ΕΝΝΟΙΕΣ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΜΕ ΤΟ SPSS

Dr. Ευθυμία Νικήτα

Θεσσαλονίκη 2012

ΠΕΡΙΕΧΟΜΕΝΑ

5. ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΔΟΚΙΜΑΣΙΕΣ

6. ΕΛΕΓΧΟΙ ΣΕ ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ

7. ΠΑΛΙΝΔΡΟΜΗΣΗ-ΣΥΣΧΕΤΙΣΗ

8. ΑΝΑΛΥΣΗ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ

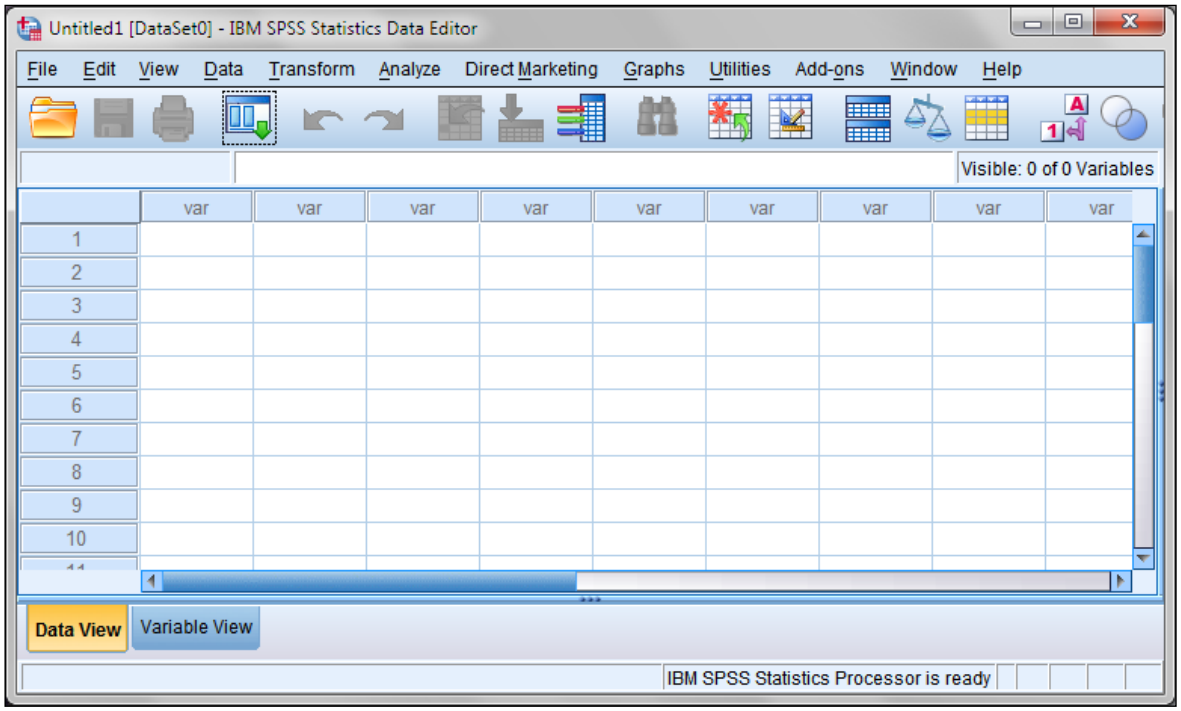
1. ΤΟ ΠΕΡΙΒΑΛΛΟΝ ΕΡΓΑΣΙΑΣ ΤΟΥ SPSS 19.0

Το στατιστικό πρόγραμμα *SPSS* (Statistical Package for the Social Sciences) είναι ένα από τα καλύτερα στατιστικά πακέτα και μπορεί να χρησιμοποιηθεί για τη στατιστική ανάλυση τόσο κοινωνικοοικονομικών δεδομένων όσο και δεδομένων των θετικών επιστημών.

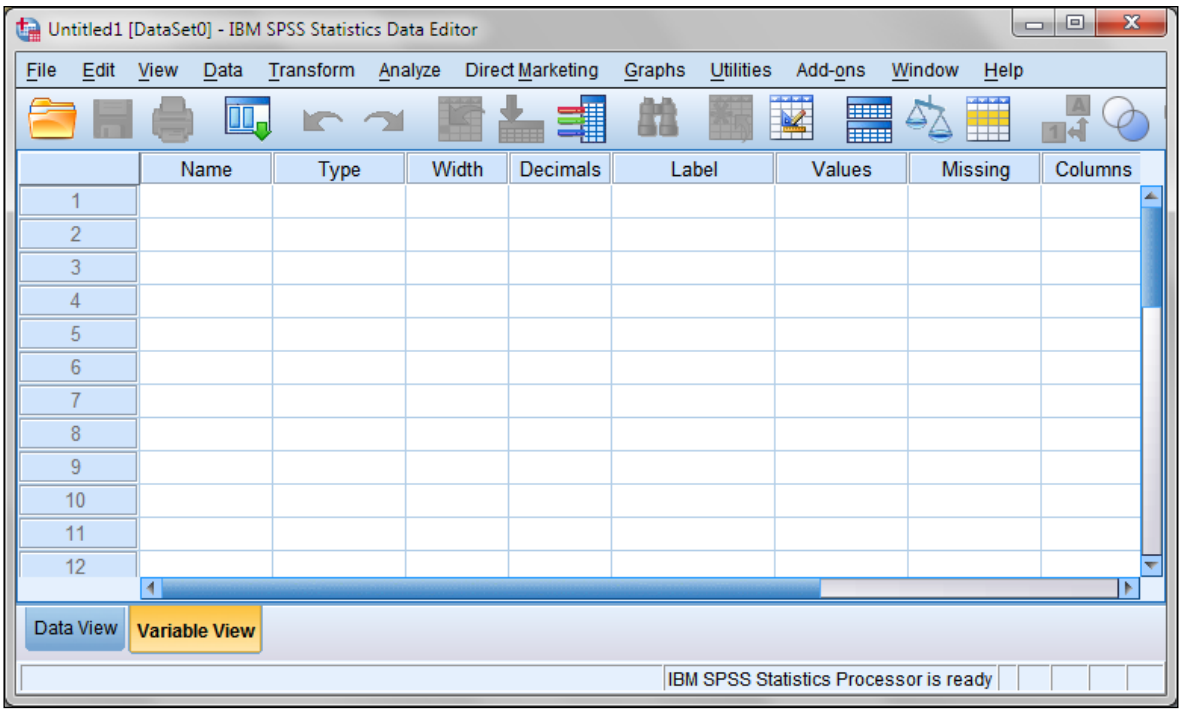
1.1 ΦΥΛΛΑ ΕΡΓΑΣΙΑΣ ΤΟΥ SPSS

Στο *SPSS* υπάρχουν δύο βασικά αρχεία: το αρχείο δεδομένων (**SPSS Data Editor**), και το αρχείο αποτελεσμάτων (**SPSS Viewer**). Ο SPSS Data Editor είναι ένα φύλλο εργασίας, στο οποίο καταχωρούμε τα δεδομένα που θέλουμε να αναλύσουμε. Ο *SPSS Data Editor* αποτελείται από δύο παράθυρα: Το *Data View* (Σχήμα 1.1) και το *Variable View* (Σχήμα 1.2). Στο πρώτο εισάγουμε τα δεδομένα που θα αναλύσουμε και στο δεύτερο ορίζουμε τα δεδομένα αυτά, δηλαδή δίνουμε επιμέρους στοιχεία για αυτά. Οι οριζόντιες γραμμές στο *Data View* ονομάζονται ***Cases*** (Περιπτώσεις) και είναι αριθμημένες με αύξουσα σειρά, ενώ οι στήλες αντιστοιχούν στις ***Variables*** (Στατιστικές Μεταβλητές).

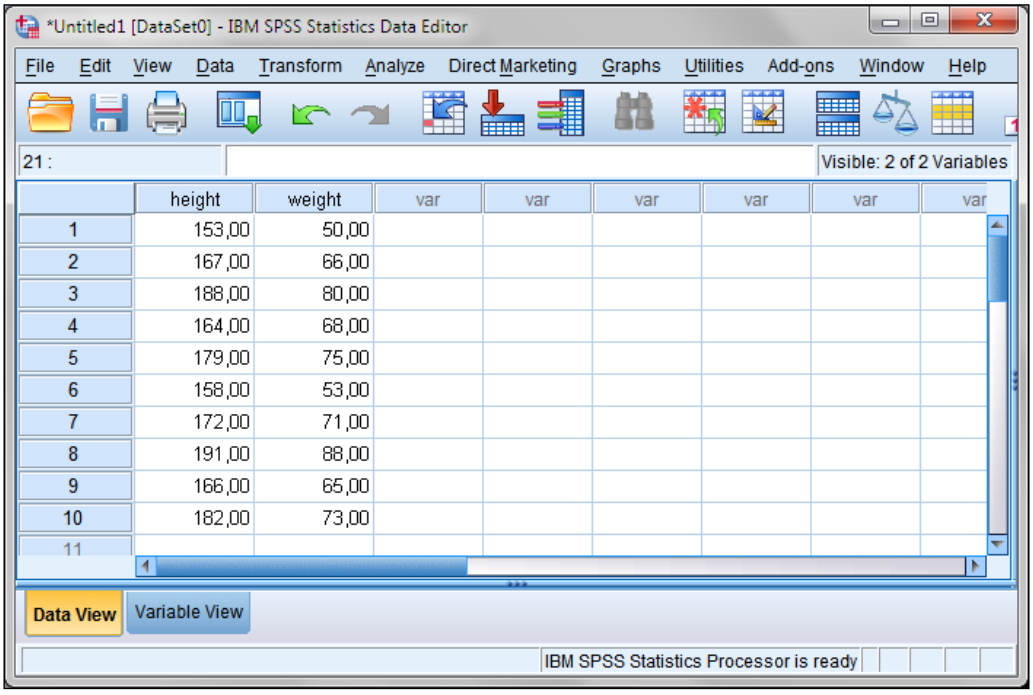
Για παράδειγμα, εάν θέλουμε να αναλύσουμε το ύψος και το βάρος 10 ατόμων, κάθε οριζόντια σειρά θα περιλαμβάνει το ύψος και το βάρος ενός ατόμου, όλα τα ύψη θα δίνονται στην ίδια στήλη και όλα τα βάρη στη διπλανή στήλη (Σχήμα 1.3).



Σχήμα 1.1. Ο *SPSS Data Editor* στο παράθυρο *Data View* για εισαγωγή δεδομένων

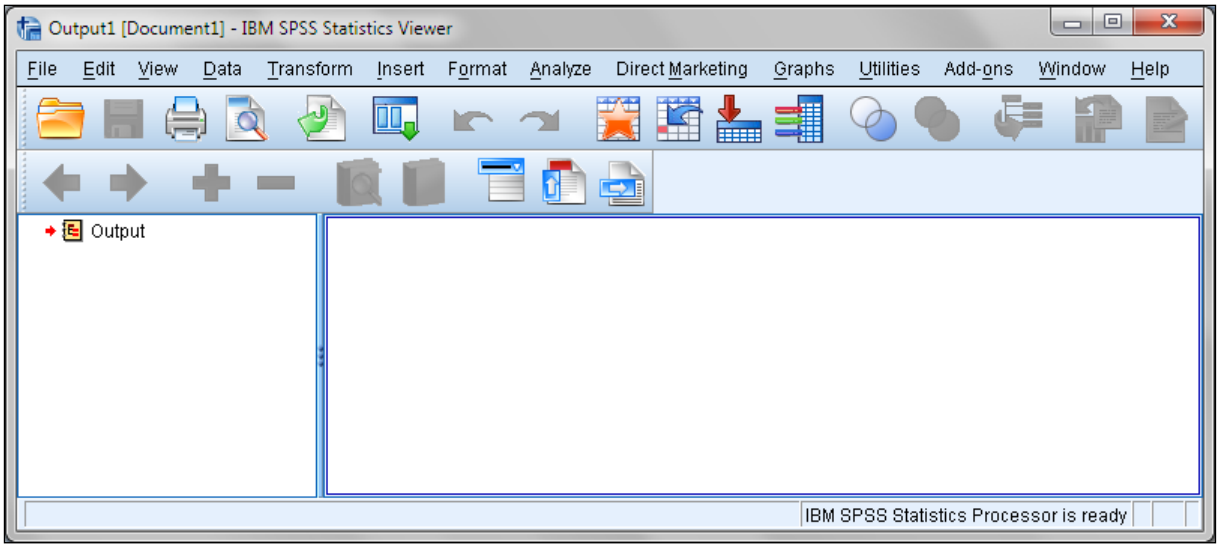


Σχήμα 1.2. Ο *SPSS Data Editor* στο παράθυρο *Variable View* για μορφοποίηση δεδομένων



Σχήμα 1.3. Ο *SPSS Data Editor* (παράθυρο *Data View*) με τα δεδομένα ύψους και βάρους 10 ατόμων

Ο *SPSS Viewer* είναι το αρχείο αποτελεσμάτων (Σχήμα 1.4). Στο αριστερό του παράθυρο, στο *Output*, εμφανίζονται οι στατιστικές πράξεις που έχουν γίνει και στο δεξιό τα στατιστικά αποτελέσματα.



Σχήμα 1.4. Ο *SPSS Viewer* για την παρουσίαση των αποτελεσμάτων

Η **γραμμή μενού** (menu bar) στον *SPSS Data Editor* περιλαμβάνει τις επιλογές: *File, Edit, View, Data, Transform, Analyze, Direct Marketing, Graphs, Utilities, Add-ons, Window, Help*. Οι ίδιες λέξεις υπάρχουν και στον SPSS Viewer, όπου όμως υπάρχουν επιπλέον και οι λέξεις *Insert* και *Format*. Οι ενέργειες που μας επιτρέπουν να κάνουμε αυτές οι επιλογές είναι οι εξής:

- File: Μπορούμε να ανοίξουμε ένα νέο αρχείο (New), ή ένα παλιό (Open), να αποθηκεύσουμε ένα αρχείο (Save), να εκτυπώσουμε (Print), κ.ο.κ.
- Edit: Μπορούμε να τροποποιήσουμε ή να αντιγράψουμε τμήματα του αρχείου δεδομένων.
- View: Μπορούμε να προσαρμόζουμε τα διάφορα στοιχεία του παραθύρου ανάλογα με τις επιλογές μας.
- Data: Μπορούμε να πραγματοποιήσουμε αλλαγές στα δεδομένα.
- Transform: Μπορούμε να πραγματοποιήσουμε αλλαγές στις μεταβλητές.
- Analyze: Πραγματοποιούμε τη στατιστική ανάλυση των δεδομένων.
- Direct Marketing: Περιέχει εφαρμογές για διαχείριση επιχειρησιακών δεδομένων.
- Graphs: Δημιουργούμε γραφικές παραστάσεις.
- Utilities: Πρόκειται για μια επιλογή γενικών χρήσεων. Για παράδειγμα, δίνονται πληροφορίες για μια μεταβλητή ή ένα αρχείο.
- Add-ons: Περιλαμβάνει πρόσθετες παροχές της IBM (εταιρείας-κατόχου του SPSS)
- Window: Μπορούμε να μεταβούμε σε κάποιο άλλο ενεργό παράθυρο.
- Help: Προσφέρει διάφορα είδη βοήθειας.

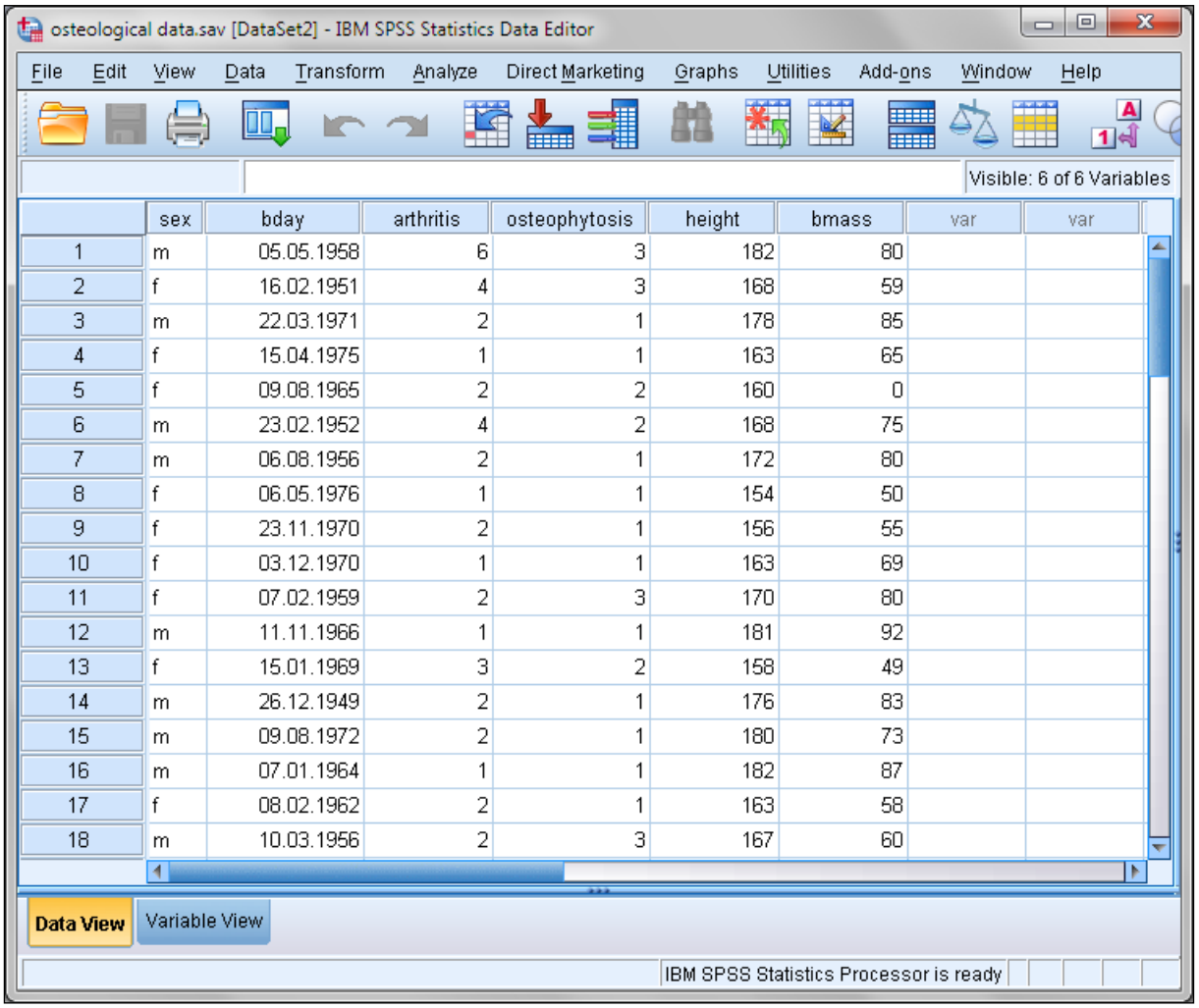
Κάτω από τη γραμμή μενού υπάρχει η **γραμμή εργαλείων** (toolbars), η οποία περιέχει με μορφή εικόνας ή σχήματος εντολές που ήδη βρίσκονται στη γραμμή μενού.

1.2 ΚΑΤΑΧΩΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟ ΦΥΛΛΟ ΕΡΓΑΣΙΑΣ

Ο απλούστερος τρόπος καταχώρησης δεδομένων σ' ένα φύλλο εργασίας είναι με απ' ευθείας πληκτρολόγηση των δεδομένων στο *Data View*. Ένα τέτοιο παράδειγμα δίνεται στο Σχήμα 1.5, όπου τα δεδομένα προέρχονται από σύγχρονη οστεολογική συλλογή από άτομα γνωστού φύλου και ηλικίας (documented

collection), είναι όμως απλοποιημένα για τις ανάγκες αυτού του βοηθήματος. Αφορούν τη σχέση φύλου, επιπέδων αρθρίτιδας στα χέρια (hand arthritis), επιπέδων οστεοφύτωσης στους οσφυϊκούς σπονδύλους (lumbar vertebrae osteophytosis), εκτιμώμενου ύψους και βάρους. Τα δεδομένα αυτά έχουν ληφθεί από το Παράρτημα Ι.

Επίσης μπορούμε να μεταφέρουμε δεδομένα από ένα φύλλο του Excel σε φύλλο του SPSS επιλέγοντας τα δεδομένα στο φύλλο του Excel, αντιγράφοντάς τα με Ctrl+C και επικολλώντας τα στο φύλλο του SPSS με Ctrl+V.



	sex	bday	arthritis	osteophytosis	height	bmass	var	var
1	m	05.05.1958	6	3	182	80		
2	f	16.02.1951	4	3	168	59		
3	m	22.03.1971	2	1	178	85		
4	f	15.04.1975	1	1	163	65		
5	f	09.08.1965	2	2	160	0		
6	m	23.02.1952	4	2	168	75		
7	m	06.08.1956	2	1	172	80		
8	f	06.05.1976	1	1	154	50		
9	f	23.11.1970	2	1	156	55		
10	f	03.12.1970	1	1	163	69		
11	f	07.02.1959	2	3	170	80		
12	m	11.11.1966	1	1	181	92		
13	f	15.01.1969	3	2	158	49		
14	m	26.12.1949	2	1	176	83		
15	m	09.08.1972	2	1	180	73		
16	m	07.01.1964	1	1	182	87		
17	f	08.02.1962	2	1	163	58		
18	m	10.03.1956	2	3	167	60		

Σχήμα 1.5. Τμήμα δεδομένων οστεολογικής μελέτης

1.3 ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ

Κάθε χαρακτηριστικό ενός πληθυσμού που μελετάμε ονομάζεται **μεταβλητή** (variable) και, όπως αναφέρθηκε, κάθε μεταβλητή εισάγεται σε μία ξεχωριστή στήλη του SPSS στο παράθυρο Data View. Οι μεταβλητές χωρίζονται σε δύο τύπος, τις **αριθμητικές** (numeric) και τις **αλφαριθμητικές** (string). Μία μεταβλητή είναι αριθμητική όταν οι τιμές της εκφράζονται με αριθμούς, ενώ στις αλφαριθμητικές εκφράζονται με χαρακτήρες, δηλαδή γράμματα του ελληνικού ή λατινικού αλφαβήτου, συνδυασμό γραμμάτων και αριθμών ή οποιοδήποτε άλλο σύμβολο. Επιπλέον χωρίζονται σε δύο βασικές κατηγορίες: Σε ποσοτικές (quantitative) και ποιοτικές ή κατηγορικές (qualitative/categorical) μεταβλητές.

Οι **ποσοτικές μεταβλητές** αντιστοιχούν σε μεγέθη που μπορούν να μετρηθούν, όπως το βάρος το μήκος, ο χρόνος, η θερμοκρασία, κτλ. Συνεπώς παίρνουν αριθμητικές τιμές και εκφράζονται με μια μονάδα μέτρησης, με την προϋπόθεση ότι υπάρχει μονάδα μέτρησης. Για παράδειγμα το βάρος έχει μονάδες μέτρησης (gr, kgr, κτλ), ενώ αντίθετα το pH δεν έχει. Στο SPSS οι ποσοτικές μεταβλητές ονομάζονται και **μεταβλητές κλίμακας** (scale). Γενικότερα, στη Στατιστική οι ποσοτικές μεταβλητές διακρίνονται σε μεταβλητές διαστήματος (interval) και σε μεταβλητές αναλογίας (ratio). Η μοναδική διαφοροποίηση ανάμεσα σε αυτές τις δύο κατηγορίες είναι ότι στις μεταβλητές διαστήματος το μηδέν ορίζεται συμβατικά με βάση κάποια κλίμακα και δεν εκφράζει την έλλειψη ποσότητας. Τυπική περίπτωση μεταβλητής διαστήματος είναι η θερμοκρασία, δεδομένου ότι θερμοκρασία 0 ^{o}C δεν σημαίνει απουσία θερμοκρασίας ή θερμότητας.

Όλες οι ποσοτικές μεταβλητές, ανάλογα με τις δυνατές τιμές που μπορούν να πάρουν, διακρίνονται σε **συνεχείς** (continuous) και σε **διακριτές** (discrete) ή **ασυνεχείς** (discontinuous) μεταβλητές. Οι συνεχείς μεταβλητές μπορούν να πάρουν οποιαδήποτε πραγματική τιμή, ενώ η διαφορά μεταξύ δύο δυνατών τιμών τους μπορεί να γίνει οσοδήποτε μικρή. Αντίθετα οι διακριτές μεταβλητές παίρνουν συγκεκριμένες τιμές, συνήθως ακέραιες, χωρίς να έχουν τη δυνατότητα να πάρουν μεταξύ αυτών των τιμών άλλες ενδιάμεσες.

Οι **ποιοτικές ή κατηγορικές μεταβλητές** δεν αντιστοιχούν σε μετρήσιμα μεγέθη αλλά εκφράζουν γενικά ποιοτικά χαρακτηριστικά του πληθυσμού. Στο SPSS και γενικότερα στη Στατιστική οι μεταβλητές αυτές χωρίζονται σε ονομαστικές (nominal) και σε σειριακές ή διατεταγμένες (ordinal).

Ονομαστικές (nominal) είναι οι τιμές μιας ποιοτικής μεταβλητής όταν δεν έχουν καμιά σειρά ή σχέση μεταξύ τους. Για παράδειγμα, μια μεταβλητή που δηλώνει το φύλο και παίρνει τις τιμές f (γυναίκα) και m (άνδρας) είναι ονομαστική. Θα μπορούσαμε αντί για f και m να χρησιμοποιούσαμε τους αριθμούς 1 και 2, αντίστοιχα. Και πάλι οι τιμές 1 και 2 θα ήταν ονομαστικές.

Σειριακές ή διατεταγμένες ή διατάξιμες (ordinal) είναι οι τιμές μιας ποιοτικής μεταβλητής όταν υποδηλώνουν μια σειριακή σχέση. Για παράδειγμα, στα οστεολογικά δεδομένα του Σχήματος 1.5 η αρθρίτιδα (arthritis) του χεριού και η οστεοφύτοση (osteophytosis) είναι ποιοτικές μεταβλητές τύπου ordinal. Η πρώτη μεταβλητή παίρνει τις τιμές 1, 2, 3, 4, 5, 6, ενώ η δεύτερη τις τιμές 1, 2 και 3. Και στις δύο περιπτώσεις η αύξουσα σειρά των αριθμών δηλώνει αντίστοιχη αύξουσα εκδήλωση της πάθησης.

Παρατήρηση 1. Η επιλογή των στατιστικών τεχνικών εξαρτάται κατά κύριο λόγο από τον τύπο των μεταβλητών που εξετάζονται.

Παρατήρηση 2. Για να εισάγουμε μια μεταβλητή που δηλώνεται με γράμματα ή με γράμματα και αριθμούς (αλφαριθμητική μεταβλητή-string) και όχι αποκλειστικά με αριθμούς σ' ένα φύλλο εργασίας, θα πρέπει πρώτα να ορίσουμε τον τύπο της στο *Variable View* (βλέπε παρακάτω).

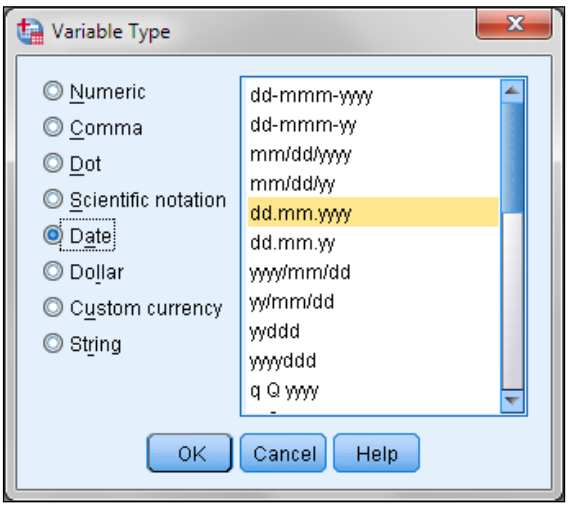
1.4 ΜΟΡΦΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

Μετά την εισαγωγή των δεδομένων, τα μορφοποιούμε από το παράθυρο *Data View* ως εξής:

- Κάνουμε κλικ στο **Variable View** και στην πρώτη στήλη (**Name**) πληκτρολογούμε τις επικεφαλίδες που θέλουμε να έχουν οι στήλες (Μεταβλητές) στο *Data View* (π.χ. sex, bday, arthritis, osteophytosis, height, bmass).

  Στη δεύτερη στήλη (**Type**) προσδιορίζουμε τον τύπο των μεταβλητών. Αν κάνουμε κλικ σε ένα κελί αυτής της στήλης, στα δεξιά του κελιού εμφανίζεται ένα μικρό ορθογώνιο. Με κλικ στο ορθογώνιο αυτό εμφανίζεται ένα παράθυρο διαλόγου που μας επιτρέπει να επιλέξουμε τον τύπο της μεταβλητής. Έχουμε τις ακόλουθες επιλογές: *Numeric, Comma, Dot, Scientific notation, Date, Dollar, Custom currency* και *String*. Παρατηρούμε ότι το SPSS εκτός από τους

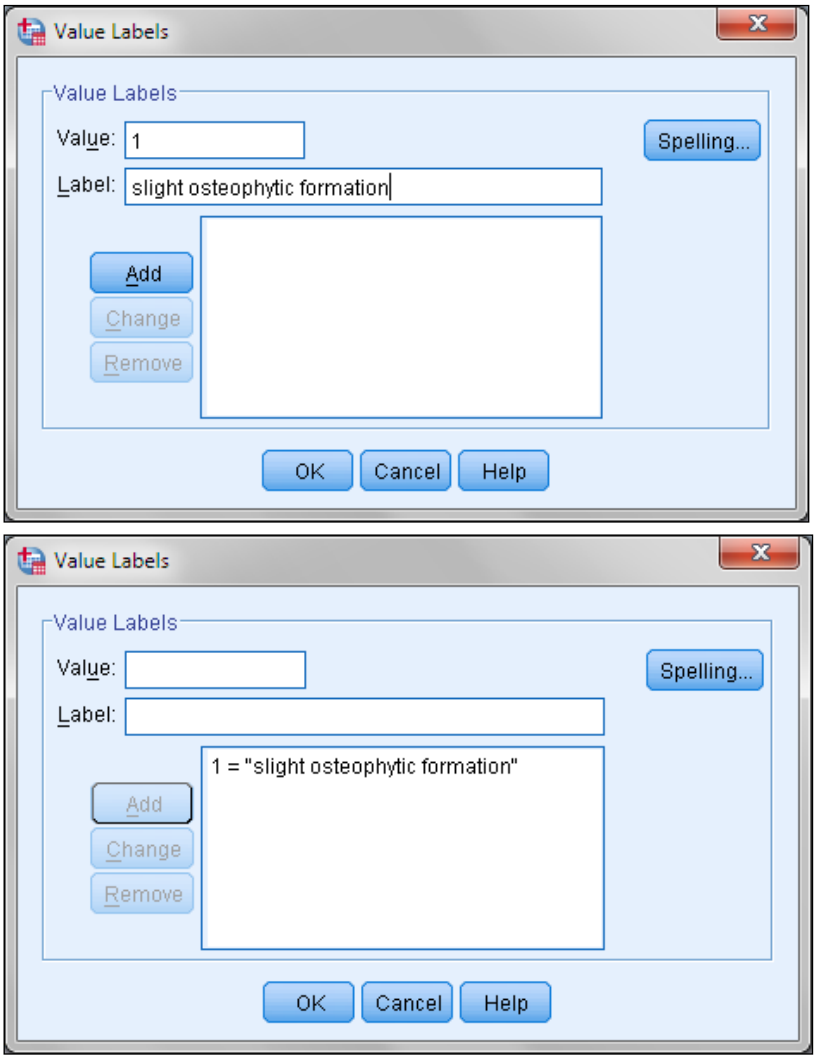
τύπους *Numeric* και *String* χρησιμοποιεί 6 επιπλέον τύπους μεταβλητών. **Comma** είναι μια αριθμητική μεταβλητή όταν οι χιλιάδες προσδιορίζονται με κόμμα ενώ τα δεκαδικά με τελεία, π.χ. 5,012.6. **Dot** είναι μια αριθμητική μεταβλητή όταν οι χιλιάδες προσδιορίζονται με τελεία και τα δεκαδικά με κόμμα, π.χ. 5.012,6. Το **scientific notation** δηλώνει ότι θα χρησιμοποιηθεί επιστημονική παρουσίαση της αριθμητικής μεταβλητής, π.χ. 9.12E2 αντί για 912 ή 9.12E-2 αντί για 0.0912. Το **Date** χρησιμοποιείται για να εισάγουμε ημερομηνίες. Στο παράδειγμα που εξετάζουμε, όταν επιλέγουμε το **Date** για τη μορφοποίηση των ημερομηνιών της μεταβλητής birthday, στη συνέχεια επιλέγουμε τη μορφή dd.mm.yyy, όπως φαίνεται στο Σχήμα 1.6. Το **Dollar** χρησιμοποιείται όταν αναφερόμαστε σε νομίσματα δολαρίου. Τέλος, σε περίπτωση άλλων νομισμάτων χρησιμοποιούμε το **Custom currency**.



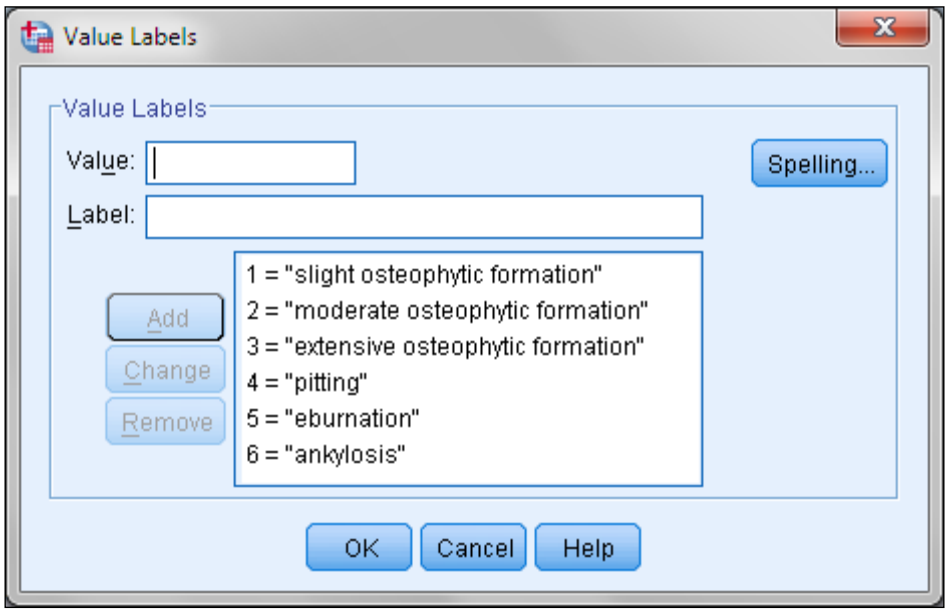
Σχήμα 1.6. Μορφοποίηση ημερομηνιών

- Στη στήλη ***Width*** καθορίζουμε πόσα γράμματα μπορεί να έχει το όνομα της μεταβλητής.
- Στη στήλη ***Decimals*** καθορίζεται ο αριθμός των δεκαδικών ψηφίων των αριθμητικών μεταβλητών.
- Στη στήλη ***Label*** (Ετικέτες) μπορούμε να δώσουμε μια σύντομη περιγραφή της κάθε μεταβλητής.
- Στη στήλη ***Values*** δίνουμε πληροφορίες για τις τιμές της μεταβλητής όταν αυτή είναι κατηγορική. Η προεπιλογή είναι *None* και αφορά κυρίως τις ποσοτικές μεταβλητές. Έστω όμως για παράδειγμα μια μεταβλητή, η arthritis,

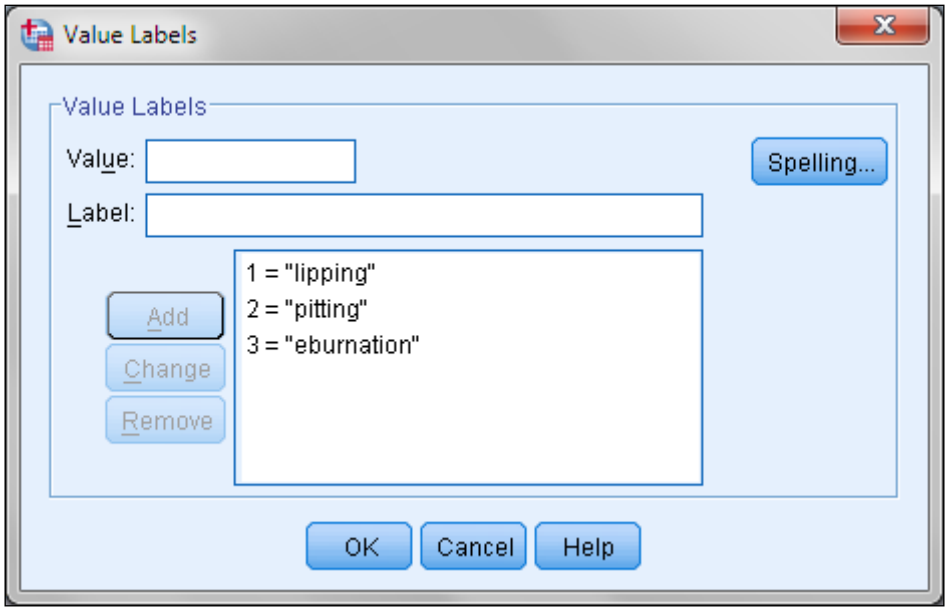
η οποία παίρνει τις τιμές 1, 2, 3, 4, 5 και 6 ανάλογα με το επίπεδο της ασθένειας, όπως διευκρινίζεται στον πίνακα του Παραρτήματος. Σ' αυτή την περίπτωση, για δική μας πληροφόρηση, κάνουμε κλικ στο *Values* που αντιστοιχεί στη μεταβλητή arthritis και κλικ στο μικρό ορθογώνιο, οπότε ανοίγει το παράθυρο διαλόγου του Σχήματος 1.7-άνω. Στο πλαίσιο *Value* πληκτρολογούμε 1, στο *Label* πληκτρολογούμε slight osteophytic formation και κάνουμε κλικ στο *Add*. Η έκφραση 1 = "slight osteophytic formation" εισέρχεται στο μεγάλο ορθογώνιο πλαίσιο (Σχήμα 1.7-κάτω). Συνεχίζουμε εισάγοντας την τιμή 2 στο *Value*, τις λέξεις moderate osteophytic formation στο *Label* και πάλι κλικ στο *Add*. Με αυτόν τον τρόπο στο τέλος θα πάρουμε την εικόνα του Σχήματος 1.8. Με ανάλογο τρόπο ορίζουμε τις τιμές της μεταβλητής osteophytosis (Σχήμα 1.9).



