισης είπαμε ότι δύο εξ' αυτών δε χρειάζονται την υπόθεση της κανονικότητας. Όταν κάνουμε έναν έλεγχο υπόθεσης χωρίς να υποθέτουμε ότι τα δεδομένα ακολουθούν κάποια γνωστή κατανομή (π.χ. κανονική) τότε λέμε ότι χρησιμοποιούμε μη παραμετρική στατιστική και διεξάγουμε μη παραμετρικούς ελέγχους. Η πιο κλασική περίπτωση χρησιμοποίησης μη παραμετρικών στατιστικών τεχνικών είναι όταν τα δεδομένα δεν ακολουθούν την κανονική κατανομή.

Ας υποθέσουμε τώρα ότι έχουμε συλλέξει ένα δείγμα από παρατηρήσεις που αφορούν μετρήσεις μίας μεταβλητής. Θα χρησιμοποιήσουμε τα δεδομένα των αυτοκινήτων όπως και προηγουμένως. Θα χρησιμοποιήσουμε τη μεταβλητή που αφορά στην ιπποδύναμη των αυτοκινήτων. Βρήκαμε ότι ο μέση ιπποδύναμη των αυτοκινήτων είναι ίση με 104.83. Εμείς ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η μέση ιπποδύναμη του πληθυσμού των αυτοκινήτων αυτών των χωρών είναι ίση με 110. Δηλαδή οι υποθέσεις (μηδενική και εναλλακτική) είναι της μορφής:

Ho: μ=110
H_1: μ≠110

Υπάρχει μία προϋπόθεση που απαιτείται για να διεξάγουμε τον έλεγχο t, είναι αυτή της κανονικότητας των δεδομένων. Δυστυχώς ο έλεγχος κανονικότητας των Kolmogorov-Smirnov απέρριψε την υπόθεση της κανονικότητας των μετρήσεων της ιπποδύναμης. Υπάρχει όμως ένα θεώρημα που μας βοηθάει να ξεπεράσουμε το πρόβλημα της παραβίασης της κανονικότητας. Αυτό είναι το κεντρικό οριακό θεώρημα, το οποίο λέει ότι καθώς το μέγεθος του δείγματος τείνει στο άπειρο, η κατανομή του δειγματικού μέσου τείνει στην κανονική κατανομή. Η θεωρία αναφέρει ότι ένα δείγμα της τάξεως των 30 παρατηρήσεων είναι ικανοποιητικό για να μας εξασφαλίσει την ισχύ του θεωρήματος. Στη συγκεκριμένη περίπτωση το δείγμα αποτελείται από 400 μετρήσεις ιπποδύναμης.

Για να διεξάγουμε τον έλεγχο t στο SPSS εργαζόμαστε ως εξής: πατάμε **Analyze→Compare Means→One-Sample T-test** για να εμφανιστεί το παράθυρο της εικόνας 46. Περνάμε τη μεταβλητή στο λευκό κουτάκι και μετά πηγαίνουμε στο λευκό κουτάκι **Test Value**. Θα διαγράψουμε το μηδέν και θα τοποθετήσουμε την τιμή την οποία θέλουμε να ελέγξουμε (η τιμή της μηδενικής υπόθεσης). Η τιμή αυτή στο παράδειγμα μας είναι ίση με 100. Δεν έχω επιλέξει το bootstrap διότι δεν είμαι σίγουρος για το τι ακριβώς κάνει σε αυτήν την περίπτωση (ελέγχους υποθέσεων γενικά). Πατάμε **OK** και το αποτέλεσμα φαίνεται στο σχήμα 24, από το οποίο ο πρώτος πίνακας που εμφανίζεται έχει παραληφθεί. Ο λόγος είναι ότι περιέχει την τιμή του μέσου, της τυπικής απόκλισης, της τυπικής απόκλισης του μέσου (τυπικό σφάλμα του μέσου) και το μέγεθος του δείγματος.



Εικόνα 46

Η πρώτη στήλη του πίνακα στο σχήμα 26 περιέχει την ονομασία της μεταβλητής, η δεύτερη περιέχει έναν αριθμό. Αναφέραμε σε προηγούμενη παράγραφο ότι όλοι οι έλεγχοι χρησιμοποιούν κάποιους μαθηματικούς τύπους. Η τιμή αυτή (2.509) είναι η τιμή του ελέγχου t για αυτή τη μεταβλητή, η οποία χρησιμοποιήθηκε για να υπολογιστεί η p-value (Sig. (2-tailed)).

One-Sample Test

	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Horsepower	**2.509**	399	**.013**	4.832	1.05	8.62

Σχήμα 26: Αποτελέσματα ελέγχου t.

Στην πρώτη γραμμή του πίνακα φαίνεται η τιμή του μέσου την οποία ελέγξαμε μέσω του ελέγχου. Η στήλη δίπλα από την p-value περιέχει τη διαφορά ανάμεσα στην τιμή της μηδενικής υπόθεσης (100) και στη μέση τιμή της μεταβλητής (4.832) και οι επόμενες δύο στήλες περιέχουν ένα 95% διάστημα εμπιστοσύνης για αυτή τη διαφορά. Εφόσον δεν περιέχει την τιμή που εξετάζουμε, απορρίπτουμε την Ho. Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι ίσο με 0.013. Εφόσον είναι μικρότερο του 0.05, καταλήγουμε στο συμπέρασμα ότι μπορούμε να απορρίψουμε τη μηδενική υπόθεση ότι ο πραγματικός μέσος είναι ίσος με 100. Αλλιώς μπορούμε να πούμε ότι η διαφορά του μέσου από την τιμή της μηδενικής υπόθεσης είναι στατιστικά σημαντική, δηλαδή ο μέσος του δείγματος διαφέρει στατιστικά σημαντικά από την τιμή που ελέγξαμε και η οποία είναι ίση με 100.

Υπάρχει ακόμα μία στήλη που λέγεται **df**. Αυτό συμβολίζει τους βαθμούς ελευθερίας μίας κατανομής (**degrees of freedom**). Σε αυτόν τον έλεγχο ο μαθηματικός τύπος (ελεγχοσυνάρτηση) δε βασίστηκε στην κανονική κατανομή αλλά στην κατανομή t-student. Αυτή η κατανομή όπως και κάποιες άλλες έχουν αυτήν την παράμετρο ως ένα βασικό χαρακτηριστικό, τη λεγόμενη βαθμούς ελευθερίας. Με τον όρο αυτό ορίζουμε το μέγιστο αριθμό μεταβλητών οι οποίες μπορούν να προσδιοριστούν αυθαίρετα κάτω από κάποιες συνθήκες έτσι ώστε να ισχύουν οι συνθήκες. Με λίγα λόγια, για μία δεδομένη τιμή του μέσου σε ένα δείγμα μεγέθους ν μπορούμε να επιλέξουμε αυθαίρετα ν-1 τιμές, αλλά η νι-οστή τιμή θα είναι τέτοια ώστε να καταλήγουμε στον ήδη γνωστό μέσο.

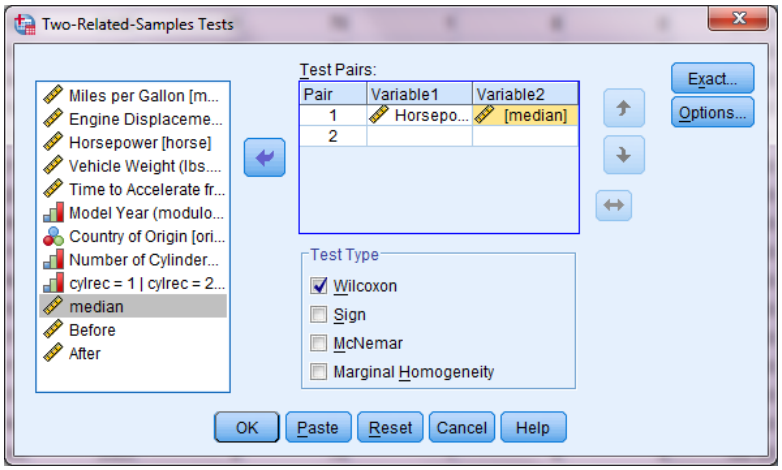
Ο αντίστοιχος μη παραμετρικός έλεγχος καλείται έλεγχος των προσημασμένων τάξεων μεγέθους του **Wilcoxon** για τη **διάμεσο** ενός πληθυσμού. Επειδή βασίζεται στις τάξεις μεγέθους των παρατηρήσεων και όχι στις παρατηρήσεις αυτές καθεαυτές δε χρειάζεται καμία προϋπόθεση ως προς την κατανομή των παρατηρήσεων. Αυτό ισχύει για όλους τους μη παραμετρικούς ελέγχους που θα δούμε. Ο έλεγχος όμως εδώ βασίζεται στη διάμεσο και όχι στο μέσο του δείγματος. Γενικά η ισχύς των μη παραμετρικών ελέγχων πλησιάζει την ισχύ των κλασικών παραμετρικών ελέγχων. Αυτός είναι και ο λόγος που η μόνη προϋπόθεση που χρειάζεται είναι η συμμετρία της κατανομής. Επομένως οι υποθέσεις (μηδενική και εναλλακτική) θα είναι λίγο διαφορετικές.

$$\mathbf{H_0: X_{0.5}=m}$$
$$\mathbf{H_1: X_{0.5}\neq m}$$

Όπου $X_{0.5}$ είναι η διάμεσος του δείγματος και m είναι η τιμή της διαμέσου την οποία θέλουμε να ελέγξουμε. Είχαμε αναφέρει στην αρχή ότι αν η κατανομή είναι συμμετρική, ο μέσος και η διάμεσος ταυτίζονται. Αυτός είναι και ο λόγος που ο έλεγχος του Wilcoxon "θέλει" την κατανομή να είναι συμμετρική. Η τιμή m την

οποία θα ελέγξουμε είναι πάλι η τιμή 100. Ένα πλεονέκτημα όμως των μη παραμετρικών ελέγχων που θα συζητήσουμε εδώ είναι ότι εφαρμόζονται και στις περιπτώσεις που οι μεταβλητές ποιοτικές διατεταγμένης κλίμακας (καλό, καλύτερο, πολύ καλό). Σε αυτήν την περίπτωση οι μέθοδοι ελέγχων υποθέσεων που θα δούμε είναι από τις πιο ισχυρές μεθόδους που υπάρχουν.

Ο έλεγχος όμως δεν προσφέρεται από το SPSS μέσω του μενού επιλογών, για αυτό θα πρέπει να γίνει μία διεργασία πρώτα. Η διεργασία έχει ως εξής: θα περάσουμε σε μία νέα στήλη την τιμή την οποία θέλουμε να ελέγξουμε τόσες φορές, όσες και οι παρατηρήσεις της μεταβλητής στην οποία θέλουμε να κάνουμε τον έλεγχο υπόθεσης. Επομένως για τα δεδομένα των αυτοκινήτων θα δημιουργήσουμε μία νέα στήλη που θα περιέχει την τιμή 100, 400 φορές. Επιλέγουμε από το μενού εντολών τα εξής: **Analyze→Non Parametric Tests→Legacy Dialogs→2 Related Samples** και θα εμφανιστεί το παράθυρο της εικόνας 47.



Εικόνα 47

Η επιλογή **Exact** είναι γνωστό πια τι δυνατότητα μας δίνει, η επιλογή **Options** μας επιτρέπει να εμφανίσουμε και κάποια περιγραφικά μέτρα. Αυτός ο έλεγχος έχει μία μικρή διαφορά ως προς το πως θα περάσουμε τις δύο μεταβλητές στο λευκό κουτάκι αριστερά. Η μία θα είναι η στήλη της οποίας τη διάμεσο ελέγχουμε και η άλλη η στήλη με την τιμή την οποία ελέγχουμε (100 για το παράδειγμα). Πρέπει πρώτα να επιλέξουμε τις δύο στήλες, όπως είναι στην εικόνα 47. Μόλις τις επιλέξουμε, στο κάτω μέρος αριστερά θα δούμε ότι έχουν εμφανιστεί τα ονόματα των στηλών-μεταβλητών. Επίσης θα ενεργοποιηθεί το βελάκι στη μέση του παραθύρου που μας επιτρέπει να περάσουμε τις μεταβλητές δεξιά. Η επιλογή **Wilcoxon** είναι προεπιλεγμένη, άρα μόλις περάσουμε το ζεύγος των στηλών δεξιά πατάμε **OK** για να εμφανιστούν οι 2 πίνακες των σχημάτων 27 και 28.

Ranks

		N	Mean Rank	Sum of Ranks
median - Horsepower	Negative Ranks	156[a]	239.57	37373.50
	Positive Ranks	227[b]	159.31	36162.50
	Ties	17[c]		
	Total	400		

a. median < Horsepower

b. median > Horsepower

c. median = Horsepower

Σχήμα 27: Πληροφορίες για τις τάξεις μεγέθους.

Ο πίνακας του σχήματος 27 περιέχει πληροφορίες σχετικά με τις τάξεις μεγέθους, οι οποίες χρησιμοποιούνται από τον έλεγχο. Ο παρακάτω πίνακας του σχήματος 28 περιέχει τις πληροφορίες σχετικά με την απόρριψη της μηδενικής υπόθεσης. Μας ενδιαφέρουν οι γραμμές της **Asymp. Sig. (2-tailed)** και της **Monte Carlo Sig (2-tailed).** Οι αριθμοί στην τελευταία στήλη είναι τα παρατηρηθέντα επίπεδα στατιστικής σημαντικότητας για τον έλεγχο (p-value). Οι αριθμοί που μας ενδιαφέρουν είναι αυτοί που βρίσκονται στις δύο γραμμές που αναφέραμε. Τόσο η p-value που υπολογίζεται μέσω του ασυμπτωτικού ελέγχου όσο και αυτή που υπολογίζεται μέσω του Monte Carlo είναι ίσες με μηδέν. Αφού είναι μικρότερες του 0.05, οδηγούμαστε στο συμπέρασμα ότι η μηδενική υπόθεση απορρίπτεται. Δηλαδή η διάμεσος του δείγματος διαφέρει στατιστικά σημαντικά από την τιμή που ελέγξαμε (100).

Test Statistics[a,c]

			median - Horsepower
Z			-.279[b]
Asymp. Sig. (2-tailed)			**.780**
Monte Carlo Sig. (2-tailed)	Sig.		**.770**
	99% Confidence Interval	Lower Bound	.759
		Upper Bound	.781
Monte Carlo Sig. (1-tailed)	Sig.		.382
	99% Confidence Interval	Lower Bound	.369
		Upper Bound	.395

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

c. Based on 10000 sampled tables with starting seed 1335104164.

Σχήμα 28: Αποτελέσματα ελέγχου του Wilcoxon.

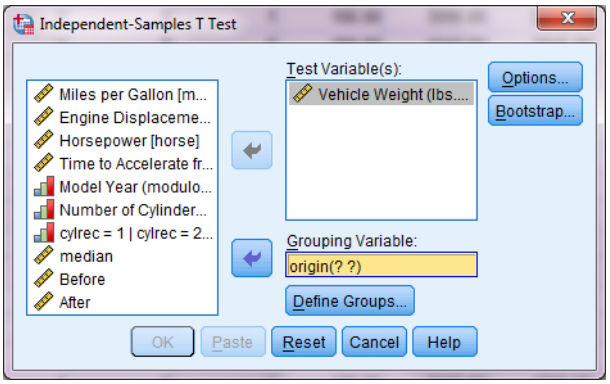
5.11 Έλεγχοι υποθέσεων για τη διαφορά των μέσων δύο ανεξάρτητων δειγμάτων (έλεγχος t και έλεγχος των Mann-Whitney-Wilcoxon)

Οι έλεγχοι αυτοί εφαρμόζονται για ελέγχους ισότητας μέσων μεταξύ δύο δειγμάτων τα οποία είναι ανεξάρτητα (η μπορούμε να υποθέσουμε ότι είναι ανεξάρτητα). Πρέπει όμως πρώτα να προηγηθεί μία διαδικασία, να καταχωρήσουμε σε μία στήλη τις μετρήσεις που αφορούν στα δύο δείγματα και σε μία άλλη στήλη να δηλώσουμε το δείγμα από το οποίο προέρχεται η κάθε μεταβλητή. Για παράδειγμα το 1 θα δηλώνει τις τιμές της μεταβλητής που προέρχονται από το πρώτο δείγμα και με 2 τις τιμές που προέρχονται από το δεύτερο δείγμα. Τα μεγέθη των δύο δειγμάτων δεν χρειάζεται να είναι ίσα. Δεν πρέπει να ξεχνάμε ότι έχουμε πάλι την υπόθεση της κανονικότητας που πρέπει να ικανοποιείται (για τον έλεγχο t), εκτός και αν έχουμε μεγάλο μέγεθος δείγματος. Αυτό που θέλουμε να ελέγξουμε είναι αν οι μέσοι των πληθυσμών από τους οποίος προέρχονται τα δείγματα διαφέρουν. Οι υποθέσεις διαμορφώνονται ως εξής:

$$\mathbf{Ho: \mu_1=\mu_2}$$
$$\mathbf{H_1: \mu_1\neq\mu_2}$$

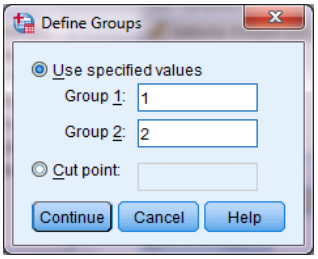
όπου μ_1 ο μέσος του πληθυσμού του πρώτου δείγματος και μ_2 ο μέσος του πληθυσμού του δεύτερου δείγματος.

Χρησιμοποιώντας τα δεδομένα των αυτοκινήτων για να εκτελέσουμε τον (παραμετρικό) έλεγχο t θα επιλέξουμε **Analyze→Compare Means→Independent-Samples T Test** και θα εμφανιστεί το παράθυρο της εικόνας 48.



Εικόνα 48

Περνάμε τη στήλη η οποία περιέχει τις μετρήσεις και για τα δύο δείγματα στο λευκό κουτάκι δεξιά (**Test variable(s):**). Και τη στήλη που δηλώνει οι παρατηρήσεις σε ποιο δείγμα ανήκουν στο κάτω λευκό κουτάκι (**Grouping Variable:**) Δεν μπορούμε όμως να πατήσουμε **OK** ακόμη. Πρέπει να δώσουμε στο SPSS να καταλάβει ότι η μεταβλητή που περιέχει τις χώρες προέλευσης των αυτοκινήτων θα χρησιμοποιηθεί για το διαχωρισμό των αυτοκινήτων ανάλογα με τη χώρα προέλευσης. Αυτό θα γίνει πατώντας **Define Groups** για να εμφανιστεί το παράθυρο της εικόνας 49.



Εικόνα 49

Έστω ότι εμείς θέλουμε να ελέγξουμε αν υπάρχουν διαφορές στα βάρη των αυτοκινήτων ανάλογα με το αν είναι αμερικάνικα (1) ή ευρωπαϊκά (2). Πρέπει να δηλώσουμε ότι στη μεταβλητή που αναφέρεται στη χώρα προέλευσης των αυτοκινήτων η τιμή 1 αναφέρεται στο ένα δείγμα και η τιμή 2 στο άλλο δείγμα. Πατάμε **Continue** για να επιστρέψουμε στο αρχικό παράθυρο, αυτό της εικόνας 46 και εκεί πατάμε **OK** για να εμφανιστούν στο Output οι παρακάτω δύο πίνακες.

Group Statistics

	Country of Origin	N	Mean	Std. Deviation	Std. Error Mean
Vehicle Weight (lbs.)	American	253	3367.33	788.612	49.580
	European	73	2431.49	490.884	57.454

Σχήμα 29: Περιγραφικά μέτρα των δύο δειγμάτων.

Independent Samples Test

		Levene's Test for Equality of Variances	t-test for Equality of Means						
								95% Confidence Interval of the Difference	
		Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Vehicle Weight (lbs.)	Equal variances assumed	**.000**	9.610	324	**.000**	935.835	97.382	744.255	1127.415
	Equal variances not assumed		12.332	189.187	**.000**	935.835	75.888	786.139	1085.531

Σχήμα 30: Αποτελέσματα ελέγχου t.

Ο πρώτος πίνακας περιέχει κάποια περιγραφικά μέτρα για τα δύο δείγματα. Ο δεύτερος πίνακας είναι αυτός που μας ενδιαφέρει. Ο έλεγχος t έχει δύο "κατευθύνσεις". Η μία κατεύθυνση είναι αυτή που δεν μπορούμε να υποθέσουμε ότι

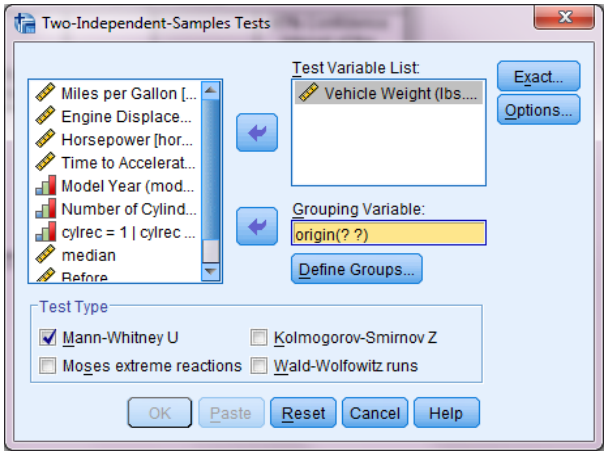
οι διακυμάνσεις των δύο δειγμάτων είναι περίπου ίσες και αυτή που μπορούμε να υποθέσουμε ότι είναι ίσες. Ο παραπάνω πίνακας έχει δύο γραμμές αποτελεσμάτων, η πρώτη αναφέρεται στην περίπτωση που μπορούμε να υποθέσουμε ισότητα των δύο διακυμάνσεων και η δεύτερη στην περίπτωση που δεν μπορούμε να υποθέσουμε ισότητα των δύο διακυμάνσεων. Ο πίνακας είναι χωρισμένος σε δύο κατηγορίες αποτελεσμάτων, η μία αφορά το **Levene** για την ισότητα των διακυμάνσεων και η άλλη περιέχει τα αποτελέσματα του ελέγχου t που επιλέξαμε να κάνουμε. Όπως αναφέραμε, ο πίνακας έχει δύο γραμμές αποτελεσμάτων, το αν θα κοιτάξουμε την πρώτη ή τη δεύτερη γραμμή αποτελεσμάτων του ελέγχου t θα μας το "πει" ο έλεγχος του Levene.

Ο έλεγχος του Levene ελέγχει την υπόθεση της ισότητας των δύο διακυμάνσεων και υπολογίζει μία p-value. Αν η p-value είναι μικρότερη του 0.05, απορρίπτεται η υπόθεση της ισότητας των διακυμάνσεων. Στην αντίθετη περίπτωση δεν απορρίπτεται. Επομένως, ανάλογα με την p-value (Sig.) του έλεγχου του Levene, κοιτάζουμε την πρώτη ή τη δεύτερη γραμμή αποτελεσμάτων. Στην προκειμένη περίπτωση η p-value είναι μικρότερη του 0.05, άρα δεν μπορούμε να υποθέσουμε ισότητα των δύο διακυμάνσεων. Επομένως θα κοιτάξω τη δεύτερη γραμμή αποτελεσμάτων του πίνακα. Η p-value για τον έλεγχο της ισότητας των δύο μέσων είναι ίση με μηδέν (Sig. (2-tailed)). Άρα η μηδενική υπόθεση απορρίπτεται, δηλαδή οι μέσοι των δύο πληθυσμών από τα οποία προήλθαν τα δύο δείγματα διαφέρουν στατιστικά σημαντικά σε επίπεδο στατιστικής σημαντικότητας α=5% πάντα. Το 95% διάστημα εμπιστοσύνης για τη διαφορά των πραγματικών μέσων είναι διαφορετικό για την περίπτωση που δεν μπορούμε να υποθέσουμε ισότητα των διακυμάνσεων.

Το αντίστοιχο μη παραμετρικό ανάλογο του ελέγχου t είναι ο έλεγχος των Mann-Whitney-Wilcoxon. Η διαδικασία την οποία πρέπει να κάνουμε για εκτελέσουμε αυτόν το μη παραμετρικό έλεγχο είναι ίδια με προηγουμένως (συνένωση των δύο μεταβλητών σε μία στήλη κ.λπ.). Οι υποθέσεις σε αυτήν την περίπτωση ορίζονται ως εξής:

Ho: Τα αυτοκίνητα των δύο χωρών δε διαφέρουν ως προς το βάρος στους
H_1: Τα αυτοκίνητα των δύο χωρών διαφέρουν ως προς το βάρος τους

Επιλέγουμε **Analyze→Non Parametric Tests→Legacy Dialogs→2 Independent Samples** και εμφανίζεται το παράθυρο της εικόνας 50. Η διαδικασία είναι ακριβώς η ίδια με προηγουμένως όσον αφορά στη στήλη που δηλώνει τα δείγματα (**Define Groups**). Πατώντας **Exact** επιλέγουμε να εμφανιστούν τα αποτελέσματα του ελέγχου Monte Carlo. Μόλις τελειώσουμε με όλες τις επιλογές πατάμε **OK** και εμφανίζονται στο Output δύο πίνακες, ο πρώτος περιέχει πληροφορίες για τις τάξεις μεγέθους και τα μεγέθη των δύο δειγμάτων και ο δεύτερος είναι αυτός που φαίνεται στο σχήμα 31.



Εικόνα 50

Ο πίνακας μοιάζει πάρα πολύ με αυτόν του σχήματος 26, όπου η τελευταία στήλη περιέχει p-value από τις οποίες κοιτάζουμε αυτές που αφορούν στην ασυμπτωτική σημαντικότητα και στη σημαντικότητα που υπολογίζεται μέσω του Monte Carlo. Και στις δύο περιπτώσεις βλέπουμε ότι η p-value είναι ίση με μηδέν. Επομένως απορρίπτουμε τη μηδενική υπόθεση, άρα το βάρος των αυτοκινήτων διαφέρει στατιστικά σημαντικά ανάμεσα στα αμερικάνικα και στα ευρωπαϊκά.

Test Statistics[a]

			Vehicle Weight (lbs.)
Mann-Whitney U			3049.500
Wilcoxon W			5750.500
Z			-8.718
Asymp. Sig. (2-tailed)			.000
Monte Carlo Sig. (2-tailed)	Sig.		.000[b]
	99% Confidence Interval	Lower Bound	.000
		Upper Bound	.000
Monte Carlo Sig. (1-tailed)	Sig.		.000[b]
	99% Confidence Interval	Lower Bound	.000
		Upper Bound	.000

a. Grouping Variable: Country of Origin

b. Based on 10000 sampled tables with starting seed 329836257.

Σχήμα 31: Αποτελέσματα ελέγχου Mann-Whitney-Wilcoxon.

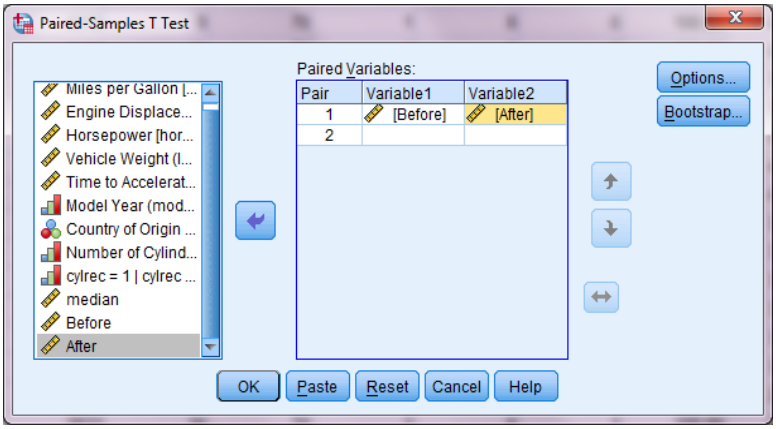
5.12 Έλεγχοι υποθέσεων για τη διαφορά των μέσων δύο εξαρτημένων δειγμάτων (έλεγχος t και έλεγχος Wilcoxon για δείγμα ζευγών παρατηρήσεων)

Είδαμε προηγουμένως πως να κάνουμε έλεγχο υποθέσεων για δύο δείγματα όταν τα δείγματα είναι ανεξάρτητα. Τι γίνεται όμως όταν τα δύο δείγματα δεν είναι, ή δεν μπορούμε να υποθέσουμε ότι είναι ανεξάρτητα; Την απάντηση τη δίνουν ο έλεγχος t

για ζεύγη παρατηρήσεων και ο έλεγχος των προσημασμένων τάξεων μεγέθους του Wilcoxon για δείγμα ζευγών παρατηρήσεων.

Κλασικό παράδειγμα εφαρμογής των δύο αυτών ελέγχων είναι στην ιατρική, όταν έχουμε μετρήσεις για κάποια άτομα πριν και μετά από μία δίαιτα και ενδιαφερόμαστε να δούμε κατά πόσο ήταν αποτελεσματική η δίαιτα ή όχι. Και τις δύο φορές μετρήσαμε το βάρος των ίδιων ατόμων, άρα γίνεται σαφές ότι τα μεγέθη των δύο δειγμάτων πρέπει να είναι απαραίτητα ίσα. Μία άλλη περίπτωση εφαρμογής αυτών των ελέγχων είναι όταν εξετάζουμε αδέρφια ως προς ένα ποσοτικό χαρακτηριστικό. Η υπόθεση της εξάρτησης των τιμών του χαρακτηριστικού ανάμεσα στα αδέρφια είναι εύλογη. Είναι προφανές ότι η μηδενική και η εναλλακτική υπόθεση είναι ίδιες με αυτέ της προηγούμενης παραγράφου.

Για τον έλεγχο t εργαζόμαστε ως εξής: επιλέγουμε **Analyze→Compare Means→Paired-Samples T Test** και θα εμφανιστεί το παράθυρο της εικόνας 51. Όπως και στην περίπτωση διεξαγωγής του ελέγχου του Wilcoxon για τη διάμεσο ενός δείγματος (βλ. Εικόνα 47) έτσι και εδώ πρέπει πρώτα να επιλέξουμε το ζεύγος μεταβλητών τους μέσους των οποίων θα χρησιμοποιήσουμε για να διεξάγουμε συμπεράσματα. Προσέξτε κάτω δεξιά ότι θα εμφανιστούν τα ονόματα των μεταβλητών. Τώρα μπορούμε να πατήσουμε το βελάκι για να τις περάσουμε στο λευκό κουτάκι δεξιά. Πατάμε **OK** για να εμφανιστούν 3 πίνακες από τους οποίους παρουσιάζουμε τους δύο τελευταίους. Ο πρώτος πίνακας περιέχει το μέγεθος των δειγμάτων, τους μέσους, τις τυπικές αποκλίσεις των μέσων και τις τυπικές αποκλίσεις κάθε δείγματος. Ενδιαφέρον παρουσιάζουν οι άλλοι δύο πίνακες που φαίνονται στα σχήματα 32 και 33 αντίστοιχα.



Εικόνα 51

Τα δεδομένα που χρησιμοποιήθηκαν στο παράδειγμα προέρχονται από ελέφαντες στους οποίους μετρήθηκε το βάρος πριν και μετά μία δίαιτα. Ο συντελεστής γραμμικής συσχέτισης που υπολογίστηκε έχει χαμηλή πλην στατιστικά σημαντική τιμή, φανερώνοντας μία γραμμική συσχέτιση μεταξύ των βαθμολογιών στις δύο μετρήσεις. Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας για τον έλεγχο της υπόθεσης ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ των δύο βαθμολογιών είναι ίσο με 0.028, γεγονός που σημαίνει ότι ο συντελεστής γραμμικής συσχέτισης είναι στατιστικά σημαντικός.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Before & After	27	**-.422**	**.028**

Σχήμα 32: Συντελεστής συσχέτισης του Pearson μεταξύ των δύο μετρήσεων.

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	Before - After	188.51852	1577.15999	303.52458	-435.38520	812.42223	.621	26	.540

Σχήμα 33: Αποτελέσματα ελέγχου t για το ζεύγος τιμών.

Η τελευταία στήλη του παρακάτω πίνακα περιέχει το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας, το οποίο είναι ίσο με μηδέν. Αφού είναι μεγαλύτερο του 0.05 συμπεραίνουμε ότι οι μέσοι των βαρών για τα δύο δείγματα, πριν και μετά δε διαφέρουν στατιστικά σημαντικά. Επίσης αν προσέξετε η διαφορά είναι το βάρος πριν μείον το βάρος μετά. Αυτό έγινε διότι στο παράθυρο της εικόνας 51 βάλαμε αριστερά το βάρος πριν και δεξιά το βάρος μετά. Οπότε για να έχουμε την ανάποδη διαφορά, απλά τοποθετούμε τις μεταβλητές με την ανάποδη σειρά.