Σχήμα 1.7. Βήματα συμπλήρωσης του πλαισίου διαλόγου Value Labels για τη μεταβλητή arthritis



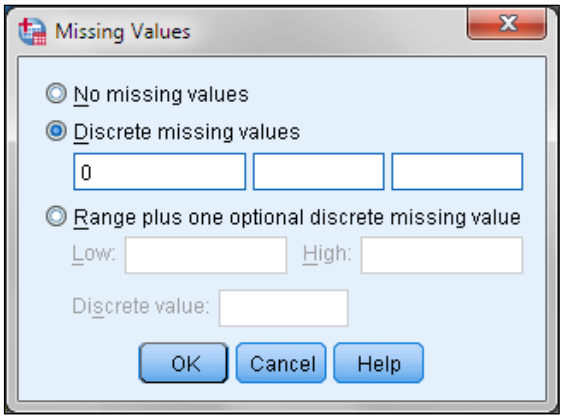
Σχήμα 1.8. Το παράθυρο διαλόγου *Value Labels* για τη μεταβλητή arthritis



Σχήμα 1.9. Το παράθυρο διαλόγου *Value Labels* για τη μεταβλητή osteophytosis

- Στο *SPSS* δεν επιτρέπεται να υπάρχουν κενά κελιά. Γι' αυτό χρησιμοποιούμε μια συγκεκριμένη τιμή (στο παράδειγμά μας την τιμή 0) για να δηλώσουμε μία απούσα τιμή. Στο παράδειγμά μας θα χρησιμοποιήσουμε την τιμή 0 για τις απούσες τιμές της μεταβλητής bmass επειδή αυτή η τιμή δεν μπορεί να υπάρχει στις φυσιολογικές τιμές της bmass. Οι απούσες τιμές εισάγονται στη

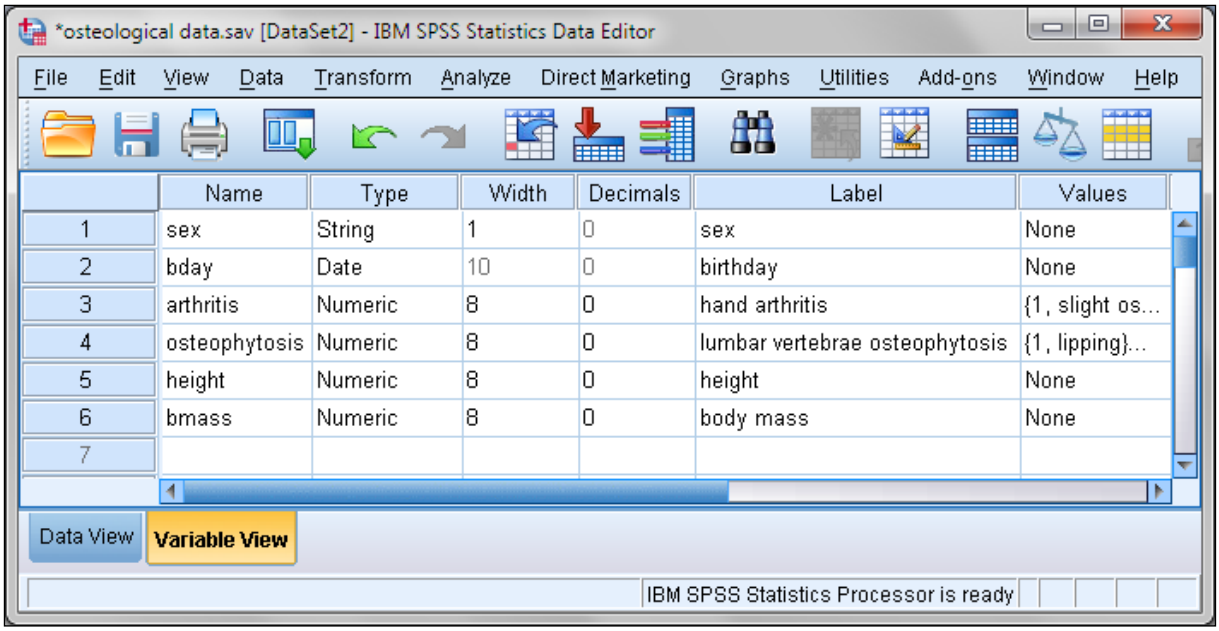
στήλη ***Missing*** *(Values)* για χρήση στη στατιστική επεξεργασία που θα επιλέξουμε. Αυτό γίνεται αν κάνουμε κλικ στο μικρό ορθογώνιο που εμφανίζεται στα κελιά αυτής της στήλης. Τότε ανοίγει το παράθυρο διαλόγου του Σχήματος 1.10. Τα τρία πλαίσια που υπάρχουν κάτω από το *Discrete missing values* δείχνουν ότι μπορούμε να χρησιμοποιήσουμε μέχρι και τρεις διαφορετικές τιμές για να δηλώσουμε απούσες τιμές. Επίσης, μπορούμε να ορίσουμε ένα εύρος απουσών τιμών, π.χ. όλες οι τιμές από 0 έως -10 να δηλώνουν πως τα κελιά είναι κενά, καθώς επίσης ένα εύρος τιμών και μία επιπλέον τιμή (π.χ. 0 έως -10, 333).



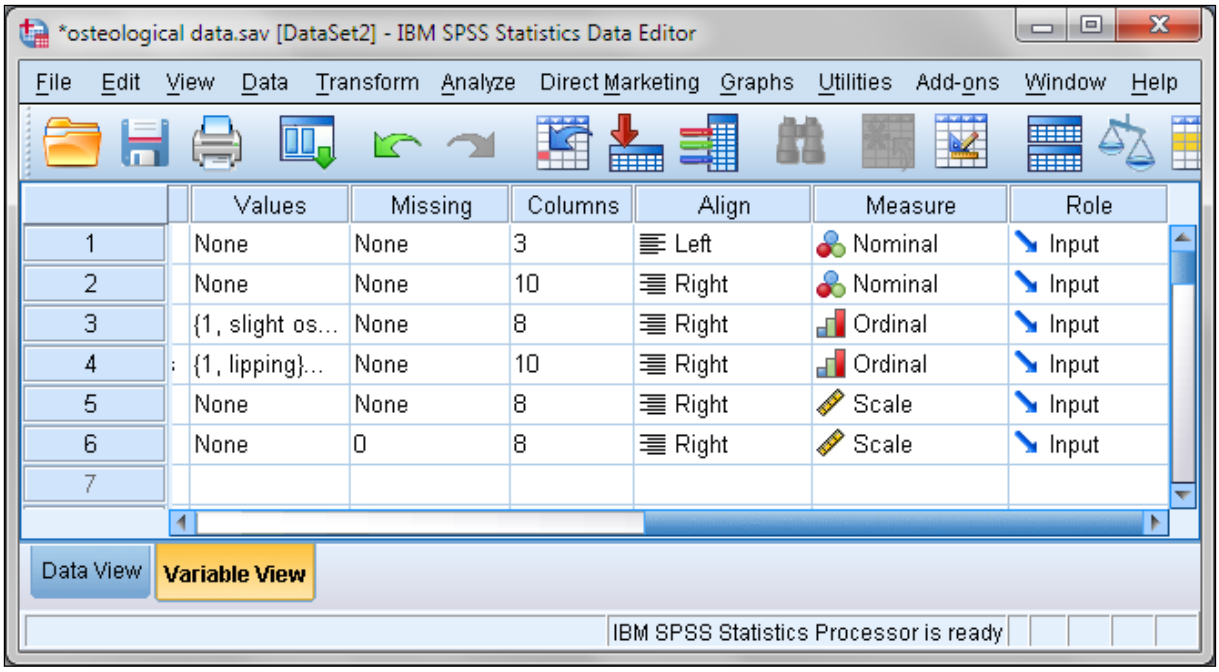
Σχήμα 1.10. Το παράθυρο διαλόγου *Missing Values*

- Στη στήλη ***Columns*** καθορίζουμε το πλάτος που θα έχει η στήλη μιας μεταβλητής (πόσα ψηφία μπορεί να πάρει η μεταβλητή).
- Η στήλη ***Align*** καθορίζει τη στοίχιση των τιμών μιας μεταβλητής στη στήλη της με επιλογές *Left* (αριστερά), *Right* (δεξιά) και *Center* (κέντρο).
- Η στήλη ***Measure*** καθορίζει αν μια μεταβλητή είναι ποσοτική (*Scale*) ονομαστική (*Nominal*) ή σειριακή (*Ordinal*).

Η μορφοποίηση των δεδομένων του Σχήματος 1.5 δίνεται στα Σχήματα 1.11 και 1.12.



Σχήμα 1.11. Πρώτο τμήμα μορφοποίησης των δεδομένων του Σχήματος 1.5



Σχήμα 1.12. Δεύτερο τμήμα μορφοποίησης των δεδομένων του Σχήματος 1.5

Παρατήρηση: Στο SPSS δεν μπορούμε να χρησιμοποιήσουμε ελληνικούς χαρακτήρες.

1.5 ΠΡΟΣΘΗΚΗ Ή ΔΙΑΓΡΑΦΗ ΝΕΩΝ ΠΕΡΙΠΤΩΣΕΩΝ ΚΑΙ ΜΕΤΑΒΛΗΤΩΝ

Προκειμένου να εισάγουμε μια νέα περίπτωση (γραμμή) μεταξύ περιπτώσεων που ήδη υπάρχουν, επιλέγουμε ένα οποιοδήποτε κελί της γραμμής που βρίσκεται ***κάτω*** από τη θέση όπου θέλουμε να εισάγουμε τη νέα γραμμή και από το *Edit* επιλέγουμε *Insert Case*. Εναλλακτικά κάνουμε κλικ στον αριθμό της γραμμής που βρίσκεται κάτω από τη θέση όπου θα εισάγουμε τη νέα γραμμή και με δεξί κλικ επιλέγουμε *Insert Case*.

Για να εισάγουμε μια νέα μεταβλητή (στήλη) μεταξύ μεταβλητών που ήδη υπάρχουν, κάνουμε κλικ σ' ένα οποιοδήποτε κελί της στήλης που βρίσκεται ***δεξιά*** από τη θέση όπου θέλουμε να εισάγουμε τη νέα στήλη και από το *Edit* επιλέγουμε *Insert Variable*. Εναλλακτικά επιλέγουμε τη στήλη που βρίσκεται δεξιά από τη θέση όπου θα εισάγουμε τη νέα στήλη και με δεξί κλικ επιλέγουμε *Insert Variable*.

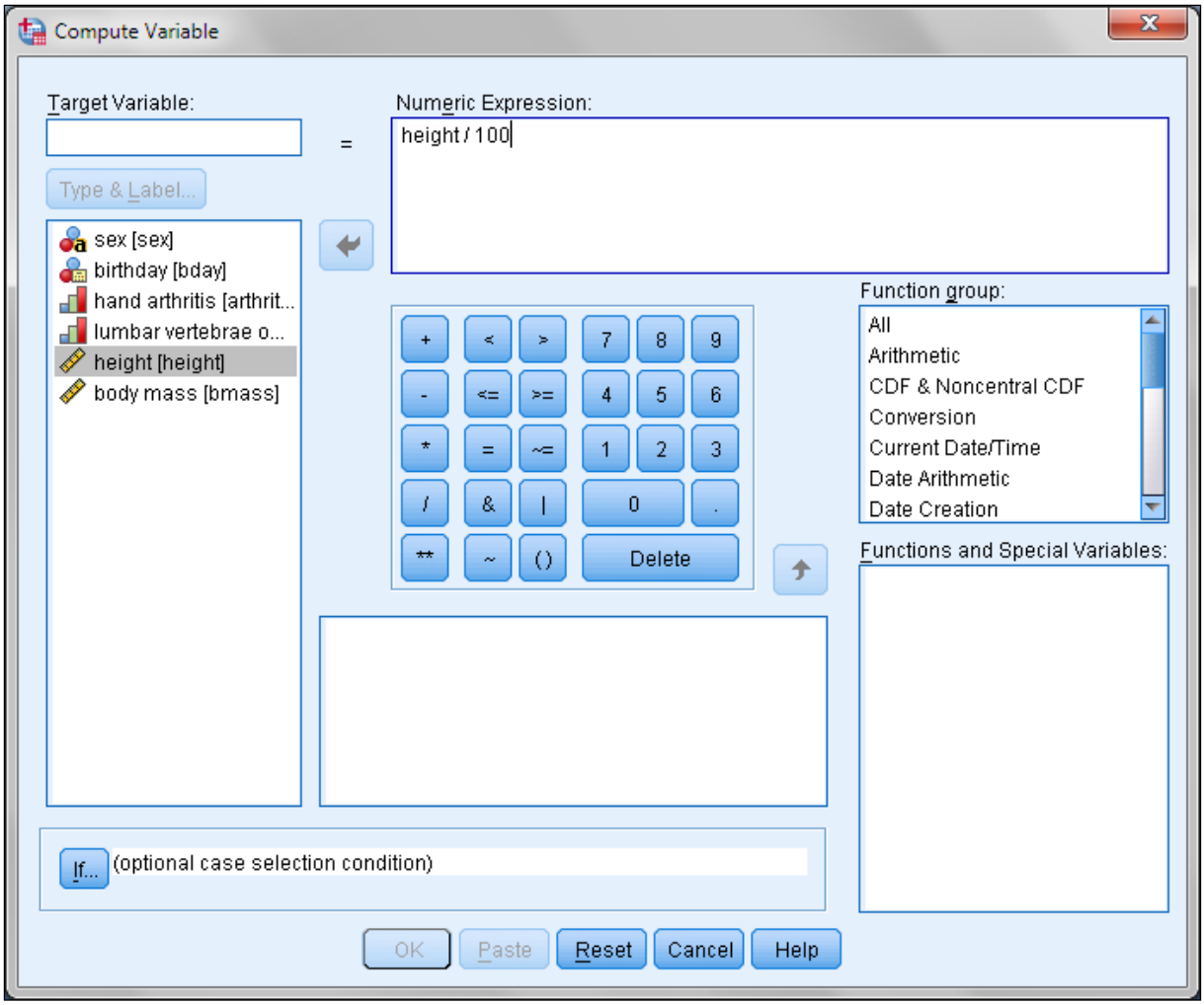
Αν η εισαγωγή της γραμμής ή στήλης έχει γίνει σε λάθος θέση, αναιρούμε την ενέργεια με *Edit* → *Undo*.

Για να διαγράψουμε μια περίπτωση (γραμμή) ή μια μεταβλητή (στήλη), επιλέγουμε τη γραμμή ή τη στήλη αυτή και με δεξί κλικ επιλέγουμε *Clear*.

1.6 ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ ΜΕΤΑΒΛΗΤΩΝ

Έστω, για παράδειγμα, ότι θέλουμε να σχηματίσουμε μια στήλη με το εκτιμώμενο ύψος σε μέτρα (m) αντί για εκατοστά (cm). Από το *Transform* → *Compute Variable* ανοίγουμε το παράθυρο διαλόγου του Σχήματος 1.13, κάνουμε κλικ στο *height (height)*, κλικ στο βέλος ▶, οπότε η μεταβλητή height εισέρχεται στο πλαίσιο *Numeric Expression*, και συνεχίζουμε πληκτρολογώντας / και 100. Στο *Target Variable* εισάγουμε τον τίτλο, έστω heightm, και στο *Type & Label* αν θέλουμε γράφουμε πιο αναλυτικά height in m. Με κλικ στο *OK* σχηματίζεται η ζητούμενη νέα μεταβλητή δίπλα στην τελευταία στήλη.

Παρατήρηση. Αν και το πλαίσιο *Compute Variable* μας δίνει πολλές δυνατότητες για να δημιουργήσουμε μια νέα μεταβλητή με βάση τις ήδη υπάρχουσες μεταβλητές, είναι πολύ πιο εύκολο αυτό να το κάνουμε σε ένα φύλλο του Excel και ακολούθως να μεταφέρουμε τη νέα μεταβλητή στο SPSS.

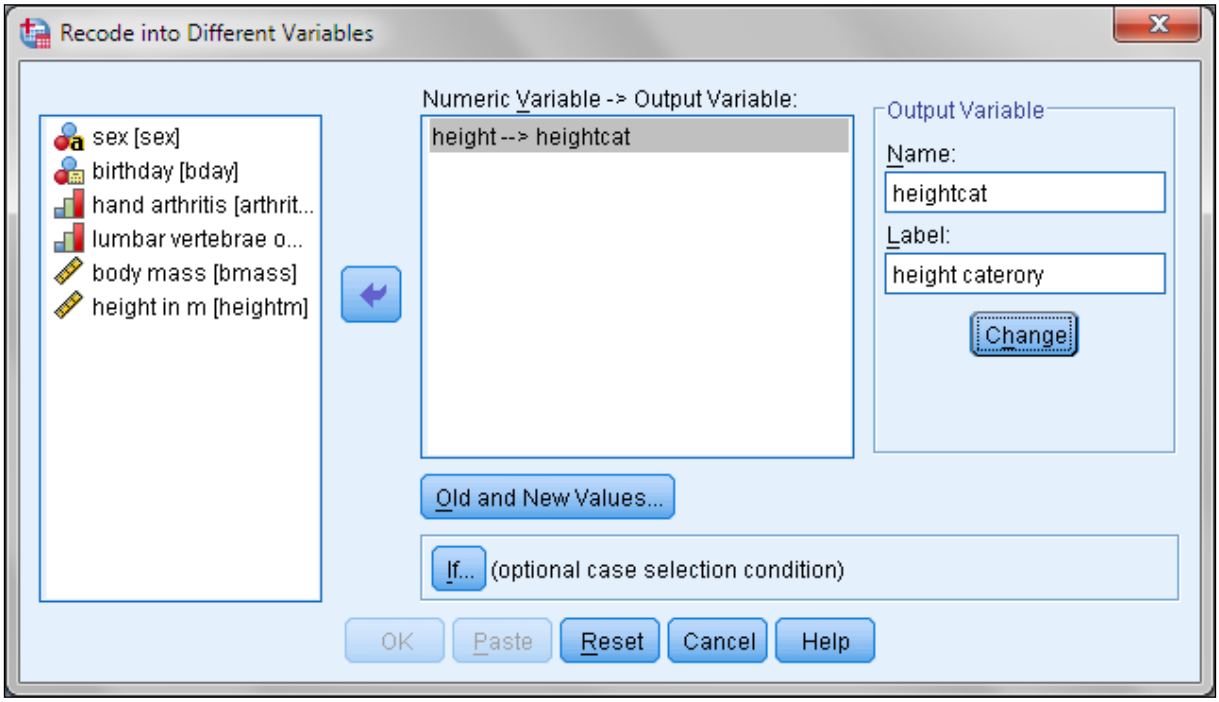


Σχήμα 1.13. Το παράθυρο διαλόγου *Compute Variable*

1.7 ΟΜΑΔΟΠΟΙΗΣΗ ΤΙΜΩΝ ΜΙΑΣ ΣΥΝΕΧΟΥΣ ΜΕΤΑΒΛΗΤΗΣ

Έστω ότι θέλουμε να χωρίσουμε τα ύψη (μεταβλητή height) σε τέσσερεις κατηγορίες: μικρότερα από 160 cm, μεταξύ 161 – 170 cm, μεταξύ 171 – 180 cm και μεγαλύτερα από 181 cm. Στις κατηγορίες αυτές αποδίδουμε τις τιμές 1, 2, 3 και 4, αντίστοιχα. Για να κάνουμε την κωδικοποίηση τιμών ακολουθούμε την παρακάτω πορεία:

Από το *Transform* → *Recode Into Different Variables* ανοίγουμε το παράθυρο *Recode into Different Variables*. Σ' αυτό επιλέγουμε τη μεταβλητή height και την εισάγουμε στο πλαίσιο *Numeric Variable* → *Output Variable* με κλικ στο ▶. Στο πλαίσιο *Name* πληκτρολογούμε το όνομα της μεταβλητής, έστω heightcat, και στο πλαίσιο *Label* πληκτρολογούμε μια ετικέτα γι' αυτήν, έστω height category. Με κλικ στο *Change* παίρνουμε την εικόνα του Σχήματος 1.14.



Σχήμα 1.14. Το παράθυρο διαλόγου *Recode into Different Variables*

Στο σημείο αυτό κάνουμε κλικ στο *Old and New Values,* οπότε ανοίγει ένα νέο παράθυρο διαλόγου, το *Recode into Different Variables: Old and New Values*. Συμπληρώνουμε το παράθυρο αυτό ακολουθώντας τα ακόλουθα βήματα:

1. Κάνουμε κλικ στο *Range, LOWEST through value* και στο πλαίσιο που υπάρχει από κάτω εισάγουμε τον αριθμό 160. Ακολούθως στο πλαίσιο *New Value* πληκτρολογούμε 1, οπότε και ενεργοποιείται το κουμπί *Add*. Με κλικ σ' αυτό εισέρχεται στο πλαίσιο *Old → New* η *έκφραση Lowest thru 160 → 1*.

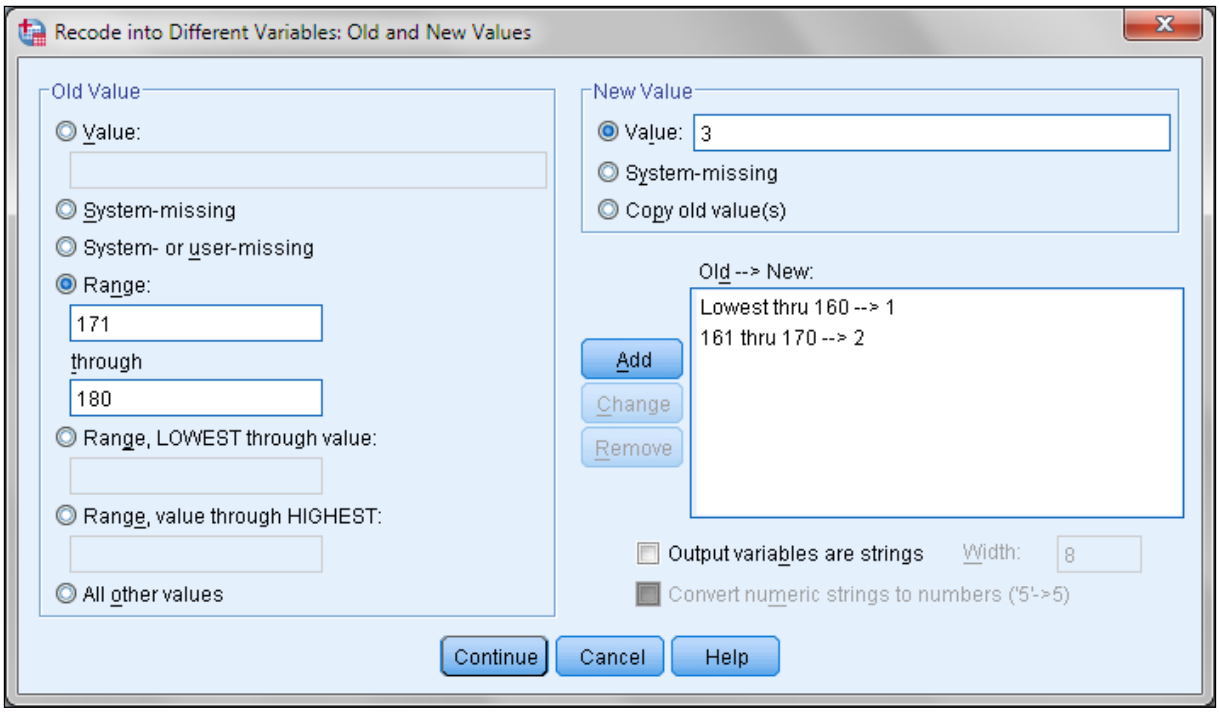
2. Συνεχίζουμε με κλικ στο *Range* και εισάγουμε τους αριθμούς 161 στο πάνω και 170 στο κάτω πλαίσιο κάτω από τη λέξη *Range*. Στο πλαίσιο *New Value* πληκτρολογούμε 2 και κάνουμε κλικ στο *Add*.

3. Επαναλαμβάνουμε το βήμα 2 με 171 και 180 αντί για 161 και 170, αντίστοιχα. Επίσης στο πλαίσιο *New Value* πληκτρολογούμε 3 και κάνουμε κλικ στο *Add* (Σχήμα 1.15).

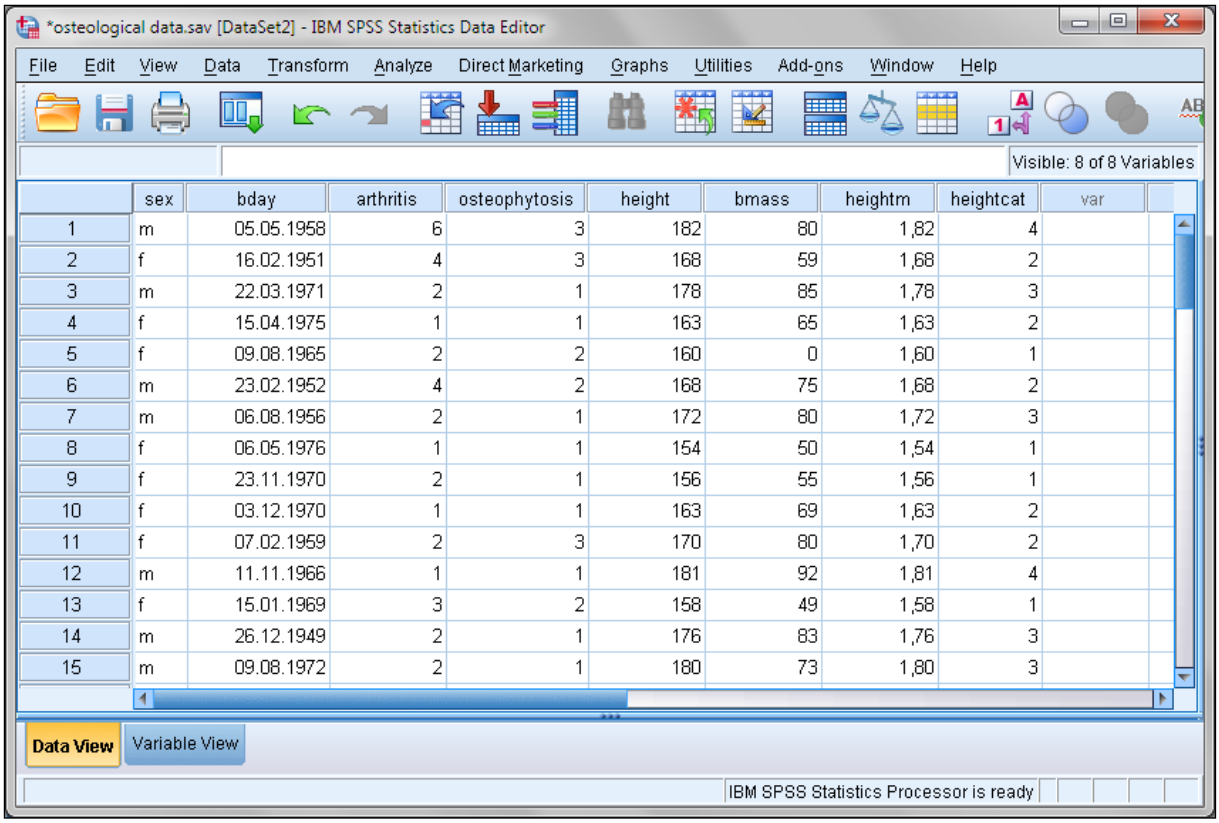
4. Κάνουμε κλικ στο *Range, value through HIGHEST* και στο πλαίσιο που υπάρχει από κάτω εισάγουμε τον αριθμό 181. Στο *New Value* πληκτρολογούμε 4 και κάνουμε κλικ στο *Add*.

5. Ολοκληρώνουμε τη διαδικασία με κλικ στο *Continue* και στο *OK*. Στη στήλη heightcat εμφανίζεται η νέα κωδικοποίηση (Σχήμα 1.16).

Παρατήρηση. Αν χρησιμοποιούσαμε την επιλογή *Into Same Variable* αντί για *Into Different Variable*, θα γινόταν αντικατάσταση της αρχικής μεταβλητής salary από τη νέα, heightcat. Γι αυτό χρειάζεται προσοχή στην επιλογή.



Σχήμα 1.15. Συμπλήρωση του *Recode into Different Variables: Old and New Values*



	sex	bday	arthritis	osteophytosis	height	bmass	heightm	heightcat	var
1	m	05.05.1958	6	3	182	80	1,82	4	
2	f	16.02.1951	4	3	168	59	1,68	2	
3	m	22.03.1971	2	1	178	85	1,78	3	
4	f	15.04.1975	1	1	163	65	1,63	2	
5	f	09.08.1965	2	2	160	0	1,60	1	
6	m	23.02.1952	4	2	168	75	1,68	2	
7	m	06.08.1956	2	1	172	80	1,72	3	
8	f	06.05.1976	1	1	154	50	1,54	1	
9	f	23.11.1970	2	1	156	55	1,56	1	
10	f	03.12.1970	1	1	163	69	1,63	2	
11	f	07.02.1959	2	3	170	80	1,70	2	
12	m	11.11.1966	1	1	181	92	1,81	4	
13	f	15.01.1969	3	2	158	49	1,58	1	
14	m	26.12.1949	2	1	176	83	1,76	3	
15	m	09.08.1972	2	1	180	73	1,80	3	

Σχήμα 1.16. Η μεταβλητή heightcat

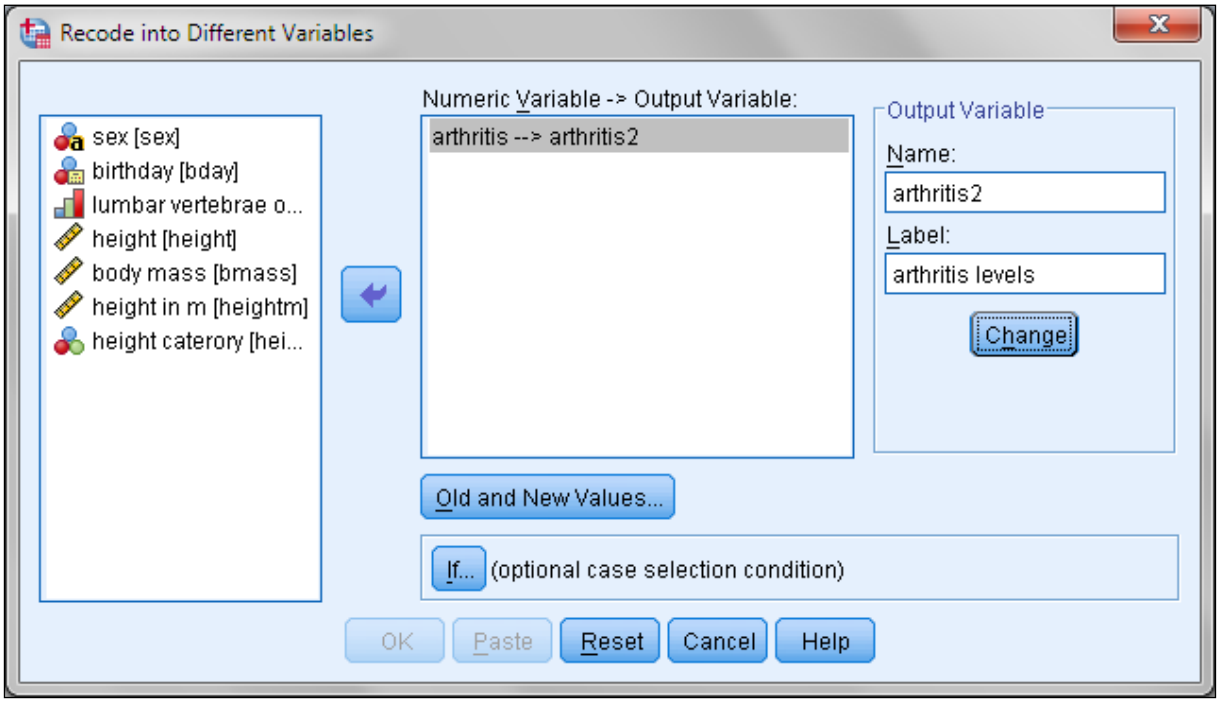
1.8 ΕΠΑΝΑΚΩΔΙΚΟΠΟΙΗΣΗ ΤΙΜΩΝ

Στο παράδειγμα που εξετάζουμε, η μεταβλητή arthritis είναι ordinal και παίρνει τις τιμές 1, 2, 3, 4, 5, 6 που υποδηλώνουν το επίπεδο ασθένειας από το 1 (slight osteophytic formation) έως το 6 (ankylosis). Έστω τώρα ότι θέλουμε να περιορίσουμε τα επίπεδα εκδήλωσης της πάθησης σε τέσσερα, ενοποιώντας τα επίπεδα 1 και 2, 3 και 4, και 6 και 7. Δηλαδή πρέπει να κάνουμε τις εξής αλλαγές:

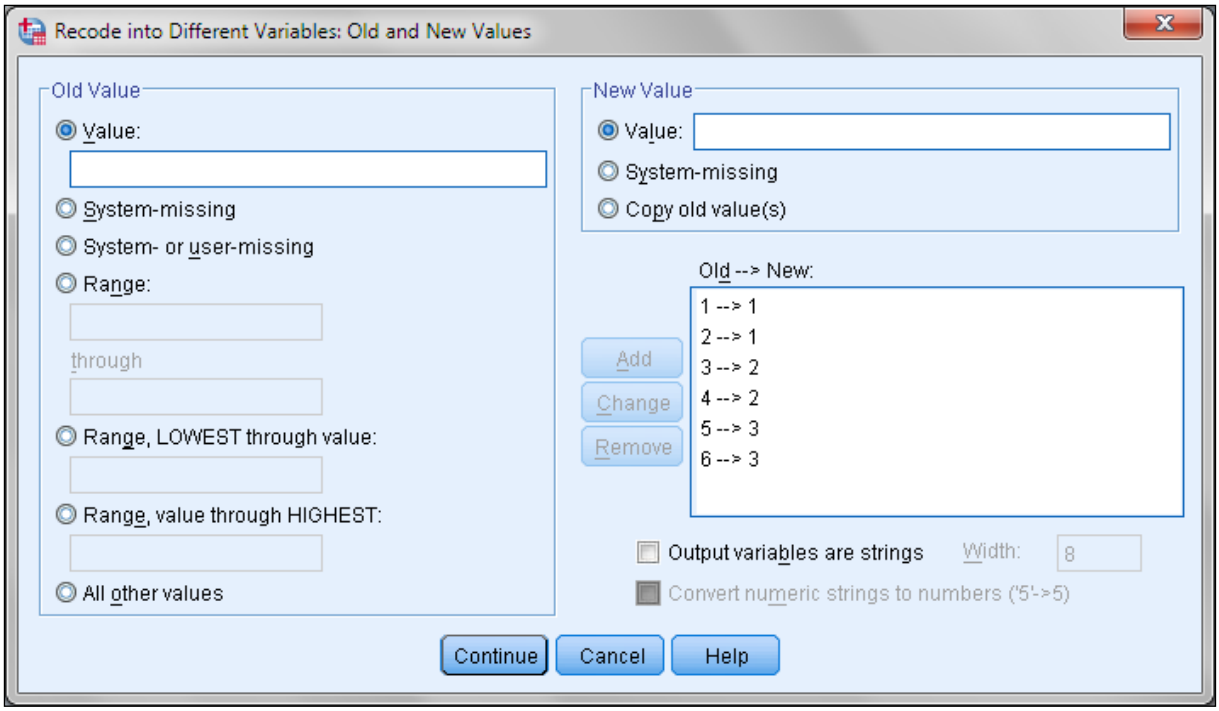
Παλιά τιμή	Νέα τιμή	Περιγραφή
1	1	Low
2	1	Low
3	2	Moderate
4	2	Moderate
5	3	High
6	3	High

Για να πετύχουμε αυτήν την επανακωδικοποίηση των τιμών ακολουθούμε πορεία ανάλογη με την προηγούμενη. Από το *Transform* → *Recode Into Different Variables* ανοίγουμε το παράθυρο *Recode into Different Variables* και κάνουμε κλικ στο Reset. Ακολούθως επιλέγουμε τη μεταβλητή arthritis, την οποία εισάγουμε στο πλαίσιο *Numeric Variable* → *Output Variable* με κλικ στο ▶. Στο πλαίσιο *Name* πληκτρολογούμε το όνομα της νέας μεταβλητής, έστω arthritis2, και στο πλαίσιο *Label* πληκτρολογούμε μια ετικέτα γι' αυτήν, έστω arthritis levels. Με κλικ στο *Change* παίρνουμε την εικόνα του Σχήματος 1.17.