Ο έλεγχος t όμως προϋποθέτει κανονικότητα των δεδομένων αν και αυτό το πρόβλημα μπορεί να ξεπεραστεί δεδομένου ότι έχουμε μεγάλο δείγμα μαθητών. Ο αντίστοιχος μη παραμετρικός έλεγχος του Wilcoxon εκτελείται στο SPSS επιλέγοντας τα εξής: **Analyze→Non Parametric Tests→Legacy Dialogs→2 Related Samples** και θα εμφανιστεί το παράθυρο της εικόνας 45. Το παράθυρο της εικόνας 45 μοιάζει πάρα πολύ με το παράθυρο της εικόνας 49. Η διαδικασία για να περάσουμε αριστερά στο κουτάκι τις δύο μεταβλητές είναι γνωστή. Επιλέγουμε να εμφανίσουμε και τα αποτελέσματα του Monte Carlo, οπότε πατώντας **OK** τα αποτελέσματα που θα εμφανιστούν στο Output παρουσιάζονται στο σχήμα 34 (ο πρώτος πίνακας παραλείπεται). Τα παρατηρηθέντα επίπεδα στατιστική σημαντικότητας είναι μηδέν, οδηγώντας μας στο συμπέρασμα ότι οι μέσοι των βαθμολογιών διαφέρουν στατιστικά σημαντικά σε α=5%.

Test Statistics[a,c]

			After - Before
Z			-.673[b]
Asymp. Sig. (2-tailed)			.501
Monte Carlo Sig. (2-tailed)	Sig.		.515
	99% Confidence Interval	Lower Bound	.502
		Upper Bound	.528
Monte Carlo Sig. (1-tailed)	Sig.		.253
	99% Confidence Interval	Lower Bound	.242
		Upper Bound	.264

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

c. Based on 10000 sampled tables with starting seed 1535910591.

Σχήμα 34: Αποτελέσματα ελέγχου του Wilcoxon για ζεύγη παρατηρήσεων.

6.1 Μια μικρή εισαγωγή στη γραμμική παλινδρόμηση

Συνήθως όταν κάποιος μιλάει για την απλή γραμμική παλινδρόμηση μιλάει και για τη σχέση της με τον γραμμικό συντελεστή συσχέτισης, αν και σε αυτές τις σημειώσεις αυτό δε συνέβη. Η απλή γραμμική παλινδρόμηση χρησιμοποιείται για να εκτιμήσει τη σχέση που υπάρχει μεταξύ μίας ανεξάρτητης μεταβλητής (X) και μίας εξαρτημένης μεταβλητής (Y). Με τον όρο εξαρτημένη μεταβλητή εννοούμε ότι οι τιμές τις εξαρτώνται από τις τιμές της ανεξάρτητης μεταβλητής X. Αυτό σημαίνει ότι η σχέση που υπάρχει μεταξύ του είναι στοχαστική ή στατιστική, αφού σε κάθε τιμή του X μπορεί να αντιστοιχούν περισσότερες από μία τιμές στην Y. Αν δεν ίσχυε αυτό, τότε θα μιλάγαμε για μαθηματικές ή συναρτησιακές σχέσεις μονοσήμαντα ορισμένες.

Για να μετρηθεί η ένταση της γραμμικής σχέσης χρησιμοποιείται ο γραμμικός συντελεστή συσχέτισης που είδαμε στην παράγραφο 3.6. Γίνεται εμφανές ότι απαραίτητη προϋπόθεση εφαρμογής της απλή γραμμικής παλινδρόμησης ή της προσαρμογής ενός απλού γραμμικού μοντέλου στις δύο αυτές μεταβλητές είναι η ύπαρξη γραμμικής σχέσης. Ένας γραφικός τρόπος για να ελέγξουμε τη γραμμικότητα της σχέσης μεταξύ δύο μεταβλητών είναι το λεγόμενο διάγραμμα διασποράς (**scatter plot**). Πριν μιλήσουμε για τους τρόπους με τους οποίους κατασκευάζουμε αυτό το διάγραμμα στο SPSS, θα αναφερθούμε στις υπόλοιπες υποθέσεις της απλής αλλά και της πολλαπλής γραμμικής παλινδρόμησης για την οποία θα μιλήσουμε παρακάτω. Θα χρησιμοποιήσουμε πάλι τα δεδομένα που αφορούν μετρήσεις στα αυτοκίνητα και ενδιαφερόμαστε να εκτιμήσουμε τη σχέση που συνδέει την ιπποδύναμη (ανεξάρτητη μεταβλητή X) με την κατανάλωση (εξαρτημένη μεταβλητή Y).

Με την απλή γραμμική παλινδρόμηση προσπαθούμε να εκτιμήσουμε τις τιμές της ανεξάρτητης μεταβλητής χρησιμοποιώντας τις τιμές της εξαρτημένης. Οι εκτιμώμενες (ή προσληφθείσες) τιμές θα είναι προφανώς διαφορετικές από τις πραγματικές τιμές της ανεξάρτητης μεταβλητής. Οι αποκλίσεις των τιμών των ανεξάρτητων μεταβλητών από τις αντίστοιχες εκτιμώμενες τιμές τους ονομάζονται κατάλοιπα (ή σφάλματα) και συμβολίζονται με e_i, όπου i είναι δείκτης (i=1,2,3,…n) και αναφέρεται στην i-οστή τιμή. Οι υποθέσεις των γραμμικών μοντέλων αναφέρονται στα κατάλοιπα. Πιο συγκεκριμένα αυτές είναι οι εξής:

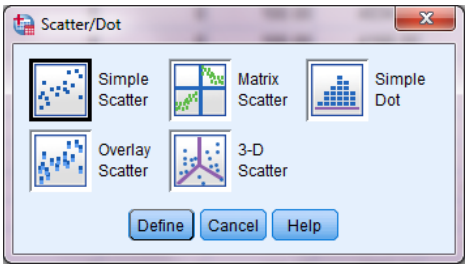
- Υπόθεση της κανονικότητας των καταλοίπων, δηλαδή ότι $\mathbf{e_i \sim N(0,\sigma^2)}$, όπου N είναι ο συμβολισμός της κανονικής κατανομής (Normal distribution) και 0 (μηδέν) και σ^2 είναι ο μέσος και η διακύμανση της κατανομής.
- Υπόθεση της ανεξαρτησίας των καταλοίπων, δηλαδή ότι $\mathbf{Cov(e_i , e_j)=0}$ εάν $\mathbf{i \neq j}$. Αυτό σημαίνει ότι θέλουμε για όλα τα ζεύγη των καταλοίπων η συνδιακύμανση τους (Covariance) να είναι μηδέν.
- Υπόθεση της ομοσκεδαστικότητας των καταλοίπων, δηλαδή $\mathbf{Cov(e_i , e_j)=\sigma^2}$ σταθερή εάν $\mathbf{i=j}$ για κάθε i. Η διακύμανση δηλαδή των καταλοίπων πρέπει να είναι σταθερή και ίση με $\mathbf{\sigma^2}$ για όλα τα κατάλοιπα.

Θα μιλήσουμε επίσης για το πως ελέγχουμε τις παραπάνω υποθέσεις και τι κάνουμε στις περιπτώσεις που δεν ικανοποιούνται.

6.2 Διαγράμματα διασποράς

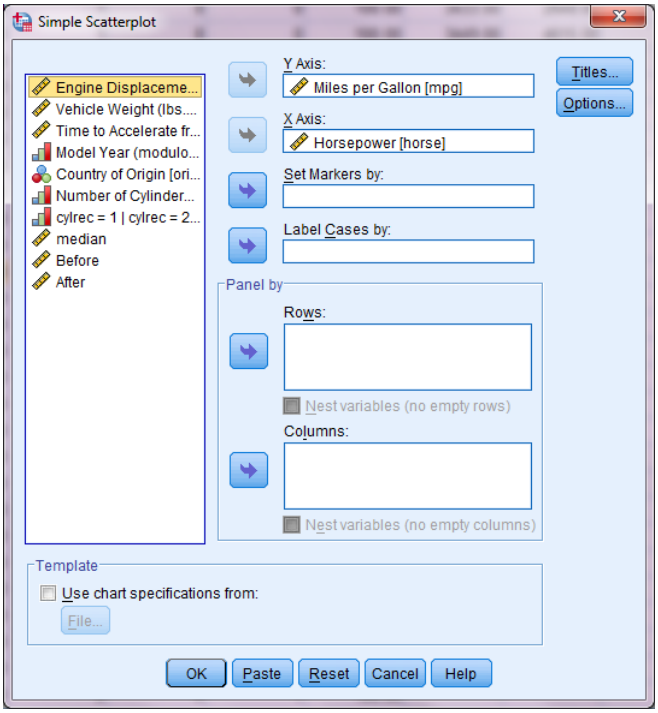
Υπάρχουν δύο τρόποι για να κατασκευάσουμε ένα διάγραμμα διασποράς. Ο πρώτος τρόπος γίνεται με τις επιλογές **Graphs→Legacy Dialogs→Scatter/Dot** για να εμφανιστεί το παράθυρο της εικόνας 52. Σε αυτό το παράθυρο θα επιλέξουμε **Simple**

Scatter και μετά **Define** για να οδηγηθούμε στο παράθυρο της εικόνας 52 στο οποίο θα ορίσουμε τις μεταβλητές που θα απεικονίζονται με το διάγραμμα διασποράς. Εκεί θα «σύρουμε» με το ποντίκι τις μεταβλητές για να τις περάσουμε στα δεξιά κουτάκια.



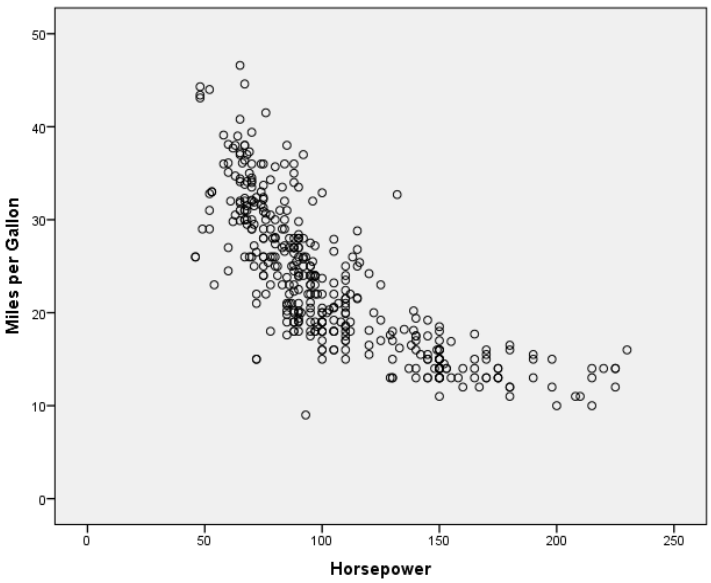
Εικόνα 52

Με την επιλογή **Titles** τοποθετούμε τίτλους στο διάγραμμα και με την επιλογή **Options** επιλέγουμε αν θέλουμε την εμφάνιση των εκλιπουσών τιμών. Υπενθυμίζουμε ότι το SPSS δεν τις χρησιμοποιεί στην ανάλυση, από προεπιλογή. Αν βάλετε μία κατηγορική μεταβλητή στο **Set Markers by…** θα βάλει χρώματα στους κύκλους του διαγράμματος ανάλογα με τις τιμές της κατηγορικής μεταβλητής.



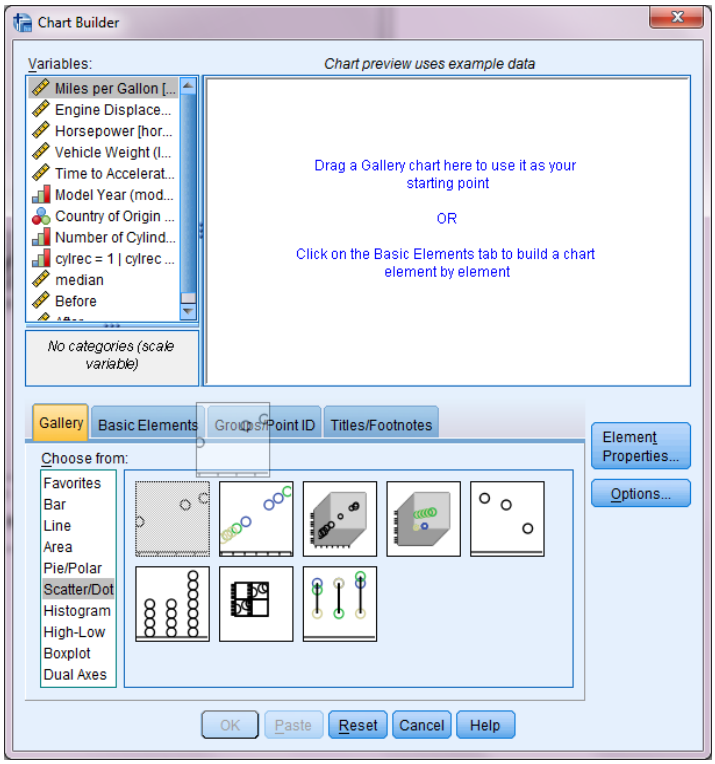
Εικόνα 53

Το διάγραμμα που θα εμφανιστεί για τις μεταβλητές που ορίσαμε παρουσιάζεται στο σχήμα 35. Παρατηρούμε ότι υπάρχει σχέση μεταξύ των δύο μεταβλητών και ότι μπορούμε να υποθέσουμε την υπόθεση της γραμμικότητας της σχέσης μεταξύ τους. Αν υπολογίσουμε το συντελεστή γραμμικής συσχέτισης θα δούμε ότι έχει υψηλή τιμή (-0.771, στατιστικά σημαντικός).



Σχήμα 35: Διάγραμμα διασποράς.

Ο δεύτερος τρόπος εμφάνισης του ίδιου διαγράμματος είναι πάλι από την επιλογή των γραφημάτων και γίνεται ως εξής: **Graphs→Chart Builder** για να εμφανιστεί το παράθυρο της εικόνας 54.

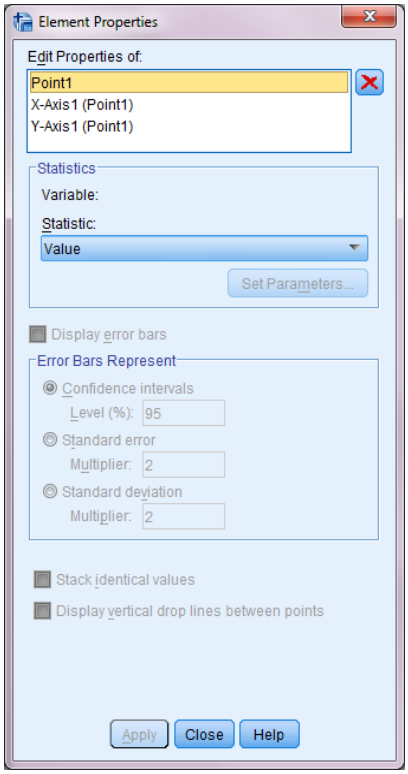


Εικόνα 54

Το παράθυρο της εικόνας 54 αυτό είναι παρόμοιο με αυτό του Excel. Αν παρατηρήσετε επιλέξαμε από την Gallery κάτω δεξιά το **Scatter/Dot** και το οποίο πρέπει να σύρουμε στο μεγάλο κουτί πάνω δεξιά (**Drag a Gallery chart to use it as**

your starting point). Μόλις περάσουμε το διάγραμμα που θέλουμε μέσα στο κουτάκι τότε θα εμφανιστεί το παράθυρο της εικόνας 55 δίπλα από αυτό της εικόνας 54. Πατώντας το κουμπάκι **Element properties** το παράθυρο της εικόνας 55 μπορεί να φύγει και να εμφανιστεί πάλι αν το ξαναπατήσουμε.

Εν συνεχεία θα περάσουμε στον κατακόρυφο και στον οριζόντιο άξονα τις μεταβλητές που θέλουμε. Το γράφημα που θα εμφανιστεί είναι ακριβώς το ίδιο με αυτό του σχήματος 35.



Εικόνα 55

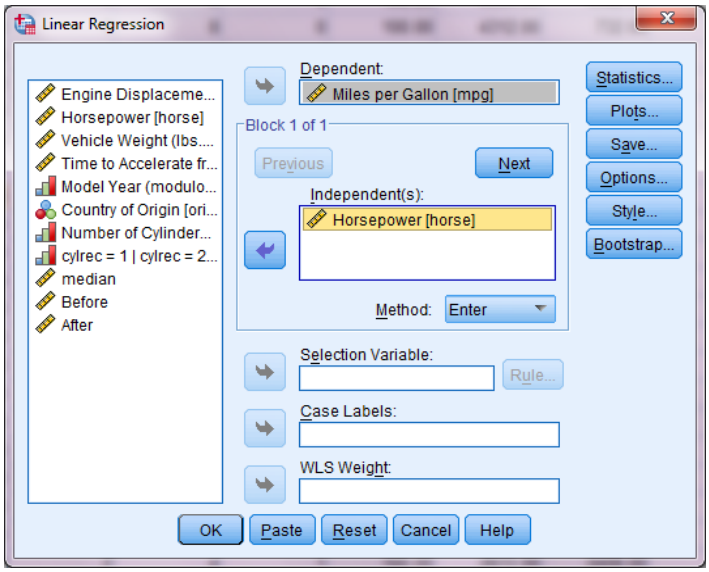
6.3 Απλή γραμμική παλινδρόμηση

Για να μας εμφανίσει το SPSS την ευθεία γραμμικής παλινδρόμησης μαζί με κάποια διαγνωστικά μέτρα επιλέγουμε από το μενού επιλογών **Analyze→Regression→Linear** και θα εμφανιστεί το παράθυρο της εικόνας 55. Στο λευκό κουτάκι με τον τίτλο **Dependent:** θα περάσουμε την εξαρτημένη μεταβλητή και στο λευκό κουτάκι με τον τίτλο **Independent(s):** θα περάσουμε την ανεξάρτητη μεταβλητή της οποίας την επίδραση πάνω στην εξαρτημένη θέλουμε να εκτιμήσουμε.

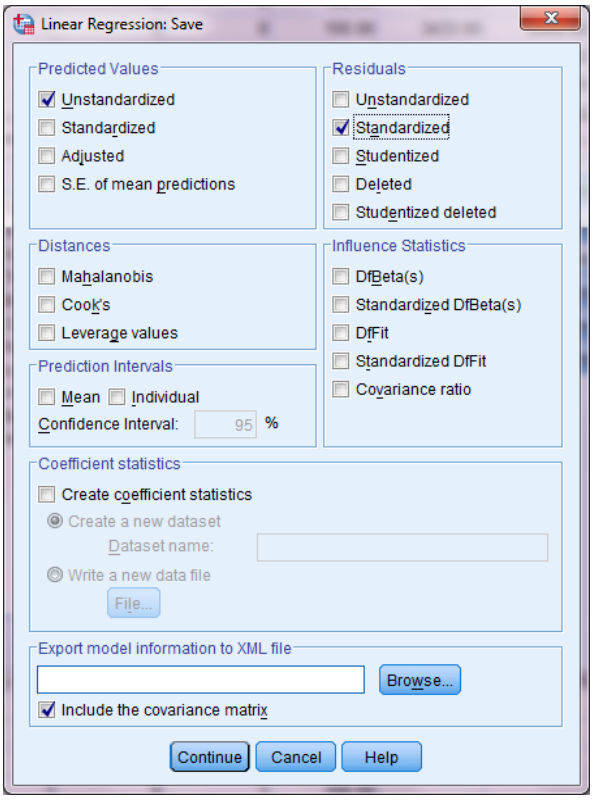
Αν γνωρίζαμε τη χρονική σειρά με την οποία έγιναν οι μετρήσεις θα μπορούσαμε να επιλέξουμε από την επιλογή **Statistics** να εμφανιστεί ο έλεγχος των **Durbin-Watson** το οποίο χρησιμοποιείται για τον έλεγχο ύπαρξης σειριακής συσχέτισης των καταλοίπων. Στην ίδια επιλογή μπορούμε να επιλέξουμε την εμφάνιση των διαγνωστικών μέτρων συγγραμμικότητας στην οποία θα αναφερθούμε όταν μιλήσουμε για την πολλαπλή γραμμική παλινδρόμηση (όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές).

Θα επιλέξουμε **Save** και θα εμφανιστεί το παράθυρο της εικόνας 57, όπου θα επιλέξουμε να σώσουμε τις μη τυποποιημένες εκτιμηθείσες τιμές και τα τυποποιημένα κατάλοιπα (τα έχουμε επιλέξει στην εικόνα 57). Θα μπορούσαμε να

ανοίξουμε το παράθυρο της επιλογής **Plots** για την κατασκευή των απαραίτητων διαγραμμάτων ελέγχων των υποθέσεων, αντί να επιλέξουμε να σώσουμε τις μη τυποποιημένες τιμές από την επιλογή **Save**.



Εικόνα 56

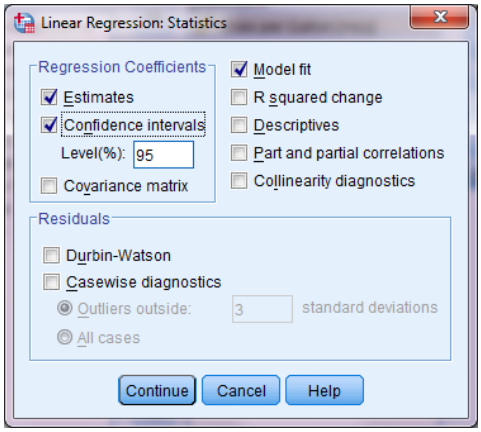


Εικόνα 57

Θα χρειαστεί όμως και σε αυτήν την περίπτωση να σώσουμε τα κατάλοιπα, για να κάνουμε τον έλεγχο κανονικότητας των Kolmogorov-Smirnov. Επιλέξαμε τα τυποποιημένα κατάλοιπα και όχι τα μη τυποποιημένα διότι μπορούμε να δούμε αν

κάποιες τιμές δίνουν κατάλοιπα με μεγάλες κατά απόλυτη τιμή τιμές. Αν είναι πάνω από 2 ή 2.5 τότε αυτό είναι ένδειξη ότι αυτές οι τιμές είναι ακραίες.

Επιλέγοντας **Statistics** στο παράθυρο της εικόνας 56 θα εμφανιστεί το παράθυρο της εικόνας 58 όπου εκεί θα «τικάρουμε» την επιλογή **Confidence intervals**.



Εικόνα 58

Πατώντας **OK** στο παράθυρο της εικόνας 56, θα εμφανιστούν τα αποτελέσματα που παρουσιάζονται στα επόμενα σχήματα.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.771**[a]	**.595**	**.594**	4.974

a. Predictors: (Constant), Horsepower

b. Dependent Variable: Miles per Gallon

Σχήμα 36: Συντελεστής προσδιορισμού.

Η τιμή **R** αναφέρεται στην **απόλυτη τιμή** του συντελεστή γραμμικής συσχέτισης. Το **R Square** είναι το τετράγωνο του συντελεστή γραμμικής συσχέτισης και ονομάζεται συντελεστής προσδιορισμού. Ο συντελεστής προσδιορισμού φανερώνει το ποσοστό της μεταβλητότητας των δεδομένων που εξηγείται από το γραμμικό μοντέλο που προσαρμόσαμε. Δηλαδή, το συγκεκριμένο μοντέλο εξηγεί το 59.5% της μεταβλητότητας των δεδομένων. Ο προσαρμοσμένος συντελεστής προσδιορισμού (**Adjusted R Square**) έχει λάβει υπόψη του και το μέγεθος του δείγματος.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14169.756	1	14169.756	572.709	**.000**[b]
	Residual	9649.237	390	24.742		
	Total	23818.993	391			

a. Dependent Variable: Miles per Gallon

b. Predictors: (Constant), Horsepower

Σχήμα 37: Πίνακας ανάλυσης διακύμανσης.

Ο έλεγχος F (βασίζεται στην F κατανομή) ελέγχει αν όλοι οι παράμετροι του μοντέλου είναι μηδέν ή αν έστω και ένας είναι διάφορος του μηδενός. Έχουμε σκιάσει δύο αριθμούς στον πίνακα της ανάλυσης διακύμανσης. Ο πρώτος (14169.756) δείχνει τη διακύμανση που εξηγείται από το μοντέλο που προσαρμόσαμε και ο δεύτερος (23818.993) δείχνει τη συνολική διακύμανση των δεδομένων. Προφανώς η διαφορά τους είναι η διακύμανση που δεν εξηγείται από το μοντέλο. Το πηλίκο των δύο αριθμών που αναφέραμε είναι στην ουσία ο συντελεστής προσδιορισμού.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**39.855**	.730		54.578	**.000**	38.419	41.290
	Horsepower	**-.157**	.007	**-.771**	-23.931	**.000**	-.170	-.145

a. Dependent Variable: Miles per Gallon

Σχήμα 38: Εκτιμήσεις παραμέτρων.

Το μοντέλο που προσαρμόσαμε στις δύο μεταβλητές (ή ευθεία ελαχίστων τετραγώνων όπως αλλιώς λέγεται) είναι της μορφής $\mathbf{y=\alpha+\beta x+e_i}$, όπου **y** είναι η εξαρτημένη μεταβλητή, **x** η ανεξάρτητη μεταβλητή και **α**, **β** οι παράμετροι του μοντέλου τις οποίες εκτιμάμε και ο όρος $\mathbf{e_i}$ αναφέρεται στο κατάλοιπο της i-οστής τιμής.

Miles per Gallon=39.855-0.157*Horsepower.

Στο σχήμα 38 εμφανίζονται σκιασμένες οι ίδιες τιμές. Η τιμή 39.855 (**Constant**) είναι η τιμή στην οποία η ευθεία ελαχίστων τετραγώνων που προσαρμόσαμε τέμνει τον κάθετο άξονα των **y'y**. Η τιμή -0.157 είναι η κλίση της ευθείας. Επίσης φανερώνει την επίδραση της ανεξάρτητης μεταβλητής στην εξαρτημένη. Για κάθε αύξηση της ανεξάρτητης μεταβλητής κατά 1 μονάδα η εκτιμώμενη μέση τιμή της εξαρτημένης μεταβλητής μειώνεται ή αυξάνεται κατά **β** μονάδες. Δηλαδή για μία αύξηση της ιπποδύναμης κατά 100 μονάδες, η μείωση της εκτιμώμενης μέσης κατανάλωσης είναι ίση με 15.7 μονάδες.

Η στήλη *Standardized* coefficients είναι απλά οι τιμές των συντελεστών που θα προέκυπταν αν είχαμε τυποποιήσει τις μεταβλητές μας στην αρχή. Βέβαια υπάρχει και τύπος που δίνει τη σχέση μεταξύ τους χωρίς να κάνουμε αυτή τη διαδικασία. Η ερμηνεία της τυποποιημένης κλίσης είναι η εξής. Αν αυξηθεί η ανεξάρτητη μεταβλητή κατά 38.522 (τυπική απόκλιση της ιπποδύναμης) μονάδες τότε η εξαρτημένη μεταβλητή θα μειωθεί κατά 0.771*7.816=6.026, όπου 7.816 είναι η τυπική απόκλιση της κατανάλωσης. Προσέξτε ότι είναι ίση με -0.771 (συντελεστής συσχέτισης).

Η τιμή της σταθεράς είναι μηδέν διότι έχουμε τυποποιήσει και την εξαρτημένη μεταβλητή όπως αναφέραμε. Η σημασία των τυποποιημένων συντελεστών θα φανεί περισσότερο στην πολλαπλή παλινδρόμηση παρακάτω.

Η στήλη *Sig.* περιέχει τα παρατηρηθέντα επίπεδα στατιστικής σημαντικότητας, τα οποία χρησιμεύουν για να βγάλουμε συμπεράσματα σχετικά με τη στατιστική σημαντικότητα των παραμέτρων **α** και **β** του μοντέλου. Οι υποθέσεις που ελέγχονται εδώ όσον αφορά στους συντελεστές α και β είναι οι εξής:

Ho: α=0	και	**Ho: β=0**
H_1: α≠0		**H_1: β≠0**

Αφού και οι δύο p-value είναι μικρότερες του 0.05 συμπεραίνουμε ότι και οι δύο μηδενικές υποθέσεις απορρίπτονται, συνεπώς και οι δύο συντελεστές είναι στατιστικά σημαντικοί άρα απαραίτητοι για το μοντέλο.

Οι δύο τελευταίες στήλες περιέχουν τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές του μοντέλου. Αν περιέχουν το μηδέν τότε μπορούμε να πούμε ότι η μηδενική υπόθεση ότι η τιμή του συντελεστή είναι μηδέν δεν απορρίπτεται. Επίσης με το διάστημα εμπιστοσύνης μπορούμε να ξέρουμε για ποιες τιμές ο έλεγχος δεν θα απορριφθεί. Για παράδειγμα, το 95% διάστημα εμπιστοσύνης για την πραγματική τιμή του **β** είναι (-0.17, -0.145). Παρατηρούμε ότι δεν περιέχει το μηδέν, άρα η υπόθεση ότι το **β** μπορεί να πάρει την τιμή 0 απορρίπτεται σε επίπεδο σημαντικότητας 5%. Επίσης όποια τιμή θέλουμε να ελέγξουμε η οποία δε βρίσκεται στο διάστημα αυτό θα απορριφθεί σε α=5%. Αν επιλέγαμε να τρέξουμε και bootstrap τότε δεν θα σωζόντουσαν τα κατάλοιπα και οι εκτιμηθείσες τιμές. Αυτό που θα κερδίζαμε θα ήταν τα διαστήματα εμπιστοσύνης για τις τιμές των παραμέτρων.

Ο τελευταίος πίνακας περιέχει κάποια περιγραφικά μέτρα για τα κατάλοιπα. Ο μέσος είναι ίσος με 0. Να υπενθυμίσουμε ότι η πρώτη υπόθεση που αφορούσε στα κατάλοιπα ήταν ότι ακολουθούν την κανονική κατανομή με μέσο 0. Το γεγονός ότι είναι μηδέν αποδεικνύεται μαθηματικά αλλά και διαισθητικά.

Για να έχουν όμως τα κατάλοιπα μηδενικό άθροισμα θα πρέπει η σταθερά να είναι μέσα στο μοντέλο. Δηλαδή αν για παράδειγμα η p-value για τη σταθερά είναι μεγαλύτερη του 0.05 (άρα δεν απορρίπτεται η υπόθεση ότι η πραγματική της τιμή είναι μηδέν) δεν πρέπει να αφαιρεθεί από το μοντέλο. Με λίγα λόγια, αφήστε τη σημαντικότητα της σταθεράς, η σημαντικότητα του συντελεστή της μεταβλητής είναι σημασίας και άξιας ελέγχου.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.64	32.61	23.45	6.020	392
Residual	-16.212	16.980	.000	4.968	392
Std. Predicted Value	-3.290	1.523	.000	1.000	392
Std. Residual	-3.259	3.414	.000	.999	392

a. Dependent Variable: Miles per Gallon

Σχήμα 39: Περιγραφικά μέτρα καταλοίπων.

Θα δούμε τώρα τρόπους με τους οποίους μπορούμε να ελέγξουμε την ικανοποίηση των υποθέσεων που αναφέραμε. Από το παράθυρο της εικόνας 56 είχαμε επιλέξει να αποθηκεύσουμε στο SPSS Data Editor τα κατάλοιπα και τις εκτιμηθείσες τιμές για την ευθεία ελαχίστων τετραγώνων που εκτιμήσαμε. Όσον αφορά τώρα στην ικανοποίηση της κανονικότητας των καταλοίπων, μπορούμε να την ελέγξουμε είτε γραφικά (P-P ή Q-Q Plots) αλλά και με τον έλεγχο των Kolmogorov-Smirnov που είδαμε στο 5^{ο} κεφάλαιο. Αν η κανονικότητα των καταλοίπων δεν μπορεί να υποτεθεί, τότε προσπαθούμε με μετασχηματισμό των εξαρτημένων μεταβλητών να οδηγηθούμε στην κανονικότητα των καταλοίπων. Σε περίπτωση που το μέγεθος του δείγματος είναι αρκετά μεγάλο μπορούμε να βασιστούμε στην ασυμπτωτική προσέγγιση της κανονικής κατανομής. Σε αντίθετη περίπτωση καταφεύγουμε σε άλλες τεχνικές, όπως εύρωστα γραμμικά μοντέλα, παλινδρόμηση του Theil ή παλινδρόμηση πάνω στις τάξεις μεγέθους των τιμών των μεταβλητών, τεχνική που είναι διαθέσιμη από το SPSS.

Για να ελέγξουμε την ανεξαρτησία και την ομοσκεδαστικότητα των καταλοίπων χρησιμοποιούμε ένα διάγραμμα διασποράς το οποίο θα περιέχει τις εκτιμηθείσες τιμές στον οριζόντιο άξονα και τα κατάλοιπα στο κάθετο άξονα. Αν τα κατάλοιπα είναι ανεξάρτητα θα περιμένουμε ένα "σύννεφο" σημείων να εμφανιστεί στο διάγραμμα. Δεν πρέπει δηλαδή να υπάρχει κάποιο "σχήμα" (pattern) στο διάγραμμα. Ένας ακόμα τρόπος είναι να υπολογίσουμε τη συνδιακύμανση των καταλοίπων με των εκτιμηθέντων τιμών. Αν υπάρχει ανεξαρτησία τότε η συνδιακύμανση θα ισούται με το μηδέν. Δυστυχώς το αντίστροφο δεν είναι πάντα αληθές, οπότε αν βρούμε ότι η συνδιακύμανση είναι μηδέν δε σημαίνει απαραίτητα ότι έχουμε και ανεξαρτησία. Αν έχουμε υποψίες παρόλα αυτά ότι η υπόθεση της ανεξαρτησίας των καταλοίπων δεν ικανοποιείται, τότε η γραμμική παλινδρόμηση δεν μπορεί να εφαρμοστεί, οπότε περνάμε σε άλλες τεχνικές. Εάν τα κατάλοιπα είναι σειριακά συσχετισμένα μπορούμε να χρησιμοποιήσουμε μονότονη παλινδρόμησης η οποία είναι διαθέσιμη από το SPSS.

Αν η υπόθεση της ομοσκεδαστικότητας ικανοποιείται, κοιτάζοντας το ίδιο διάγραμμα θα πρέπει υποθέτοντας δύο παράλληλες στον οριζόντιο άξονα, γραμμές το "σύννεφο" των σημείων να βρίσκεται ανάμεσα στις δύο παράλληλες ευθείες. Με άλλα λόγια το εύρος των σημείων στο κατακόρυφο άξονα πρέπει να είναι σταθερό καθώς "κινούμαστε" στον οριζόντιο άξονα. Αν το εύρος μεγαλώνει ή μικραίνει καθώς μετακινούμαστε δεξιά του οριζόντιου άξονα (σχηματίζοντας ένα "χωνί") τότε δεν μπορούμε να υποθέσουμε ομοσκεδαστικότητα των καταλοίπων. Η αντιμετώπιση αυτού του προβλήματος γίνεται με μετασχηματισμό των ανεξάρτητων ή και των εξαρτημένων μεταβλητών. Αν δε λυθεί το πρόβλημα τότε μπορούμε να δοκιμάσουμε μη παραμετρικές μεθόδους παλινδρόμησης (Theil, παλινδρόμηση στις τάξεις μεγέθους). Να τονίσουμε όμως και σε αυτές τις περιπτώσεις η ετεροσκεδαστικότητα

αποτελεί πρόβλημα (ίσως όχι τόσο σοβαρό όπως στις παραμετρικές μεθόδους). Άλλος τρόπος αντιμετώπισης είναι η χρησιμοποίηση γενικευμένων γραμμικών μοντέλων.

Στο παράδειγμα με τα αυτοκίνητα υπήρχαν κάποιες ενδείξεις στο διάγραμμα διασποράς των καταλοίπων με τις εκτιμηθείσες τιμές ότι η υπόθεση της ομοσκεδαστικότητας δεν ικανοποιείται. Εφαρμόζοντας λογαριθμικό μετασχηματισμό στις τιμές της ανεξάρτητης μεταβλητής και εκτιμώντας την εξίσωση της ευθείς παλινδρόμησης πάνω στη μετασχηματισμένη μεταβλητή το διάγραμμα βελτιώθηκε σημαντικά.

6.4 Πολλαπλή γραμμική παλινδρόμηση

Όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές και θέλουμε να εξετάσουμε την επίδραση τους σε μία εξαρτημένη μεταβλητή χρησιμοποιούμε την πολλαπλή γραμμική παλινδρόμηση. Να τονίσουμε ότι όταν χρησιμοποιούμε το όρο "γραμμική", εννοούμε "γραμμική" ως προς τις παραμέτρους του μοντέλου (**α , β**). Άρα η συνάρτηση της ευθείας ελαχίστων τετραγώνων για την περίπτωση της πολλαπλής γραμμικής παλινδρόμησης θα είναι της μορφής:

$$\mathbf{Y=a+b_1X_1+b_2X_2+\ldots b_pX_p+e_i}\,,$$

όπου με **p** συμβολίζουμε το πλήθος των ανεξάρτητων μεταβλητών και ο όρος $\mathbf{e_i}$ αναφέρεται στο κατάλοιπο της i-οστής τιμής.

Οι υποθέσεις που πρέπει να ικανοποιούνται είναι οι ίδιες με την απλή γραμμική παλινδρόμηση. Μία απαραίτητη προϋπόθεση η οποία είναι απαραίτητη γενικά σε όλα τα μοντέλα με περισσότερες εκ της μίας ανεξάρτητων μεταβλητών είναι η έλλειψη συγγραμμικότητας. Η συγγραμικότητα είναι ένα σοβαρό πρόβλημα για την πολλαπλή γραμμική παλινδρόμηση. Όταν μία ανεξάρτητη μεταβλητή συσχετίζεται με μία άλλη ανεξάρτητη, δηλαδή μέσω της μίας μπορούμε να εκτιμήσουμε τις τιμές της άλλης τότε μιλάμε για πρόβλημα συγγραμμικότητας. Επομένως η ύπαρξη και των δύο μεταβλητών στο μοντέλο δεν είναι δυνατή.

Σκεφτείτε για παράδειγμα την περίπτωση στην οποία έχουμε δύο ανεξάρτητες μεταβλητές, το βάρος και το ύψος και ενδιαφερόμαστε να δούμε πως επιδρούν πάνω σε μία εξαρτημένη μεταβλητή. Είναι προφανές ότι υπάρχει σχέση μεταξύ βάρους και ύψους. Επομένως δε χρειάζεται να γνωρίζουμε και τις δύο μεταβλητές, αφού η γνώση της μίας είναι αρκετή (μέσω της μίας μπορούμε να εκτιμήσουμε τις τιμές της άλλης). Η τοποθέτηση "άχρηστων" μεταβλητών στο μοντέλο μπορεί φαινομενικά να είναι καλή αλλά ουσιαστικά οδηγεί στο λεγόμενο πρόβλημα της υπερ-προσαρμογής του μοντέλου. Με το να κρατήσουμε δηλαδή και τις δύο μεταβλητές που αναφέραμε σε ένα μοντέλο, φαινομενικά το βελτιώνουμε αλλά ουσιαστικά το χειροτερεύουμε. Οπότε ή αφαιρούμε μία εκ των δύο ή χρησιμοποιούμε άλλες τεχνικές, π.χ. κεντροποίηση των τιμών των μεταβλητών, πριν την πολλαπλή γραμμική παλινδρόμηση ή άλλες τεχνικές αντί της παλινδρόμησης. Ένα μέτρο διάγνωσης που προσφέρεται από το SPSS είναι το **VIF**, το οποίο θα το δούμε παρακάτω. Άλλος τρόπος είναι η ύπαρξη υψηλών τιμών του συντελεστή γραμμικής συσχέτισης, το Added Variable Plot και η παλινδρόμηση ανάμεσα σε ζεύγη ανεξάρτητων μεταβλητών, για τις οποίες υποψιαζόμαστε συγγραμμικότητα.

Ο τρόπος με τον οποίο διεξάγουμε πολλαπλή γραμμική παλινδρόμηση στο SPSS είναι ίδιος με προηγουμένως. Ανοίγοντας το παράθυρο της εικόνας 56 θα περάσουμε δύο ανεξάρτητες μεταβλητές στο λευκό κουτάκι (**Independent(s):**) αντί

για μία, έστω ότι περνάμε ακόμα την μεταβλητή που αναφέρεται στην επιτάχυνση των αυτοκινήτων. Η επιπλέον επιλογή είναι να επιλέξουμε την επιλογή **Statistics** για να εμφανιστεί το παράθυρο της εικόνας 58. Επιλέξαμε την επιλογή που αφορά στα διαγνωστικά μέτρα της συγγραμμικότητας (**Collinearity diagnostics**). Αν επιλέξουμε **Confidence Intervals** θα εμφανιστούν τα 95% διαστήματα εμπιστοσύνης για τις εκτιμήσεις των παραμέτρων **α** και **β**. Πατώντας **Continue** και μετά **OK** θα εμφανιστούν τα αποτελέσματα στο Output.

Η λογική με την οποία υπολογίζονται οι συντελεστές προσδιορισμού στον πίνακα του σχήματος 40 είναι ίδια με την περίπτωση της απλή γραμμικής παλινδρόμησης.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.783[a]**	**.613**	**.611**	4.867

a. Predictors: (Constant), Time to Accelerate from 0 to 60 mph (sec), Horsepower

b. Dependent Variable: Miles per Gallon

Σχήμα 40: Συντελεστής προσδιορισμού

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14605.572	2	7302.786	308.331	**.000[b]**
	Residual	9213.422	389	23.685		
	Total	23818.993	391			

a. Dependent Variable: Miles per Gallon

b. Predictors: (Constant), Time to Accelerate from 0 to 60 mph (sec), Horsepower

Σχήμα 41: Πίνακας ανάλυσης διακύμανσης.

Ο παρακάτω πίνακας του σχήματος 42 περιέχει τις εκτιμήσεις του μοντέλου. Είναι ίδιος με την περίπτωση της απλής γραμμικής παλινδρόμησης. Το μοντέλο δηλαδή που προσαρμόστηκε στα δεδομένα αυτά είναι το εξής:

Miles per Gallon=50.812 -0.184*Horsepower-0.528*Time to Accelerate

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	**50.812**	**2.652**		19.157	**.000**	45.597	56.027		
	Horsepower	**-.184**	**.009**	-.902	-20.595	**.000**	-.202	-.166	.519	**1.927**
	Time to Accelerate from 0 to 60 mph (sec)	**-.528**	**.123**	-.188	-4.290	**.000**	-.769	-.286	.519	**1.927**

a. Dependent Variable: Miles per Gallon

Σχήμα 42: Εκτιμήσεις παραμέτρων πολλαπλής παλινδρόμησης.

Όλοι οι συντελεστές είναι στατιστικά σημαντικοί και η ερμηνεία που θα δώσουμε σε αυτούς τους συντελεστές είναι παρόμοια με την περίπτωση της μίας ανεξάρτητης μεταβλητής. Η σταθερά (50.812) είναι η τιμή στην οποία η ευθεία (ελαχίστων τετραγώνων) τέμνει τον κατακόρυφο άξονα συντεταγμένων. Ο συντελεστής της ιπποδύναμης (-0.184) δείχνει τη μείωση στην αναμενόμενη μέση τιμή της κατανάλωσης αν αυξήσουμε την ιπποδύναμη κατά μία μονάδα, δεδομένου ότι κρατάμε την επιτάχυνση σταθερή. Είναι δηλαδή η κύρια επίδραση της ιπποδύναμης πάνω στην κατανάλωση. Ο συντελεστής της επιτάχυνσης (-0.528) αναφέρεται στην κύρια επίδραση της επιτάχυνσης πάνω στην κατανάλωση. Για κάθε μονάδα αύξησης της επιτάχυνσης η αναμενόμενη μέση κατανάλωση μειώνεται κατά 0.528 μονάδες δοθέντος ότι έχουμε κρατήσει την επίδραση της ιπποδύναμης σταθερή.

Οι τυποποιημένες τιμές των συντελεστών εδώ έχουν σημασία διότι μπορεί να γίνει μία σχετική σύγκριση μεταξύ τους. Η τυπική απόκλιση της εξαρτημένης μεταβλητής είναι ίση με 7.816, της ιπποδύναμης με 38.522 και του χρόνου επιτάχυνσης με 2.821. Οι τιμές των τυποποιημένων συντελεστών είναι -0.902 για την ιπποδύναμη και -0.188 για το χρόνο επιτάχυνσης. Αν αυξηθεί η ιπποδύναμη κατά 38.522 μονάδες, τότε η εξαρτημένη μεταβλητή αναμένεται να μειωθεί κατά 0.902*7.816=7.05 μονάδες, ενώ αν αυξηθεί ο χρόνος επιτάχυνσης κατά 2.821 μονάδες η αναμενόμενη μείωση της εξαρτημένης μεταβλητής είναι ίση με 0.188*7.816=1.47 μονάδες. Επομένως η σχετική επίδραση της ιπποδύναμης είναι μεγαλύτερη από το χρόνο επιτάχυνσης.

Οι δύο τελευταίες στήλες του πίνακα του σχήματος 42 αναφέρονται σε διαγνωστικά συγγραμμικότητας, όπως και ο παρακάτω πίνακας. Το **VIF** (Variation Inflation Factor) είναι μέτρο διάγνωσης συγγραμμικότητας. Τιμές μεγαλύτερες του δύο αποτελούν ένδειξη ότι έχουμε πρόβλημα συγγραμμικότητας. Οι τιμή της **Tolerance** για μία τιμή φανερώνει το ποσοστό της διακύμανσης της μεταβλητής που εξηγείται από τις υπόλοιπες ανεξάρτητες μεταβλητές του μοντέλου. Πιο συγκεκριμένα ισχύει ότι το ποσοστό αυτό είναι ίσο με (1-Tolerance)%. Τιμές της Tolerance μικρότερες του 0.5 αποτελούν ένδειξη του προβλήματος. Βλέπουμε ότι οι τιμές για τις δύο ανεξάρτητες μεταβλητές του μοντέλου είναι σε οριακές τιμές. Η στήλη του παρακάτω πίνακα **VIF** αποτελεί ένα ακόμα διαγνωστικό του προβλήματος.

Τιμές μεγαλύτερες του 15 φανερώνουν πιθανό πρόβλημα συγγραμμικότητας και τιμές άνω του 30 σοβαρό πρόβλημα συγγραμμικότητας.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Horsepower	Time to Accelerate from 0 to 60 mph (sec)
1	1	2.871	1.000	.00	.01	.00
	2	.124	4.820	.00	.31	.05
	3	.005	23.139	1.00	.69	.95

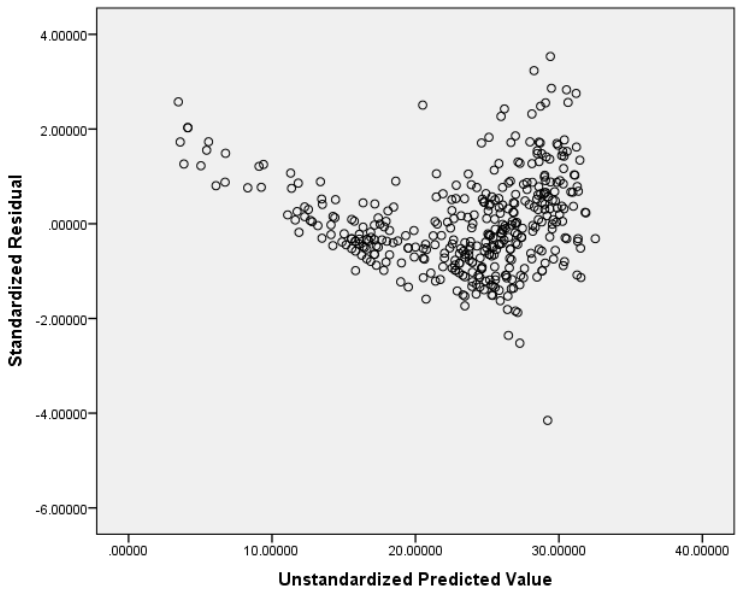
a. Dependent Variable: Miles per Gallon

Σχήμα 43: Διαγνωστικά συγγραμμικότητας.

Οι έλεγχοι των υποθέσεων για την ισχύ του μοντέλου είναι ίδιες με την περίπτωση της μίας ανεξάρτητης μεταβλητής και ελέγχονται με τις ίδιες μεθόδους.

6.5 Παραβίαση των υποθέσεων στη γραμμική παλινδρόμηση

Είπαμε προηγουμένως πως οι υποθέσεις της γραμμικής παλινδρόμησης είναι η ομοσκεδαστικότητα, η ανεξαρτησία και η κανονικότητα των καταλοίπων. Θα εκτελέσουμε την παλινδρόμηση δύο μεταβλητών, της κατανάλωσης (mpg) πάνω στον κυβισμό (cu.inches). Κατασκευάζοντας το διάγραμμα διασποράς των τυποποιημένων καταλοίπων με τις εκτιμηθείσες τιμές το αποτέλεσμα είναι το εξής:

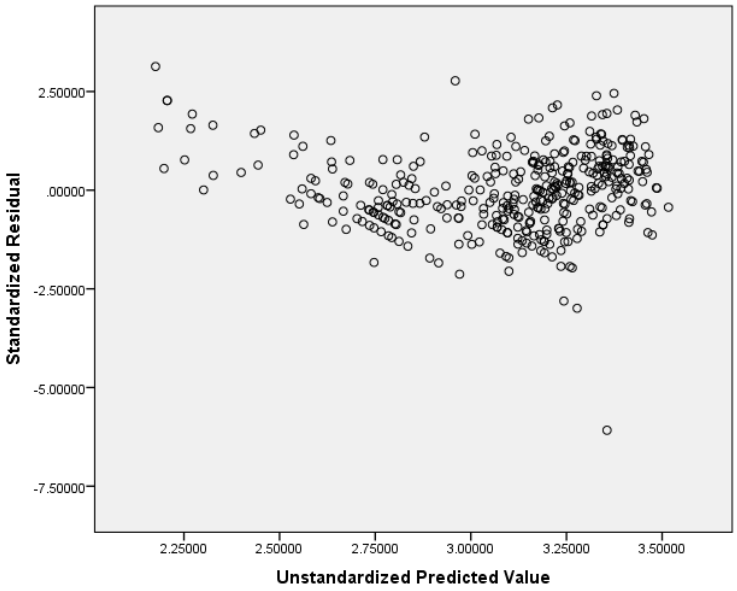


Σχήμα 44: Διάγραμμα διασποράς.

Παρατηρούμε από το παραπάνω διάγραμμα ότι καθώς κινούμαστε προς τα δεξιά, το εύρος των καταλοίπων απλώνεται, δημιουργώντας ένα "χωνί".

Συμπεραίνουμε ότι μάλλον κάτι δεν πάει καλά όσον αφορά την υπόθεση της σταθερής διακύμανσης των καταλοίπων. Διεξάγοντας τον μη παραμετρικό έλεγχο κανονικότητας των Kolmogorov-Smirnov για τα κατάλοιπα παίρνουμε μία p-τιμή (Sig.) ίση με 0.018, χαμηλότερη του 0.05. Άρα υπάρχουν ενδείξεις ότι η κατανομή των καταλοίπων διαφέρει στατιστικά σημαντικά από την κανονική κατανομή. Δύο υποθέσεις της γραμμικής παλινδρόμησης παραβιάζονται.

Ένας τρόπος αντιμετώπισης είναι να μετασχηματίσουμε τις τιμές της εξαρτημένης μεταβλητής (mpg), χρησιμοποιώντας το φυσικό λογάριθμο. Πηγαίνοντας στο παράθυρο της εικόνας 10, θα επιλέξουμε την **Arithmetic** στο **Function group** και θα πατήσουμε το βελάκι να "πάει επάνω". Μετά επιλέγουμε τη μεταβλητή που θέλουμε (mpg) και την περνάμε δεξιά μέσα στη συνάρτηση του λογαρίθμου (LN(mpg)). Στο **Target variable** συμπληρώνουμε με ένα όνομα για τη νέα μεταβλητή. Τώρα θα τρέξουμε την απλή γραμμική παλινδρόμηση της λογαριθμοποιημένης μεταβλητής (κατανάλωση) πάνω στον κυβισμό. Κατασκευάζοντας το διάγραμμα καταλοίπων με τις εκτιμηθείσες τιμές θα πάρουμε το παρακάτω αποτέλεσμα



Σχήμα 45: Διάγραμμα διασποράς.

Παρατηρούμε ότι το σχήμα του διαγράμματος δεν είναι πια ένα χωνί, αλλά το εύρος των καταλοίπων έχει σταθεροποιηθεί καθώς κινούμαστε από τα αριστερά προς τα δεξιά. Ένα κυκλάκι που φαίνεται στο κάτω μέρος και των δύο διαγραμμάτων είναι ένα ακραίο σημείο. Ο έλεγχος κανονικότητας υπολόγισε μία p-τιμή (Sig.) ίση με 0.160. Άρα υπάρχουν ενδείξεις ότι η υπόθεση της κανονικότητας των καταλοίπων ικανοποιείται. Προσοχή όμως, ότι συμπεράσματα προκύψουν και ο σχολιασμός των παραμέτρων του μοντέλου αναφέρεται στις λογαριθμοποιημένες τιμές της εξαρτημένης μεταβλητής. Για να επιστρέψουμε από το λογάριθμο πίσω στις κανονικές τιμές, μπορούμε απλά να χρησιμοποιήσουμε τη συνάρτηση **Exp** στην επιλογή **Compute**. Εδώ χρησιμοποιήσαμε το φυσικό λογάριθμο, αν είχαμε χρησιμοποιήσει το λογάριθμο με βάση το 10 θα χρησιμοποιούσαμε άλλο τρόπο για να επιστρέψουμε στις αρχικές τιμές.

Παρατηρήστε επιπλέον ότι μία τιμή έχει κατάλοιπο μεγαλύτερο από 5 (σε απόλυτη τιμή) και φαίνεται στο διάγραμμα του σχήματος 45 κάτω δεξιά. Αυτή η τιμή

δείχνει ότι η παρατήρηση στην οποία αντιστοιχεί είναι ακραίο σημείο. Αυτό το λέω σα σχόλιο απλά. Το τι κάνουμε σε αυτές τις περιπτώσεις δε θα το πούμε εδώ. Αλλά για τον ενδιαφερόμενο αναγνώστη θα πούμε ότι μπορεί να κοιτάξει για εύρωστες στατιστικές μεθόδους (robust statistics).

6.7 Μέθοδοι πολλαπλής παλινδρόμησης

Όταν τρέξαμε την πολλαπλή παλινδρόμηση με τις δύο ανεξάρτητες μεταβλητές χρησιμοποιήσαμε το παράθυρο της εικόνας 56. Κάτω από το λευκό κουτάκι στο οποίο περάσαμε τις ανεξάρτητες μεταβλητές υπάρχει η μέθοδος της παλινδρόμησης που επιθυμούμε να χρησιμοποιήσουμε (**Method**). Είναι προεπιλεγμένη η επιλογή **Enter**. Την πολλαπλή γραμμική παλινδρόμηση με τις δύο ανεξάρτητες μεταβλητές την τρέξαμε χρησιμοποιώντας αυτή τη μέθοδο. Ας δούμε όμως τι είναι και τι κάνουν αυτές οι μέθοδοι. Είδαμε ότι όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές ελέγχουμε την ύπαρξη συγγραμμικότητας μεταξύ των μεταβλητών κοιτάζοντας το δείκτη **VIF.**

Τι γίνεται όμως αν δύο ανεξάρτητες μεταβλητές δεν είναι συγγραμμικές (όπως οριακά στο παράδειγμα) αλλά παρόλα αυτά μία από τις δύο πρέπει να φύγει; Υπάρχουν και άλλα κριτήρια που χρησιμοποιούνται για τη διαπίστωση της καταλληλότητας μίας μεταβλητής να χρησιμοποιηθεί στο μοντέλο. Αν δηλαδή υπάρχει μία μεταβλητή ήδη στο μοντέλο, μήπως η χρησιμοποίηση και μίας δεύτερης δεν προσφέρει παραπάνω πληροφορία;

Η μέθοδος **Enter** τρέχει την πολλαπλή γραμμική παλινδρόμηση χρησιμοποιώντας όλες τις ανεξάρτητες μεταβλητές. Θα γίνει προφανές μετά ότι θα πρέπει να είναι η λιγότερο συχνά χρησιμοποιούμενη.

Η μέθοδος **Forward** κάνει κάτι καλύτερο, ελέγχει με βάση ένα κριτήριο ποια είναι η καλύτερη μεταβλητή που πρέπει να "εισέλθει" πρώτη στο μοντέλο. Αν η "καλύτερη" μεταβλητή δεν ικανοποιεί το κριτήριο για να εισέλθει στο μοντέλο, τότε καμία μεταβλητή δεν θα εισέλθει στο μοντέλο. Η μεταβλητή που επιλέγεται εισέρχεται στο μοντέλο. Η μέθοδος αυτή συνεχίζει μετά ψάχνοντας την καλύτερη μεταβλητή που πρέπει να συμπεριληφθεί στο μοντέλο δοθέντος ότι είναι ήδη μία μεταβλητή μέσα στο μοντέλο. Αν η δεύτερη μεταβλητή ικανοποιεί το κριτήριο εισέρχεται στο μοντέλο και ούτω καθεξής.

Η μέθοδος **Backward** κάνει την αντίστροφη διαδικασία. Ξεκινάει με το μοντέλο που περιέχει όλες τις ανεξάρτητες μεταβλητές μέσα και αρχίζει να βγάζει μεταβλητές που δεν ικανοποιούν κάποια κριτήρια μέχρι να καταλήξει σε ένα μοντέλο στο οποίο όλες οι ανεξάρτητες μεταβλητές ικανοποιούν κάποιο κριτήριο. Και οι δύο αυτές μέθοδοι έχουν το ίδιο μειονέκτημα. Αν μία μεταβλητή μπει στο μοντέλο ή βγει από το μοντέλο, τότε δε γυρίζει πίσω. Δηλαδή μία μεταβλητή που μπήκε δεύτερη η Τρίτη μπορεί να κάνει μία μεταβλητή που μπήκε προηγούμενα στο μοντέλο να είναι άχρηστη, αλλά επειδή μπήκε πιο νωρίς στο μοντέλο, τώρα δεν μπορεί να βγει.

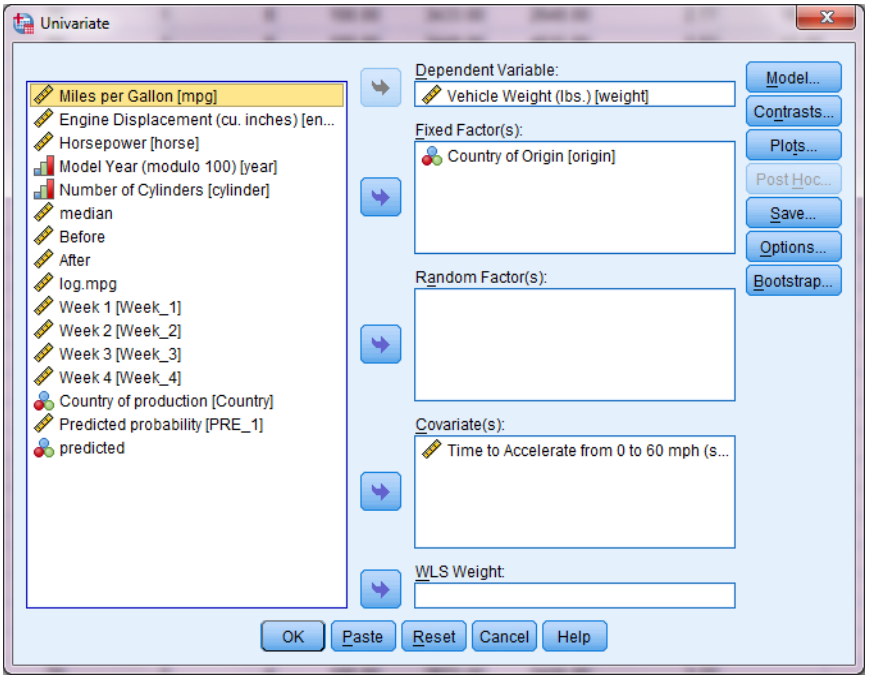
Η μέθοδος **Stepwise** κάνει μία πιο σύνθετη διαδικασία, ένα συνδυασμό και των δύο τελευταίων μεθόδων αντιμετωπίζοντας το μειονέκτημα αυτών. Βρίσκει την καλύτερη μεταβλητή για να εισέλθει στο μοντέλο και αφού εισέλθει στο μοντέλο κάνει έλεγχο μήπως πρέπει να βγει από το μοντέλο. Συνεχίζει την ίδια διαδικασία για όλες τις ανεξάρτητες μεταβλητές. Δηλαδή αφού μπουν κάποιες μεταβλητές κάνει έλεγχο μήπως κάποια/ες πρέπει να αφαιρεθούν από το μοντέλο. Δηλαδή στην ουσία ξεκινάει με **Forward** μέθοδο και συνεχίζει με **Backward** μέθοδο. Είναι προφανές ότι αυτή είναι η καλύτερη μέθοδος γραμμικής παλινδρόμησης.

Όλες όμως οι μέθοδοι έχουν ένα προφανές μειονέκτημα, δεν ελέγχουν τις υποθέσεις του γραμμικού μοντέλου, αυτό είναι δική μας δουλειά. Δοκιμάστε να τρέξετε την πολλαπλή γραμμική παλινδρόμηση που τρέξαμε στο παράδειγμα με τις δύο ανεξάρτητες μεταβλητές και θα δείτε ότι δε χρειάζεται να κρατήσετε και τις δύο μεταβλητές. Το SPSS θα εμφανίσει και ένα πίνακα με μία ανεξάρτητη μεταβλητή που δε χρειάζεται (**Excluded Variables**).

6.8 Πολλαπλή γραμμική παλινδρόμηση με κατηγορική(ές) μεταβλητή(ές)

Ας δούμε τώρα τι γίνεται αν έχουμε μία κατηγορική μεταβλητή μέσα στο μοντέλο. Στο προηγούμενο παράδειγμα είχαμε δύο συνεχείς μεταβλητές. Τώρα θα αντικαταστήσουμε την ιπποδύναμη με τη χώρα προέλευσης. Εάν λοιπόν έχουμε έστω και μία κατηγορική μεταβλητή στο μοντέλο μας δεν θα πάμε από το μενού της παλινδρόμησης που περιγράψαμε προηγουμένως αλλά θα ακολουθήσουμε άλλο μονοπάτι[**].

Θα επιλέξουμε **Analyze→General Linear Model→Univariate** και θα εμφανιστεί το παράθυρο της εικόνας 59. Εκεί θα βάλουμε την εξαρτημένη μεταβλητή στο πάνω κουτάκι, την κατηγορική (και γενικά τις κατηγορικές) στο κουτάκι **Fixed Factor(s):** και τη συνεχής ανεξάρτητη μεταβλητή (και γενικά όλες τις συνεχείς μεταβλητές) στο κουτάκι **Covariate(s):**. Αν «κλικάρουμε»στην επιλογή **Save** θα εμφανιστεί ένα παράθυρο παρόμοιο με αυτό της εικόνας 57. Εκεί μπορούμε να σώσουμε τα τυποποιημένα κατάλοιπα και τις εκτιμηθείσες τιμές όπως και προηγουμένως.

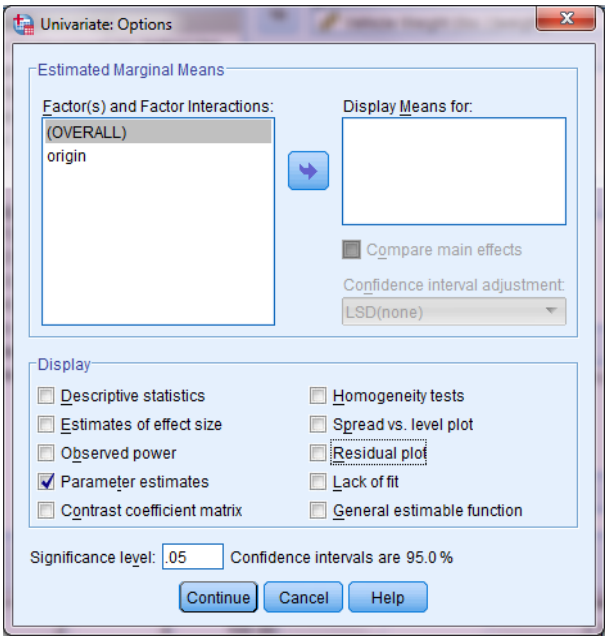


Εικόνα 59

Εμείς όμως θα επιλέξουμε το **Options** μόνο και στο παράθυρο της εικόνας 60 θα «τικάρουμε» την επιλογή **Parameter estimates**. Πατάμε **Continue** για να επιστρέψουμε στο βασικό παράθυρο της εικόνας 59 και μετά **OK**.

[**]Θα μπορούσαμε να πάμε μέσω της επιλογής **Generalized Linear Models** αλλά εκεί είναι λίγο πιο πολύπλοκη η διαδικασία

Θα εμφανιστούν δύο βασικοί πίνακες, των σχημάτων 46 και 47. Ας πάμε πρώτα στον πίνακα του σχήματος 44. Εδώ βλέπουμε τη στατιστική σημαντικότητα όλων των μεταβλητών. Βλέπουμε δύο γραμμές (αφού έχουμε δύο μεταβλητές) η στατιστική σημαντικότητα της κάθε μίας φαίνεται στην τελευταία στήλη (**Sig.**). Η p-value και για τις δύο περιπτώσεις είναι μικρότερη του 0.001, άρα οι μεταβλητές είναι στατιστικά σημαντικές. Παρατηρήστε ότι στη στήλη **df** η επιτάχυνση (**accel**) έχει την τιμή 1 (ένας βαθμός ελευθερίας), ενώ η χώρα προέλευσης (**origin**) έχει την τιμή 2 (2 βαθμοί ελευθερίας).



Εικόνα 60

Tests of Between-Subjects Effects

Dependent Variable: Vehicle Weight (lbs.)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	130219752.616[a]	3	43406584.205	110.687	.000
Intercept	189749495.215	1	189749495.215	483.862	.000
accel	24827862.557	**1**	24827862.557	63.311	**.000**
origin	74950891.835	**2**	37475445.917	95.562	**.000**
Error	157254719.359	401	392156.407		
Total	3872185850.000	405			
Corrected Total	287474471.975	404			

a. R Squared = **.453** (Adjusted R Squared = **.449**)

Σχήμα 46: Πίνακας ανάλυσης διακύμανσης.

Στον πίνακα του σχήματος 47 βλέπουμε τις εκτιμηθείσες τιμές των συντελεστών. Α επιτάχυνση έχει αρνητικό πρόσημο, άρα επιδρά αρνητικά στο βάρος του αυτοκινήτου (μάλλον το ανάποδο ισχύει, αλλά ας το παρακάμψουμε αυτό τώρα).