Κάνουμε κλικ στο *Old and New Values* και συμπληρώνουμε το παράθυρο διαλόγου *Recode into Different Variables: Old and New Values* ως εξής: Στο *Old Value* εισάγουμε την παλιά τιμή, στο *New Value* τη νέα και κάνουμε κλικ στο *Add*. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να πάρουμε την εικόνα του Σχήματος 1.18. Τότε κάνουμε κλικ στο *Continue* και στο *OK*, οπότε στη στήλη arthritis2 εμφανίζεται η νέα κωδικοποίηση. Καλό είναι η στήλη αυτή να μορφοποιηθεί κατάλληλα και στο *Value Labels* να καταχωρηθούν οι ετικέτες της νέας μεταβλητής.



Σχήμα 1.17. Το παράθυρο διαλόγου *Recode into Different Variables*



Σχήμα 1.18. Επανακωδικοποίηση των τιμών της arthritis στο παράθυρο διαλόγου *Recode into Different Variables: Old and New Values*

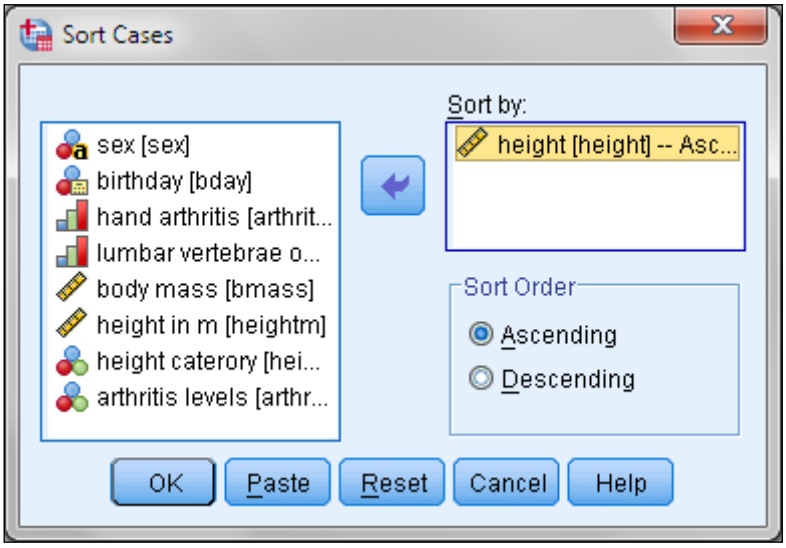
Παρατήρηση. Όταν χρησιμοποιούμε επανειλημμένα το παράθυρο *Recode into Different Variables* είναι απαραίτητο να αφαιρέσουμε την παλιά κωδικοποίηση από το πλαίσιο *Old → New*. Αυτό γίνεται αν επιλέξουμε το στοιχείο που θέλουμε να αφαιρέσουμε και κάνουμε κλικ στο *Remove* ή με κλικ στο *Reset*.

1.9 ΤΑΞΙΝΟΜΗΣΗ ΤΩΝ ΤΙΜΩΝ ΜΙΑΣ ΜΕΤΑΒΛΗΤΗΣ

Η ταξινόμηση των τιμών μιας μεταβλητής γίνεται εύκολα από το *Data → Sort Cases*. Στο παράθυρο *Sort Cases* που ανοίγει κάνουμε κλικ στη μεταβλητή της οποίας οι τιμές θα ταξινομηθούν, κλικ στο βέλος ▶ και επιλέγουμε αν η ταξινόμηση θα γίνει με αύξουσα (*Ascending*) ή φθίνουσα (*Descending*) σειρά (Σχήμα 1.19). Ολοκληρώνουμε με κλικ στο *OK*. Η ταξινόμηση των τιμών μιας μεταβλητής οδηγεί σε αντίστοιχες ανακατατάξεις στις τιμές όλων των μεταβλητών, δεδομένου ότι όλες οι περιπτώσεις (γραμμές) ανακατατάσσονται ακολουθώντας την ταξινόμηση της μεταβλητής που έχουμε επιλέξει.

Εναλλακτικά η ταξινόμηση των τιμών μιας μεταβλητής μπορεί να γίνει αν κάνουμε κλικ στο όνομα της μεταβλητής, οπότε και επιλέγεται όλη η στήλη της

μεταβλητής, και με δεξί κλικ επιλέξουμε *Sort Ascending* ή *Sort Descending*.

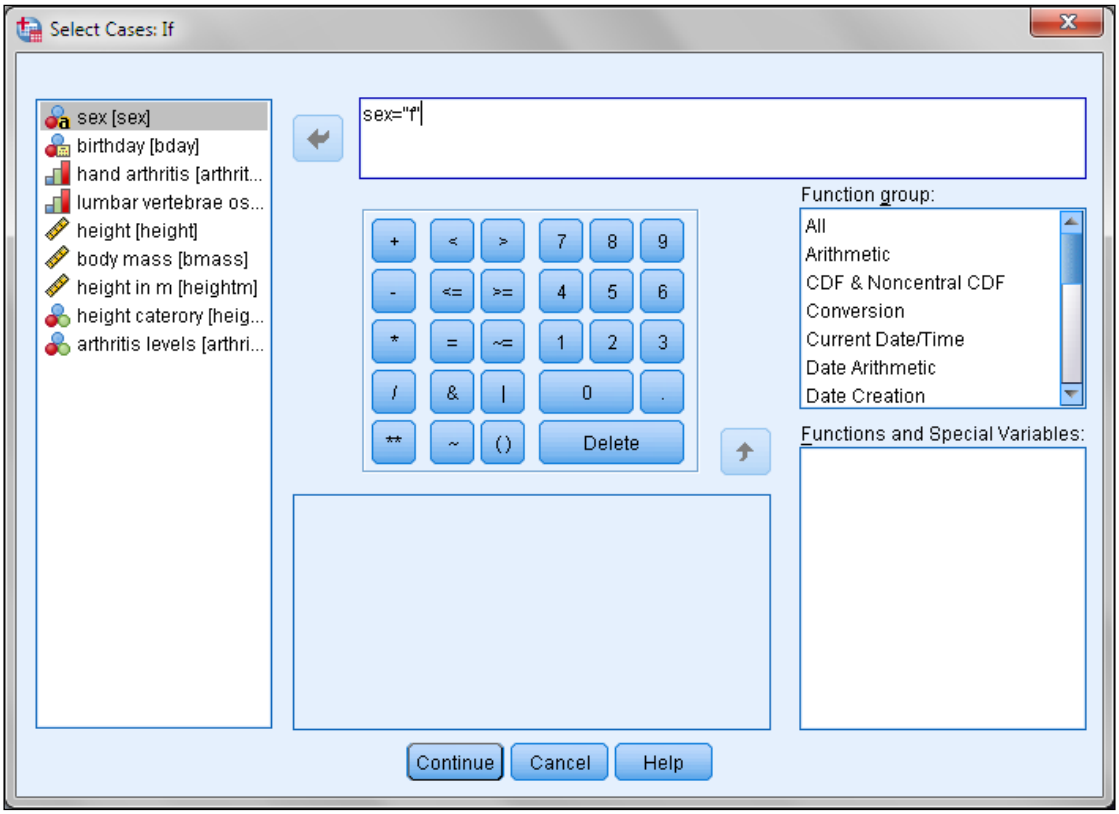


Σχήμα 1.19. Το παράθυρο διαλόγου *Sort Cases*

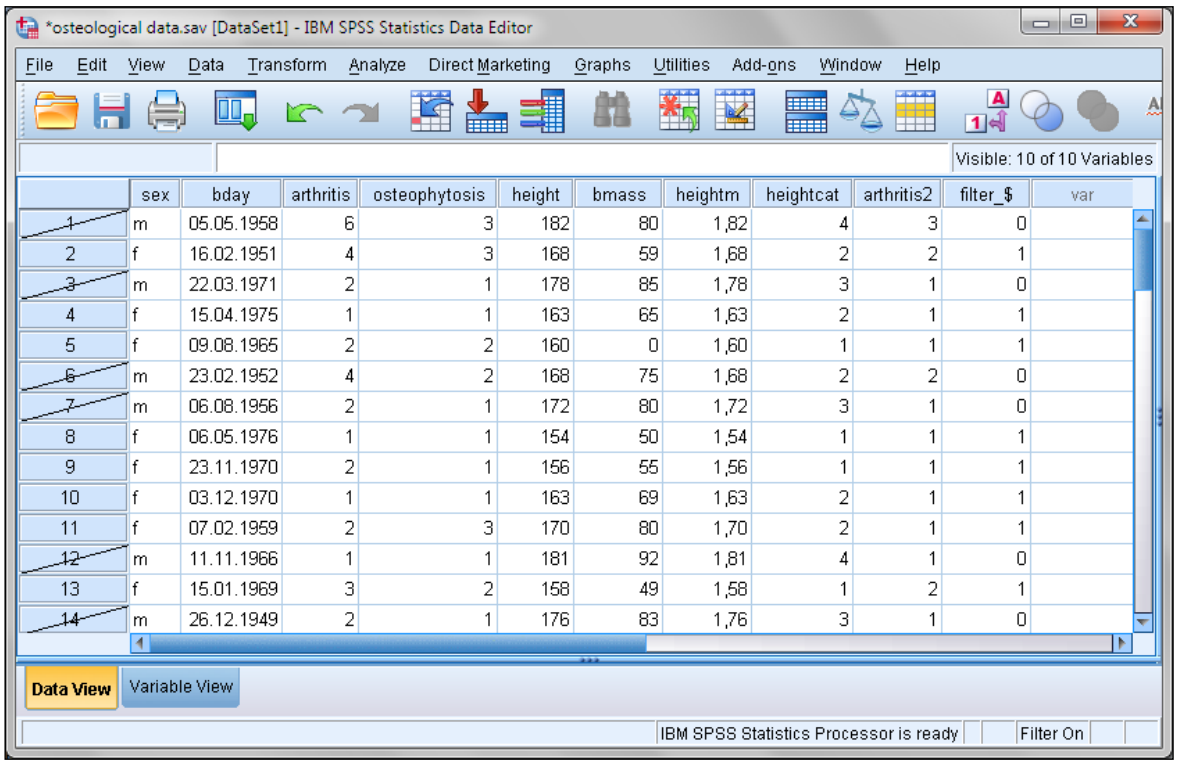
1.10 ΕΠΙΛΟΓΗ ΠΕΡΙΠΤΩΣΕΩΝ

Η διαδικασία αυτή εφαρμόζεται όταν θέλουμε να μελετήσουμε ένα υποσύνολο των δεδομένων μας. Για παράδειγμα, έστω ότι θέλουμε να μελετήσουμε τα στατιστικά στοιχεία μόνο των γυναικών στα δεδομένα του Σχήματος 1.5. Τότε από το *Data* → *Select Cases* ανοίγουμε το αντίστοιχο παράθυρο διαλόγου και κάνουμε κλικ στο *If condition is satisfied* και κλικ *στο If…* Στο νέο παράθυρο που ανοίγει επιλέγουμε τη μεταβλητή sex και τη μεταφέρουμε στο πλαίσιο που βρίσκεται δεξιά με κλικ στο βέλος ▶. Συνεχίζουμε με κλικ στο κουμπί = και πληκτρολογούμε "f". Θα πάρουμε την εικόνα του Σχήματος 1.20.

Με κλικ στο *Continue* και στο *OK* το πρόγραμμα επιλέγει μόνο τις περιπτώσεις όπου sex = "f". Αυτό φαίνεται από το ότι στο αρχικό αρχείο οι περιπτώσεις sex = "m" έχουν διαγραφεί, όπως φαίνεται από τη διαγραφή των αριθμών των περιπτώσεων sex = "m" στο Σχήμα 1.21. Επιπλέον, έχει προστεθεί μια τελευταία στήλη με όνομα filter_$ και τιμές 0 όταν sex = "m" και 1 όταν sex = "f". Από αυτό το σημείο κι έπειτα, αν ζητήσουμε μια οποιαδήποτε στατιστική επεξεργασία, αυτή περιορίζεται μόνο στις περιπτώσεις όπου sex = "f".



Σχήμα 1.20. Το παράθυρο διαλόγου *Select Cases: If*



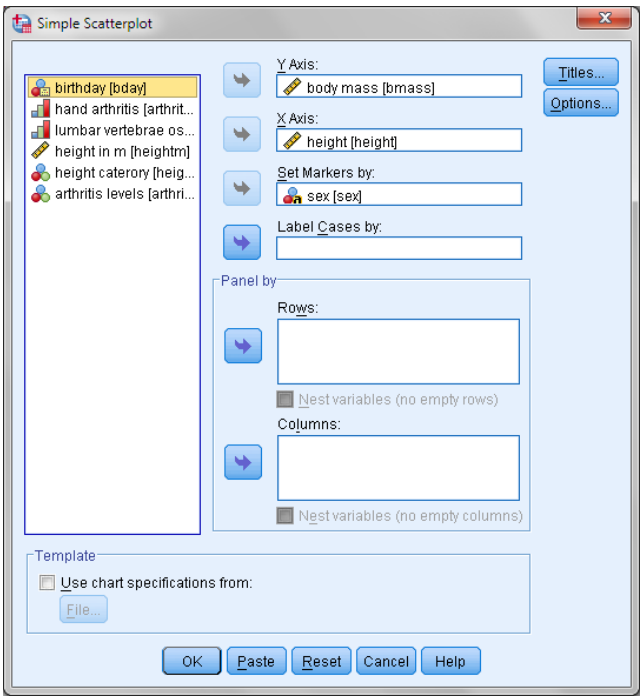
	sex	bday	arthritis	osteophytosis	height	bmass	heightm	heightcat	arthritis2	filter_$	var
1	m	05.05.1958	6	3	182	80	1,82	4	3	0	
2	f	16.02.1951	4	3	168	59	1,68	2	2	1	
3	m	22.03.1971	2	1	178	85	1,78	3	1	0	
4	f	15.04.1975	1	1	163	65	1,63	2	1	1	
5	f	09.08.1965	2	2	160	0	1,60	1	1	1	
6	m	23.02.1952	4	2	168	75	1,68	2	2	0	
7	m	06.08.1956	2	1	172	80	1,72	3	1	0	
8	f	06.05.1976	1	1	154	50	1,54	1	1	1	
9	f	23.11.1970	2	1	156	55	1,56	1	1	1	
10	f	03.12.1970	1	1	163	69	1,63	2	1	1	
11	f	07.02.1959	2	3	170	80	1,70	2	1	1	
12	m	11.11.1966	1	1	181	92	1,81	4	1	0	
13	f	15.01.1969	3	2	158	49	1,58	1	2	1	
14	m	26.12.1949	2	1	176	83	1,76	3	1	0	

Σχήμα 1.21. Αρχείο αρχικών δεδομένων από τα οποία έχουν διαγραφεί οι περιπτώσεις sex = "m".

Για να απενεργοποιήσουμε την επιλογή *Select Cases* πηγαίνουμε από το *Data* → *Select Cases* στο αντίστοιχο παράθυρο διαλόγου και κάνουμε κλικ ή στο *All Cases* ή στο *Reset*. Επίσης, καλό είναι να διαγράψουμε και τη στήλη filter_$, αφού πρώτα την επιλέξουμε και ακολούθως πατήσουμε το πλήκτρο *Delete*.

1.11 ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

Το SPSS παρέχει τη δυνατότητα δημιουργίας πολλών διαφορετικών γραφικών παραστάσεων για την απεικόνιση των δεδομένων μας. Για παράδειγμα, έστω ότι θέλουμε να δούμε πως μεταβάλλεται το βάρος με το ύψος των ατόμων στο παράδειγμα που εξετάζουμε. Για το σκοπό αυτό, ακολουθούμε την πορεία *Graphs* → *Legacy Dialogs* → *Scatter/Dot*. Στο πλαίσιο διαλόγου που ανοίγει επιλέγουμε *Simple Scatter* και κάνουμε κλικ στο Define, οπότε εμφανίζεται το παράθυρο διαλόγου του Σχήματος 1.22. Στο παράθυρο αυτό κάνουμε κλικ στο εικονίδιο της μεταβλητής bmass και ακολούθως κάνουμε κλικ στο βελάκι δίπλα από το πλαίσιο Y Axis, ώστε η μεταβλητή bmass να εισαχθεί στο πλαίσιο που αφορά τον άξονα των y. Με τον ίδιο τρόπο εισάγουμε τη μεταβλητή height στο πλαίσιο X Axis και την sex στο πλαίσιο Set Markers by.

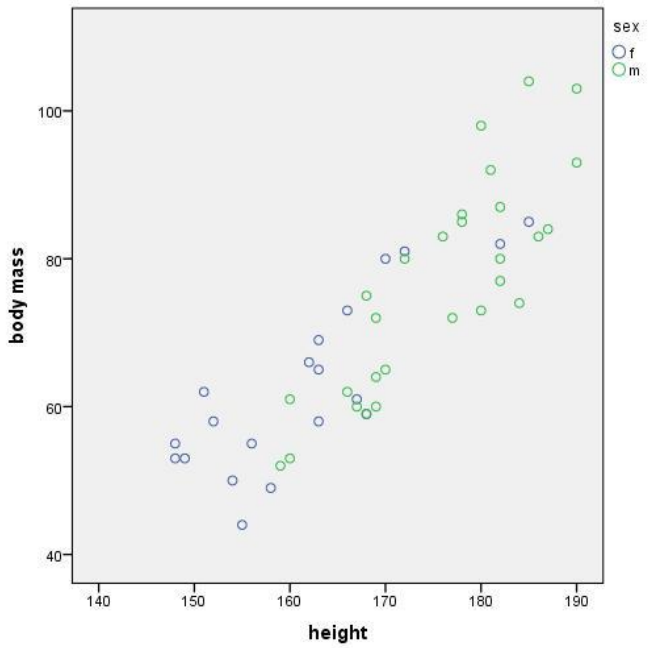


Σχήμα 1.22. Εισαγωγή δεδομένων στα πλαίσια X Axis, Y Axis και Set Markers by

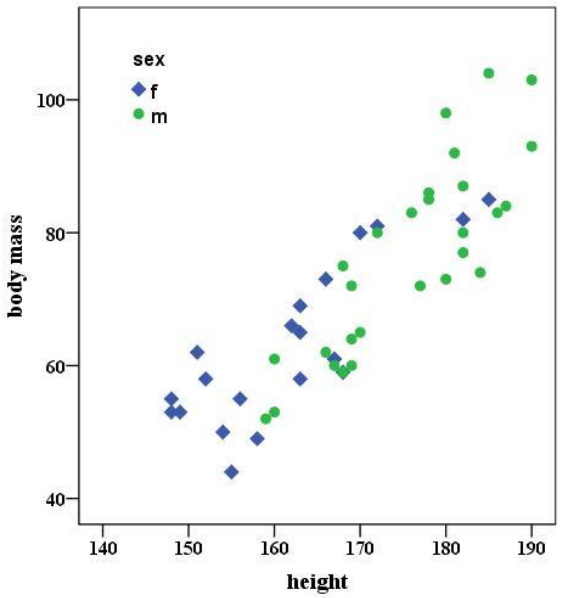
Με κλικ στο OK παίρνουμε τη γραφική παράσταση του Σχήματος 1.23. Συνήθως αυτές οι γραφικές παραστάσεις απαιτούν μορφοποίηση έτσι ώστε οι αριθμοί και οι τίτλοι των αξόνων να έχουν το κατάλληλο μέγεθος, οι κλίμακες και τα σύμβολα να έχουν τα επιθυμητά χαρακτηριστικά.