Στο κάτω μέρος του πίνακα του σχήματος 46 βλέπουμε και την τιμή του συντελεστή προσδιορισμού και την τιμή του διορθωμένου συντελεστή προσδιορισμού.

Η χώρα προέλευσης έχεις δύο τιμές όμως, παρότι είναι μία μεταβλητή. Επειδή η μεταβλητή είναι κατηγορική, όχι αριθμητική, δεν μπορούμε να κάνουμε μαθηματικές πράξεις. Δεν μπορούμε να πούμε Αμερική+Ευρώπη=Ιαπωνία για παράδειγμα. Για αυτό το λόγο δημιουργούμε τις λεγόμενες ψευδομεταβλητές. Η κατηγορική μας μεταβλητή έχει 3 τιμές; τότε θέλουμε 2 ψευδομεταβλητές. Γενικά θέλουμε μία λιγότερη. Δείτε για παράδειγμα τις εδώ δημιουργηθέντες μεταβλητές.

D_1=1 αν το αυτοκίνητο είναι Αμερικάνικο και 0 ειδάλλως.
D_2=1 αν το αυτοκίνητο είναι Ευρωπαϊκό και 0 ειδάλλως.

Όπως βλέπετε δε χρειαζόμαστε πληροφορία για το αν το αυτοκίνητο είναι Ιαπωνικό ή όχι. Αν δεν είναι Αμερικάνικο, η τιμή της D_1 είναι 0. Αν δεν είναι ούτε Ευρωπαϊκό τότε και η τιμή της D_2 είναι 0. Άρα, είναι Ιαπωνικό. Στον πίνακα του σχήματος 45 η τελευταία τιμή της στήλης **B** είναι 0. Να θυμίσουμε ότι **origin=1** είναι τα Αμερικάνικα αυτοκίνητα, **origin=2** είναι τα Ευρωπαϊκά και **origin=3** είναι τα Ιαπωνικά.

Τι σημαίνουν όμως οι τιμές της στήλης **B** για τα Αμερικάνικα και τα Ευρωπαϊκά αυτοκίνητα; Η τιμή για τα Αμερικάνικα αυτοκίνητα είναι 1031.578. Αυτό σημαίνει ότι κατά μέσο όρο τα Αμερικάνικα αυτοκίνητα είναι βαρύτερα από τα Ιαπωνικά κατά 1031.578 λίβρες. Αυτή η διαφορά είναι στατιστικά σημαντική. Τα Ευρωπαϊκά αυτοκίνητα είναι βαρύτερα κατά 270.0098 λίβρες κατά μέσο όρο και αυτή η διαφορά είναι στατιστικά σημαντική. Άρα η μέση διαφορά στο βάρος μεταξύ Αμερικάνικων και Ευρωπαϊκών αυτοκινήτων είναι ίση με 1031.578-270.098=761.48 λίβρες, αλλά δεν βλέπουμε αν αυτή η διαφορά είναι στατιστικά σημαντική.

Το SPSS έχει θέσει ως επίπεδο αναφοράς την μεγαλύτερη τιμή της χώρα προέλευσης (την τιμή 3). Αν θέλουμε να αλλάξουμε την τιμή αναφοράς θα πρέπει να αλλάξουμε την κωδικοποίηση και να ορίσουμε την τιμή αναφοράς που θέλουμε να έχει τη μεγαλύτερη τιμή. Φυσικά υπάρχει και άλλος τρόπος να δούμε τη στατιστική σημαντικότητα μεταξύ Ευρωπαϊκών και Αμερικάνικων αυτοκινήτων από το μενού που είμαστε τώρα, αλλά προτιμήσαμε να το αφήσουμε προς το παρόν.

Parameter Estimates

Dependent Variable: Vehicle Weight (lbs.)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	3710.404	199.979	18.554	.000	3317.265	4103.543
accel	**-92.083**	11.573	-7.957	**.000**	-114.834	-69.332
[origin=1]	**1031.578**	81.983	12.583	**.000**	870.408	1192.747
[origin=2]	**270.098**	101.944	2.649	**.008**	69.686	470.509
[origin=3]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Σχήμα 47: Πίνακας με τις εκτιμήσεις των παραμέτρων της παλινδρόμησης.

Το μοντέλο της γραμμικής παλινδρόμησης γράφεται τώρα ως εξής:

Weight=3710.404-92.083*accel+1031.578*D_1+270.098*D_2

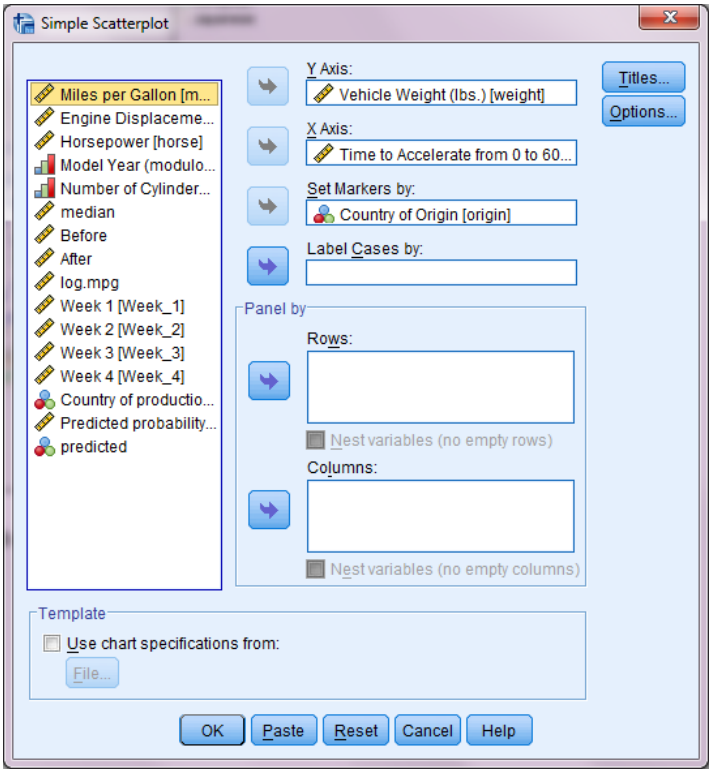
Δείτε τον πίνακα 4 για να δείτε πως αλλάζει το μοντέλο ανάλογα με τη χώρα προέλευσης. Στην ουσία η χώρα προέλευσης οδηγεί σε ένα μοντέλο με διαφορετική σταθερά. Άρα έχουμε ένα μοντέλο που αποτελείται από τρεις παράλληλες ευθείες γραμμές.

Χώρα προέλευσης	Μοντέλο	Μοντέλο
Αμερικάνικα	**3710.404-92.083*accel+1031.578**	**4741.982-92.083*accel**
Ευρωπαϊκά	**3710.404-92.083*accel+270.098**	**3980.502-92.083*accel**
Ιαπωνικά	**3710.404-92.083*accel**	**3710.404-92.083*accel**

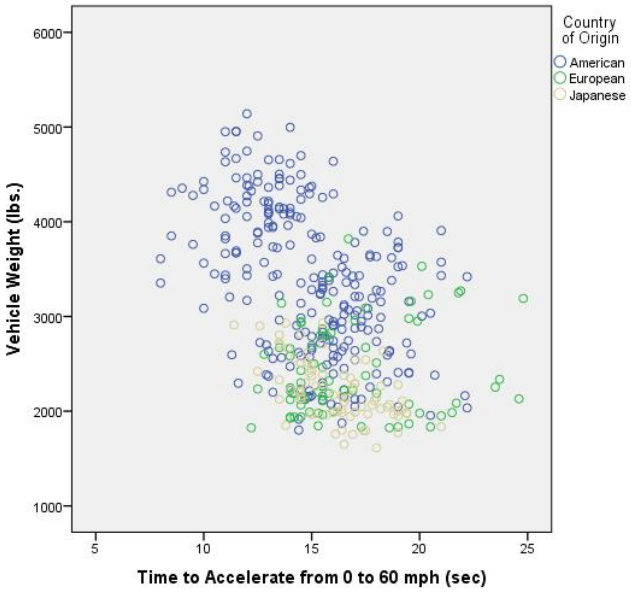
Πίνακας 4: Μοντέλο παλινδρόμησης ανάλογα με τη χώρα προέλευσης.

Ας δούμε όμως ένα άλλο θέμα εξίσου σημαντικό. Αφού έχουμε δύο μεταβλητές, μία κατηγορική και μία συνεχής μπορούμε να φτιάξουμε ένα γράφημα για να εξηγήσουμε αυτό που θέλουμε να πούμε. Η επιτάχυνση επιδρά στο βάρος του αυτοκινήτου αρνητικά όπως είπαμε. Αν αυξήσουμε το χρόνο, που χρειάζεται ένα αυτοκίνητο μέχρι να πιάσει τα 60 μίλια την ώρα (περίπου 96 χιλιόμετρα την ώρα), κατά ένα δευτερόλεπτο τότε αναμένουμε το βάρος του αυτοκινήτου να μειωθεί κατά 92 λίβρες περίπου ασχέτως της χώρας προέλευσης. Δηλαδή την ίδια μείωση αναμένουμε και για τα Αμερικάνικα, και για τα Ευρωπαϊκά και για τα Ιαπωνικά αυτοκίνητα. Μήπως είναι πιο λογικό να υποθέσουμε ότι η επίδραση της επιτάχυνσης στο βάρος του αυτοκινήτου διαφέρει από χώρα σε χώρα; Μήπως δηλαδή οι τρεις ευθείες γραμμές δεν πρέπει να είναι παράλληλες, αλλά να έχουν διαφορετική κλίση η κάθε μία;

Ας πάμε να κατασκευάσουμε ένα διάγραμμα διασποράς να δούμε τι γίνεται. Στο παράθυρο της εικόνας 61 (ίδιο με της εικόνας 53) θα βάλουμε στην εξαρτημένη μεταβλητή (**Y Axis:**) το βάρος, στην ανεξάρτητη μεταβλητή (**X Axis:**) επιτάχυνση και στο κουτάκι **Set Markers by:** την κατηγορική μεταβλητή, τη χώρα προέλευσης.



Εικόνα 61

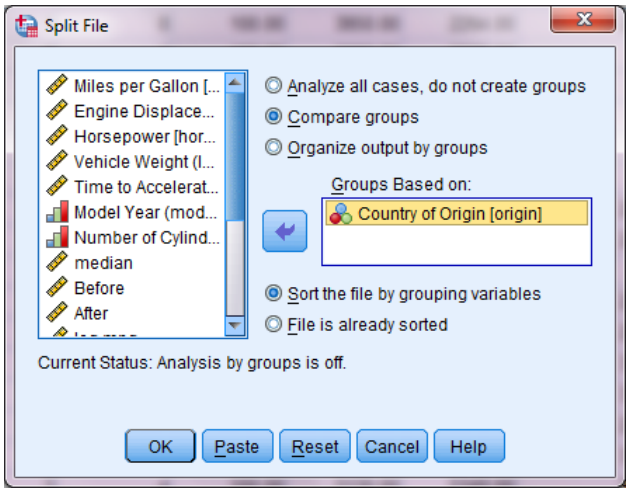


Σχήμα 48: Διάγραμμα διασποράς μεταξύ επιτάχυνσης και βάρους αυτοκινήτου ανάλογα με τη χώρα προέλευσης.

Αν παρατηρήσετε προσεκτικά στο διάγραμμα διασπορά του σχήματος 48 θα δείτε ότι όσο ανεβαίνει η επιτάχυνση το βάρος των Αμερικάνικων αυτοκινήτων μειώνεται. Το ίδιο ισχύει και για τα Ιαπωνικά αυτοκίνητα, αλλά η κλίση δεν είναι και τόσο μεγάλη, ενώ το βάρος των Ευρωπαϊκών φαίνεται να ψιλοαυξάνει. Μπορούμε να υπολογίσουμε το συντελεστή συσχέτισης του Pearson μεταξύ του βάρους και της

επιτάχυνσης για κάθε χώρα ξεχωριστά για να δούμε κατά πόσο ισχύει αυτό που είπαμε ότι βλέπουμε γραφικά; Η απάντηση είναι ναι, μπορούμε.

Θα πάμε ως εξής: **Data→Split File** και θα εμφανιστεί το παράθυρο της εικόνας 62. Εκεί θα επιλέξουμε το **Compare groups** και θα «ανοίξει» το λευκό κουτάκι από κάτω. Εκεί θα βάλουμε την κατηγορική μεταβλητή με βάση τις τιμές της οποία θέλουμε να χωρίσουμε τα δεδομένα μας. Πατάμε **OK** και θα δούμε ότι η σειρά των δεδομένων έχει αλλάξει. Έχουν ταξινομηθεί ανάλογα με την τιμή της χώρας προέλευσης (της κατηγορικής μεταβλητής γενικά).



Εικόνα 62

Τώρα μπορούμε να υπολογίσουμε το συντελεστή συσχέτισης του Pearson (παράθυρο εικόνας 36) και το αποτέλεσμα φαίνεται στον πίνακα του σχήματος 49. Παρατηρούμε ότι γενικά συντελεστής συσχέτισης μεταξύ των δύο μεταβλητών δεν έχει υπολογιστεί (το πάνω μέρος του πίνακα είναι κενό). Ο συντελεστής συσχέτισης μεταξύ των δύο μεταβλητών για τα Αμερικάνικα αυτοκίνητα είναι αρνητικός και στατιστικά σημαντικός (p-value<0.001). Το ίδιο ισχύει και για τα Ιαπωνικά αυτοκίνητα. Ο συντελεστή συσχέτισης όμως μεταξύ των δύο αυτών μεταβλητών για τα Ευρωπαϊκά αυτοκίνητα είναι θετικός (όπως είδαμε και στο διάγραμμα) πλην όμως μη στατιστικά σημαντικός (p-value=0.227).

Άρα βλέπουμε ότι η επίδραση της επιτάχυνσης στο βάρος των αυτοκινήτων ίσως να μην είναι ίδια τελικά για όλες τις χώρες προέλευσης. Αυτό μάλλον πρέπει να το εξετάσουμε πιο πολύ βάζοντας αυτήν την πληροφορία στο μοντέλο μας. Πρώτα όμως πρέπει να επανενώσουμε τα τρία σετ που φτιάξαμε, να πούμε δηλαδή στο SPSS ότι όλες οι χώρες είναι πάλι μαζί και άρα ένα δείγμα. Θα πάμε στο παράθυρο της εικόνας 62 και θα επιλέξουμε το **Analyze all cases, do not create groups** και μετά **OK**.

Relations

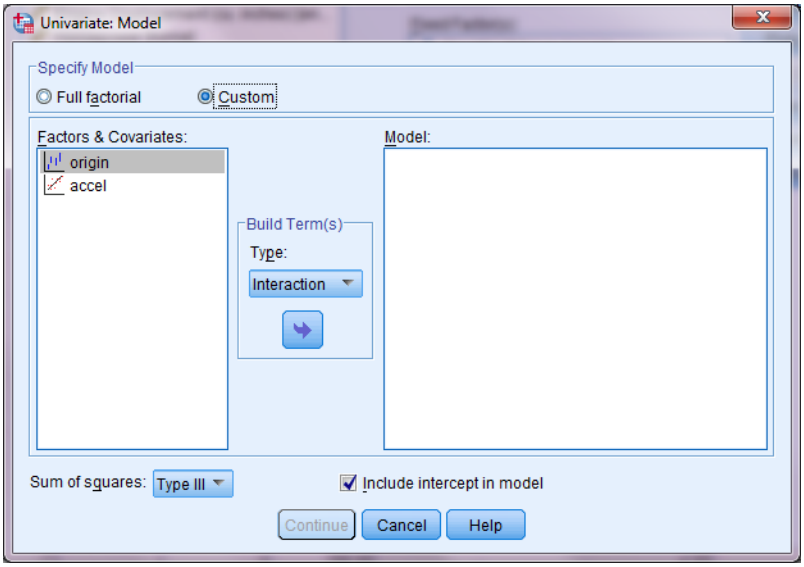
Country of Origin			Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)
.	Vehicle Weight (lbs.)	Pearson Correlation	.[a]	.[a]
		Sig. (2-tailed)		.
		N	1	1
	Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	.[a]	.[a]
		Sig. (2-tailed)	.	
		N	1	1
American	Vehicle Weight (lbs.)	Pearson Correlation	1	**-.462**[**]
		Sig. (2-tailed)		**.000**
		N	253	253
	Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	-.462[**]	1
		Sig. (2-tailed)	.000	
		N	253	253
European	Vehicle Weight (lbs.)	Pearson Correlation	1	**.143**
		Sig. (2-tailed)		**.227**
		N	73	73
	Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	.143	1
		Sig. (2-tailed)	.227	
		N	73	73
Japanese	Vehicle Weight (lbs.)	Pearson Correlation	1	**-.568**[**]
		Sig. (2-tailed)		**.000**
		N	79	79
	Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	-.568[**]	1
		Sig. (2-tailed)	.000	
		N	79	79

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

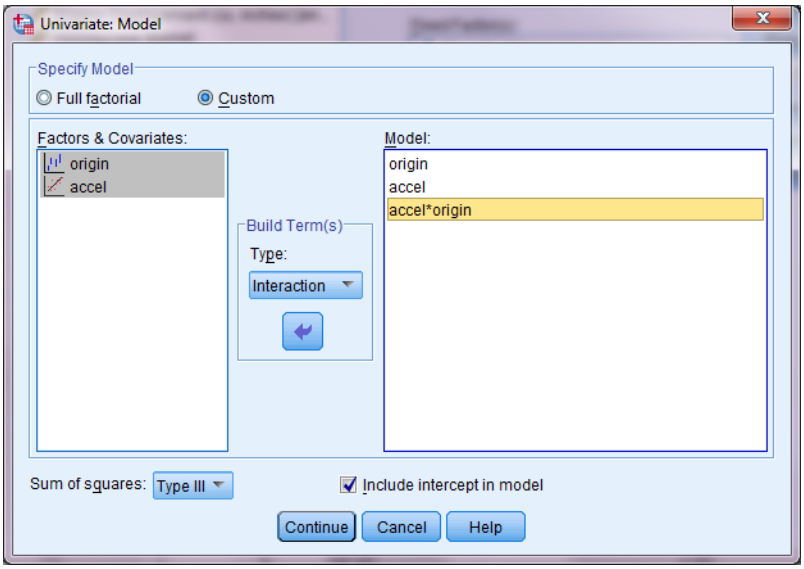
a. Cannot be computed because at least one of the variables is constant.

Σχήμα 49: Συντελεστής συσχέτισης του Pearson ανάλογα με τη χώρα προέλευσης.

Τώρα στο παράθυρο της εικόνας 59 θα επιλέξουμε το **Model** και θα εμφανιστεί το παράθυρο της εικόνας 63. Εκεί θα επιλέξουμε το **Custom**. Τώρα θα πρέπει να περάσουμε τις μεταβλητές δεξιά. Επιλέγουμε μία και τις περνάμε δεξιά (με το βελάκι), μετά τις επιλέγουμε και τις δύο μαζί και τις περνάμε δεξιά (με το βελάκι πάλι). Το αποτέλεσμα που θα πρέπει να δείτε φαίνεται στο παράθυρο της εικόνας 64.



Εικόνα 63



Εικόνα 64

Τώρα πατάμε **OK**, γυρνάμε στο παράθυρο της εικόνας 59 και μετά **OK**. Οι πίνακες των σχημάτων 46 και 47 έχουν αλλάξει τώρα και οι νέοι πίνακες παρουσιάζονται στα σχήματα 50 και 51. Στον πίνακα του σχήματος 48, βλέπουμε μία επιπλέον γραμμή, **origin*accel**. Αυτή είναι η αλληλεπίδραση μεταξύ των δύο μεταβλητών. Είναι στατιστικά σημαντική, άρα η επίδραση της επιτάχυνσης στο βάρος των αυτοκινήτων διαφέρει στατιστικά σημαντικά από χώρα σε χώρα.

Tests of Between-Subjects Effects

Dependent Variable: Vehicle Weight (lbs.)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	141762714.490[a]	5	28352542.898	77.637	.000
Intercept	84345510.350	1	84345510.350	230.962	.000
origin	19792402.611	**2**	9896201.306	27.099	**.000**
accel	7400787.631	**1**	7400787.631	20.265	**.000**
origin * accel	11542961.874	**2**	5771480.937	15.804	**.000**
Error	145711757.485	399	365192.375		
Total	3872185850.000	405			
Corrected Total	287474471.975	404			

a. R Squared = .493 (Adjusted R Squared = .487)

Σχήμα 50: Πίνακας ανάλυσης διακύμανσης με αλληλεπίδραση μέσα.

Παρατηρήστε επίσης ότι οι συντελεστές (στήλη με τα **B**) στον πίνακα του σχήματος 51 έχουν αλλάξει σε σχέση με αυτούς του πίνακα του σχήματος 47 και λογικό είναι, αφού αλλάξαμε λίγο το μοντέλο προσθέτοντας ένα επιπλέον όρο, αυτόν της αλληλεπίδρασης. Επίσης, η τιμή του διορθωμένου συντελεστή συσχέτισης αυξήθηκε από 0.449 σε 0.487.

Parameter Estimates

Dependent Variable: Vehicle Weight (lbs.)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	3726.324	570.111	6.536	.000	2605.528	4847.120
[origin=1]	1582.223	606.326	2.610	**.009**	390.231	2774.216
[origin=2]	-1687.326	698.821	-2.415	**.016**	-3061.158	-313.494
[origin=3]	0[a]	.	.	.	.	.
accel	-93.067	35.001	-2.659	**.008**	-161.877	-24.258
[origin=1] * accel	-36.968	37.547	-.985	**.325**	-110.782	36.847
[origin=2] * accel	116.400	42.244	2.755	**.006**	33.351	199.448
[origin=3] * accel	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Σχήμα 51: Πίνακας με τις εκτιμήσεις των παραμέτρων της παλινδρόμησης με αλληλεπίδραση μέσα.

Ας δούμε το μοντέλο πως γράφεται πια με την αλληλεπίδραση μέσα:

Weight=3726+1582*D_1-1687*D_2 -93*accel-36*D_1*accel+116*D_2*accel,

όπου D_1 και D_2 είναι οι δύο ψευδομεταβλητές που δηλώνουν τη χώρα προέλευσης. Ο πίνακας 4 γράφεται τώρα ως εξής.

Χώρα προέλευσης	Μοντέλο	Μοντέλο
Αμερικάνικα	**3726-93*accel+1582-36*accel**	**5308-129*accel**
Ευρωπαϊκά	**3726-93*accel-1687+116*accel**	**2039+23*accel**
Ιαπωνικά	**3726-93*accel**	**3726-93*accel**

Πίνακας 5: Μοντέλο παλινδρόμησης ανάλογα με τη χώρα προέλευσης και με την αλληλεπίδραση μέσα.

Από τον πίνακα 5 βλέπουμε ότι ανάλογα με τη χώρα προέλευσης έχουμε και διαφορετική ευθεία παλινδρόμησης. Τα πιο ελαφριά αυτοκίνητα φαίνεται να είναι τα Ευρωπαϊκά (2039 λίβρες), μετά έρχονται τα Ιαπωνικά (3726 λίβρες) και τα πιο βαριά είναι τα Αμερικάνικα (5308). Η επιτάχυνση επιδρά αρνητικά στα Ιαπωνικά και στα Αμερικάνικά αυτοκίνητα και θετικά στα Ευρωπαϊκά. Επίσης, αν παρατηρήσουμε τον πίνακα του σχήματος 51, θα δούμε ότι η αρνητική επίδραση της επιτάχυνσης για τα Αμερικάνικά αυτοκίνητα δε διαφέρει στατιστικά σημαντικά από την επίδραση για τα Ιαπωνικά αυτοκίνητα (p-value=0.325). Η θετική όμως επίδραση της επιτάχυνσης για τα Ευρωπαϊκά αυτοκίνητα διαφέρει στατιστικά σημαντικά από την αρνητική επίδραση για τα Ιαπωνικά αυτοκίνητα (p-value=0.006).

7.1 Ανάλυση διακύμανσης κατά ένα παράγοντα (One-way ANOVA)

Στο τέταρτο κεφάλαιο είδαμε τους ελέγχους υποθέσεων για την περίπτωση ενός και δύο ανεξάρτητων και μη δειγμάτων. Στην περίπτωση που έχουμε δείγματα που προέρχονται από τρεις πληθυσμούς τους μέσους των οποίων θέλουμε να συγκρίνουμε χρησιμοποιούμε την τεχνική της ανάλυσης διακύμανσης κατά ένα παράγοντα. Μία εναλλακτική μέθοδος είναι χρησιμοποιήσουμε τον έλεγχο t που είδαμε σε όλα τα πιθανά ζεύγη δειγμάτων. Με αυτόν τον τρόπο όμως μειώνουμε αισθητά την πιθανότητα να ισχύουν και οι δύο έλεγχοι ταυτόχρονα. Η ανάλυση διακύμανσης διατηρεί την πιθανότητα αυτή σταθερή με ένα μόνο έλεγχο. Άλλη έκφραση του προβλήματος είναι ο έλεγχος ισότητας των μέσων για μία ποσοτική μεταβλητή με παράγοντα διαφοροποίησης μία κατηγορική μεταβλητή με τρία επίπεδα. Άρα ελέγχουμε αν η κατηγορική μεταβλητή ή παράγοντας επηρεάζει την ποσοτική μεταβλητή.

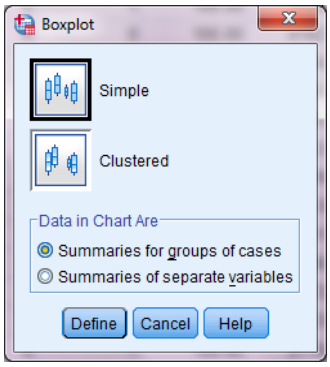
Οι υποθέσεις όμως εφαρμογής της μεθόδου είναι πιο αυστηρές σε σχέση με την περίπτωση των δύο δειγμάτων. Είναι ίδιες με την περίπτωση της απλής γραμμικής παλινδρόμησης, δηλαδή κανονικότητα, ανεξαρτησία και ομοσκεδαστικότητα των καταλοίπων. Όταν λέμε ομοσκεδαστικότητα των καταλοίπων εννοούμε ότι τα κατάλοιπα που δημιουργούνται να έχουν ίσες διασπορές για κάθε επίπεδο του παράγοντα. Η τυχαιότητα εννοείται στη στατιστική αλλιώς δεν υπάρχει νόημα διεξαγωγής των στατιστικών ελέγχων που αναφέραμε. Στην περίπτωση που δεν ισχύουν οι υποθέσεις για τα κατάλοιπα υπάρχουν στατιστικές τεχνικές (οι οποίες προσφέρονται από το SPSS) και μας βοηθάνε να εξαγάγουμε συμπεράσματα. Μπορούμε όμως και να χρησιμοποιήσουμε κάποιο είδος μετασχηματισμού πάνω στην ποσοτική ή εξαρτημένη μεταβλητή. Τα αποτελέσματα της ανάλυσης διακύμανσης κατά ένα παράγοντα δε χάνουν την ισχύ τους όταν έχουμε μικρές αποκλίσεις από την κανονικότητα.

Αν όμως δεν μπορούμε να υποθέσουμε κανονικότητα των καταλοίπων καλό θα ήταν είτε να μετασχηματίσουμε τις τιμές της εξαρτημένης μεταβλητής είτε να χρησιμοποιήσουμε τον έλεγχο των **Kruskal-Wallis**, τον αντίστοιχο μη παραμετρικό έλεγχο της ανάλυσης διακύμανσης κατά ένα παράγοντα. Στην ουσία πρόκειται για την ανάλυση διακύμανσης κατά ένα παράγοντα βασισμένο στις τάξεις μεγέθους των τιμών της εξαρτημένης μεταβλητής. Η μηδενική υπόθεση όμως που ελέγχεται με αυτόν τον έλεγχο αφορά στην ισότητα των διαμέσων και η υπόθεση που κάνουμε για τη χρήση του ελέγχου είναι ότι οι κατανομές των τιμών της εξαρτημένης μεταβλητής, που δημιουργούνται για κάθε επίπεδο του παράγοντα έχουν το ίδιο σχήμα. Στην περίπτωση που δεν ισχύει η περίπτωση της ομοσκεδαστικότητας αλλά η υπόθεση της κανονικότητας ικανοποιείται, μπορούμε να χρησιμοποιήσουμε τον έλεγχο του **Welch** ή τον έλεγχο των **Brown-Forsythe**, τα οποία είναι ανθεκτικά σε περιπτώσεις ετεροσκεδαστικότητας. Θα χρησιμοποιήσουμε τα δεδομένα των αυτοκινήτων για να ελέγξουμε κατά πόσο τα βάρη διαφέρουν από χώρα σε χώρα. Οι μηδενικές υποθέσεις ορίζονται ως εξής:

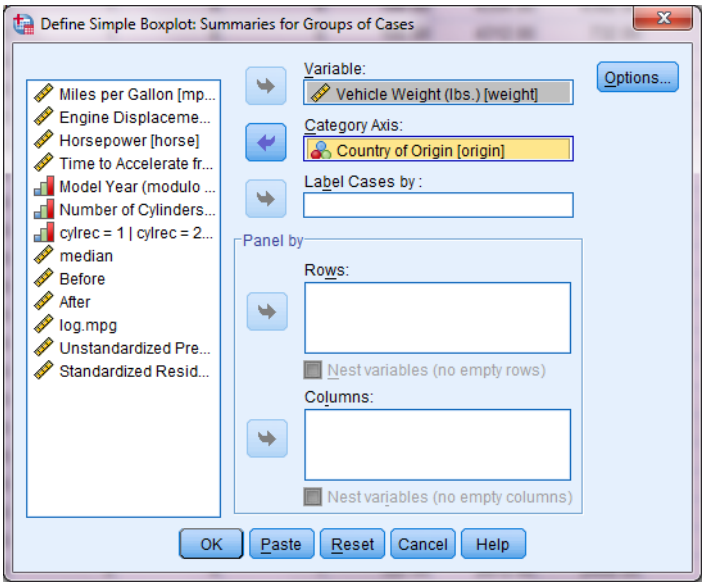
Ho: $\mu_1=\mu_2=\mu_3$
H_1: ένα τουλάχιστον ζεύγος μέσων διαφέρει

Πριν όμως διεξάγουμε τον έλεγχο της ανάλυσης διακύμανσης καλό θα ήταν να δούμε γραφικά πως κατανέμονται τα βάρη των αυτοκινήτων στις τρεις χώρες προέλευσης. Αυτό θα γίνει με την κατασκευή ενός γραφήματος, του λεγόμενου **Box Plot**. Για την κατασκευή του διαγράμματος επιλέγουμε **Graphs→Legacy**

Dialogs→Boxplot και στο παράθυρο της εικόνας 65 που θα εμφανιστεί επιλέγουμε το πρώτο εικονίδιο (**Simple**) και μετά **Define** για να οδηγηθούμε στο παράθυρο της εικόνας 66.

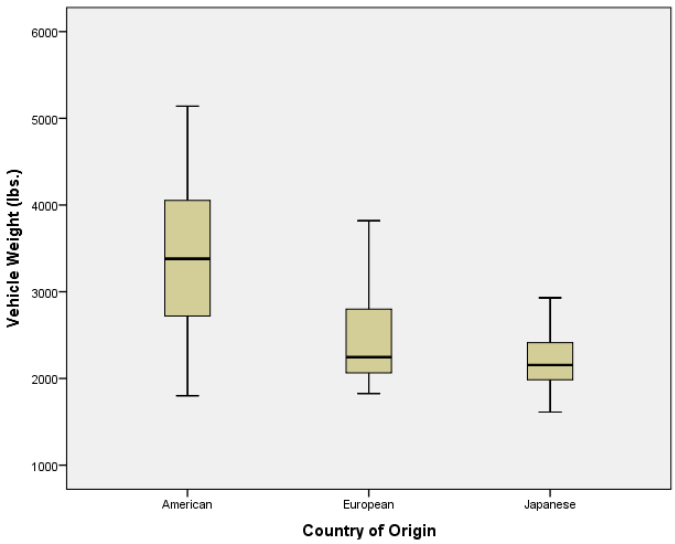


Εικόνα 65



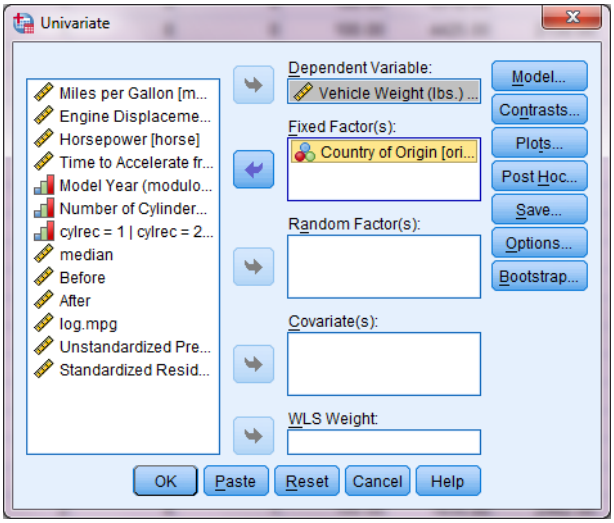
Εικόνα 66

Περνάμε την εξαρτημένη μεταβλητή στο πρώτο λευκό κουτάκι (**Variable:**) και τον παράγοντα στο δεύτερο λευκό κουτάκι (**Category Axis:**) και μετά πατάμε **OK** για να εμφανιστεί το διάγραμμα του σχήματος 52. Η οριζόντια γραμμή που φαίνεται μέσα σε κάθε ορθογώνιο είναι η διάμεσος και όχι ο μέσος. Τα ορθογώνια που κατασκευάστηκαν για κάθε χώρα ξεχωριστά έχουν μήκος το οποίο υπολογίζεται με βάση τους λεγόμενους "φράχτες". Τα άκρα του ορθογωνίου ονομάζονται "εσωτερικοί φράχτες". Πάνω και κάτω από κάθε ορθογώνιο υπάρχουν κάθετες γραμμές. Τα άκρα των γραμμών ονομάζονται "εξωτερικοί φράχτες". Σημεία που βρίσκονται έξω από το ορθογώνιο αλλά εντός των εσωτερικών φραχτών ονομάζονται ήπια ακραία σημεία, ενώ σημεία που βρίσκονται έξω από τους εξωτερικούς φράχτες ονομάζονται εξαιρετικά ακραία σημεία. Με το όρο ακραίο σημείο στη στατιστική εννοούμε μία παρατήρηση (ή τιμή) η οποία απέχει περισσότερο από δύο ή τρεις τυπικές αποκλίσεις από τη μέση τιμή.