Για να μορφοποιήσουμε μια γραφική παράσταση κάνουμε διπλό κλικ επάνω της, οπότε ανοίγει ο επεξεργαστής γραφικών παραστάσεων (**Chart Editor**) ώστε να κάνουμε τις μετατροπές που θέλουμε. Για να φύγουμε από τον επεξεργαστή κάνουμε κλικ στο εικονίδιο στην πάνω δεξιά γωνία. Όταν είμαστε στον επεξεργαστή γραφικών παραστάσεων και κάνουμε κλικ σ' έναν από τους αριθμούς της κλίμακας ενός άξονα, επιλέγονται όλοι οι αριθμοί και ταυτόχρονα ανοίγει ένα πλαίσιο διαλόγου στο οποίο μπορούμε να αλλάξουμε τη γραμματοσειρά, το μέγεθος, το χρώμα, αλλά και τον αριθμό των δεκαδικών ψηφίων, καθώς επίσης την κλίμακα του άξονα και τη μορφή του. Κάθε φορά που κάνουμε μια αλλαγή πρέπει να πατάμε το κουμπί Apply.

Με τον ίδιο τρόπο αν κάνουμε κλικ σ' ένα σύμβολο που παριστάνει τα δεδομένα των ανδρών στη γραφική παράσταση, επιλέγονται όλα τα σύμβολα και αν ξανακάνουμε κλικ στο ίδιο σύμβολο επιλέγονται μόνο τα δεδομένα των ανδρών. Από το πλαίσιο που εμφανίζεται μπορούμε να τα μορφοποιήσουμε κατάλληλα αλλάζοντας τον τύπο, το μέγεθος και το χρώμα τους (Σχήμα 1.24).



Σχήμα 1.23. Γραφική παράσταση των μεταβλητών body mass και height



Σχήμα 1.24. Μορφοποιημένη γραφική παράσταση

Ιδιαίτερη προσοχή απαιτεί η λεζάντα. Αν την επιλέξουμε, μπορούμε να τη μεταφέρουμε οπουδήποτε μέσα στο γράφημα. Αν όμως αυξήσουμε το μέγεθος των γραμμάτων ενδέχεται να εξαφανιστεί ένα μέρος της. Τότε πρέπει με προσοχή να την ξανα-επιλέξουμε και να αυξήσουμε με το ποντίκι τις διαστάσεις της.

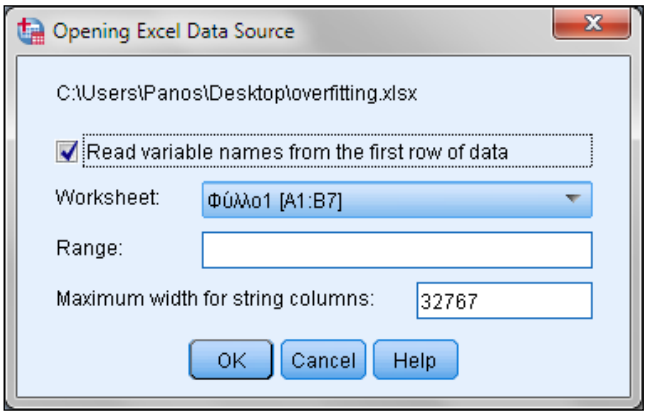
Πιο εξειδικευμένες γραφικές παραστάσεις θα εξεταστούν στα επόμενα κεφάλαια.

1.12 ΑΝΟΙΓΜΑ ΑΡΧΕΙΩΝ

Για να ανοίξουμε ένα αποθηκευμένο αρχείο του SPSS ακολουθούμε τις γνωστές στα Windows διαδικασίες ανοίγματος ενός αρχείου: Κάνουμε διπλό κλικ πάνω στο εικονίδιό του ή πηγαίνουμε *File* → *Open*, στη γραμμή μενού.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η περίπτωση να ανοίξουμε ένα αρχείο του Excel ως αρχείο του SPSS. Για το σκοπό αυτό ακολουθούμε την πορεία *File* → *Open* → *Data* και στο παράθυρο διαλόγου *Open File* που ανοίγει επιλέγουμε στο *Look in* το φάκελο ή γενικότερα τη διεύθυνση στην οποία είναι το αρχείο. Στο *Files of Type* ορίζουμε τον τύπο του αρχείου, δηλαδή Excel (*.xls, *.xlsx, *.xlsm), και επιλέγουμε με κλικ το αρχείο που θέλουμε να ανοίξουμε. Με κλικ στο *Open* ανοίγει το παράθυρο του Σχήματος 1.25. Αν στο αρχείο του Excel η πρώτη γραμμή έχει ετικέτες (τίτλους), τότε κάνουμε κλικ στο *Read variable names from the first row of data*. Επίσης επιλέγουμε ποιο φύλλο εργασίας θα ανοίξει ως

έγγραφο του SPSS και κάνουμε κλικ στο *OK*. Με τον τρόπο αυτό ολόκληρο το φύλλο εργασίας που επιλέξαμε μετατρέπεται σε έγγραφο του SPSS.



Σχήμα 1.25. Το παράθυρο διαλόγου *Opening Excel Data Source*

Όπως ήδη έχουμε αναφέρει, μπορούμε μεμονωμένες στήλες του Excel να τις μεταφέρουμε άμεσα σε ένα αρχείο του SPSS με τις γνωστές εντολές *Copy - Paste*. Δηλαδή, επιλέγουμε μια ή περισσότερες στήλες από ένα φύλλο εργασίας του Excel, κάνουμε *Ctrl+C*, μεταφερόμαστε στο έγγραφο του SPSS, κάνουμε κλικ στο πρώτο κελί μιας στήλης του φύλλου εργασίας του SPSS και ολοκληρώνουμε τη μεταφορά των δεδομένων με *Ctrl+V*. Η ίδια διαδικασία ισχύει και αντίστροφα (για μεταφορά δεδομένων από το SPSS στο Excel).

1.13 ΑΠΟΘΗΚΕΥΣΗ ΑΡΧΕΙΩΝ

Για να αποθηκεύσουμε ένα αρχείο του SPSS κάνουμε *κλικ* στο *File* της γραμμής μενού, επιλέγουμε το *Save as (Αποθήκευση ως)* και συμπληρώνουμε κατάλληλα το παράθυρο διαλόγου που θα εμφανιστεί. Το αρχείο δεδομένων, ο SPSS Data Editor, αποθηκεύεται με την προέκταση **.sav**. Αντίθετα, ένα αρχείο αποτελεσμάτων, SPSS Viewer, αποθηκεύεται με την προέκταση **.spo**.

Τα αρχεία του SPSS μπορούν να αποθηκευτούν και ως αρχεία του Excel. Ένα αρχείο δεδομένων αποθηκεύεται ως αρχείο του Excel αν στο παράθυρο διαλόγου *Save Data As*, που ανοίγει μέσω της διαδρομής *File* → *Save as*, επιλέξουμε το *Excel 97 through 2003 (*.xls)* ή *Excel 2007 through 2010 (*.xlsx)* στο πλαίσιο *Save as type*. Εισάγουμε το όνομα του αρχείου στο πλαίσιο *File name* και κάνουμε κλικ στο *Save*. Από το αρχείο του SPSS αποθηκεύεται μόνο το περιεχόμενο του φύλλου εργασίας που βρίσκεται στο *Data View*.

2. ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Η Στατιστική επιστήμη επιχειρεί να εξαγάγει συμπεράσματα χρησιμοποιώντας εμπειρικά-πειραματικά δεδομένα. Η **περιγραφική στατιστική** είναι ο κλάδος της στατιστικής που αναπτύσσει μεθόδους για τη συνοπτική και αποτελεσματική παρουσίαση δεδομένων με τη χρήση α) αριθμητικών περιγραφικών μέτρων, β) πινάκων συχνοτήτων και γ) μεθόδων γραφικής παρουσίασης δεδομένων.

2.1 ΠΛΗΘΥΣΜΟΣ, ΔΕΙΓΜΑ, ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Ο **πληθυσμός** είναι το σύνολο όλων των στοιχείων υπό μελέτη. Για παράδειγμα, εάν ενδιαφερόμαστε για τα λίθινα εργαλεία σε μία αρχαιολογική θέση, ο πληθυσμός αποτελείται από όλα τα λίθινα εργαλεία που έχουν βρεθεί εκεί.

Το **δείγμα** είναι ένα υποσύνολο του πληθυσμού το οποίο επιλέγουμε να αναλύσουμε. Για παράδειγμα, εάν τα λίθινα εργαλεία που έχουν βρεθεί σε έναν οικισμό είναι υπερβολικά πολλά και δεν προλαβαίνουμε να τα εξετάσουμε όλα, μπορούμε να επιλέξουμε τυχαία ένα ποσοστό και αυτό θα αποτελέσει το δείγμα μας. Εάν η επιλογή του δείγματος έγινε τυχαία (δηλαδή με τέτοιο τρόπο ώστε κάθε στοιχείο του πληθυσμού να έχει την ίδια πιθανότητα να βρίσκεται στο δείγμα) μιλάμε για ένα **τυχαίο δείγμα**. Αντίθετα, μπορεί να μελετήσουμε μόνο τις λεπίδες, οπότε δεν μπορούμε να μιλήσουμε για τυχαίο δείγμα εφόσον επιλέγουμε ένα συγκεκριμένο υποσύνολο του πληθυσμού μας. Και σε αυτή την περίπτωση όμως, είναι δυνατό όλες οι λεπίδες του οικισμού να αποτελέσουν τον πληθυσμό και από αυτές εμείς να μελετήσουμε μόνο ορισμένες, επιλέγοντας τυχαία ένα υποσύνολο (τυχαίο δείγμα) ή διακρίνοντας ορισμένες λεπίδες με συγκεκριμένα χαρακτηριστικά (μη τυχαίο δείγμα).

Το μέγεθος του δείγματος παίζει καθοριστικό ρόλο στην αξιοπιστία των στατιστικών αποτελεσμάτων. Το δείγμα πρέπει να είναι μεγάλο κυρίως όταν α) υπάρχει ανομοιογένεια στον πληθυσμό, β) επιθυμούμε μεγάλη ακρίβεια αποτελεσμάτων και γ) χρησιμοποιούμε πολύπλοκες στατιστικές αναλύσεις.

Η διαδικασία δημιουργίας ενός δείγματος ονομάζεται **δειγματοληψία**. Επειδή το δείγμα αποτελεί ένα ποσοστό του πληθυσμού, χρειάζεται ιδιαίτερη προσοχή και μεθοδικότητα προκειμένου αυτό να είναι αντιπροσωπευτικό του πληθυσμού. Αν το δείγμα δεν είναι αντιπροσωπευτικό του πληθυσμού, τότε

ανεξάρτητα από το μέγεθός του, η στατιστική ανάλυση θα οδηγήσει σε λανθασμένα συμπεράσματα.

2.2 ΑΡΙΘΜΗΤΙΚΑ ΠΕΡΙΓΡΑΦΙΚΑ ΜΕΤΡΑ

Τα αριθμητικά μέτρα χαρακτηρίζουν διάφορες ιδιότητες των δειγμάτων. Τα βασικά μέτρα ομαδοποιούνται σε τρεις κατηγορίες: Μέτρα θέσης, διασποράς και σχήματος κατανομής (Πίνακας 2.1). Τα μέτρα θέσης δίνουν πληροφορίες που σχετίζονται με τη θέση των δεδομένων του δείγματος, τα μέτρα διασποράς ελέγχουν πόσο διασκορπισμένα είναι τα δεδομένα, ενώ τα μέτρα σχήματος κατανομής αφορούν το σχήμα της κατανομής των δεδομένων, δηλαδή πόσο συμμετρικά ή ασύμμετρα κατανέμονται οι τιμές ενός δείγματος γύρω από κάποια τιμή.

Πίνακας 2.1. Βασικά μέτρα ιδιοτήτων δείγματος

Μέτρα θέσης	**Μέτρα διασποράς**
Μέση τιμή (Mean) Διάμεσος (Median) Κορυφή (Mode) Πρώτο τεταρτημόριο (First quartile) Τρίτο τεταρτημόριο (Third quartile)	Διασπορά (Variance) Τυπική απόκλιση (Standard deviation) Τυπική απόκλιση μέσου (Standard error of mean) Μέγιστη τιμή (maximum) Ελάχιστη τιμή (Minimum) Ενδοτεταρτημοριακό εύρος (Interquartile range)
Μέτρα σχήματος κατανομής	
Συντελεστής ασυμμετρίας (Skewness)	Συντελεστής κυρτότητας (Kurtosis)

Η μέση τιμή (**mean** ή **average** value) του δείγματος είναι η τιμή γύρω από την οποία βρίσκονται συγκεντρωμένες οι τιμές του δείγματος και ορίζεται από τη σχέση:

$$\bar{x} = (x_1 + x_2 + \ldots + x_m)/m$$

Όπου x_1, x_2, …, x_m είναι οι μετρήσεις και m το μέγεθος του δείγματος. Για παράδειγμα, εάν εξετάζουμε το ύψος των μαθητών μίας τάξης τριάντα ατόμων και μετρήσουμε όλους τους μαθητές, τότε το δείγμα μας συμπίπτει με τον πληθυσμό και ισούται με 30 (άρα m=30), x_1 είναι το ύψος του πρώτου μαθητή, x_2 το ύψος του δεύτερου… x_{30} το ύψος του τριακοστού μαθητή.

Η διασπορά ή διακύμανση (**variance**) μας δείχνει τη διασπορά των τιμών ενός δείγματος γύρω από τη μέση του τιμή και ορίζεται από τη σχέση:

$$\text{Var(x)} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_m-\bar{x})^2}{m-1}$$

Αν οι τιμές της διασποράς είναι υψηλές τότε οι τιμές του δείγματος ποικίλουν σημαντικά σε σχέση με την μέση τιμή. Για παράδειγμα, εάν βρούμε υψηλές τιμές διασποράς στο δείγμα ύψους των μαθητών, σημαίνει πως κάποιοι μαθητές ήταν πολύ ψηλοί και κάποιοι πολύ κοντοί. Στην αντίθετη περίπτωση, το ύψος όλων των μαθητών ήταν παρόμοιο και κοντά στη μέση τιμή του δείγματος.

Η τυπική απόκλιση (**standard deviation**) είναι η τετραγωνική ρίζα της διασποράς και επίσης εκφράζει την απόκλιση των μετρήσεων από τη μέση τιμή $\bar{x}$.

Η διάμεσος (**median**) είναι η "μεσαία" τιμή ενός δείγματος, δηλαδή οι μισές τιμές του δείγματος είναι μικρότερες ή ίσες με τη διάμεσο και οι υπόλοιπες μισές μεγαλύτερες ή ίσες με τη διάμεσο. Προκειμένου να εντοπίσουμε τη διάμεσο, πρέπει να διατάξουμε τις τιμές του δείγματος κατά αύξουσα σειρά (από τη μικρότερη τιμή προς τη μεγαλύτερη). Για παράδειγμα, έστω το δείγμα ύψους πέντε μαθητών Δ = {1.53, 1.65, 1.78, 1.84, 1.86}. Η διάμεσος είναι η τιμή 1.78. Αντίθετα στο δείγμα Δ = {1.53, 1.65, 1.78, 1.80, 1.84, 1.86}, η διάμεσος υπολογίζεται από τη σχέση (1.78 + 1.80)/2=1.79, δηλαδή προκύπτει από το ημι-άθροισμα των δύο μεσαίων τιμών. Η διάμεσος δεν επηρεάζεται από ακραίες τιμές (π.χ. εάν στο δείγμα μας υπήρχε ένας μαθητής με ύψος 1.35 ή με ύψος 2.10). Έτσι, για την περιγραφή δεδομένων που εμφανίζουν ακραίες τιμές προτιμάται ως μέτρο θέσης από τη μέση τιμή, η οποία επηρεάζεται πολύ από ακραίες τιμές.

Η **κορυφή** (**mode**) είναι η μέτρηση με τη μεγαλύτερη συχνότητα σ' ένα δείγμα. Για παράδειγμα, στο δείγμα Δ = {2, 3, 5, 3, 6, 2, 4, 3} η κορυφή είναι η τιμή 3.

Πρώτο, τρίτο τεταρτημόριο (First, third quartile) και ενδοτεταρτημοριακό εύρος (Interquartile range): Κάθε δείγμα έχει τρία τεταρτημόρια (quartiles). Το πρώτο τεταρτημόριο (Q_1) είναι η τιμή του δείγματος

για την οποία ισχύει ότι το 25% των τιμών του δείγματος είναι μικρότερες ή ίσες με αυτή. Το τρίτο τεταρτημόριο (Q_3) είναι η τιμή του δείγματος για την οποία ισχύει ότι το 75% των τιμών του δείγματος είναι μικρότερες ή ίσες με αυτή. Η διαφορά Q_3-Q_1 ισούται με το ενδοτεταρτημοριακό εύρος.

Η τυπική απόκλιση μέσου (Standard error of mean) υπολογίζεται από τη σχέση

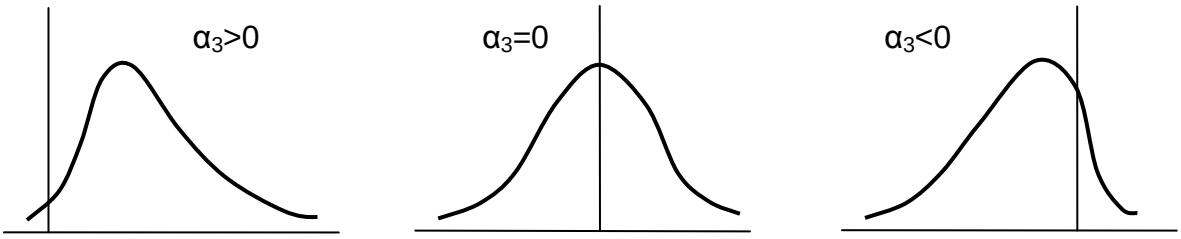
$$s_m = \frac{s}{\sqrt{m}}$$

όπου s είναι η τυπική απόκλιση και m το μέγεθος του δείγματος.