Σχήμα 52: Box plot.

Για να διεξάγουμε τον έλεγχο της ανάλυσης διακύμανσης στο SPSS εργαζόμαστε ως εξής: **Analyze→General Linear Model→Univariate** και θα εμφανιστεί το παράθυρο της εικόνας 67. Είναι το ίδιο παράθυρο με της εικόνας 59, η ίδια σχεδόν διαδικασία θα ακολουθηθεί και εδώ, αλλά θα την περιγράψουμε σαν να ήταν η πρώτη φορά.

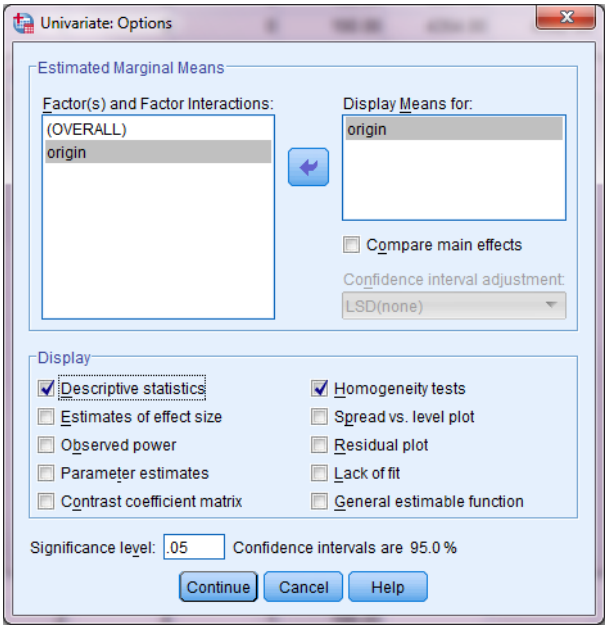


Εικόνα 67

Στο δεξιά πάνω λευκό κουτάκι (**Dependent Variable:**) θα περάσουμε την εξαρτημένη μεταβλητή, το βάρος των αυτοκινήτων δηλαδή. Στο λευκό κουτάκι που βρίσκεται ακριβώς από κάτω (**Fixed Factor(s):**) θα περάσουμε την κατηγορική μεταβλητή της οποία τα επίπεδα αντιστοιχούν στις τρεις χώρες προέλευσης των αυτοκινήτων.

Πατώντας την επιλογή **Save** θα εμφανιστεί ένα παράθυρο παρόμοιο με αυτό της εικόνας 57 στο οποίο θα επιλέξουμε να σώσουμε τα κατάλοιπα και τις εκτιμηθείσες τιμές, όπως και κάναμε και στην περίπτωση της γραμμικής παλινδρόμησης, για να ελέγξουμε την υπόθεση της κανονικότητας (και της

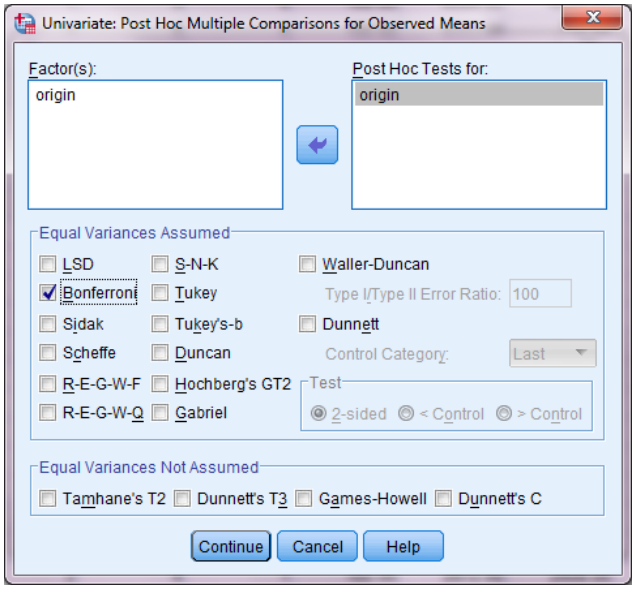
ανεξαρτησίας). Έπειτα θα πατήσουμε την επιλογή **Options** και θα εμφανιστεί το παράθυρο της εικόνας 68.



Εικόνα 68

Στο παράθυρο της εικόνας 68 εμφανίζονται κάποιες επιλογές από τις οποίες θα επιλέξουμε μόνο τα περιγραφικά μέτρα (**Descriptive statistics**) και τον έλεγχο ισότητας διασπορών του Levene (**Homogeneity tests**). Πατάμε **Continue** και γυρίζουμε στο παράθυρο της εικόνας 61.

Πατώντας **Post Hoc** από το παράθυρο της εικόνας 61 θα οδηγηθούμε στο παράθυρο της εικόνας 69. Περνάμε τον παράγοντα στο δεξιό κουτάκι και επιλέγουμε τον έλεγχο του **Bonferroni** για την περίπτωση που η υπόθεση της ισότητας των διασπορών ικανοποιείται (αν επιλέξουμε το **Compare main effects** στο παράθυρο της εικόνας 68 μπορούμε να επιλέξουμε αυτούς και άλλους δύο ελέγχους) και τον έλεγχο **Tamhane's T2** για την περίπτωση που η υπόθεση της ομοσκεδαστικότητας δεν είναι εύλογη. Πατώντας **Continue** και μετά **OK** κάποιοι από τους πίνακες που θα εμφανιστούν στο Output φαίνονται στα επόμενα σχήματα.



Εικόνα 69

Descriptive Statistics

Dependent Variable: Vehicle Weight (lbs.)

Country of Origin	Mean	Std. Deviation	N
American	3367.33	788.612	253
European	2431.49	490.884	73
Japanese	2221.23	320.497	79
Total	2975.09	843.546	405

Σχήμα 53: Περιγραφικά μέτρα για τα επίπεδα του παράγοντα.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Vehicle Weight (lbs.)

F	df1	df2	Sig.
44.145	2	402	**.000**

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + origin

Σχήμα 54: Έλεγχος του Levene για την ισότητα των διακυμάνσεων.

Για τον πίνακα του σχήματος 53 δεν έχουμε να πούμε και πολλά, είναι κάποια περιγραφικά μέτρα. Ο πίνακας του σχήματος 54 όμως είναι πολύ σημαντικός. Ο έλεγχος του Levene χρησιμοποιείται για τον έλεγχο της υπόθεσης της ισότητας των διασπορών για όλα τα επίπεδα του παράγοντα. Η μηδενική και η εναλλακτική υπόθεση για αυτόν τον έλεγχο παρουσιάζονται παρακάτω.

Ho: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

H_1: τουλάχιστον μία διακύμανση διαφέρει από τις υπόλοιπες

Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας για τον έλεγχο του Levene είναι ίσο με μηδέν. Άρα συμπεραίνουμε ότι η παραπάνω μηδενική υπόθεση απορρίπτεται, δηλαδή η υπόθεση της ισότητας των διακυμάνσεων δεν ικανοποιείται.

Ο πίνακας στο σχήμα 55 περιέχει το αποτέλεσμα του ελέγχου F. Το σκιασμένο παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι ίσο με μηδέν. Επομένως συμπεραίνουμε ότι η μηδενική υπόθεση ότι τα βάρη των αυτοκινήτων δε διαφέρουν από χώρα σε χώρα απορρίπτεται. Δεν πρέπει όμως να ξεχνάμε ότι η υπόθεση της ισότητας των διασπορών δεν ικανοποιείται, οπότε καλό θα ήταν να διεξάγουμε και το έλεγχο του Welch ή και των Brown-Forsythe οι οποίοι είναι ανθεκτικοί σε τέτοιες περιπτώσεις, αλλιώς μπορούμε με ένα μετασχηματισμό στις τιμές της εξαρτημένης μεταβλητής (βάρη αυτοκινήτων) να προσπαθήσουμε να σταθεροποιήσουμε τις διακυμάνσεις. Θα δούμε παρακάτω πως γίνονται οι ανθεκτικοί έλεγχοι στο SPSS. Κατασκευάζοντας ένα διάγραμμα διασποράς των καταλοίπων σε σχέση με τις εκτιμηθείσες τιμές θα δείτε ότι οι διακυμάνσεις αυξάνονται, το διάγραμμα έχει το σχήμα ενός "χωνιού".

Tests of Between-Subjects Effects

Dependent Variable: Vehicle Weight (lbs.)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	105391890.059[a]	2	52695945.030	116.342	.000
Intercept	2122151191.716	1	2122151191.716	4685.263	.000
origin	105391890.059	2	52695945.030	116.342	**.000**
Error	182082581.916	402	452941.746		
Total	3872185850.000	405			
Corrected Total	287474471.975	404			

a. R Squared = .367 (Adjusted R Squared = .363)

Σχήμα 55: Αποτέλεσμα ελέγχου F.

Δείτε τον παρακάτω πίνακα πολλαπλών ελέγχων (σχήμα 56). Δείχνει ποια ζεύγη μέσων διαφέρουν. Το πρώτο μισό του πίνακα περιέχει τους ελέγχους που έγιναν με την τεχνική του Bonferroni, η οποία βασίζεται στην κλασική ανάλυση διακύμανσης. Το δεύτερο μισό του πίνακα περιέχει τους ελέγχους που έγιναν με την τεχνική του Tamhane, η οποία εφαρμόζεται στις περιπτώσεις που δεν ισχύει η υπόθεση της ομοσκεδαστικότητας. Παρατηρήστε ότι η τεχνική του Bonferroni δεν ανιχνεύει τη διαφορά βάρους ανάμεσα στα ευρωπαϊκά και στα Ιαπωνικά αυτοκίνητα ενώ η τεχνική του Tamhane τη βρίσκει στατιστικά σημαντική.

Multiple Comparisons

Dependent Variable: Vehicle Weight (lbs.)

	(I) Country of Origin	(J) Country of Origin	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	American	European	935.83*	89.415	**.000**	720.88	1150.79
		Japanese	1146.10*	86.739	**.000**	937.58	1354.63
	European	American	-935.83*	89.415	.000	-1150.79	-720.88
		Japanese	210.27	109.262	.165	-52.40	472.93
	Japanese	American	-1146.10*	86.739	.000	-1354.63	-937.58
		European	-210.27	109.262	.165	-472.93	52.40
Tamhane	American	European	935.83*	75.888	.000	753.02	1118.65
		Japanese	1146.10*	61.306	**.000**	998.92	1293.28
	European	American	-935.83*	75.888	**.000**	-1118.65	-753.02
		Japanese	210.27*	67.832	**.007**	46.05	374.48
	Japanese	American	-1146.10*	61.306	.000	-1293.28	-998.92
		European	-210.27*	67.832	.007	-374.48	-46.05

Based on observed means.

The error term is Mean Square(Error) = 452941.746.

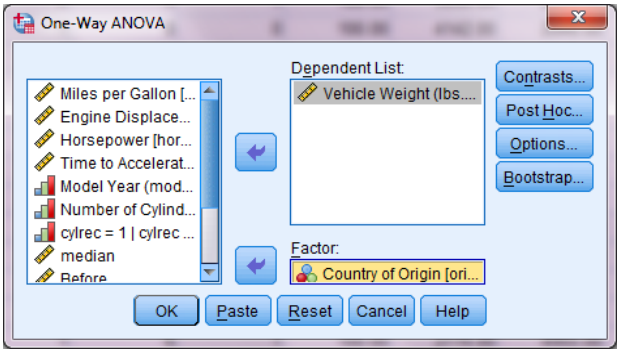
*. The mean difference is significant at the .05 level.

Σχήμα 56: Αποτελέσματα πολλαπλών ελέγχων.

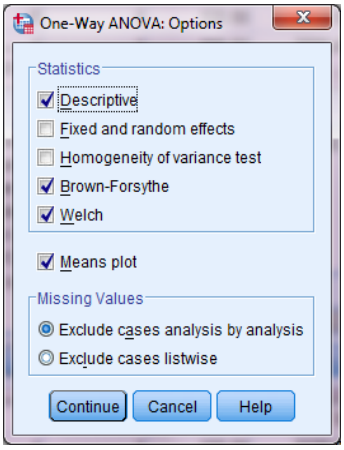
Υπάρχουν δύο τρόποι για να αντιμετωπίσουμε το πρόβλημα της μη ισότητας των διακυμάνσεων των καταλοίπων κατά μήκος των τριών δειγμάτων. Ο ένας είναι να μετασχηματίσουμε τις τιμές της εξαρτημένης μεταβλητής, π.χ. λογαριθμικά. Αυτόν τον τρόπο προσέγγισης τον είδαμε στην πολλαπλή γραμμική παλινδρόμηση. Ο δεύτερος είναι να περάσουμε σε πιο ανθεκτικές, στις υποθέσεις του μοντέλου, μεθόδους όπως είναι η ανάλυση διακύμανσης με τη μέθοδο του Welch και των Brown-Forsythe.

7.2 Ανάλυση διακύμανσης με τη μέθοδο του Welch και των Brown-Forsythe

Η ανάλυση διακύμανσης κατά ένα παράγοντα μπορεί να διεξαχθεί και με άλλο τρόπο από την επιλογή Analyze. Επιλέγοντας **Analyze→Compare Means→One-way ANOVA** και θα εμφανιστεί το παράθυρο της εικόνας 70. Στο οποίο περνάμε την εξαρτημένη μεταβλητή και τον παράγοντα στα δεξιά λευκά κουτάκια κατά τα γνωστά. Επιλέγοντας **Post Hoc** θα εμφανιστεί το παράθυρο της εικόνας 63 στο οποίο θα επιλέξουμε τη διεξαγωγή των πολλαπλών ελέγχων μόνο μέσω του Tamhane. Πατώντας **Options** θα εμφανιστεί το παράθυρο της εικόνας 71.



Εικόνα 70



Εικόνα 71

Στο παράθυρο της εικόνας 71 θα επιλέξουμε όλες τις επιλογές πλην των **Fixed and random effects** και **Homogeneity of variance test** (γνωρίζουμε ότι οι διακυμάνσεις διαφέρουν, οπότε δε θέλουμε να εμφανίσει πάλι τον έλεγχο του Levene). Πατώντας Continue επιστρέφουμε στο παράθυρο της εικόνας 70 και πατώντας OK τα αποτελέσματα φαίνονται παρακάτω. Εμφανίζεται και ο πίνακας της ανάλυσης διακύμανσης που είδαμε προηγουμένως (δεν τον παρουσιάζουμε πάλι εδώ).

Descriptives

Vehicle Weight (lbs.)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
American	253	3367.33	788.612	49.580	3269.68	3464.97	1800	5140
European	73	2431.49	490.884	57.454	2316.96	2546.02	1825	3820
Japanese	79	2221.23	320.497	36.059	2149.44	2293.02	1613	2930
Total	405	2975.09	843.546	41.916	2892.69	3057.49	1613	5140

Σχήμα 57: Περιγραφικά μέτρα.

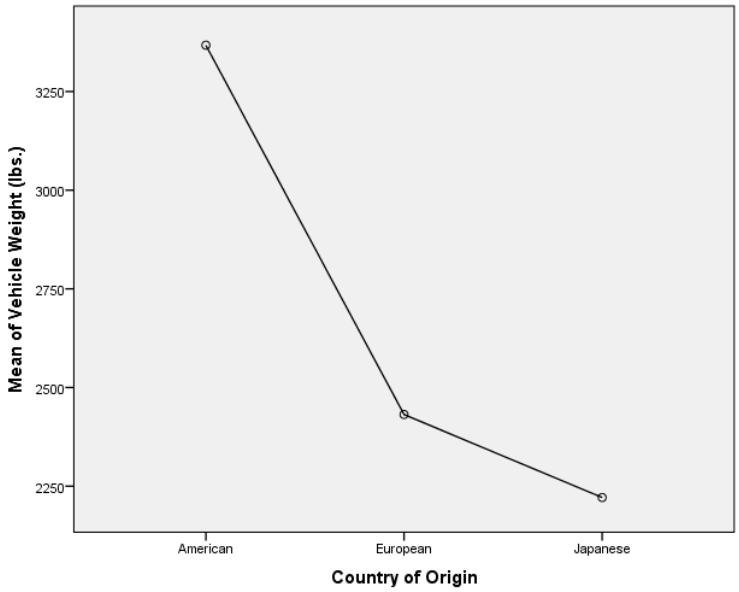
Robust Tests of Equality of Means

Vehicle Weight (lbs.)

	Statistic[a]	df1	df2	Sig.
Welch	178.311	2	192.960	**.000**
Brown-Forsythe	205.193	2	311.916	**.000**

a. Asymptotically F distributed.

Σχήμα 58: Αποτελέσματα των ανθεκτικών ελέγχων.



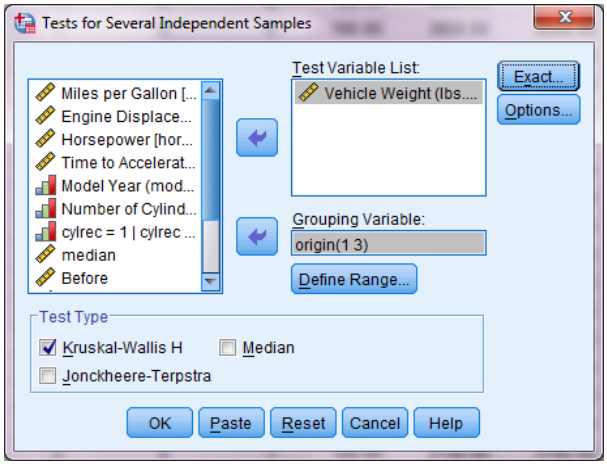
Σχήμα 59: Διάγραμμα των μέσων.

Στον πίνακα του σχήματος 57 βλέπουμε κάποια περιγραφικά μέτρα για τα βάρη των αυτοκινήτων κάθε χώρας ξεχωριστά μαζί με τα 95% διαστήματα εμπιστοσύνης για τα μέσα βάρη της κάθε χώρας. Εδώ παρεμβάλλεται ο πίνακας της ANOVA τον οποίο έχουμε ήδη δει. Ο πίνακας του σχήματος 58 περιέχει τα αποτελέσματα των ανθεκτικών ελέγχων. Τα παρατηρηθέντα επίπεδα στατιστικής σημαντικότητας και για τους δύο ελέγχους είναι μηδέν. Το συμπέρασμα είναι ίδιο με προηγουμένως, τα βάρη διαφέρουν στατιστικά σημαντικά από χώρα σε χώρα. Το επόμενο σχήμα στο Output είναι ο πίνακας με τους πολλαπλούς ελέγχους του Tamhane τους οποίους είδαμε προηγουμένως. Τελευταίο εμφανίζεται το διάγραμμα των μέσων το οποίο θα μπορούσαμε να είχαμε κατασκευάσει και με τον πρώτο τρόπο. Βλέπουμε ότι τα αμερικάνικα αυτοκίνητα είναι σαφώς βαρύτερα από τα άλλα.

7.3 Μη παραμετρική ανάλυση διακύμανσης (έλεγχος των Kruskal-Wallis)

Είδαμε πως αντιμετωπίζουμε το πρόβλημα της μη ικανοποίησης των υποθέσεων. Τι γίνεται όμως στην περίπτωση που εξακολουθούμε να έχουμε πρόβλημα; Τότε διεξάγουμε ανάλυση διακύμανσης βασισμένη στις τάξεις μεγέθους των τιμών της εξαρτημένης μεταβλητής. Η τεχνική προτάθηκε από τον William Kruskal και τον W. Allen Wallis και στην ουσία αποτελεί τη γενίκευση του ελέγχου των Mann-Whitney-Wilcoxon για τρία ή περισσότερα δείγματα. Οι μηδενική και η εναλλακτική υπόθεση

όμως θα αλλάξουν μορφή και θα αναφέρονται στην ισότητα ή όχι των διαμέσων των δειγμάτων. Η διεξαγωγή του ελέγχου στο SPSS γίνεται ως εξής:
Analyze→Non Parametric Tests→Legacy Dialogs→K Independent Samples και θα εμφανιστεί το παράθυρο της εικόνας 72.



Εικόνα 72

Στο παράθυρο αυτό περνάμε την εξαρτημένη μεταβλητή στο λευκό κουτάκι (**Test Variable List:**) και την κατηγορική μεταβλητή που δηλώνει τον παράγοντα στο κάτω λευκό κουτάκι (**Grouping Variable:**). Μετά πατάμε **Define Range** για να ορίσουμε τα επίπεδα του παράγοντα (τις τιμές της κατηγορικής μεταβλητής). Από την επιλογή **Exact** επιλέγουμε την εμφάνιση των αποτελεσμάτων που βασίζονται στην τεχνική Monte Carlo. Μετά πατάμε **OK** και το αποτέλεσμα φαίνεται στον πίνακα του σχήματος 60.

Test Statistics[a,b]

			Vehicle Weight (lbs.)
Chi-Square			161.198
df			2
Asymp. Sig.			**.000**
Monte Carlo Sig.	**Sig.**		**.000[c]**
	99% Confidence Interval	Lower Bound	.000
		Upper Bound	.000

a. Kruskal Wallis Test

b. Grouping Variable: Country of Origin
c. Based on 10000 sampled tables with starting seed 1993510611.

Σχήμα 60: Αποτέλεσμα του ελέγχου των Kruskal-Wallis.

Τα παρατηρηθέντα επίπεδα στατιστικής σημαντικότητας για τον έλεγχο των Kruskal-Wallis είναι ίσα με μηδέν. Επομένως η μηδενική υπόθεση απορρίπτεται, άρα οι διάμεσοι των βαρών των αυτοκινήτων διαφέρουν στατιστικά σημαντικά μεταξύ τους ως προς τις χώρες προέλευσης σε α=5%.

7.4 Ανάλυση διακύμανσης κατά δύο παράγοντες (Two-way ANOVA)

Τι γίνεται όμως αν έχουμε δύο κατηγορικές μεταβλητές τη στατιστική σημαντικότητα των οποίων θέλουμε να εξετάσουμε; Είδαμε για παράδειγμα πως η χώρα προέλευσης των αυτοκινήτων επηρεάζει το βάρος τους, ή ότι το βάρος των αυτοκινήτων δεν είναι το ίδιο από χώρα σε χώρα, αλλά σε κάποιες διαφέρει στατιστικά σημαντικά. Τι γίνεται όμως αν έχουμε και μία δεύτερη κατηγορική μεταβλητή. Αν θέλουμε να δούμε αν το βάρος π.χ. διαφέρει στατιστικά σημαντικά όσον αφορά στη χρονιά κατασκευής[††]; Άρα θέλουμε να δούμε αν το βάρος των αυτοκινήτων διαφέρει ως προς τη χώρα αλλά και ως προς τη χρονιά κατασκευής. Ή με άλλα λόγια να δούμε αν η χώρα προέλευσης και η χρονιά κατασκευής των αυτοκινήτων επιδρούν στατιστικά σημαντικά στο βάρος τους.

Η απάντηση είναι η ανάλυση διακύμανσης κατά δύο παράγοντες. Πάμε στο παράθυρο της εικόνας 67 και περνάμε και τη δεύτερη κατηγορική μεταβλητή στο ίδιο κουτάκι (**Fixed Factor(s):**) κάτω από την πρώτη μεταβλητή. Μετά κάνουμε τα ίδια με προηγουμένως και βλέπουμε τα αποτελέσματα στο Output του SPSS. Θα εστιάσουμε στον πίνακα της ανάλυσης διακύμανσης που παρουσιάζεται στο σχήμα 61 παρακάτω.

Tests of Between-Subjects Effects

Dependent Variable: Vehicle Weight (lbs.)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	142265071.835[a]	38	3743817.680	9.436	.000
Intercept	1676175041.842	1	1676175041.842	4224.796	.000
origin	79036410.203	2	39518205.101	99.606	**.000**
year	6316334.243	12	526361.187	1.327	**.201**
origin * year	15671418.820	24	652975.784	1.646	**.030**
Error	145209400.140	366	396746.995		
Total	3872185850.000	405			
Corrected Total	287474471.975	404			

a. R Squared = .495 (Adjusted R Squared = .442)

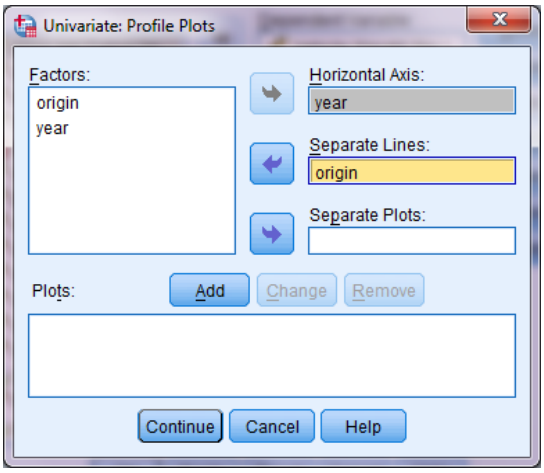
Σχήμα 61: Πίνακας ανάλυσης διακύμανσης για την ανάλυση διακύμανσης κατά δύο παράγοντες.

Παρατηρούμε ότι η p-value για τη χώρα προέλευσης (**origin**) και για τη χρονιά (**year**) είναι μικρότερη του 0.05. Άρα η χώρα προέλευσης επηρεάζει το βάρος των αυτοκινήτων αλλά όχι η χρονιά κατασκευής τους. Ή διαφορετικά το βάρος των αυτοκινήτων δεν είναι το ίδιο ανάλογα με τη χώρα αλλά δε διαφέρει στατιστικά σημαντικά από χρονιά σε χρονιά. Οι πολλαπλοί έλεγχοι είναι όπως και προηγουμένως, οπότε είμαστε σχεδόν στα ίδια.

[††]Η χρονιά κατασκευής είναι προφανώς αριθμητική, αλλά θα τη θεωρήσουμε κατηγορική για το παράδειγμα μας.

Βλέπουμε τώρα στον πίνακα του σχήματος 61 και μία επιπλέον γραμμή. Αυτή η γραμμή είναι η αλληλεπίδραση μεταξύ χώρας προέλευσης και αριθμού κυλίνδρων (**origin*year**), η οποία όμως είναι και στατιστικά σημαντική. Άρα θα πρέπει να την εξετάσουμε, να δούμε τι είναι. Μπορεί σε μία επόμενη ανάλυση να είναι πάλι στατιστικά σημαντική και να πρέπει να την κρατήσουμε μέσα. Θα πρέπει να ξέρουμε τι είναι. Θα δούμε γραφικά τι είναι ώστε να την καταλάβουμε καλύτερα.

Πάμε πρώτα να απαντήσουμε στο δεύτερο ερώτημα. Τι είναι αυτή η αλληλεπίδραση. Στο παράθυρο της εικόνας 61 (αφού έχουμε βάλει και τις δύο μεταβλητές μέσα, στο δεξιά κουτάκι) θα επιλέξουμε **Options** και θα εμφανιστεί το παράθυρο της εικόνας 73.

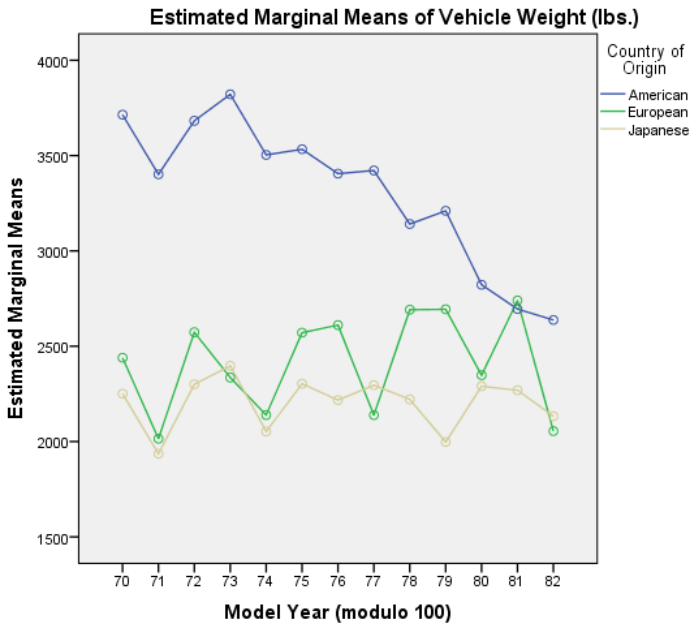


Εικόνα 73

Εκεί όπως βλέπετε στο παράθυρο της εικόνα 73 θα περάσουμε τις δύο μεταβλητές στα δύο πάνω δεξιά κουτάκια. Η σειρά δεν έχει σημασία, απλά θα είναι διαφορετικό το γράφημα, η επεξήγηση όμως ίδια. Θα πατήσουμε **Add** και μετά **Continue**. Αν δεν πατήσουμε **Add** θα εμφανιστεί ένα μηνυματάκι που θα μας προειδοποιεί ότι κάναμε θα χαθεί και επομένως δεν θα κατασκευαστεί το διάγραμμα που θέλουμε να δούμε. Πατώντας **OK** στο παράθυρο της εικόνας 67 θα εμφανιστεί το διάγραμμα του σχήματος 62.

Βλέπουμε ότι υπάρχουν τρεις γραμμές, μία για κάθε χώρα. Κάθε γραμμή έχει τόσα σημεία (κυκλάκια) όσα και τα επίπεδα ή τιμές της κατηγορικής μεταβλητής (χρόνια δηλαδή εδώ). Κάθε κυκλάκι είναι η μέση τιμή του βάρους των αυτοκινήτων που αντιστοιχεί στη χώρα και στη χρονιά κατασκευής.

Τι παρατηρούμε εδώ; Παρατηρούμε ότι τα Αμερικάνικα αυτοκίνητα καθώς περνούν τα χρόνια τείνουν να μειώνουν το βάρος των αυτοκινήτων τους (μπλε γραμμή). Τα Ιαπωνικά αυτοκίνητα από την άλλη φαίνεται να κρατούν το βάρος των αυτοκινήτων τους σταθερό. Το μέσο βάρος των δε Ευρωπαϊκών μία πάει πάνω, μία κάτω. Οι γραμμές δηλαδή δεν ακολουθούν παράλληλες ή περίπου παράλληλες πορείες. Αυτό είναι ένδειξη αλληλεπίδρασης την οποία είδαμε στον πίνακα του σχήματος 61 ότι είναι στατιστικά σημαντική. Δηλαδή, ο τρόπος με τον οποίο αλλάζει η μέση τιμή της εξαρτημένης μεταβλητής για έναν παράγοντα (χρονιά κατασκευής) δεν είναι ο ίδιος για τα επίπεδα όλως των παράγοντα. Αλλιώς μεταβάλλεται το μέσο βάρος για τα Αμερικάνικα, αλλιώς για Ιαπωνικά και αλλιώς για Ευρωπαϊκά. Υπάρχει δηλαδή μία αλληλεπίδραση μεταξύ χώρας προέλευσης και χρονιάς κατασκευής.



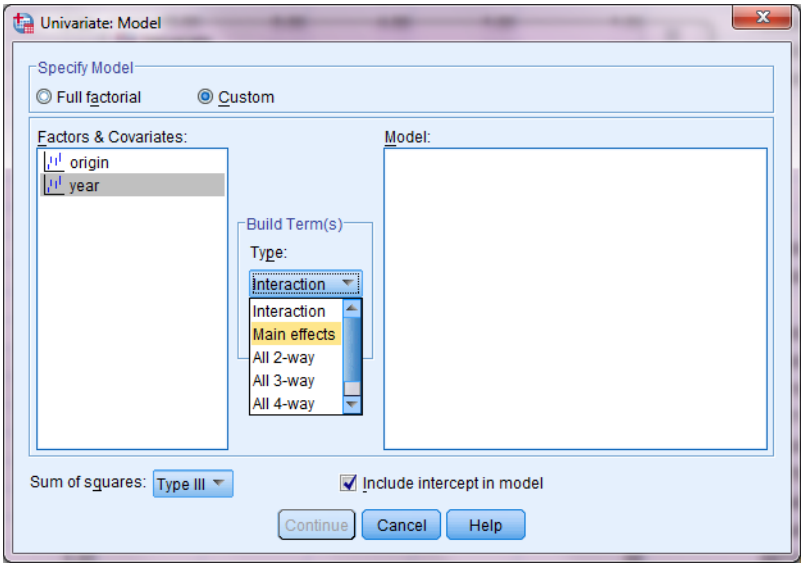
Σχήμα 62: Γραφική αναπαράσταση αλληλεπίδρασης μεταξύ δύο κατηγορικών μεταβλητών.