Η μέγιστη τιμή (maximum) είναι η μέγιστη τιμή του δείγματος. Για παράδειγμα, στο δείγμα με τα ύψη μαθητών Δ = {1.53, 1.65, 1.78, 1.84, 1.86} η μέγιστη τιμή είναι 1.86.

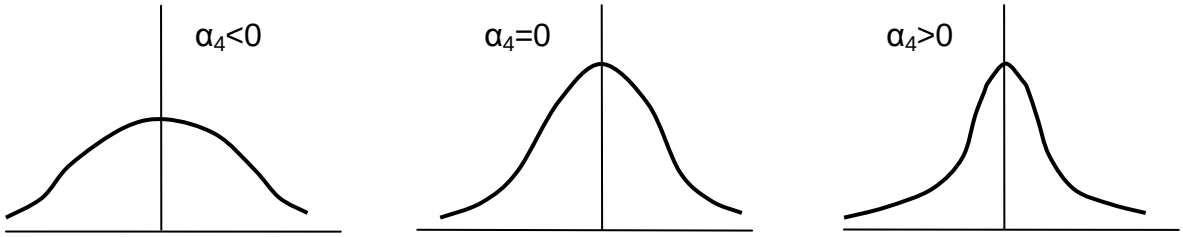
Η ελάχιστη τιμή (minimum) είναι η ελάχιστη τιμή του δείγματος. Για παράδειγμα, στο παραπάνω δείγμα η ελάχιστη τιμή είναι 1.53.

Συντελεστές ασυμμετρίας (Skewness) και **κυρτότητας** (Kurtosis). Όπως θα δούμε στο Κεφάλαιο 3, τα δεδομένα ενός δείγματος ακολουθούν κάποια κατανομή που μπορεί να είναι συμμετρική ή ασύμμετρη. Οι δείκτες που χρησιμοποιούνται για το σχήμα της κατανομής είναι οι συντελεστές ασυμμετρίας και κυρτότητας. Ο συντελεστής ασυμμετρίας συμβολίζεται συνήθως με α_3 και ο συντελεστής κυρτότητας με α_4. Όταν $\alpha_3 = 0$ το σχήμα της κατανομής είναι συμμετρικό ως προς τη μέση τιμή, αν $\alpha_3 < 0$ η κατανομή είναι ασύμμετρη προς τα αριστερά, δηλαδή οι περισσότερες τιμές βρίσκονται αριστερά της κορυφής, ενώ αν $\alpha_3 > 0$ η κατανομή είναι ασύμμετρη προς τα δεξιά (Σχήμα 2.1).



Σχήμα 2.1. *Κατανομές με διαφορετική ασυμμετρία*

Όταν $\alpha_4 = 0$ η κορυφή της κατανομής μοιάζει με αυτή της τυπικά κανονικής κατανομής (Κεφάλαιο 3). Όταν $\alpha_4 < 0$ η κατανομή πλατειάζει, ενώ όταν $\alpha_4 > 0$ η κατανομή έχει οξεία κορυφή (Σχήμα 2.2).



Σχήμα 2.2. *Κατανομές με διαφορετική κυρτότητα*

Τα αριθμητικά περιγραφικά μέτρα υπολογίζονται ιδιαίτερα εύκολα στο SPSS, όπως θα δούμε παρακάτω.

2.3 ΠΙΝΑΚΕΣ ΣΥΧΝΟΤΗΤΩΝ

Έστω x_1, x_2, …, x_m οι τιμές μιας μεταβλητής x σ' ένα δείγμα. Ονομάζουμε **συχνότητα** της τιμής x_i τον φυσικό αριθμό v_i που δείχνει πόσες φορές επαναλαμβάνεται η τιμή x_i στο δείγμα. Αν $v = v_1 + v_2 + … + v_m$, τότε ο λόγος

$$f_i = \frac{v_i}{v}$$

ονομάζεται **σχετική συχνότητα** της τιμής x_i. Για παράδειγμα, έστω το δείγμα Δ={2, 5, 3, 5, 8, 9, 6, 2, 5, 8, 7}. Η συχνότητα της τιμής 5 ισούται με:

$f_5= \frac{3}{11}$ επειδή εμφανίζεται τρεις φορές σε ένα σύνολο 11 τιμών.

Όταν το πλήθος των τιμών του δείγματος είναι μεγάλο και κυρίως όταν η μεταβλητή x είναι συνεχής, δηλαδή μπορεί να πάρει μια οποιαδήποτε τιμή στο πεδίο ορισμού της (για παράδειγμα, το ύψος και το βάρος είναι συνεχείς μεταβλητές, ενώ τα διακοσμητικά μοτίβα της κεραμικής δεν είναι συνεχής μεταβλητή), οι συχνότητες ορίζονται σε μια περιοχή τιμών που ονομάζεται **κλάση**. Συγκεκριμένα αν x_{min} και x_{max} είναι η ελάχιστη και η μέγιστη τιμή της μεταβλητής x στο δείγμα, διαιρούμε το διάστημα x_{max} - x_{min} σε k υποδιαστήματα που ονομάζονται **κλάσεις** και σε κάθε κλάση υπολογίζουμε το σύνολο των τιμών του δείγματος που ανήκουν σ' αυτή. Η ποσότητα αυτή είναι η συχνότητα της κλάσης.

Για παράδειγμα, έχουμε το δείγμα ηλικιών των υπαλλήλων μιας εταιρείας:

Δ={28, 36, 22, 41, 27, 50, 32, 29, 42, 29, 25, 38, 36, 45, 27, 29, 32, 39, 47, 33, 53, 33, 31, 40, 20, 34, 37, 29, 33, 27, 39, 37, 44, 26, 43, 26, 36, 34, 49, 36, 26, 31, 28, 59, 30, 28, 30, 34, 28, 24}

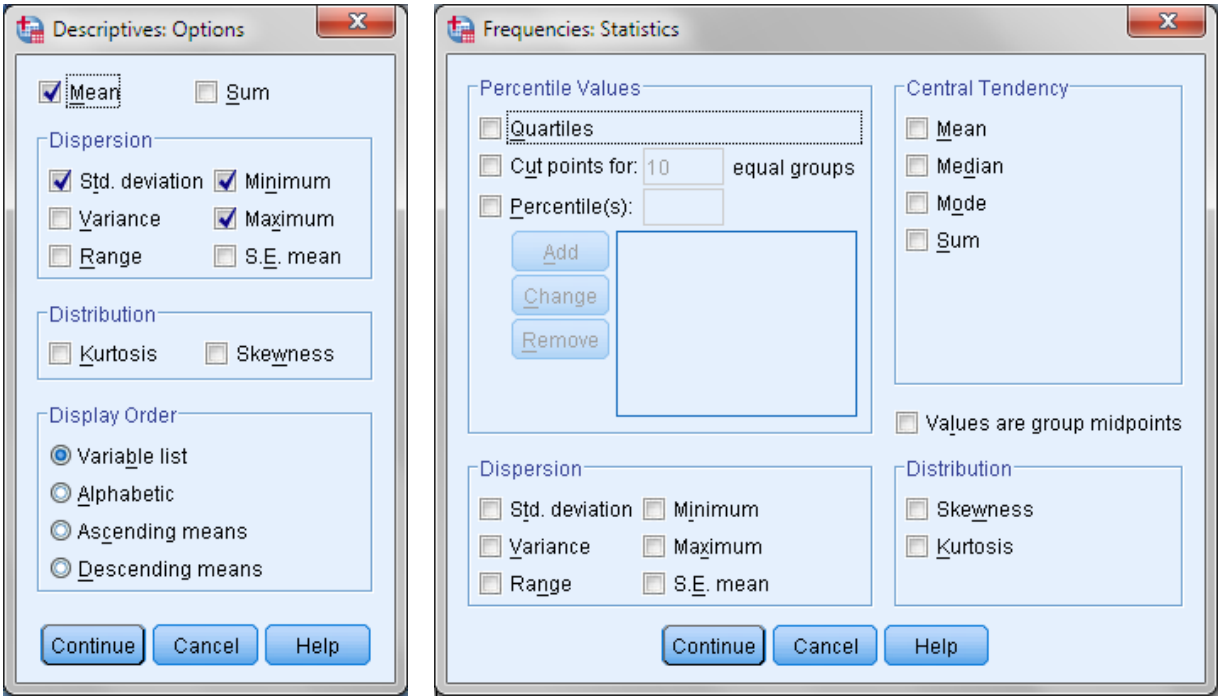
και θέλουμε να το ομαδοποιήσουμε σε 8 κλάσεις. Στο δείγμα αυτό η μικρότερη παρατήρηση είναι το 20 και η μεγαλύτερη το 59. Οπότε εφαρμόζοντας τον παραπάνω τύπο έχουμε: $\frac{59-20}{8} \approx 5$. Έτσι, το πλάτος κάθε κλάσης είναι 5, οπότε οι κλάσεις μας θα είναι οι εξής: 20-25, 25-30, 30-35, 35-40, 40-45, 45-50, 50-55, 55-60. Στην πρώτη κλάση ανήκουν οι τιμές από 20 έως 24, στη δεύτερη οι τιμές από 25 έως 29 κ.ο.κ. Συνεπώς, καταλήγουμε με τον Πίνακα 2.2.

Πίνακας 2.2. Κλάσεις και αντίστοιχες συχνότητες του παραδείγματος

Κλάσεις	**Συχνότητα**
20-25	3
25-30	15
30-35	12
35-40	9
40-45	5
45-50	3
50-55	2
55-60	1

2.4 ΥΠΟΛΟΓΙΣΜΟΣ ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΤΡΩΝ ΣΥΝΕΧΩΝ ΜΕΤΑΒΛΗΤΩΝ

Για να υπολογίσουμε με το SPSS τα στατιστικά μέτρα μιας scale μεταβλητής (ποσοτική, συνεχής μεταβλητή) ακολουθούμε τη διαδικασία: *Analyze* → *Descriptive Statistics* → *Descriptives* και στο παράθυρο διαλόγου *Descriptives* επιλέγουμε τη μεταβλητή που θέλουμε να αναλύσουμε κάνοντας κλικ στη μεταβλητή αυτή και κλικ στο βέλος ▶. Με κλικ στο *Options* ανοίγει το παράθυρο *Descriptives: Options* όπου μπορούμε να επιλέξουμε τα μέτρα που θέλουμε να υπολογιστούν (Σχήμα 2.3-αριστερά). Παρατηρούμε ότι αυτά είναι σχετικά λίγα. Περισσότερα στατιστικά μέτρα μπορούμε να υπολογίσουμε μέσω της διαδικασίας *Analyze* → *Descriptive Statistics* → *Frequencies*. Στη διαδικασία αυτή κάνουμε κλικ στο *Statistics* και επιλέγουμε τα μέτρα που θέλουμε από το παράθυρο *Frequencies: Statistics* (Σχήμα 2.3-δεξιά).

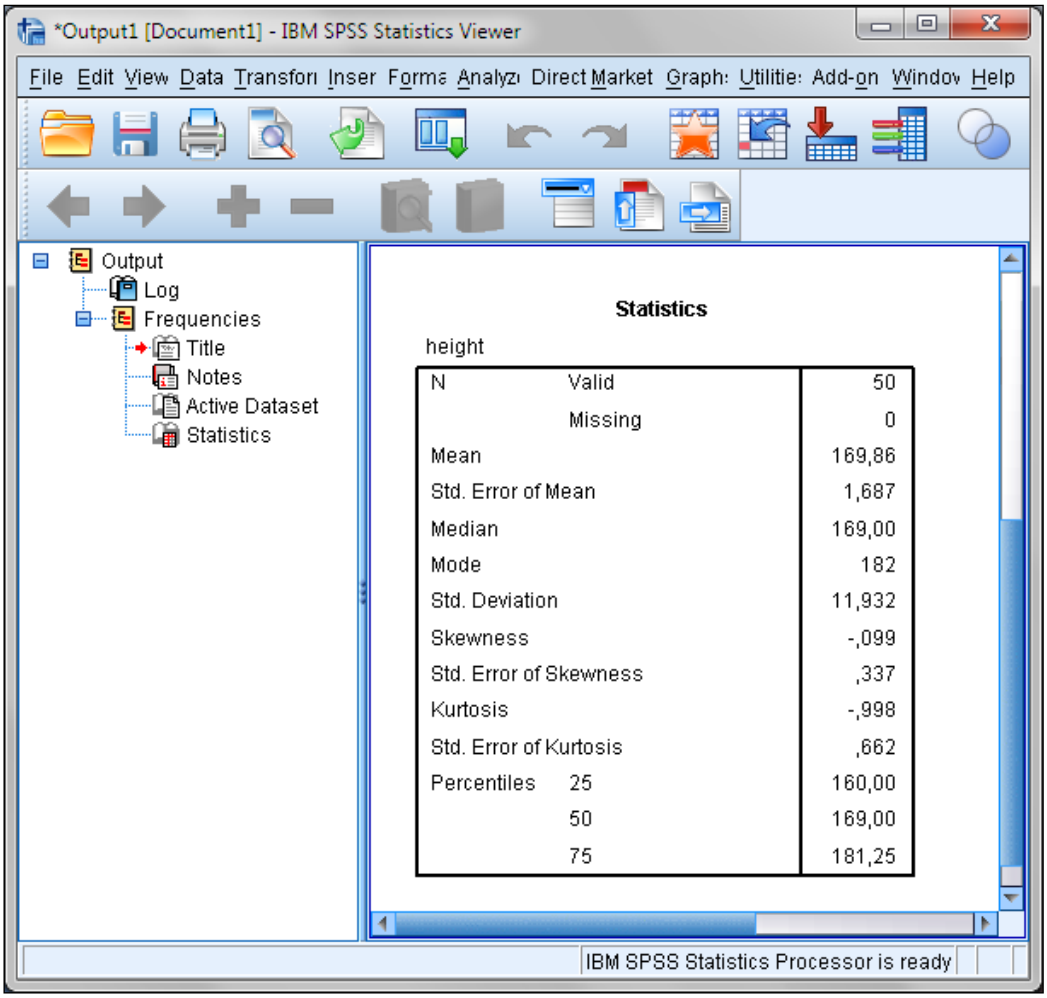


Σχήμα 2.3. Τα παράθυρα διαλόγου *Desciptives: Options* και Frequencies:Statistics

Παράδειγμα

Να υπολογιστεί η μέση τιμή, η τυπική απόκλιση, η τυπική απόκλιση του μέσου (S.E. mean), τα σημεία Quartiles, η διάμεσος, η κορυφή, η κύρτωση και η ασυμμετρία της μεταβλητής height των δεδομένων του αρχείου osteological data.sav.

◆ Ανοίγουμε το αρχείο osteological data.sav και ακολουθούμε τη διαδικασία *Analyze* → *Descriptive Statistics* → *Frequencies*, επειδή οι ποσότητες που θέλουμε να υπολογίσουμε βρίσκονται όλες στο παράθυρο διαλόγου *Frequencies: Statistics*. Έτσι στο παράθυρο Frequencies που ανοίγει κάνουμε κλικ στη μεταβλητή salary και κλικ στο βέλος ►. Επίσης, απενεργοποιούμε την επιλογή *Display frequency tables* και με κλικ στο *Statistics* επιλέγουμε τις ποσότητες που θέλουμε να υπολογιστούν. Στη συνέχεια κάνουμε κλικ στο *Continue* και *OK*. Τα αποτελέσματα που παίρνουμε δίνονται στο Σχήμα 2.4. Στο σχήμα αυτό τα σημεία quartiles είναι τα percentiles 25 και 75.



Σχήμα 2.4. Παράθυρο αποτελεσμάτων

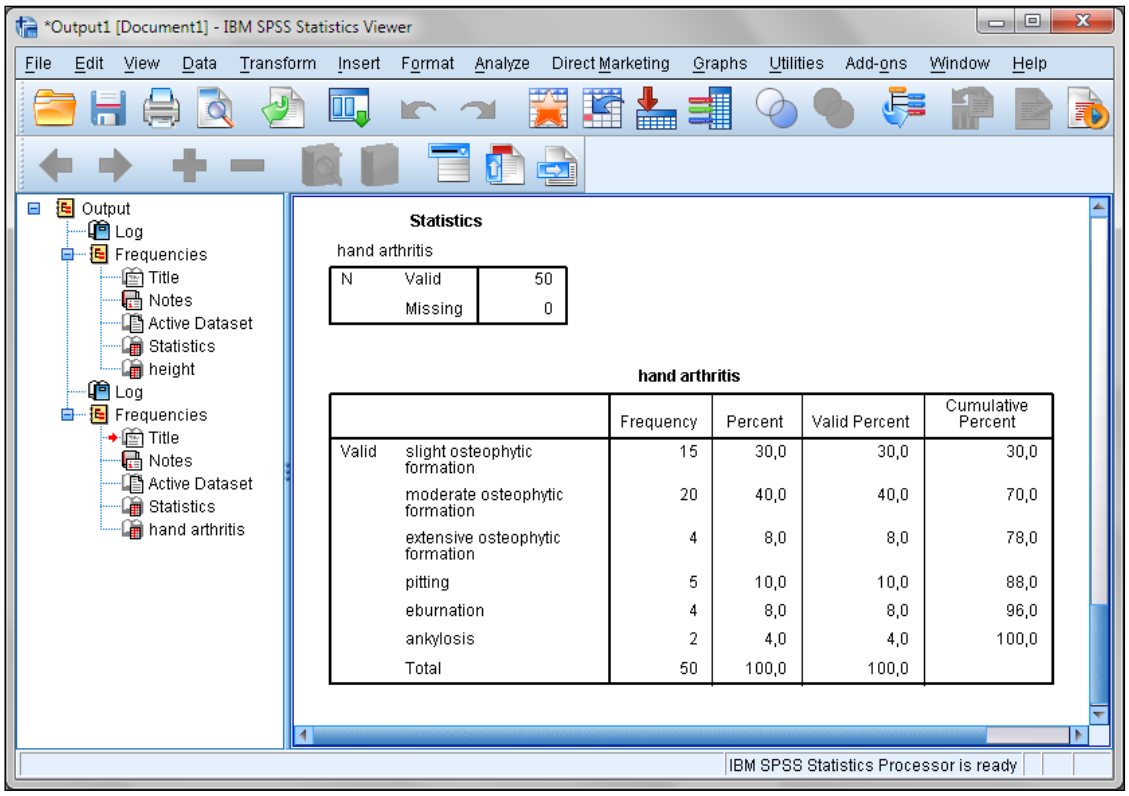
2.5 ΥΠΟΛΟΓΙΣΜΟΣ ΣΥΧΝΟΤΗΤΩΝ

Όταν μια μεταβλητή είναι nominal ή ordinal δεν μπορούν να υπολογιστούν τα παραπάνω στατιστικά μέτρα. Σ' αυτή την περίπτωση υπολογίζονται μόνο συχνότητες, δηλαδή τα ποσοστά εμφάνισης των διαφόρων τιμών της μεταβλητής. Η πορεία που ακολουθούμε είναι: *Analyze* → *Descriptive Statistics* → *Frequencies*. Επιλέγουμε τη μεταβλητή που μελετάμε, ενεργοποιούμε την επιλογή *Display frequency tables* και με κλικ στο *Statistics* απενεργοποιούμε όλες τις επιλογές.