Προσέξτε επίσης ότι η χρονιά κατασκευής δεν είναι στατιστικά σημαντική μεταβλητή όπως είπαμε και προηγουμένως, αλλά η αλληλεπίδραση της με τη χώρα προέλευσης είναι. Άρα δεν μπορούμε να βγάλουμε τη χρονιά κατασκευής από το μοντέλο και να αφήσουμε μέσα την αλληλεπίδραση της με τη χώρα προέλευσης[‡‡]. Να τονίσουμε επίσης ότι γενικά τα Αμερικάνικα αυτοκίνητα είναι κατά μέσο όρο πιο βαριά και από τα Ιαπωνικά αλλά και από τα Ευρωπαϊκά.

Ας δούμε τώρα πως μπορούμε να βγάλουμε την αλληλεπίδραση από το μοντέλο στην περίπτωση που δεν είναι στατιστικά σημαντική. Αφού έχουμε τρέξει το μοντέλο (οι επιλογές στις αναλύσεις δεν έχουν σβηστεί), θα ξαναπάμε στο παράθυρο της εικόνας 61 και θα επιλέξουμε το **Model** για να εμφανιστεί το παράθυρο της εικόνας 68. Εκεί θα επιλέξουμε το **Custom** για να «φωτιστεί» το παράθυρο. Στη μεσαία επιλογή θα επιλέξουμε το **Main effects** όπως βλέπουμε και την εικόνα 74. Μετά θα περάσουμε τις μεταβλητές δεξιά και θα πατήσουμε **Continue** για να επιστρέψουμε στο παράθυρο της εικόνας 67 όπου θα πατήσουμε **OK** και θα πάρουμε τον πίνακα του σχήματος 63.

Παρατηρήστε τώρα ότι και οι δύο μεταβλητές είναι στατιστικά σημαντικοί. Άρα η χρονιά κατασκευής επιδρά στο μέσο βάρος των αυτοκινήτων στατιστικά σημαντικά. Δηλαδή, το μέσο βάρος των αυτοκινήτων διαφέρει στατιστικά σημαντικά για κάποιες χρονιές. Η τεχνική των πολλαπλών ελέγχων του Bonferroni ή του Tamhane θα φανούν χρήσιμες εδώ και πάλι για να ανιχνεύσουμε τα στατιστικά σημαντικά ζεύγη όσον αφορά στις χρονιές κατασκευής.

[‡‡] Μαθηματικά μπορούμε αν κάνουμε κάποιες αλλαγές, αλλά στατιστικά και επεξηγηματικά δεν έχει και πολύ νόημα.



Εικόνα 74

Tests of Between-Subjects Effects

Dependent Variable: Vehicle Weight (lbs.)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	126593653.014[a]	14	9042403.787	21.920	.000
Intercept	2085380027.486	1	2085380027.486	5055.284	.000
origin	85308626.329	2	42654313.165	103.401	**.000**
year	21201762.955	12	1766813.580	4.283	**.000**
Error	160880818.961	390	412514.920		
Total	3872185850.000	405			
Corrected Total	287474471.975	404			

a. R Squared = .440 (Adjusted R Squared = .420)

Σχήμα 63: Πίνακας ανάλυσης διακύμανσης κατά δύο παράγοντες.

7.6 Ανάλυση διακύμανσης για εξαρτημένα δείγματα

Είδαμε τι γίνεται όταν τα δείγματα είναι ανεξάρτητα μεταξύ τους. Τι γίνεται όμως όταν τα δείγματα δεν είναι ανεξάρτητα? Είδαμε μία περίπτωση με δύο δείγματα στο 5^{ο} κεφάλαιο όπου είχαμε τον έλεγχο t για δύο εξαρτημένα δείγματα. Εδώ θα δούμε τι μπορούμε να κάνουμε όταν έχουμε παραπάνω από δύο εξαρτημένα δείγματα. Η τεχνική στα αγγλικά λέγεται *Repeated Measures ANOVA*, ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις.

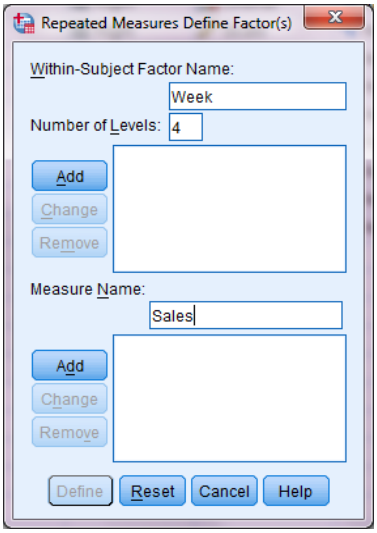
Θα χρησιμοποιήσουμε δεδομένα από το SPSS πάλι, τα **testmarket.sav**. Άλλαξα τη μορφή τους ώστε να είναι όπως στον πίνακα 4. Πρέπει δηλαδή τα δείγματα να είναι το κάθε ένα σε μία στήλη. Στην περίπτωση των ανεξάρτητων δειγμάτων (ανάλυση διακύμανσης και έλεγχος μέσων δύο ανεξάρτητων δειγμάτων) όλες οι μετρήσεις ήταν σε μία στήλη και είχαμε μία άλλη στήλη που έδειχνε το

δείγμα στο οποίο ανήκει η κάθε μέτρηση. Στην περίπτωση όμως δύο εξαρτημένων δειγμάτων χρησιμοποιήσαμε δύο στήλες.

Προϊόν	Week1	Week2	Week3	Week4
1	18	14	12	6
2	19	12	8	4
3	14	10	6	2
4	16	12	10	4
5	12	8	6	2
6	18	10	5	1
7	16	10	8	4
8	18	8	4	1
9	16	12	6	2
10	19	16	10	8
11	16	14	10	9
12	16	12	8	8

Πίνακας 6: Δεδομένα επαναλαμβανόμενων μετρήσεων.

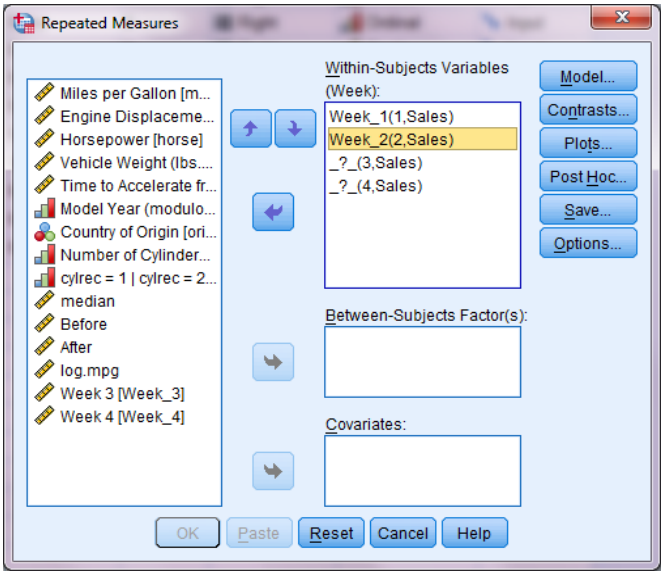
Ο πίνακας 4 περιέχει κάποιες τιμές για μία αγορά κατά μήκος 4 εβδομάδων. Ο σκοπός είναι να δούμε αν υπάρχει διαφορά στις μέσες τιμές ανάμεσα στις 4 εβδομάδες. Για να διεξάγουμε την ανάλυση στο SPSS επιλέγουμε **Analyze→General Linear Model→Repeated Measures** και θα εμφανιστεί το παράθυρο της εικόνας 75. Στο πάνω τετραγωνάκι (**Within-Subject Factor Name:**) δώσαμε ένα όνομα στον παράγοντα για τον οποίο ενδιαφερόμαστε. Στο δεύτερο κουτάκι (**Number of Levels:**) βάζουμε τον αριθμό των επαναλαμβανόμενων μετρήσεων που έχουμε, στην περίπτωση μας έχουμε μετρήσεις για 4 εβδομάδες. Στο κουτάκι **Measure Name:** μπορούμε να δώσουμε ένα όνομα στις μετρήσεις, δηλαδή σε τι αναφέρονται. Εμείς βάλαμε πωλήσεις.



Εικόνα 75

Πατώντας **Add** και μετά **Define** και θα μεταφερθούμε στο παράθυρο της εικόνας 76. Εδώ θα περάσουμε στα δεξιά τα τέσσερα δείγματα (στήλες) που έχουμε.

Αν επιλέξουμε **Plots** μπορούμε να ζητήσουμε ένα διάγραμμα με τους μέσους των δειγμάτων. Η επιλογή **Options** είναι ίδια με αυτή στο παράθυρο της εικόνας 62 που είδαμε στην ανάλυση διακύμανσης. Εκεί θα επιλέξουμε πάλι την εμφάνιση κάποιων περιγραφικών για τα τέσσερα δείγματα και τις εκτιμήσεις των παραμέτρων. Αν θέλουμε μπορούμε να επιλέξουμε να σώσουμε τα τυποποιημένα κατάλοιπα για έλεγχο κανονικότητας. Μετά γυρίζουμε στο παράθυρο της εικόνας 68 και πατάμε **OK**. Θα εμφανιστούν πολλοί πίνακες αλλά εμείς εδώ παρουσιάζουμε μόνο τον έλεγχο σφαιρικότητας του Mauchly (σχήμα 64) και τους ελέγχους για την ισότητα των μέσων (σχήμα 65).



Εικόνα 76

Ο έλεγχος σφαιρικότητας ελέγχει μία διαφορετική υπόθεση σχετικά με τις διακυμάνσεις των καταλοίπων. Η p-value για τον έλεγχο αυτό είναι μεγαλύτερη του 0.05 στην περίπτωση μας (*Sig*=0.112). Καλό είναι να ικανοποιείται αυτή η υπόθεση. Στον πίνακα του σχήματος 65 έχουμε 4 διαφορετικούς ελέγχους. Ο πρώτος υποθέτει οι η υπόθεση της σφαιρικότητας ισχύει. Εμείς είμαστε σε αυτήν την περίπτωση τώρα. Αν παραβιάζεται η υπόθεση αυτή τότε μπορούμε να χρησιμοποιήσουμε τους τρεις επόμενους ελέγχους.

Όσον αφορά στην υπόθεση της κανονικότητας μπορούμε να σώσουμε τα κατάλοιπα και να ελέγξουμε για κάθε στήλη καταλοίπων την κανονικότητα.

Βλέπουμε ότι η υπόθεση της ισότητας των μέσων απορρίπτεται σε α=5%. Άρα κάποιος ή κάποιοι μέσοι διαφέρουν μεταξύ τους. Μπορούμε σε αυτήν την περίπτωση να κάνουμε πολλούς ελέγχους t για ζεύγη παρατηρήσεων. Αυτό όμως δεν είναι σωστό, διότι η διακύμανση δεν είναι ίδια για όλα τα δείγματα όπως έχουμε υποθέσει. Αυτό θέλει ψάξιμο στη βιβλιογραφία να δω τι κάνουμε σε αυτές τις περιπτώσεις και αν το IBM SPSS 22 παρέχει κάτι.

Mauchly's Test of Sphericity[a]

Measure: Sales

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Week	.398	8.957	5	**.112**	.622	.744	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept

Within Subjects Design: Week

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Σχήμα 64: Έλεγχος σφαιρικότητας του Mauchly.

Tests of Within-Subjects Effects

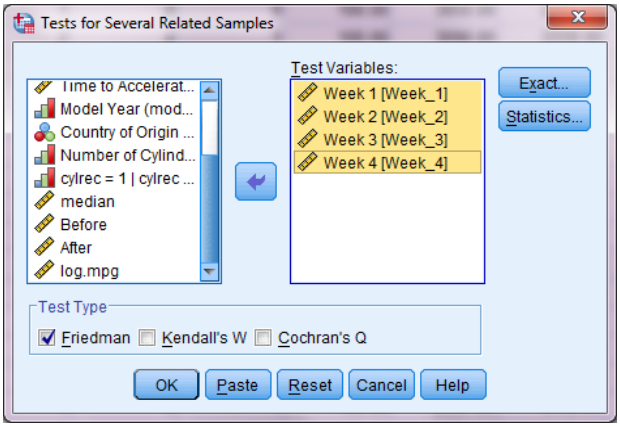
Measure: Sales

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	991.500	3	330.500	127.561	**.000**
	Greenhouse-Geisser	991.500	1.865	531.709	127.561	**.000**
	Huynh-Feldt	991.500	2.231	444.504	127.561	**.000**
	Lower-bound	991.500	1.000	991.500	127.561	**.000**
Error(Week)	Sphericity Assumed	85.500	33	2.591		
	Greenhouse-Geisser	85.500	20.512	4.168		
	Huynh-Feldt	85.500	24.536	3.485		
	Lower-bound	85.500	11.000	7.773		

Σχήμα 65: Έλεγχοι για την ισότητα των μέσων.

7.7 Μη παραμετρική ανάλυση διακύμανσης για εξαρτημένα δείγματα

Ας δούμε τώρα ένα αντίστοιχο μη παραμετρικό έλεγχο για την περίπτωση που έχουμε περισσότερα από 2 εξαρτημένα δείγματα. Ο έλεγχος όμως αναφέρεται στις διαμέσους των δειγμάτων και όχι στις μέσες τιμές τους. Είναι όμως πιο απλός και πιο σύντομος. Για να τον διεξάγουμε επιλέγουμε **Analyze→Non Parametric Tests→Legacy Dialogs→K Related samples** και θα εμφανιστεί το παράθυρο της εικόνας 77 στο οποίο θα περάσουμε δεξιά τις τέσσερις στήλες του SPSS. Πατώντας **Exact** μπορούμε να ζητήσουμε υπολογισμό της p-value μέσω Monte-Carlo. Τα αποτελέσματα του ελέγχου φαίνονται στον πίνακα του σχήματος 66.



Εικόνα 77

Test Statistics[a]

N			12
Chi-Square			35.723
df			3
Asymp. Sig.			**.000**
Monte Carlo Sig.	Sig.		**.000**
	99% Confidence Interval	Lower Bound	.000
		Upper Bound	.000

a. Friedman Test

Σχήμα 66: Αποτέλεσμα του ελέγχου του Friedman.

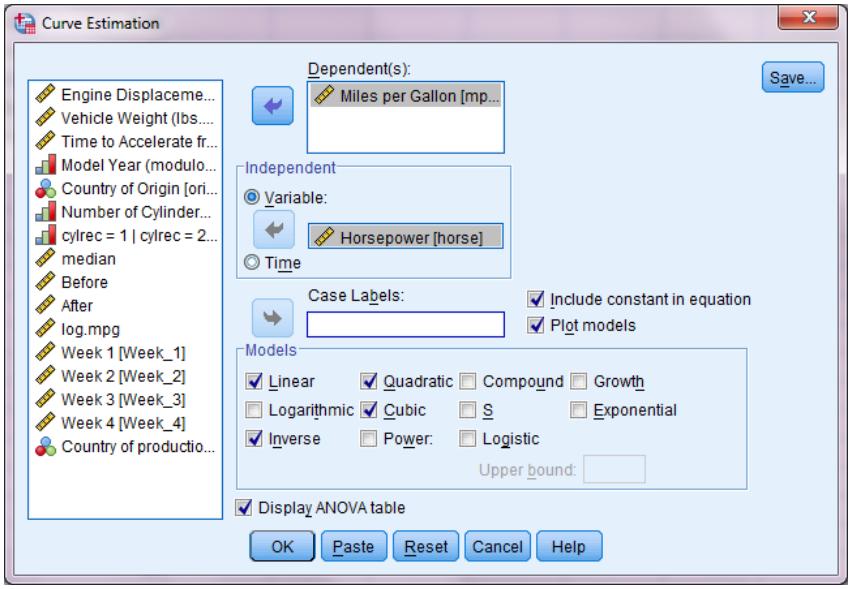
Το συμπέρασμα είναι ίδιο με το προηγούμενο, η υπόθεση της ισότητας των μέσων απορρίπτεται σε α=5%.

8.1 Προχωρημένη απλή παλινδρόμηση (μία ανεξάρτητη μεταβλητή)

Θα δούμε τώρα την απλή παλινδρόμηση με άλλο μάτι, πιο προχωρημένο. Όταν κάνουμε γραμμική παλινδρόμηση, υποθέτουμε ότι η σχέση μεταξύ της ανεξάρτητης μεταβλητής και της εξαρτημένης μεταβλητής είναι γραμμική. Δείτε για παράδειγμα το διάγραμμα διασποράς του σχήματος 35. Προφανώς η σχέση μεταξύ της κατανάλωσης και της ιπποδύναμης δεν είναι γραμμική. Για να γίνει όμως πιο κατανοητό τι εννοούμε πάμε να δούμε το εξής: **Analyze→Regression→Curve Estimation** για να εμφανιστεί το παράθυρο της εικόνας 78. Έχουμε περάσει τις μεταβλητές που μας ενδιαφέρουν στα δεξιά κουτάκια όπως βλέπετε εξαρτημένη και ανεξάρτητη. Έχουμε «τικάρει» τις επιλογές **Quadratic**, **Cubic**, **Inverse** και **Display ANOVA table**. Ας υποθέσουμε ότι με Y συμβολίζουμε την εξαρτημένη μεταβλητή και με X την ανεξάρτητη. Δείτε τον πίνακα 5 για να καταλάβετε τα μοντέλα που έχουμε επιλέξει.

Μοντέλο	**Μαθηματική έκφραση**
Γραμμικό (Linear)	$Y=\alpha+\beta X$
Τετραγωνικό (Quadratic)	$Y=\alpha+\beta X+\gamma X^2$
Κυβικό (Cubic)	$Y=\alpha+\beta X+\gamma X^2+\delta X^3$
Αντίστροφο (Inverse)	$1/Y=\alpha+\beta X$ ή $Y=1/(\alpha+\beta X)$

Πίνακας 7: Μοντέλα και μαθηματική έκφραση τους.



Εικόνα 78

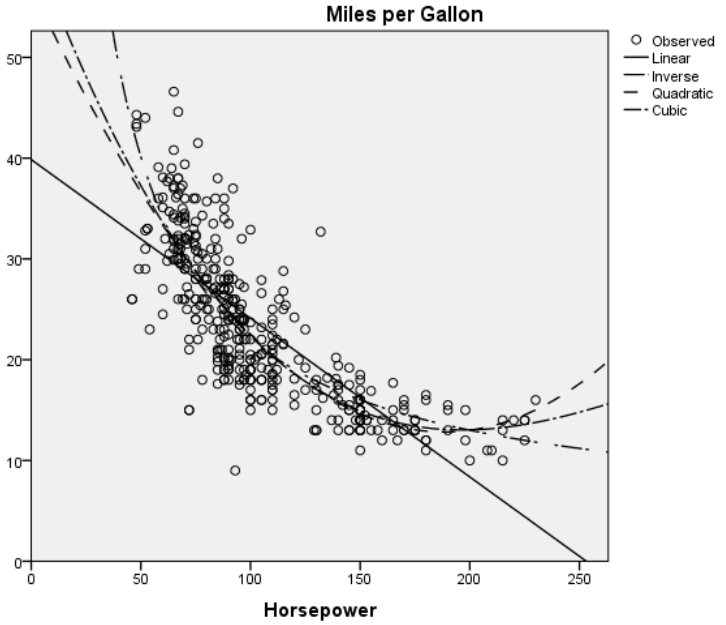
Πατώντας **OK** θα πάρουμε πολλούς πίνακες και το διάγραμμα του σχήματος Α. Παρατηρούμε ότι η γραμμή δεν ταιριάζει και πολύ στα δεδομένα. Οι καμπύλες γραμμές ταιριάζουν πιο πολύ. Παρακάτω στον πίνακα 8 παρουσιάζουμε τους συντελεστές προσδιορισμού και τις προσαρμοσμένες (λαμβάνοντας υπόψιν τον αριθμό των παρατηρήσεων και των παραμέτρων του μοντέλου) τιμές αυτών.

Μοντέλο	R	R Square	Adjusted R Square	Std. Error of the Estimate
Γραμμικό	.771	.595	**.594**	4.974
Αντίστροφο	.812	.659	**.658**	4.562
Τετραγωνικό	.824	.679	**.677**	4.437
Κυβικό	.824	.679	**.677**	4.436

Πίνακας 8: Συντελεστές προσδιορισμού των μοντέλων.

Ας κοιτάξουμε τις τιμές των προσαρμοσμένων συντελεστών προσδιορισμού του κάθε μοντέλου. Ο συντελεστή προσδιορισμού υπενθυμίζουμε είναι το ποσοστό της μεταβλητότητας (διακύμανσης) της εξαρτημένης μεταβλητής που εξηγείται από τη γνώση της ανεξάρτητης μεταβλητής και πιο συγκεκριμένα από το μοντέλο. Το γραμμικό μοντέλο έχει την χαμηλότερη τιμή. Το τετραγωνικό και το κυβικό μοντέλο έχουν την ίδια τιμή. Αυτό έγινε διότι αν δούμε και τα υπόλοιπα πινακάκια (δεν τα έβαλα για να μη γεμίζουμε σελίδες) θα δούμε ότι ο κυβικός όρος δεν είναι στατιστικά σημαντικός. Δηλαδή, τι έχουμε το τετραγωνικό τι έχουμε το κυβικό μοντέλο, δεν αλλάζει κάτι. Άρα με βάση αυτό θα προτιμούσαμε το τετραγωνικό μοντέλο. Αν όμως η διαφορά μεταξύ τετραγωνικού και γραμμικού μοντέλου ήταν μικρή (π.χ. από 0.595 πηγαίναμε στο 0.62) τότε ίσως δε συνέφερε να κρατήσουμε το τετραγωνικό.

Αυτός ο τρόπος βέβαια επιλογής μοντέλου είναι απλός. Υπάρχουν πιο αποτελεσματικοί τρόποι, αλλά δεν θα το συνεχίσουμε άλλο. Αυτή η παράγραφος ήταν περισσότερο για εξοικείωση και για εξάσκηση, για εκπαιδευτικούς λόγους δηλαδή.



Σχήμα 67: Διάγραμμα διασποράς με τις γραμμές των διάφορων μοντέλων

8.2 Λογιστική παλινδρόμηση για δίτιμη εξαρτημένη μεταβλητή

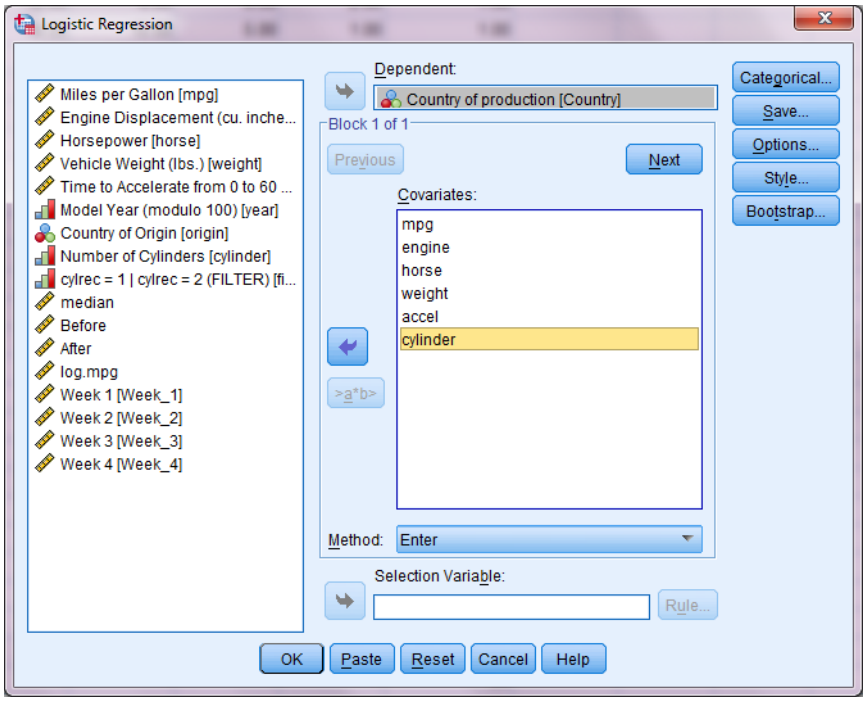
Έστω ότι θέλουμε να κάνουμε παλινδρόμηση (απλή ή πολλαπλή) και η εξαρτημένη μεταβλητή παίρνει δύο τιμές μόνο, 0 ή 1, ναι ή όχι, επιτυχία ή αποτυχία. Σε αυτήν την περίπτωση η κλασσική παλινδρόμηση που είδαμε προηγουμένως δεν μπορεί να εφαρμοστεί. Γιατί απλά ή εξαρτημένη μεταβλητή δεν ακολουθεί την κανονική κατανομή, δεν είναι ούτε καν συνεχής μεταβλητή. Πρέπει να παίξουμε με λογιστική παλινδρόμηση. Έστω Y η δίτιμη εξαρτημένη μεταβλητή (Y=0 ή Y=1) και X_1 και X_2 δύο ανεξάρτητες μεταβλητές. Το μοντέλο της λογιστικής παλινδρόμησης γράφεται ως εξής

$$\frac{\log Y}{1-Y} = a + \beta_1 X_1 + \beta_2 X_2 \quad \text{ή} \quad Y = \frac{e^{a+\beta_1 X_1+\beta_2 X_2}}{1+e^{a+\beta_1 X_1+\beta_2 X_2}}$$

Άρα το μοντέλο είναι γραμμικό όταν βάλουμε το λογάριθμο μέσα. Διαφορετικά η εξαρτημένη μεταβλητή Y σχετίζεται κατευθείαν με τις ανεξάρτητες μεταβλητές μέσω της εκθετικής συνάρτησης και άρα όχι γραμμικά. Να πούμε επίσης ότι ο λογαριθμικός μετασχηματισμός που βλέπουμε στα αριστερά ονομάζεται logit (λογιστικός μετασχηματισμός), εξού και logistic regression ή λογιστική παλινδρόμηση. Και η ονομασία οφείλεται στη λογιστική παλινδρόμηση, οπότε καμία σχέση με λογιστές η παλινδρόμηση.

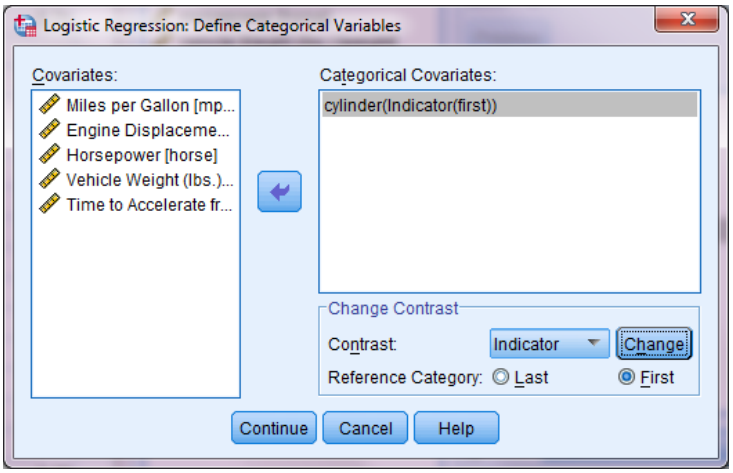
Ας δούμε ένα παράδειγμα με τα δεδομένα των αυτοκινήτων. Έχουμε στη διάθεση μας διάφορες μεταβλητές όπως είδαμε και γνωρίζουμε για το κάθε αυτοκίνητο αν είναι Ευρωπαϊκό ή όχι. Θέλουμε να δούμε το αν το αυτοκίνητο είναι Ευρωπαϊκής κατασκευής ή όχι από τι εξαρτάται και πως.

Θα επιλέξουμε **Analyze→Regression→Binary Logistic** και θα εμφανιστεί το παράθυρο της εικόνας 79.



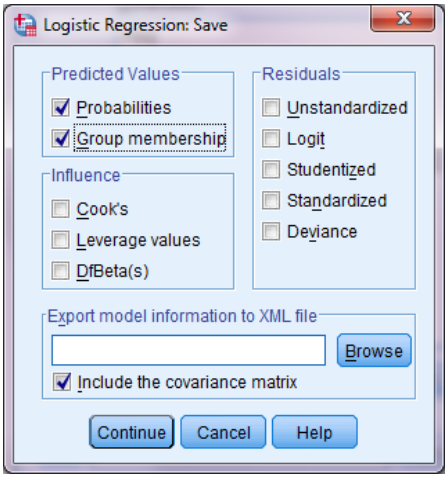
Εικόνα 79

Όπως βλέπετε περάσαμε την εξαρτημένη δίτιμη μεταβλητή στο πάνω κουτάκι και τις ανεξάρτητες μεταβλητές στο μεγάλο λευκό κουτάκι. Να πούμε ότι η μεταβλητή **cylinder** είναι κατηγορική μεταβλητή. Είναι ο αριθμός των κυλίνδρων του κάθε αυτοκινήτου, η οποία είναι προφανώς αριθμητική, αλλά όχι συνεχής και με λίγες τιμές. Άρα μπορούμε να την θεωρήσουμε κατηγορική και ως τέτοια θα την θεωρήσουμε. Δεν είναι απαραίτητο και δεν θα το συνιστούσαμε γενικά αλλά θα το κάνουμε εδώ για να δείξουμε την περίπτωση αυτή. Να τονίσουμε ότι αυτή η «κατηγορική» μεταβλητή δεν είναι και η πιο κατάλληλη όπως είδαμε και στην ανάλυση διακύμανσης κατά δύο παράγοντες, αλλά θα τη χρησιμοποιήσουμε για το παράδειγμα μας. Πατώντας **Categorical** θα πάμε στο παράθυρο της εικόνας 80.



Εικόνα 80

Στο παράθυρο της εικόνας 80 θα περάσουμε την/τις κατηγορική/ές μεταβλητή/ές δεξιά στο λευκό κουτάκι. Έπειτα θα «κλικάρουμε» το **First** (δεν είναι απαραίτητο) και μετά **Change** και μετά **Continue**. Η κατηγορική μεταβλητή εδώ παίρνει τις τιμές 3, 4, 5, 6 και 8. Με αυτόν τον τρόπο έχουμε θέσει ως τιμή ή κατηγορία αναφοράς την πρώτη τιμή το 3. Όπως θα δούμε αργότερα στις εκτιμήσεις των παραμέτρων, η σύγκριση θα γίνει σε σχέση με την τιμή 3. Πατώντας **Save** στο παράθυρο της εικόνας 79 θα πάμε στο παράθυρο της εικόνας 81. Εκεί θα επιλέξουμε τα **Probabilities** και **group membership**.

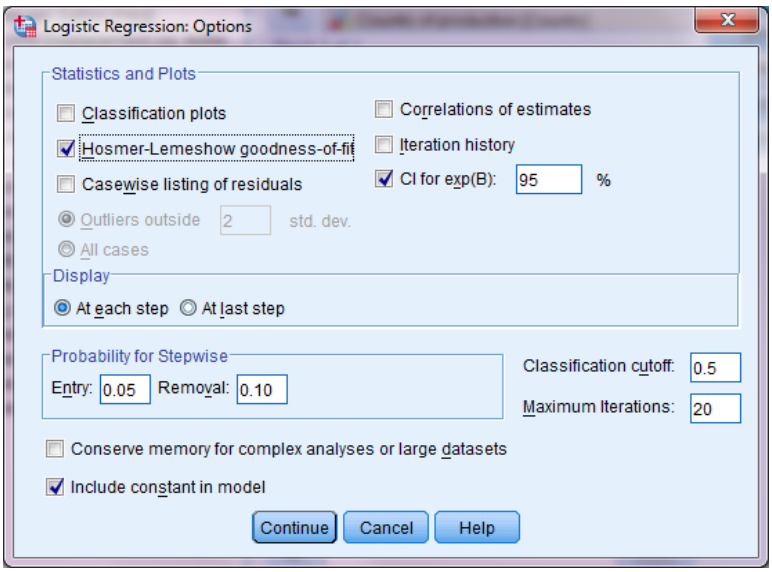


Εικόνα 81

Η λογιστική παλινδρόμηση θα εκτιμήσει την πιθανότητα της επιτυχίας, του ναι, την πιθανότητα η εξαρτημένη μεταβλητή να πάρει την τιμή 1 γενικά. Επίσης έχει έναν αυτόματο κανόνα, αν η εκτιμώμενη πιθανότητα είναι 0.5 και άνω τότε η εκτιμώμενη τιμή της εξαρτημένης μεταβλητής είναι 1. Θα δούμε πως μπορούμε να το αλλάξουμε αυτό με τη χρήση της καμπύλης ROC πιο μετά.

Πατώντας **Options** στο παράθυρο της εικόνας 79 μεταφερθούμε στο παράθυρο της εικόνας 82. Εκεί θα επιλέξουμε τα **Hosmer-Lemeshow goodness-of-fit** και **CI for exp(B)**.

Αν θέλετε να επιλέξετε bootstrap έχετε υπόψιν σας ότι δεν θα σωθούν οι εκτιμώμενες πιθανότητες και οι εκτιμώμενες τιμές της εξαρτημένης μεταβλητής. Θα σας δώσει όμως εκτιμήσεις της μεροληψίας των εκτιμηθέντων παραμέτρων καθώς και τα 95% διαστήματα εμπιστοσύνης αυτών. Εμείς εδώ δεν το επιλέξαμε.



Εικόνα 82

Θα εμφανιστούν διάφοροι πίνακες, αλλά αυτοί που μας ενδιαφέρουν είναι κάτω από το

Block 1: Method = Enter

Αρχής γενομένης με τον πίνακα του σχήματος 68.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	324.380	9	.000
	Block	324.380	9	.000
	Model	324.380	**9**	**.000**

Σχήμα 68: Έλεγχος στατιστικής σημαντικής σημαντικότητας του μοντέλου συνολικά.

Ο πίνακας του σχήματος 69 είναι κάτι σαν τον πίνακα της ανάλυσης διακύμανσης των σχημάτων 37 και 41 που είδαμε στην παλινδρόμηση. Κοιτάζει δηλαδή τη

στατιστική σημαντικότητα όλως των μεταβλητών μαζί. Το νούμερο 9 δηλώνει τον αριθμό των παραμέτρων που έχουμε στο μοντέλο μας. Μετρήστε τους συντελεστές στον πίνακα του σχήματος πλην της σταθεράς (**Constant**) για να πειστείτε.

Ο πίνακας του σχήματος 69 μας δίνει δύο ψευδό-R^2 συντελεστές. Οι τιμές τους κυμαίνονται και εδώ από 0 έως 1 και υψηλές τιμές δείχνουν καλή προσαρμογή του μοντέλου.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	193.344[a]	**.564**	**.768**

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Σχήμα 69: Ψευδό-R^2 συντελεστές.

Ο πίνακας του σχήματος 70 μας δείχνει το αποτέλεσμα του ελέγχου καλής προσαρμογής των Hosmer και Lemeshow. Η p-value είναι 0.209 εδώ. Γενικά, τιμές της p-value μεγαλύτερες του 0.05 είναι επιθυμητές εδώ. Αυτό σημαίνει ότι ο έλεγχος καλής προσαρμογής δεν απορρίπτεται, άρα το μοντέλο προσαρμόζει στα δεδομένα στατιστικά επαρκώς.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10.872	8	**.209**

Σχήμα 70: Έλεγχος καλής προσαρμογής των Hosmer και Lemeshow.

Classification Table[a]

			Predicted		
			Country of production		
	Observed		Non European	European	Percentage Correct
Step 1	Country of production	Non European	**131**	16	89.1
		European	30	**214**	87.7
	Overall Percentage				88.2

a. The cut value is **.500**

Σχήμα 71: Σχέση εκτιμηθέντων και παρατηρηθέντων τιμών.

Ο πίνακας του σχήματος 71 είναι ένας πίνακας επαλήθευσης. Στο παράθυρο της εικόνας 73 επιλέξαμε να σωθούν οι εκτιμηθείσες τιμές της εξαρτημένης μεταβλητής. Το SPSS έκανε ένα πίνακα διπλής εισόδου με βάση τις παρατηρηθέντες και τις εκτιμηθέντες τιμές.

Παρατηρούμε ότι από τα 131+16=147 μη Ευρωπαϊκά αυτοκίνητα τα 131 (89.1%) προβλέφθηκαν σωστά ως μη Ευρωπαϊκά. Από τα 244+30=274 Ευρωπαϊκά αυτοκίνητα τα 214 (87.7%) προβλέφθηκαν σωστά ως Ευρωπαϊκά. Συνολικά δηλαδή για τα 131+214=345 από τα 391 (88.2%) αυτοκίνητα είχαμε σωστή πρόβλεψη. Παρατηρήστε στο κάτω μέρος του πίνακα του σχήματος 71 το μηνυματάκι. Όπως είπαμε και προηγουμένως, αν η εκτιμώμενη πιθανότητα είναι από 0.5 και άνω η εκτιμηθείσα τιμή της εξαρτημένης μεταβλητής θεωρείται 1. Αν πάμε στο παράθυρο του SPSS και δούμε τις μεταβλητές θα δούμε δύο επιπλέον στήλες, τις εκτιμηθείσες πιθανότητες και τις αντίστοιχες εκτιμηθείσες τιμές της εξαρτημένης μεταβλητής. Αν κάνουμε cross-tabulation με τις παρατηρηθείσες και τις εκτιμηθείσες τιμές της εξαρτημένης μεταβλητής θα πάρουμε τον πίνακα του σχήματος 71.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	mpg	**-.023**	.041	.326	1	**.568**	**.977**	**.902**	**1.058**
	engine	**.129**	.019	47.548	1	**.000**	**1.138**	**1.097**	**1.180**
	horse	**-.044**	.024	3.431	1	**.064**	**.957**	**.913**	**1.003**
	weight	**-.004**	.001	9.966	1	**.002**	**.996**	**.994**	**.999**
	accel	**.047**	.112	.178	1	**.673**	**1.048**	**.842**	**1.305**
	cylinder			10.444	4	**.034**			
	cylinder(1)	**14.627**	19594.566	.000	1	.999	2251666.315	.000	.
	cylinder(2)	**-9.612**	27633.141	.000	1	1.000	.000	.000	.
	cylinder(3)	**11.666**	19594.566	.000	1	1.000	116577.006	.000	.
	cylinder(4)	**20.350**	19811.641	.000	1	.999	688783159.077	.000	.
	Constant	**-17.767**	19594.566	.000	1	**.999**	.000		

a. Variable(s) entered on step 1: mpg, engine, horse, weight, accel, cylinder.

Σχήμα 72: Πίνακας εκτιμηθέντων παραμέτρων του μοντέλου της λογιστικής παλινδρόμησης.

Ο πίνακας του σχήματος 72 περιέχει τις εκτιμήσεις των συντελεστών της παλινδρόμησης. Έχουμε 10 συντελεστές στο σύνολο, 9 για τις ανεξάρτητες μεταβλητές και 1 για τη σταθερά. Οι συντελεστές της στήλης **B** αναφέρονται στο λογιστικό μετασχηματισμό της εξαρτημένης (δίτιμης) μεταβλητής που είδαμε πιο πριν και όχι την εξαρτημένη μεταβλητή κατευθείαν. Οι συντελεστές της στήλη **Exp(B)** αναφέρονται στην εξαρτημένη (δίτιμη) μεταβλητή. Η κατανάλωση (**mpg**), ο η ιπποδύναμη (**horse weight**) και το βάρος (**weight**) επηρεάζουν αρνητικά το λογιστικό μετασχηματισμό και την εξαρτημένη μεταβλητή γενικότερα. Δηλαδή θα λέγαμε ότι καθώς αυξάνεται το βάρος, η κατανάλωση και η ιπποδύναμη μάλλον μειώνουμε τις πιθανότητες να μιλάμε για ένα Ευρωπαϊκό αυτοκίνητο. Ο κυβισμός (**engine**) και η επιτάχυνση (**accel**) όμως επιδρούν θετικά στο να έχουμε Ευρωπαϊκό αυτοκίνητο. Η στήλη με τις p-value (**Sig.**) μας λέει πως η επιτάχυνση (p-value=0.673), η ιπποδύναμη (p-value=0.064) και η κατανάλωση (p-value=0.568) δεν είναι στατιστικά σημαντικές μεταβλητές, αφού η p-value τους είναι μεγαλύτερη του 0.05.

Αν πάμε στη στήλη **Exp(B)** θα δούμε ότι εδώ έχουμε το εκθετικό των **B**. Αν το **B** έχει αρνητική τιμή, τότε το εκθετικό του θα έχει τιμή μικρότερη της μονάδας. Αν το **B** έχει θετική τιμή, τότε το εκθετικό του θα έχει θετική τιμή. Τι σημαίνει όμως η τιμή **1.138** π.χ. για τον κυβισμό; Ας πούμε για παράδειγμα ότι έχουμε ένα αυτοκίνητο Α με κυβισμό 3500 κυβικών ιντσών και ένα αυτοκίνητο Β 3501 κυβικών ιντσών. Τα odds $\left(\frac{P(Y=1)}{P(Y=0)}\right)$ του αυτοκινήτου Β να είναι Ευρωπαϊκό είναι **1.138** φορές τα odds $\left(\frac{P(Y=1)}{P(Y=0)}\right)$ του αυτοκινήτου Α να είναι Ευρωπαϊκό. Δηλαδή το **1.138** είναι το odds ratio μεταξύ δύο αυτοκινήτων που διαφέρουν ως προς τον κυβισμό τους κατά μία κυβική ίντσα. Το **Exp(B)** είναι toodds ratio γενικά. Αν δείτε το διάστημα εμπιστοσύνης για το odds ratio θα δείτε ότι δεν περιλαμβάνει τη μονάδα. Άρα η επίδραση του κυβισμού στην εξαρτημένη μεταβλητή είναι στατιστικά σημαντική.

Το νούμερο **0.957** αναφέρεται στην ιπποδύναμη και δείχνει ότι όσο αυξάνει η ιπποδύναμη μειώνεται η πιθανότητα το αυτοκίνητο να είναι Ευρωπαϊκό. Αυτό σημαίνει ότι αν αυξήσουμε την ιπποδύναμη κατά έναν ίππο, τα odds του αυτοκινήτου να είναι Ευρωπαϊκό μειώνονται κατά $\frac{1-0.957}{0.957}\% = 4.49\%$. Αυτή όμως η επίδραση δεν είναι στατιστικά σημαντική όπως είπαμε και προηγουμένως και το βλέπουμε αυτό και από το διάστημα εμπιστοσύνης το οποίο περιέχει τη μονάδα.

Όσον αφορά τους κυλίνδρους τώρα που εδώ τους θεωρήσαμε κατηγορική μεταβλητή για να κάνουμε τη δουλειά μας μπορούμε να πούμε τα εξής. Πρώτον, αφού είναι κατηγορική μεταβλητή πρέπει α δημιουργήσουμε ψευδομεταβλητές οι οποίες θα είναι αριθμητικές. Πόσες όμως; Πόσα επίπεδα έχει η κατηγορική μεταβλητή; Εδώ 5, άρα θέλουμε 4 ψευδομεταβλητές. Οι ψευδομεταβλητές παίρνουν τιμές 0 και 1.

D_1=1 αν έχουμε 4 κυλίνδρους και 0 διαφορετικά.
D_2=1 αν έχουμε 5 κυλίνδρους και 0 διαφορετικά.
D_3=1 αν έχουμε 6 κυλίνδρους και 0 διαφορετικά.
D_4=1 αν έχουμε 4 κυλίνδρους και 0 διαφορετικά.

Αν έχουμε αυτοκίνητο με 3 κυλίνδρους, τότε $D_1=D_2=D_3=D_4=0$. Το νούμερο **14.627** μας λέει ότι η αναμενόμενη διαφορά στο λογιστικό μετασχηματισμό μεταξύ αυτοκινήτου με 3 κυλίνδρους και αυτοκινήτου με 4 κυλίνδρους είναι ίση με 14.627. Δηλαδή αν αυξήσουμε τους κυλίνδρους από 3 σε 4 μάλλον αυξάνονται οι πιθανότητες να έχουμε Ευρωπαϊκό αυτοκίνητο. Αυτή διαφορά όμως δεν είναι στατιστικά σημαντική (p-value=0.999). Το odds ratio σε αυτήν την περίπτωση είναι ίσο με **2251666.315**. Άρα τα odds ενός αυτοκινήτου με 4 κυλίνδρους να είναι Ευρωπαϊκό είναι πολύ μεγαλύτερα από τα αντίστοιχα odds ενός αυτοκινήτου με 3 κυλίνδρους. Αυτή όμως η διαφορά δεν είναι στατιστικά σημαντική. Η ίδια ερμηνεία μπορεί να δοθεί και για τους συντελεστές των άλλων κυλίνδρων. Είναι όλοι συγκρινόμενοι πάντα με τους 3 κυλίνδρους.

Αν δούμε τα p-value για τους συντελεστές των κυλίνδρων θα δούμε ότι είναι όλα πολύ υψηλά και επομένως οι όποιες διαφορές στα αυτοκίνητα των 4, 5,6 και 8 κυλίνδρων σε σχέση με τα αυτοκίνητα 3 κυλίνδρων δεν είναι στατιστικά σημαντικές. Άρα να βγάλουμε την πληροφορία για τους κυλίνδρους έξω από το μοντέλο διότι δεν είναι στατιστικά σημαντική; Έχουμε φτιάξει 4 ψευδομεταβλητές που αντιστοιχούν σε μία κατηγορική μεταβλητή. Μήπως λοιπόν πρέπει να έχουμε μία p-value αντί για 4 ή παραπάνω για να αποφασίσουμε αν ο αριθμός των κυλίνδρων είναι στατιστικά

σημαντικός; Η απάντηση είναι ναι, μία p-value θέλουμε και την έχουμε. Κάτω από το **accel** και πάνω από το **cylinder(1)** υπάρχει μία γραμμή που αναφέρεται στο **cylinder**, την ανεξάρτητη κατηγορική μεταβλητή. Η στήλη **df** για αυτήν την μεταβλητή έχει αριθμό 4, διότι 4 ψευδομεταβλητές χρησιμοποιήσαμε. Το p-value δίπλα είναι ίσο με **0.034<0.05**, άρα η μεταβλητή κύλινδροι είναι στατιστικά σημαντική. Το μοντέλο μας γράφεται ως εξής:

$$\frac{\log Y}{1-Y} = -17.767 - 0.023mpg + 0.129engine + 14.627D_1 \ldots + 20.350D_4$$

ή

$$P(Y=1) = \frac{e^{-17.767-0.023mpg+0.129engine+14.627D_1\ldots+20.350D_4}}{1+e^{-17.767-0.023mpg+0.129engine+14.627D_1\ldots+20.350D_4}}$$

Η πάνω γραμμή είναι μία ευθεία, η κάτω όμως όχι. Άρα το μοντέλο είναι ένα γενικευμένο γραμμικό, με την έννοια ότι ενώ η εξαρτημένη μεταβλητή δε συνδέεται γραμμικά με τις παραμέτρους των ανεξάρτητων μεταβλητών, αλλά μόλις βάλουμε το λογιστικό μετασχηματισμό, τότε έχουμε γραμμική σχέση. Η γραμμική σχέση όμως είναι μεταξύ της μετασχηματισμένης εξαρτημένης μεταβλητής και των παραμέτρων.

Όπως είπαμε και προηγουμένως αν η πιθανότητα είναι πάνω από 0.5 τότε το SPSS θέτει ως εκτιμηθείσα τιμή το 1. Μπορούμε να το αλλάξουμε αυτό και να επιλέξουμε άλλο κατώφλι πάνω από το οποίο μία πιθανότητα θα οδηγεί στην τιμή 1 για την εκτιμηθείσα τιμή; Η απάντηση είναι ναι. Για αυτό το σκοπό θα χρειαστούμε την καμπύλη χαρακτηριστικού λειτουργικού δέκτη (ROC curve).

Θα πάμε στο παράθυρο της εικόνας 43 (παράγραφος 5.9) και θα περάσουμε στο πάνω κουτάκι τη στήλη με τις εκτιμηθείσες πιθανότητες από το μοντέλο μας και στο κάτω κουτάκι την παρατηρηθείσα τιμή της εξαρτημένης μεταβλητής και θα ορίσουμε ότι η τιμή 1 είναι η τιμή ενδιαφέροντος, και θα «τικάρουμε»όλες τις επιλογές όπως κάναμε και στο παράθυρο της εικόνας 43. Θα πούμε εδώ κάποια πράγματα που δεν είπαμε στην παράγραφο 5.9.

Η ευαισθησία (sensitivity) και η ειδικότητα (specificity) ορίζονται ως

$$Ευαισθησία = \frac{TP}{TP+FN} \quad \text{και} \quad Ειδικότητα = \frac{TN}{TN+FP}.$$

		Predicted country of production		Ειδικότητα πάνω γραμμή, ευαισθησία κάτω γραμμή
		Non European	European	
True country of production	Non European	**TN=131**	**FP=16**	$\frac{131}{131+16} = 0.89$
	European	**FN=30**	**TP=214**	$\frac{214}{214+30} = 0.87$

Σχήμα 73: Σχέση εκτιμηθέντων και παρατηρηθέντων τιμών (σχήμα 62).

Οι συμβολισμοί είναι ως εξής: TP (true positive)=αληθής πρόβλεψη επιτυχίας, FN (false negative)=ψευδής πρόβλεψη αποτυχίας, FP (false positive)=ψευδή πρόβλεψη επιτυχίας και TN (true negative)=αληθής πρόβλεψη αποτυχίας. Στο παράδειγμα μας επιτυχία είναι όταν το αυτοκίνητο είναι Ευρωπαϊκό. Άρα η ευαισθησία είναι η πιθανότητα θετικής ανίχνευσης της επιτυχίας όταν υπάρχει. Η ειδικότητα είναι η πιθανότητα σωστής ανίχνευσης της αποτυχίας όταν υπάρχει. Παρατηρήστε ότι αυτά τα ποσοστά παρουσιάζονται και στον πίνακα του σχήματος 73. Εμείς θέλουμε αυτά δύο νούμερα να είναι ταυτόχρονα υψηλά, ή την ευαισθησία υψηλή και το νούμερο 1-ειδικότητα χαμηλό. Άρα στην καμπύλη ROC (βλέπε σχήμα 24, θέλουμε να είμαστε ψηλά στον κατακόρυφο άξονα και αριστερά στον οριζόντιο).

Coordinates of the Curve

Test Result Variable(s): Predicted probability

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity	Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
.0000000	1.000	.993	…	…	…
.0000000	1.000	.986	.4447682	.881	.129
.0000000	1.000	.980	.4512554	.881	.122
.0000000	1.000	.973	.4536793	.881	.116
.0000000	1.000	.966	.4752330	.877	.116
.0393930	1.000	.918	.5643528	.865	.088
.0408885	1.000	.912	.5653434	.865	.082
.0426940	1.000	.905	.5656916	.865	.075
.0459369	1.000	.898	.5721451	.861	.075
.0485806	1.000	.891	.5797437	.861	.068
.0494644	1.000	.884	.5864937	.861	.061
...	...	...	.6765436	.852	.041
...	...	...	**.6790903**	**.848**	**.041**
...	...	...	...	...	...

Σχήμα 74:Κάποιες ενδεικτικές τιμές της καμπύλης ROC

Ας πάμε τώρα στον πίνακα του σχήματος 65 όπου έχουμε παραλείψει αρκετές τιμές για λόγους χώρου. Η πρώτη στήλη δίνει το κατώφλι για την εκτιμώμενη πιθανότητα. Οι άλλες δύο στήλες είναι η ευαισθησία και η ειδικότητα. Παρατηρούμε ότι αν αυξήσουμε το κατώφλι, πέφτει σιγά η ευαισθησία αλλά πέφτει και η 1-ειδικότητα. Θέλουμε κάτι στη μέση. Θα πάρουμε για παράδειγμα την τιμή 0.68 να είναι το κατώφλι μας.

Θα πάμε στο SPSS και θα φτιάξουμε μία νέα στήλη. Αν η εκτιμώμενη πιθανότητα είναι από 0.68 και άνω η εκτιμώμενη τιμή της εξαρτημένης μεταβλητής θα είναι 1, ειδάλλως 0. Αν δε θυμάστε πως γίνεται αυτό, δείτε τα παράθυρα των εικόνων 12-14. Μετά θα κάνουμε ένα cross-tabulation μεταξύ της πραγματικής δίτιμης μεταβλητής και της εκτιμηθείσας δίτιμης μεταβλητής. Παρατηρούμε από τον πίνακα του σχήματος 75 ότι τα αποτελέσματα είναι καλύτερα. Το εκτιμώμενο συνολικό ποσοστό σωστής κατάταξης από (131+214)/391=0.882 πήγε στο (141+207)/391==0.89. Παρατηρήστε επίσης ότι στην κάτω γραμμή, κατατάξαμε 7

Ευρωπαϊκά αυτοκίνητα ως μη Ευρωπαϊκά με αυτό το κατώφλι, ενώ στην πάνω γραμμή τα μη Ευρωπαϊκά που ανιχνεύτηκαν ως μη Ευρωπαϊκά αυξήθηκαν. Ίσως σε άλλες περιπτώσεις να παίρναμε σημαντικά υψηλότερο ποσοστό συνολικής διακύμανσης. Εδώ φάνηκε ότι χάνουμε λίγο από κάπου και κερδίζουμε λίγο από κάπου αλλού.

			predicted	
			Non European	European
Country of production	Non European	Count	**142**	5
		% within Country of production	**96.6%**	3.4%
	European	Count	37	**207**
		% within Country of production	15.2%	**84.8%**

Σχήμα 75: Νέα σχέση εκτιμηθέντων και παρατηρηθέντων τιμών (βλέπε σχήμα 71).

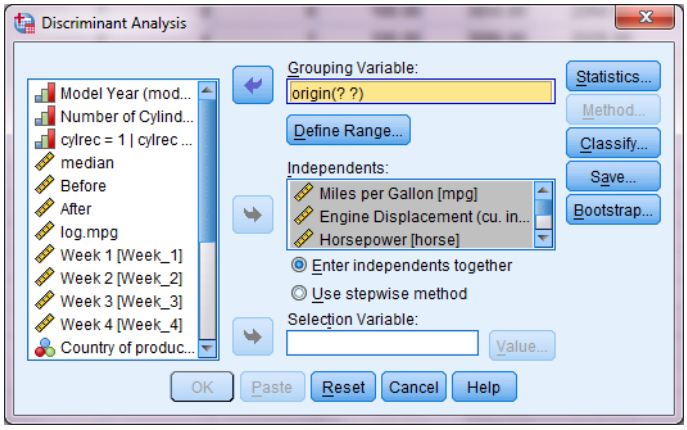
8.3 Διαχωριστική ανάλυση

Τώρα θα δούμε τη λεγόμενη διαχωριστική ανάλυση (discriminant analysis) ή διακριτική ανάλυση όπως αλλιώς είναι γνωστή. Ας υποθέσουμε ότι έχουμε μετρήσεις από μία ή περισσότερες μεταβλητές και επίσης ξέρουμε και τις διάφορες ομάδες του δείγματος. Δηλαδή, μπορεί να έχουμε την ηλικία, το βάρος, το ύψος του μισθού, τα χρόνια εκπαίδευσης και τα χρόνια προϋπηρεσίας για κάποιους εργαζομένους σε μία εταιρεία και επίσης ξέρουμε και το φύλο τους. Σκοπός είναι να δούμε αν μπορούμε να διαχωρίσουμε τους άντρες από τις γυναίκες με βάση τις μεταβλητές που έχουμε στη διάθεση μας. Να τονίσουμε ότι οι μεταβλητές που θα χρησιμοποιήσουμε πρέπει να είναι αυστηρά ποσοτικές, αριθμητικές και συγκεκριμένα συνεχείς ή έστω διακριτές με πολλές τιμές όμως. Δηλαδή γνώση της ομάδας αίματος, των πολιτικών πεποιθήσεων και γενικά κατηγορικές μεταβλητές δεν επιτρέπονται.

Μία από τις πολλές μεθόδους με την οποία μπορούμε να δούμε πόσο διαφορετικά είναι τα δύο φύλα είναι η διαχωριστική ανάλυση. Η μέθοδος αυτή έχει δύο βασικές κατηγορίες, τη γραμμική (linear) και την τετραγωνική (quadratic) διαχωριστική ανάλυση. Στη γραμμική διαχωριστική ανάλυση προσπαθούμε με γραμμές ή με «τοίχους» αν θέλετε να διαχωρίσουμε τις ομάδες (προσοχή τις ξέρουμε ήδη τις ομάδες, διαφορετικά δεν μπορούμε να την κάνουμε αυτήν την ανάλυση). Η τετραγωνική διαχωριστική ανάλυση μας επιτρέπει να διαχωρίσουμε τις ομάδες με καμπύλες γραμμές ή «καμπύλους τοίχους», όχι ευθείες γραμμές. Εμείς εδώ στο SPSS θα επικεντρωθούμε στη γραμμική περίπτωση διότι η δεύτερη περίπτωση (τετραγωνική διαχωριστική ανάλυση) δεν προσφέρεται.

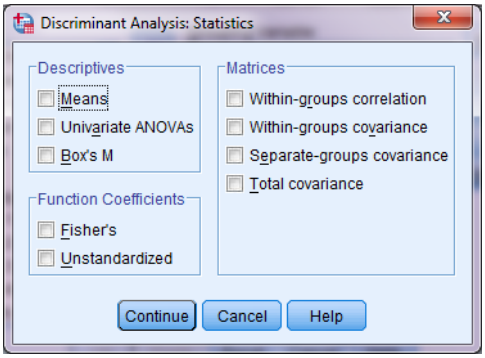
Θα χρησιμοποιήσουμε πάλι τα δεδομένα των αυτοκινήτων. Ξέρουμε ότι τα αυτοκίνητα είναι είτε Ευρωπαϊκά, είτε Αμερικάνικα, είτε Ιαπωνικά. Με βάση λοιπόν κάποιες πληροφορίες που έχουμε για αυτά θα προσπαθήσουμε να τα διαχωρίσουμε. Γνωρίζουμε το βάρος, την ιπποδύναμη, την επιτάχυνση, την κατανάλωση, τον κυβισμό και τον αριθμό των κυλίνδρων τους. Επειδή ο αριθμός των κυλίνδρων μπορεί να θεωρηθεί και κατηγορική μεταβλητή δεν θα τους χρησιμοποιήσουμε ως πληροφορία. Ο αριθμός των κυλίνδρων μπορεί να πάρει κάποιες συγκεκριμένες τιμές, 3, 4, 5, 6 ή 8 κύλινδροι, επομένως είναι λίγες τιμές.

Για να διεξάγουμε τη διαχωριστική ανάλυση επιλέγουμε ως εξής: **Analyze→Classify→Discriminant** και θα εμφανιστεί το παράθυρο της εικόνας 83.



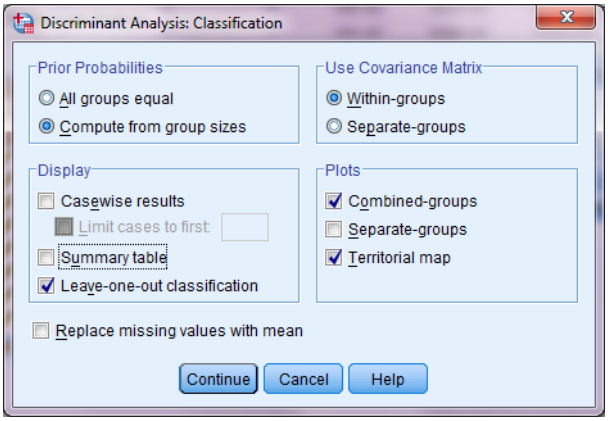
Εικόνα 83

Η μεταβλητή που δηλώνει την ομάδα (τη χώρα προέλευσης σε μας εδώ) θα πάει στο πάνω κουτάκι (**Grouping Variable:**). Θα επιλέξουμε **Define Range** για να δηλώσουμε πόσες ομάδες έχουμε. Εμείς εδώ έχουμε 3, οι οποίες είναι αριθμημένες ως 1, 2 και 3, άρα θα βάλουμε εκεί 1 και 3. Όπως βλέπουμε και στην εικόνα Α θα περάσουμε τις συνεχείς μεταβλητές στο μεγάλο κουτάκι δεξιά (**Independents:**). Κάτω από τις ανεξάρτητες μεταβλητές μας δίνεται η επιλογή να βάλουμε όλες τις μεταβλητές μέσα ή να αφήσουμε το SPSS να κάνει επιλογή ποιες μεταβλητές πρέπει να μπουν. Αυτή τη λογική την είδαμε στην πολλαπλή παλινδρόμηση. Εμείς θα το αφήσουμε ως έχει για τώρα (ο ενδιαφερόμενος αναγνώστης μπορεί να το ψάξει αν επιθυμεί) και θα επιλέξουμε το **Statistics** για να εμφανιστεί το παράθυρο της εικόνας 84.



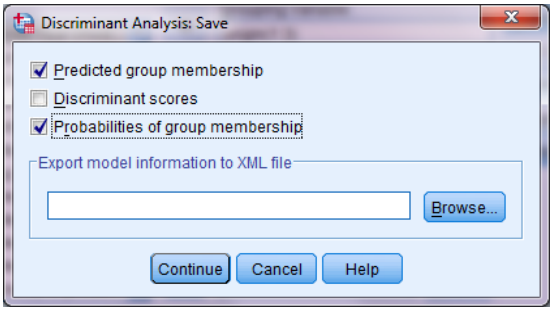
Εικόνα 84

Αν θέλουμε, μπορούμε να επιλέξουμε το **Means** για να δούμε τους μέσους των μεταβλητών για κάθε ομάδα ξεχωριστά. Αν στο παράθυρο της εικόνας 83 πατήσουμε το **Classify** θα εμφανιστεί το παράθυρο της εικόνας 85. Εκεί θα επιλέξουμε το **Compute from group sizes** και τα **Leave-one-out cross validation**, **Combined-groups** και **Territorial map** όπως βλέπετε στην εικόνα 85. Νόμιζα ότι η επιλογή **Separate-Groups** πάνω δεξιά οδηγεί στην τετραγωνική διαχωριστική ανάλυση αλλά τελικά δεν οδηγεί.



Εικόνα 85

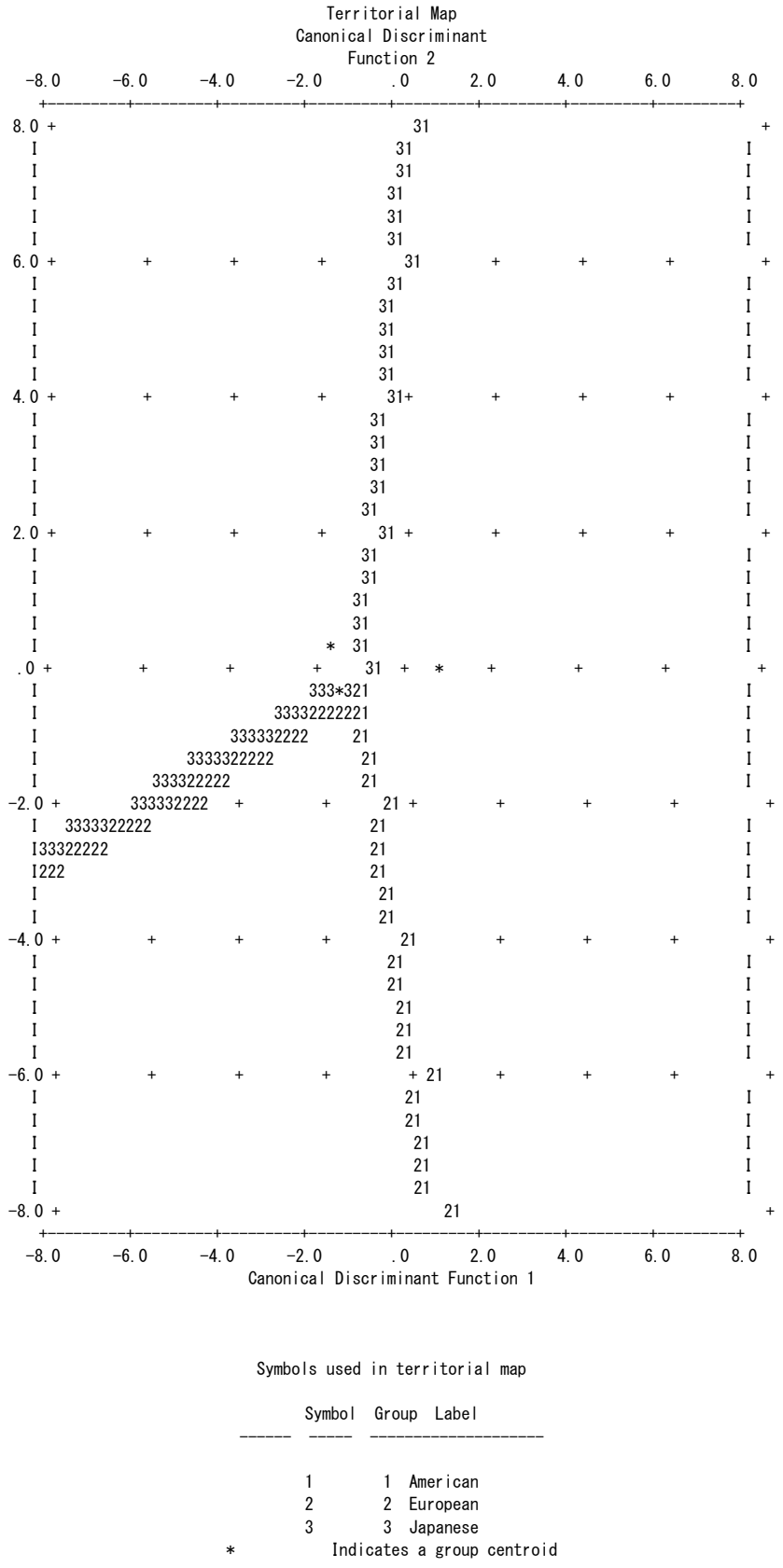
Πατάμε Continue και γυρνάμε στο παράθυρο της εικόνας Α όπου εκεί θα επιλέξουμε το **Save** και θα πάμε στο παράθυρο της εικόνας 86. Εκεί θα επιλέξουμε τα **Predicted group membership** και **Probabilities of group membership**. Το bootstrap δεν θα το χρησιμοποιήσουμε εδώ διότι δεν κάνει αυτά που θα θέλαμε να κάνει.



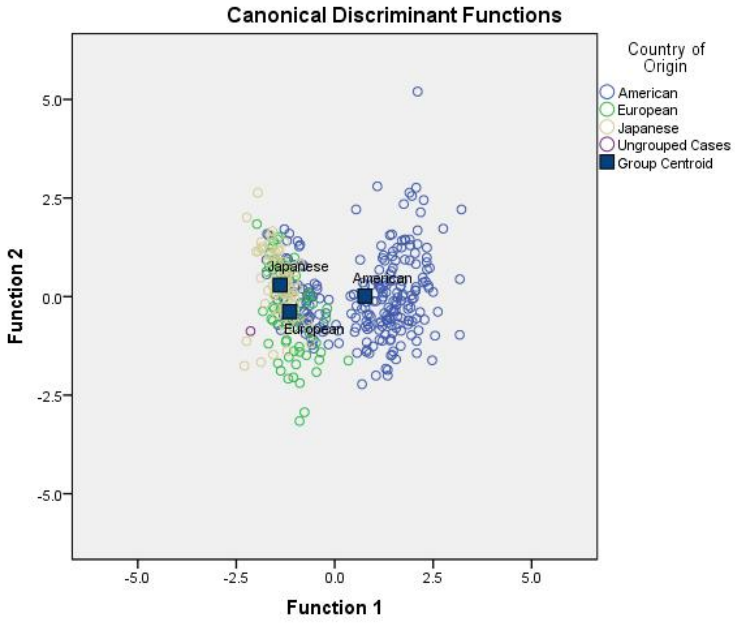
Εικόνα 86

Τελειώσαμε με τις επιλογές, οπότε τώρα πατάμε **OK** στο παράθυρο της εικόνας 83. Θα πάρουμε κάποιους πίνακες με κάτι νούμερα που δεν καταλαβαίνουμε τι είναι, μετά κάτι διαγράμματα και μετά πάλι ένα πίνακα. Το πρώτο γράφημα (σχήμα 76) είναι το λεγόμενο **Territorial Map** (το επιλέξαμε στο παράθυρο της εικόνας 85). Είναι ένα γράφημα με 3 αριθμούς συνολικά (επειδή έχουμε 3 ομάδες). Είναι 3 ευθείες γραμμές αφού κάναμε γραμμική διαχωριστική ανάλυση. Άρα βλέπουμε ότι οι τρεις ομάδες χωρίζονται με τρεις γραμμές. Δεξιά του γραφήμματος είναι η ομάδα 1 (Αμερικάνικα αυτοκίνητα). Κάτω αριστερά είναι η ομάδα 2 (Ευρωπαϊκά αυτοκίνητα) και αριστερά και πάνω είναι τα Ιαπωνικά αυτοκίνητα. Άρα το γράφημα μας βοηθάει να δούμε τις γραμμές που χωρίζουν τις τρεις ομάδες.

Το επόμενο γράφημα που βλέπουμε είναι αυτό στο σχήμα 77. Εδώ βλέπουμε με χρώμα τις τρεις ομάδες, δεξιά τα Αμερικάνικα αυτοκίνητα, αριστερά και κάτω τα Ευρωπαϊκά και αριστερά και πάνω τα Ιαπωνικά αυτοκίνητα. Ελπίζω τώρα η σύνδεση μεταξύ των δύο γραφημάτων αλλά και η εικόνα που παρουσιάζει το καθένα να είναι πιο ξεκάθαρα.



Σχήμα 76: Διάγραμμα περιοχής των ομάδων



Σχήμα 77: Γράφημα με τις ομάδες

Τέλος, το τελευταίο πινακάκι (Σχήμα 78) έχει και αυτό τη σημασία του. Ο πίνακας είναι χωρισμένος σε δύο μέρη.

Ας δούμε πρώτα το πάνω μέρος (**Original**). Αφού ολοκληρώθηκε ο αλγόριθμος, παίρνει κάθε μία παρατήρηση ξεχωριστά και με βάση τις μεταβλητές την κατατάσει σε μία ομάδα. Αυτό το κάνει για όλες τις παρατηρήσεις και στο τέλος φτιάχνει αυτόν τον πίνακα. Η πρώτη γραμμή έχει τα νούμερα 203, 13 και 28, σύνολο 244. Αυτό σημαίνει ότι από τα 244 Αμερικάνικα αυτοκίνητα, ο αλγόριθμος κατέταξε σωστά τα 203. Τα 13 θεωρήθηκαν Ευρωπαϊκά και τα 28 θωρήθηκαν Ιαπωνικά. Στη δεύτερη γραμμή έχουμε 69 Ερυωπαϊκά αυτοκίνητα εκ των οποίων τα 27 κατατάχθηκαν σωστά, τα 12 θεωρήθηκαν Αμερικάνικα και τα 29 Ιαπωνικά. Ακριβώς από κάτω είναι τα αντίστοιχα ποσοστά. Αυτά είναι τα λάθη στην κατάταξη. Αν π.χ. μπορούσαμε να κάνουμε τετραγωνική διαχωριστική ανάλυση αυτά λάθη να ήταν ενδεχομένως μικρότερα. Άρα έχουμε 203 και 27 και 60 αυτοκίνητα (στη διαγώνιο του πίνακα του σχήματος 78) τα οποία κατατάχθηκαν σωστά (και ένα που δεν γνωρίζουμε την προέλευση του και το οποίο κατατάχθηκε ως Ευρωπαϊκό). Άρα 290 από τα 391 (74.2%) αυτοκίνητα κατατάχθηκαν σωστά.

Ας δούμε τώρα το κάτω μέρος του πίνακα.Εδώ τα νούμερα είναι λίγο διαφορετικά. Ο αλγόριθμος πάει στην πρώτη παρατήρηση και την αφαιρεί. Έπειτα κάνει τη διαχωιρστική ανάλυση με βάση τις υπόλοιπες παρατηρήσεις. Μετά προβλέπει την ομάδα αυτής της τιμής. Αυτό γίνεται για όλες τις παρατηρήσεις. Αυτό σα μέθοδος επαλήθευσης του αλγόριθμου είναι πιο σωστή, διότι η κάθε παρατήρηση της οποίας η ομάδα προβλέπεται δεν είχε συμμετοχή στον αλγόριθμο. Άρα το συνολικό εκτιμώμενο ποσοστό σωστής κατάταξης (72.9% τώρα) είναι πιο σωστό, πιο αμερόληπτο, πιο αντικειμενικό, πιο αντιπροσωπευτικό.

Αν πάμε τώρα στις στήλες με τα δεδομένα θα δούμε 4 επιπλέον στήλες. Η πρώτη στήλη είναι η προβλεφθείσα ομάδα της κάθε ομάδας με βάση όμως τον πρώτο αλγόριθμο που μόλις περιγράψαμε. Δηλαδή η κάθε παρατήρηση της οποίας η ομάδα προβλέπεται έχει συμμετάσχει στον αλγόριθμο. Αν κάνουμε ένα cross-tabulation (βλέπε παράθυρο εικόνας 39) τότε θα πάρουμε το πάνω κομμάτι του πίνακα του

σχήματος 78. Οι επόμενες τρεις στήλες δίνουν την πιθανότητα κάθε παρατήρησης να ανήκει στην κάθε μία ομάδα. Άρα για κάθε παρατήρηση έχουμε τρεις πιθανότητες, είτε να ανήκει στην ομάδα 1, είτε στην ομάδα 2 είτε στην ομάδα 3.

Classification Results[a,c]

			Predicted Group Membership			
		Country of Origin	American	European	Japanese	Total
Original	Count	American	**203**	**13**	**28**	**244**
		European	**12**	**27**	**29**	**68**
		Japanese	**3**	**16**	**60**	**79**
		Ungrouped cases	**0**	**1**	**0**	**1**
	%	American	83.2	5.3	11.5	100.0
		European	17.6	39.7	42.6	100.0
		Japanese	3.8	20.3	75.9	100.0
		Ungrouped cases	.0	100.0	.0	100.0
Cross-validated[b]	Count	American	**201**	**14**	**29**	**244**
		European	**12**	**26**	**30**	**68**
		Japanese	**3**	**18**	**58**	**79**
	%	American	82.4	5.7	11.9	100.0
		European	17.6	38.2	44.1	100.0
		Japanese	3.8	22.8	73.4	100.0

a. **74.2**% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. **72.9**% of cross-validated grouped cases correctly classified.

Σχήμα 78: Αποτελέσματα επαλήθευσης ομάδων

Βιβλιογραφία

Ελληνική

- Αδαμίδη Ε. (2012). *Καμπύλες Λειτουργικού Χαρακτηριστικού Δέκτη και Στατιστική Ανάλυση Πραγματικών Ιατρικών Δεδομένων*. Διπλωματική Εργασία, ΕΜΠ.
- Δαφέρμος Β. (2005). Κοι*νωνική στατιστική με το SPSS*. Εκδόσεις Ζήτη.
- Δαφέρμος Β. (2002). *Επαναληπτικές στατιστικές μετρήσεις στις κοινωνικές επιστήμες*. Εκδόσεις Leader Books. Αθήνα.
- Καλαματιανού Α.. (2003). *Κοινωνική Στατιστική, Μέθοδοι Μονοδιάστατης Ανάλυσης*. Εκδόσεις Παπαζήση.
- Μαθηματικά και στοιχεία Στατιστικής Γ΄ ενιαίου λυκείου. Σχολικό εγχειρίδιο.
- Μπερσίμης Σ. (2007). *Διδακτικές σημειώσεις προγράμματος επιμόρφωσης στην ιατρική στατιστική*.
- Ντζούφρας Ι. (2005). *Εισαγωγή στη Βιοστατιστική και την Επιδημιολογία*. Πανεπιστημιακές σημειώσεις.
- Ξεκαλάκη Ε. (2001). *Μη Παραμετρική Στατιστική*. Αθήνα.
- Ξεκαλάκη Ε. (2004). *Τεχνικές Δειγματοληψίας*. Εκδόσεις Οικονομικού Πανεπιστημίου Αθηνών.
- Πανάρετος Ι. & Ξεκαλάκη Ε. (2000). *Εισαγωγή στη στατιστική σκέψη, Τόμος 1, Περιγραφική στατιστική*. Αθήνα.
- Πανάρετος Ι. & Ξεκαλάκη Ε. (2000). *Εισαγωγή στη στατιστική σκέψη, Τόμος 2, Εισαγωγή στις πιθανότητες και στη στατιστική συμπερασματολογία*. Αθήνα.
- Πανάρετος Ι. (20001). *Γραμμικά μοντέλα με έμφαση στις εφαρμογές*. Αθήνα.
- Παυλόπουλος Β. (2008). *Μοντέλα Ανάλυσης Διακύμανσης* .
- Ρούσσος Π. & Ευσταθίου Γ. (2008). *Σύντομο Εγχειρίδιο SPSS 16.0* .
- Βιτωράτου Σ. & Λύκου Α. *Σημειώσεις σεμιναρίου SPSS* .
- Τσαγρής Μιχαήλ (2006). *Ανάλυση Διακύμανσης στο SPSS*. Διδακτικές σημειώσεις.
- Χαλικιάς Ι. (2003). *Στατιστική: Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις* (2η Έκδοση). Rosili, Αθήνα.
- Ψιλούτσικου Μ. (2005). *Σημειώσεις για το μάθημα Ποσοτικές Μέθοδοι ΙΙ*.

Ξένη

- Charter R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology* 130(3):290-304.
- Casella G. and Berger R.L. (2002). *Statistical Inference* (2nd Ed.). Pacific Grove: Duxbury.
- Draper, N. R., & Smith, H. (1981). Applied regression analysis (2nd Ed.). New York: John Wiley.
- Edwards A. (1948). "Note on the "correction for continuity" in testing the significance of the difference between correlated proportions". *Psychometrika* 13(3): 185-187.
- Fleiss J.L. (1981). *Statistical methods for rates and proportions* (2nd Ed.). New York: John Wiley.

- Efron B. & Tibshirani R.J. (1993). *An introduction to the Bootstrap*. New York: Chapman & Hall/CRC.
- Efron B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistic*s 7(1): 1-26.
- Johnson R.A. and Wichern D.W. (2007). *Applied Multivariate Statistical Analysis* (6th Ed.). New Jersey: Prentice Hall.
- Kleinbaum, David G. and Mitchel Klein (2002). *Logistic regression: a self-learning text* (2nd Ed.). New York: Springer.
- Kuritz, S. J., J. R. Landis, and G. G. Koch. 1988. A general overview of Mantel-Haenszel methods: Applications and recent developments. *Annual Review of Public Health* 9(1): 123-160.
- Maxwell A.E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry* 116(535): 651-655.
- McNemar Q. (1947). "Note on the sampling error of the difference between correlated proportions or percentages". *Psychometrika* 12(2): 153-157.
- Montgomery D.C. (2001). *Design and Analysis of Experiments* (5th Edition). John and Wiley and Sons Inc.
- Schwab J.A. (2007). Solving Homework Problems in Data Analysis I (Lecture notes).
- Stuart A.A. (1955) A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42(3/4): 412-416
- University of Sheffield. Department of Probability and Statistics. Methods for Data Analysis (Lecture notes).
- Williams R. (2004). Advanced Social Statistics I (Lecture notes).

Λεξικό στατιστικών όρων

A

acceptance region	περιοχή αποδοχής
acceptance sampling	δειγματοληψία αποδοχής
accessibility sampling	δειγματοληψία προσιτότητας
additive model	προσθετικό μοντέλο
algorithm	αλγόριθμος
alpha	άλφα
alternative hypothesis	εναλλακτική υπόθεση
analysis of covariance	ανάλυση συνδιακύμανσης
analysis of variance	ανάλυση διακύμανσης
antithetic random variables	αντίθετες ή αντιθετικές τυχαίες μεταβλητές
approximation	προσέγγιση
area sampling	δειγματοληψία κατά περιοχές
arithmetic mean	αριθμητικός μέσος
assumption	υπόθεση
asymptotic	ασυμπτωτικός
autocorrelation	αυτοσυσχέτιση
autocorrelation function	συνάρτηση αυτοσυσχέτισης
average	μέσος όρος
axis	άξονας

B

backward elimination procedure	μέθοδος αποκλεισμού μεταβλητών
bar chart	ραβδόγραμμα
Bayes factor	παράγοντας του Bayes
Bayes' theorem	θεώρημα του Bayes
Bayesian inference	συμπερασματολογία κατά Bayes
Bayesian information criterion	κριτήριο πληροφορίας του Bayes
Bayesian statistics	μπεϋζιανή στατιστική ή στατιστική κατά Bayes
beta	βήτα
beta distribution	κατανομή βήτα
beta function	συνάρτηση βήτα
beta-binomial distribution	κατανομή βήτα-διωνυμική
bias	μεροληψία
bimodal distribution	δικόρυφη κατανομή
binary data	δυαδικά δεδομένα
binomial coefficient	διωνυμικός συντελεστής
binomial distribution	διωνυμική κατανομή
binomial test	διωνυμικός έλεγχος
biostatistics	βιοστατιστική
bivariate distribution	διμεταβλητή κατανομή
box plot	διάγραμμα πλαισίου απολήξεων ή θηκόγραμμα
branching processes	κλαδωτές ανελίξεις
Buffon's needle	βελόνα του Buffon

C

canonical correlation analysis	ανάλυση κανονικών συσχετίσεων
capability index	δείκτης ικανότητας
capture-recapture methods	μέθοδοι σύλληψης και επανασύλληψης
categorical variable	κατηγορική μεταβλητή
categorical data analysis	ανάλυση κατηγορικών δεδομένων
causality	αιτιότητα
censored data	λογοκριμμένα δεδομένα
census	απογραφή
central limit theorem	καντρικό οριακό θεώρημα
characteristic function	χαρακτηριστική συνάρτηση
chi-square distribution	χι-τετράγωνο κατανομή
classification	κατάταξη
clinical trials	κλινικές δοκιμές
cluster analysis	ανάλυση σε ομάδες
clustered bar chart	ομαδοποιημένο ραβδόγραμμα
clustered sampling	δειγματοληψία κατά ομάδες
coefficient	συντελεστής
coefficient of determination	συντελεστής προσδιορισμού
coefficient of variation	συντελεστής μεταβλητότητας
cohort studies	μελέτες κοορτής
collinearity	συγγραμμικότητα
compound Poisson process	σύνθετη διαδικασία Poisson
concordant pair	αρμονικό ζεύγος
conditional distribution	δεσμευμένη κατανομή
conditional expectation	δεσμευμένη αναμονή
conditional probability	δεσμευμένη πιθανότητα
confidence band	ζώνη εμπιστοσύνης
confidence interval	διάστημα εμπιστοσύνης
confirmatory factor analysis	επιβεβαιωτική παραγοντική ανάλυση
confounding factor	συγχυτικός παράγοντας
conjugate distribution	συζυγής κατανομή
consistency	συνέπεια
contingency coefficient	συντελεστής συνάφειας
contingency table	πίνακας συνάφειας
continuity correction	διόρθωση συνέχειας
continuous distribution	συνεχής κατανομή
continuous random variable	συνεχής τυχαία μεταβλητή
continuous variable	συνεχής μεταβλητή
contrasts	διαφορές
control chart	διάγραμμα ελέγχου
control-cases studies	μελέτες μαρτύρων ασθενών
convolution	συνέλιξη
correlation	συσχέτιση
correlation coefficient	συντελεστής συσχέτισης
correlation matrix	πίνακας συσχετίσεων
correlogram	κορελόγραμμα
correspondence analysis	ανάλυση αντιστοιχιών
covariance	συνδιακύμανση

covariate	συμμεταβλητή
coverage probability	πιθανότητα κάλυψης
credibility interval	διάστημα αξιοπιστίας
credibility region	περιοχή αξιοπιστίας
critical value	κριτική τιμή
cross-sectional studies	διατμηματικές μελέτες
cumulative distribution function	αθροιστική συνάρτηση κατανομής
curve	καμπύλη

D

data	δεδομένα
data analysis	ανάλυση δεδομένων
data mining	εξώρυξη δεδομένων
deciles	δεκατημόρια
degree of association	βαθμός συνάφειας ή σύνδεσης
degrees of freedom	βαθμοί ελευθερίας
demography	δημογραφία
dendrogram	δεντρόγραμμα
dependent variable	εξαρτημένη μεταβλητή
descriptive statistics	περιγραφικά στατιστική
design matrix	πίνακας σχεδιασμού
design of experiments	σχεδιασμός πειραμάτων
diffusion process	πρότυπα διάχυσης
discordant pair	δυσαρμονικό ζεύγος
discrete distribution	διακριτή κατανομή
discrete random variable	διακριτή τυχαία μεταβλητή
discriminant analysis	διαχωριστική ή διακριτική ανάλυση
dispersion	διασπορά
distribution	κατανομή
dot plot	σημειόγραμμα
dummy variable	ψευδομεταβλητή

E

efficient estimator	αποτελεσματικός εκτιμητής
efficiency	αποτελεσματικότητα
epidemiology	επιδημιολογία
ergodic theorem	εργοδικό θεώρημα
ergodicity	εργοδικότητα
error	λάθος, σφάλμα
estimation	εκτίμηση
estimator	εκτιμητής
estimated value	εκτιμηθείσα τιμή
event	γεγονός
exact test	ακριβής έλεγχος
expectation	αναμονή
expected value	αναμενόμενη τιμή
experimental units	πειραματικές μονάδες
explanatory variable	επεξηγηματική μεταβλητή

exponential distribution	εκθετική κατανομή
exponential family	εκθετική οικογένεια

F

factor	παράγοντας
factor analysis	παραγοντική ανάλυση
factor loadings	επιβαρύνσεις των παραγόντων
factorial design	παραγοντικός σχεδιασμός
false negative case	ψευδή θετική περίπτωση
false positive case	ψευδή αρνητική περίπτωση
finite	πεπερασμένος
first quartile	πρώτο τεταρτημόριο
fixed effects	σταθερές επιδράσεις
fixed effects model	μοντέλο σταθερών επιδράσεων
follow-up studies	μελέτες παρακολούθησης
forecasting methods	μέθοδοι προβλέψεων
forward procedure	μέθοδος προοδευτικής προσθήκης μεταβλητών
frame	πλαίσιο
frequency	συχνότητα
frequency polygon	πολύγωνο συχνοτήτων

G

game theory	θεωρία παιγνίων
gamma distribution	κατανομή γάμα
gamma function	συνάρτηση γάμα
general linear model	γενικό γραμμικό μοντέλο
generalized linear models	γενικευμένα γραμμικά μοντέλα
geometric distribution	γεωμετρική κατανομή
geometric mean	γεωμετρικός μέσος
goodness of fit test	έλεγχος καλής προσαρμογής
graph	γράφημα

H

harmonic mean	αρμονικός μέσος
hazard function	συνάρτηση κινδύνου
hazard rate	ρυθμός κινδύνου
hazard rate function	συνάρτηση βαθμού κινδύνου
heterogeneity	ετερογένεια
heteroscedasticity	ετεροσκεδαστικότητα
hierarchical model	ιεραρχικό μοντέλο
highest posterior density	περιοχή υψίστης a-posteriori πυκνότητας
histogram	ιστόγραμμα
homogeneity	ομοιογένεια
homoscedasticity	ομοσκεδαστικότητα
hypergeometric distribution	υπεργεωμετρική κατανομή
hypothesis testing	έλεγχος υπόθεσης

I

independent variable	ανεξάρτητη μεταβλητή
index number	αριθμοδείκτης
inertia	αδράνεια
infinite	άπειρος
information	μέτρο πληροφορίας
information matrix	πίνακας πληροφορίας
informative prior distribution	πληροφοριακή prior κατανομή
interaction	αλληλεπίδραση
interquartile range	ενδοτεταρτημοριακό εύρος
interval estimation	εκτίμηση σε διάστημα
interval scale	κλίμακα διαστήματος
irreducible Markov chain	αδιαχώριστη αλυσίδα Markov

J

joint distribution	από κοινού κατανομή
judgemental or purposive sampling	δειγματοληψία κρίσης ή σκοπιμότητας

K

kurtosis	κύρτωση

L

lack of memory	έλλειψη μνήμης
latent variables	λανθάνουσες μεταβλητές
latin squares	λατινικά τετράγωνα
law of large numbers	νόμος των μεγάλων αριθμών
least squares estimators	εκτιμητές ελαχίστων τετραγώνων
leptokurtic distribution	λεπτόκυρτη κατανομή
linear model	γραμμικό μοντέλο
location parameter	παράμετρος θέσης
logistic model	λογιστικό μοντέλο
logistic regression	λογιστική παλινδρόμηση
logistics distribution	λογιστική κατανομή
loglinear model	λογαριθμικό μοντέλο
lognormal distribution	λογαριθμοκανονική κατανομή
longitudinal data analysis	ανάλυση διαμήκων δεδομένων
longitudinal studies	διαμήκεις μελέτες
loss function	συνάρτηση απώλειας

M

main effects	κύριες επιδράσεις
marginal distribution	περιθώρια κατανομή
market research	έρευνα αγοράς
Markov chains	αλυσίδες Markov ή Μαρκοβιανές αλυσίδες
maximum	μέγιστο

maximum likelihood	μέγιστη πιθανοφάνεια
mean	μέσος
mean absolute deviation	μέση απόλυτη απόκλιση
mean square error	μέσο τετραγωνικό σφάλμα
measure theory	θεωρία μέτρου
measures of association	μέτρα συνάφειας
median	διάμεσος
mesokurtic distribution	μεσόκυρτη κατανομή
minimum	ελάχιστο
missing values	εκλειπούσες τιμές
mixed effects model	μοντέλο μικτών επιδράσεων
mode	κορυφή
model	μοντέλο
moment	ροπή
moment generating function	ροπογεννήτρια συνάρτηση
monotone regression	μονότονη παλινδρόμηση
mortality (death) rate	ρυθμός θνησιμότητας
moving average	κινητός μέσος
multicollinearity	πολυσυγγραμμικότητα
multidimensional scaling techniques	πολυδιάστατες τεχνικές κλιμακοποίησης
multilevel models	πολυεπίπεδα μοντέλα
multinomial distribution	πολυωνυμική κατανομή
multiple comparisons	πολλαπλοί έλεγχοι
multiple correspondence analysis	πολλαπλή ανάλυση αντιστοιχιών
multiple linear regression	πολλαπλή γραμμική παλινδρόμηση
multivariate	πολυμεταβλητός
multivariate analysis of variance	πολυμεταβλητή ανάλυση διακύμανσης
multivariate case	πολυμεταβλητή περίπτωση

N

negative binomial distribution	αρνητική διωνυμική κατανομή
negative confounder	αρνητικός συγχυτικός παράγοντας
negative predicted value	αρνητική προβλεπτική τιμή
nested models	φωλιασμένα μοντέλα
nominal scale	ονομαστική κλίμακα
nominal variable	ονομαστική μεταβλητή
non-informative prior distribution	μη πληροφοριακή prior κατανομή
non-linear model	μη-γραμμικό μοντέλο
nonparametric statistics	μη παραμετρική στατιστική
normal distribution	κανονική κατανομή
nuisance factor	ενοχλητικός παράγοντας
null hypothesis	μηδενική υπόθεση

O

observations	παρατηρήσεις
observed significance level (p-value)	παρατηρηθέν επίπεδο σημαντικότητας (p-τιμή)

odds ratio	λόγος συμπληρωματικών πιθανοτήτων ή κλάσμα λόγου πιθανοτήτων
one-sided test	μονόπλευρος έλεγχος
one-way ANOVA	ανάλυση διακύμανσης κατά ένα παράγοντα
operating characteristic curve	χαρακτηριστική λειτουργική καμπύλη
opinion poll	σφυγμομέτρηση κοινής γνώμης
ordered data	διατεταγμένα δεδομένα
ordinal scale	κλίμακα διάταξης
ordinary least squares method	μέθοδος ελαχίστων τετραγώνων
orthogonal contrasts	ορθογώνιες διαφορές
outliers	ακραίες τιμές
overparameterization	υπερπαραμετροποίηση

P

parameter	παράμετρος
parametric statistics	παραμετρική στατιστική
partial correlation coefficient	μερικός συντελεστής συσχέτισης
permutation	αναδιάταξη
pie chart	κυκλικό διάγραμμα ή διάγραμμα πίτας
pivotal quantity	αντιστρεπτή ποσότητα
platykurtic distribution	πλατύκυρτη κατανομή
plot	γράφημα
point estimation	σημειακή εκτίμηση
point process	σημειακή διαδικασία
population	πληθυσμός
population characteristic	χαρακτηριστικό του πληθσμού
positive confounder	θετικός συγχυτικός παράγοντας
positive predicted value	θετική προβλεπτική τιμή
posterior or a-posteriori distribution	posterior κατανομή
predicted value	προβλεφθείσα τιμή
prediction interval	διάστημα πρόβλεψης
prevalence	επιπολασμός
principal component analysis	ανάλυση κύριων συνιστωσών
principal coordinate analysis	ανάλυση κύριων συντεταγμένων
prior or a-priori distribution	prior κατανομή
probability	πιθανότητα
probability density function	συνάρτηση πυκνότητας πιθανότητας
probability generating function	πιθανογεννήτρια συνάρτηση
probability sampling	δειγματοληψία κατά πιθανότητα
prospective studies	προοπτικές μελέτες

Q

qualitative	ποιοτικός
quantitative	ποσοτικός
queues	ουρές
quota sampling	δειγματοληψία με προκαθορισμένα ποσοστά

R

	τυχαίες επιδράσεις
random effects	
random effects model	μοντέλο τυχαίων επιδράσεων
random factor	τυχαίος παράγοντας
random processes	τυχαίες διαδικασίες
random sample	τυχαίο δείγμα
random variable	τυχαία ματαβλητή
random walk	τυχαίος περίπατος
randomization	τυχαιοποίηση
randomized complete block design	τυχαιοποιημένοι πλήρως σχεδιασμοί κατά μπλοκ
range	εύρος
ranks	τάξεις μεγέθους
ratio scale	κλίμακα λόγου
recurrent state	επαναληπτική κατάσταση
regression analysis	ανάλυση παλινδρόμησης
regression coefficients	συντελεστές παλινδρόμησης
regressors	παλινδρομητές
rejection region	περιοχή απόρριψης
rejection sampling	δειγματοληψία απόρριψης
relative risk	σχετικός κίνδυνος
reliability	αξιοπιστία
reliability coefficient	συντελεστής αξιοπιστίας
reliability function	συνάρτηση αξιοπιστίας
renewal process	ανανεωτική διαδικασία
renewal theory	θεωρία ανανέωσης
repeated measures	επαναλαμβανόμενες μετρήσεις
replication	επαναληψιμότητα
residuals	κατάλοιπα
response surface	επιφάνεια απόκρισης
response variable	απαντητική μεταβλητή
retrospective studies	αναδρομικές μελέτες
ridge regression	αμφικλινής παλινδρόμηση
robust regression	εύρωστη παλινδρόμηση
runs test	έλεγχος ροών

S

sample	δείγμα
sample surveys	δειγματοληπτικές έρευνες
sampling distribution	δειγματική κατανομή
sampling frame	δειγματοληπτικό πλαίσιο
sampling mean	δειγματικός μέσος
sampling techniques	δειγματοληπτικές τεχνικές
sampling theory	θεωρία δειγματοληψίας
sampling units	δειγματοληπτικές μονάδες
saturated (full) model	κορεσμένο μοντέλο
scale	κλίμακα
scale parameter	παράμετρος κλίμακας
scatter plot	διάγραμμα διασποράς
seasonality	εποχικότητα

semiparametric	ημιπαραμετρικός
sensitivity analysis	ανάλυση ευαισθησίας
serial correlation coefficient	σειριακός συντελεστής συσχέτισης
shape parameter	παράμετρος σχήματος
sign test	προσημικός έλεγχος
significance level	επίπεδο σημαντικότητας
simple linear regression	απλή γραμμική παλινδρόμηση
simple random sampling	απλή τυχαία δειγματοληψία
simulation	προσομοίωση
size effect	επίδραση μεγέθους
skewed distribution	ασύμμετρη κατανομή
skewness	ασυμμετρία
specification limits	όρια προδιαγραφών
specificity	ειδικότητα
sphericity	σφαιρικότητα
split-half reliability	αξιοπιστία ημίκλαστου
stable process	σταθερή διεργασία
standard deviation	τυπική απόκλιση
standard error	τυπικό σφάλμα
standardization	τυποποίηση
standardized values	τυποποιημένες τιμές
state space	χώρος καταστάσεων
stationary distribution	στάσιμη κατανομή
stationary process	στάσιμη διαδικασία
statistical	στατιστικός (επίθετο)
statistical inference	στατιστική συμπερασματολογία
statistically significant	στατιστικά σημαντικός
statistician	στατιστικός (ο/η, ιδιότητα)
statistics	στατιστική
stem-and-leaf plot	διάγραμμα μίσχου-φύλου
stepwise regression	βηματική παλινδρόμηση
stochastic models	στοχαστικά μοντέλα
stochastic processes	στοχαστικές ανελίξεις
stratified analysis	στρωματοποιημένη ανάλυση
stratified randomization	στρωματοποιημένη τυχαιοποίηση
stratified sampling	στρωματοποιημένη δειγματοληψία
study population	υπό μελέτη πληθυσμός
subjective probability	υποκειμενική πιθανότητα
sufficiency	επάρκεια
sum	άθροισμα
sum of products	άθροισμα γινομένων
sum of squares	άθροισμα τετραγώνων
survival analysis	ανάλυση επιβίωσης
survival function	συνάρτηση επιβίωσης
symmetric distribution	συμμετρική κατανομή
systematic sampling	συστηματική δειγματοληψία

T

target population	αντικειμενικός πληθυσμός
third quartile	τρίτο τεταρτημόριο

time series	χρονολογικές σειρές
transformation	μετασχηματισμός
transient state	μεταβατική κατάσταση
transition matrix	πίνακας μετάβασης
transition probability	πιθανότητα μετάβασης
trend	τάση
trimmed mean	περικομμένος μέσος
two-sided test	αμφίπλευρος έλεγχος
two-way ANOVA	ANOVA κατά δύο παράγοντες

U

unbiased estimator	αμερόληπτος εκτιμητής
uniform distribution	ομοιόμορφη κατανομή
unimodal distribution	μονοκόρυφη κατανομή
univariate case	μονομεταβλητή περίπτωση
universe	ολότητα
unstable process	μη σταθερή διεργασία

V

variability	μεταβλητότητα
variable	μεταβλητή
variance	διακύμανση
variance-covariance matrix	πίνακας διακύμανσης-συνδιακύμανσης

W

waiting time	χρόνος αναμονής
weighted least squares method	μέθοδος σταθμισμένων ελαχίστων τετραγώνων
weighted mean	σταθμισμένος μέσος