Παράδειγμα

Να υπολογιστούν οι συχνότητες της μεταβλητής arthritis στο αρχείο osteological data.sav.

◆ Με ανοικτό το αρχείο osteological data.sav ακολουθούμε τη διαδικασία *Analyze* → *Descriptive Statistics* → *Frequencies* και στο παράθυρο διαλόγου *Frequencies* κάνουμε κλικ στη μεταβλητή hand arthritis και κλικ στο βέλος ▶. Ενεργοποιούμε την επιλογή *Display frequency tables* και στο *Statistics* απενεργοποιούμε όλες τις επιλογές. Το αρχείο αποτελεσμάτων που παίρνουμε δίνεται στο Σχήμα 2.5. Από τον πίνακα αποτελεσμάτων παρατηρούμε για παράδειγμα ότι το 70% των ατόμων είχαν αρθρίτιδα στα δύο πρώτα επίπεδα, ενώ ankylosis μόνο το 4%.



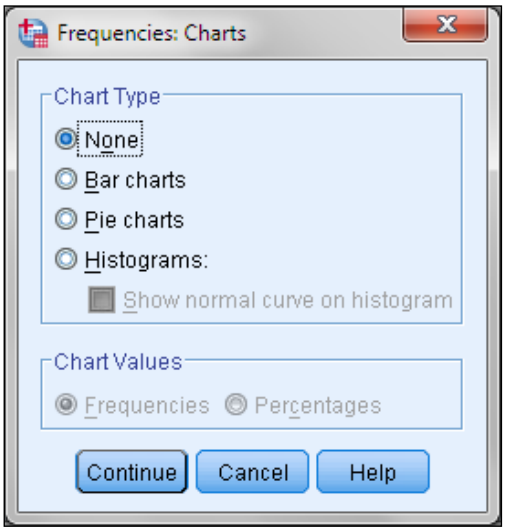
Σχήμα 2.5. Παράθυρο αποτελεσμάτων

2.6 ΜΕΘΟΔΟΙ ΓΡΑΦΙΚΗΣ ΠΑΡΟΥΣΙΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Υπάρχουν αρκετοί τύποι γραφικών παραστάσεων για την παρουσίαση στατιστικών δεδομένων. Οι πιο βασικοί είναι τα α) ραβδογράμματα (bar charts), β) κυκλικά διαγράμματα (pie charts), γ) ιστογράμματα (histograms) και δ) θηκογράμματα (boxplots). Οι δύο πρώτοι τύποι γραφικών παραστάσεων

χρησιμοποιούνται όταν η μεταβλητή είναι ποιοτική (scale), ενώ οι δύο τελευταίοι τύποι όταν έχουμε ποσοτικά δεδομένα (nominal και ordinal μεταβλητές).

Η περιγραφική στατιστική σχετίζεται μόνο με ιστογράμματα (histograms), ραβδογράμματα (bars) και κυκλικά διαγράμματα (pie). Οι γραφικές αυτές παραστάσεις εμφανίζονται στο αρχείο αποτελεσμάτων αν στο παράθυρο διαλόγου *Frequencies* κάνουμε κλικ στο *Charts* και επιλέξουμε το κατάλληλο γράφημα (Σχήμα 2.6). Εναλλακτικά, γραφικές παραστάσεις μπορούν να γίνουν μέσω της διαδικασίας *Graphs* → *Bar (Pie or Histogram)*, όπως εξετάζεται στο παρακάτω παράδειγμα.



Σχήμα 2.6. Το παράθυρο διαλόγου *Frequencies:Charts*

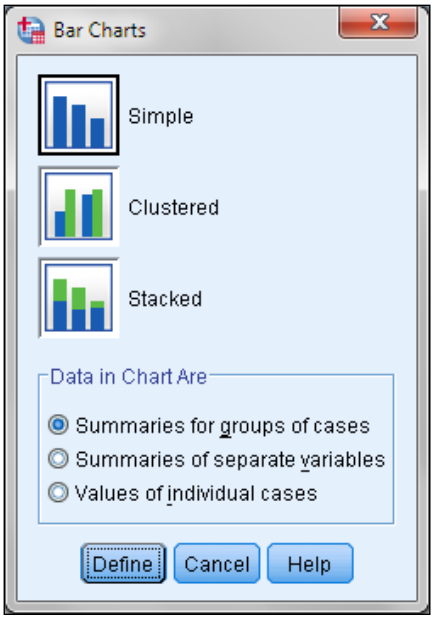
α) Ραβδόγραμμα (bar chart)

Το ραβδόγραμμα σχηματίζεται με βάση τον πίνακα συχνοτήτων μιας ποιοτικής μεταβλητής x. Στον οριζόντιο άξονα τοποθετούνται ισαπέχοντα τα στοιχεία του δείγματος και σε κάθε στοιχείο αντιστοιχεί μια ορθογώνια στήλη με ύψος ίσο με τη συχνότητα του στοιχείου.

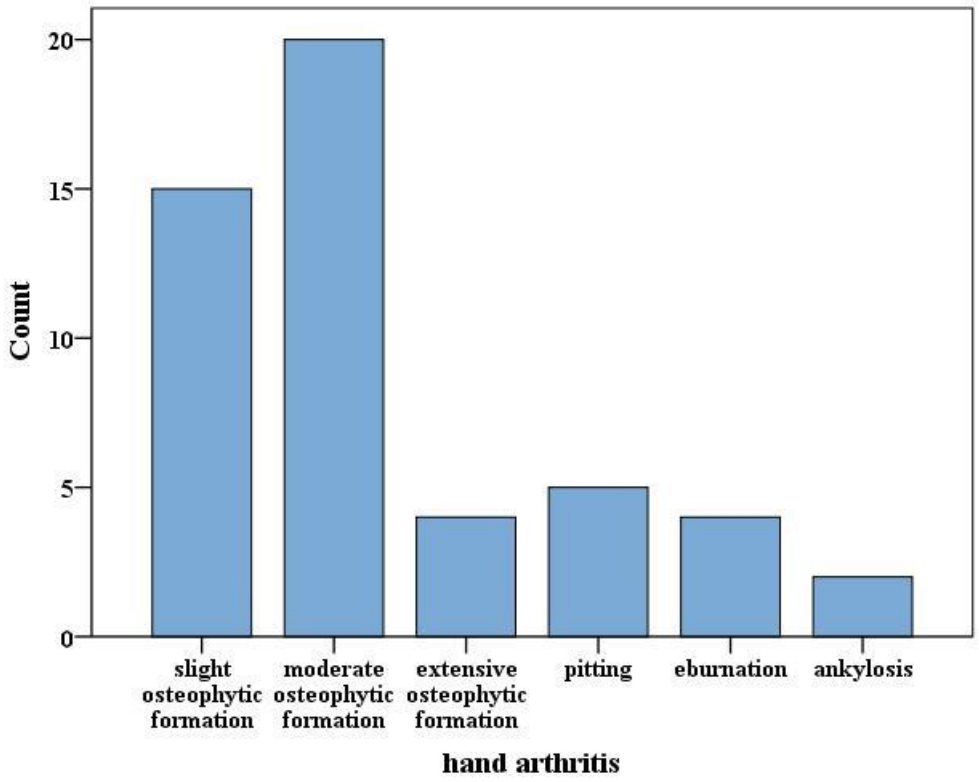
Παράδειγμα

Να γίνει το ραβδόγραμμα της μεταβλητής hand arthritis του αρχείου osteological data.sav.

◆ α) Για να κατασκευάσουμε το ραβδόγραμμα της μεταβλητής arthritis ακολουθούμε τη διαδικασία *Graphs* → *Legacy Dialogs* → *Bar* και στο παράθυρο διαλόγου επιλέγουμε *Simple* και κάνουμε κλικ στο *Define* (Σχήμα 2.7). Στο νέο παράθυρο διαλόγου που ανοίγει επιλέγουμε τη μεταβλητή arthritis την οποία στέλνουμε στο πλαίσιο *Category Axis*. Επίσης από το *Bars Represent* επιλέγουμε τον άξονα των y. Συνήθως επιλέγουμε το *N of cases* ή το *% of cases*. Στην πρώτη περίπτωση το ύψος της κάθε ράβδου θα είναι ανάλογο του αριθμού των περιπτώσεων που αναπαριστά, ενώ στη δεύτερη περίπτωση ανάλογο του εκατοστιαίου ποσοστού των περιπτώσεων που αναπαριστά. Με κλικ στο *OK* παίρνουμε το Σχήμα 2.8.



Σχήμα 2.7. Το παράθυρο διαλόγου *Bar Charts*

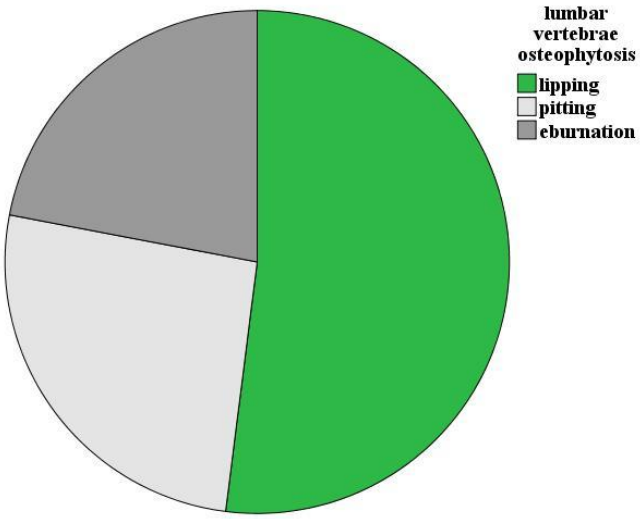


Σχήμα 2.8. Το ραβδόγραμμα της μεταβλητής hand arthritis

β) Κυκλικό διάγραμμα (piechart)

Το διάγραμμα αυτό είναι ένας κυκλικός δίσκος χωρισμένος σε τομείς. Κάθε τομέας εκφράζει ένα στοιχείο του δείγματος και έχει εμβαδό ανάλογο προς τη συχνότητα του στοιχείου. Για το κυκλικό γράφημα εργαζόμαστε ανάλογα. Από το *Graphs* → *Legacy Dialogs* → *Pie* στο παράθυρο διαλόγου επιλέγουμε *Summaries of groups of cases* και κάνουμε κλικ στο *Define*. Το νέο παράθυρο διαλόγου συμπληρώνεται όπως και στην προηγούμενη περίπτωση, δηλαδή μεταφέρουμε τη μεταβλητή που θέλουμε να μελετήσουμε στο πλαίσιο *Define Slices by* και επιλέγουμε συνήθως το *N of cases*.

Στο Σχήμα 2.9 δίνεται το κυκλικό διάγραμμα που αντιστοιχεί στη μεταβλητή lumbar vertebrae osteophytosis.



Σχήμα 2.9. Κυκλικό διάγραμμα της μεταβλητής lumbar vertebrae osteophytosis

γ) Ιστόγραμμα (histogram)

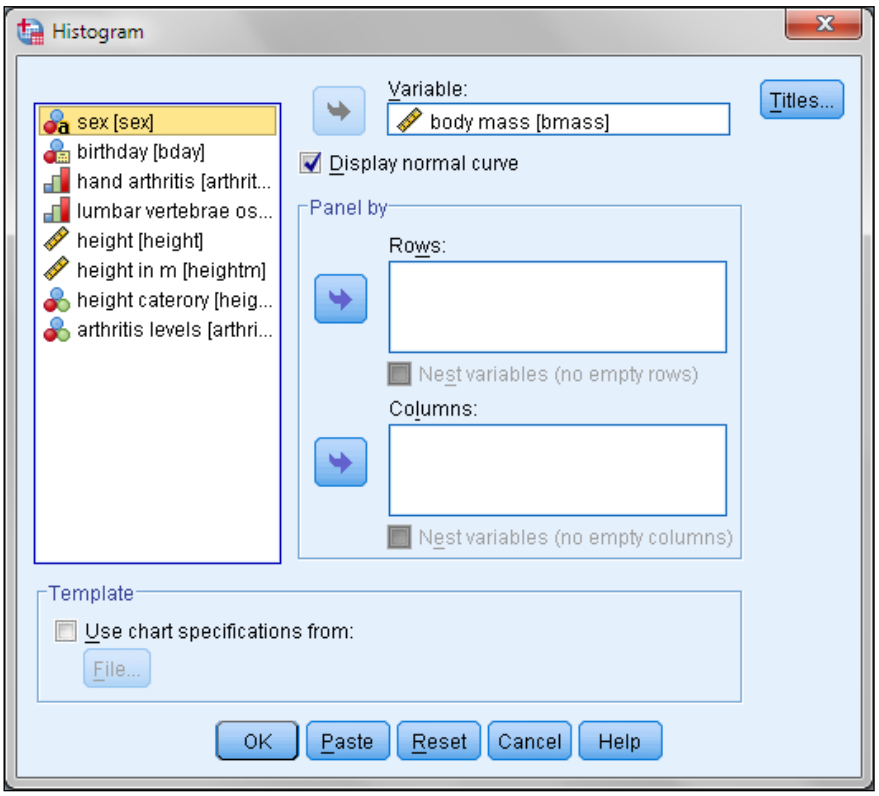
Είναι αντίστοιχο του ραβδογράμματος, μόνο που στον οριζόντιο άξονα τοποθετούμε όχι τα στοιχεία του δείγματος αλλά τις κλάσεις που δημιουργήσαμε. Έχει προταθεί ο αριθμός των κλάσεων να κυμαίνεται μεταξύ 5 και 25 ανάλογα με το μέγεθος του δείγματος. Μια άλλη πρόταση είναι ο αριθμός των κλάσεων να είναι ίσος με την τετραγωνική ρίζα των τιμών του δείγματος. Στο SPSS όμως ο αριθμός των κλάσεων υπολογίζεται από το πρόγραμμα ταυτόχρονα με την δημιουργία του ιστογράμματος.

Παράδειγμα

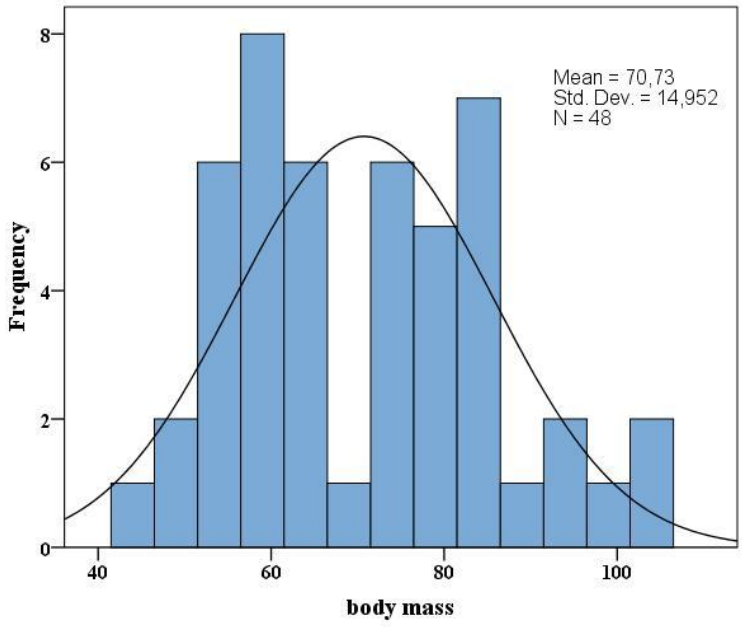
Να γίνει το ιστόγραμμα της μεταβλητής body mass του αρχείου osteological data.sav.

◆ Τονίζεται και πάλι ότι τα ιστογράμματα αφορούν scale μεταβλητές, ενώ τα ραβδογράμματα και τα κυκλικά γραφήματα nominal και ordinal μεταβλητές. Για να κατασκευάσουμε το ιστόγραμμα της μεταβλητής body mass ακολουθούμε τη διαδικασία *Graphs → Legacy Dialogs → Histogram.* Στο παράθυρο διαλόγου επιλέγουμε τη μεταβλητή body mass και ενεργοποιούμε την επιλογή *Display normal curve* (Σχήμα 2.10). Η έννοια και η σημασία της κανονικής καμπύλης,

δηλαδή της καμπύλης της κανονικής κατανομής, θα εξετασθεί στο επόμενο κεφάλαιο. Με κλικ στο *OK* παίρνουμε το ιστόγραμμα του Σχήματος 2.11.



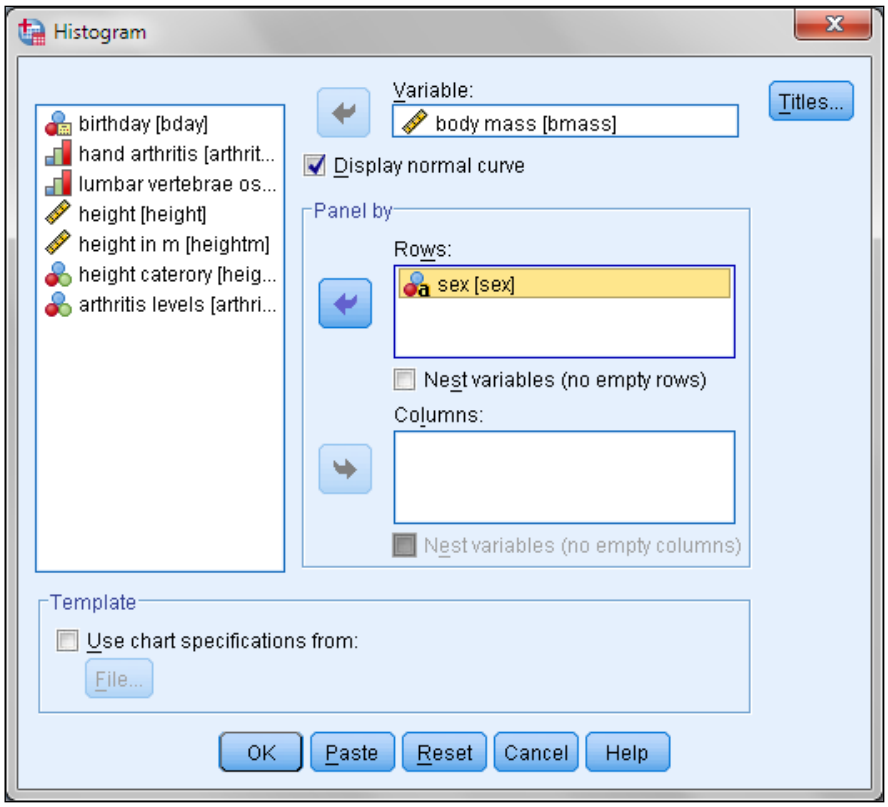
Σχήμα 2.10. Το παράθυρο διαλόγου *Histogram*



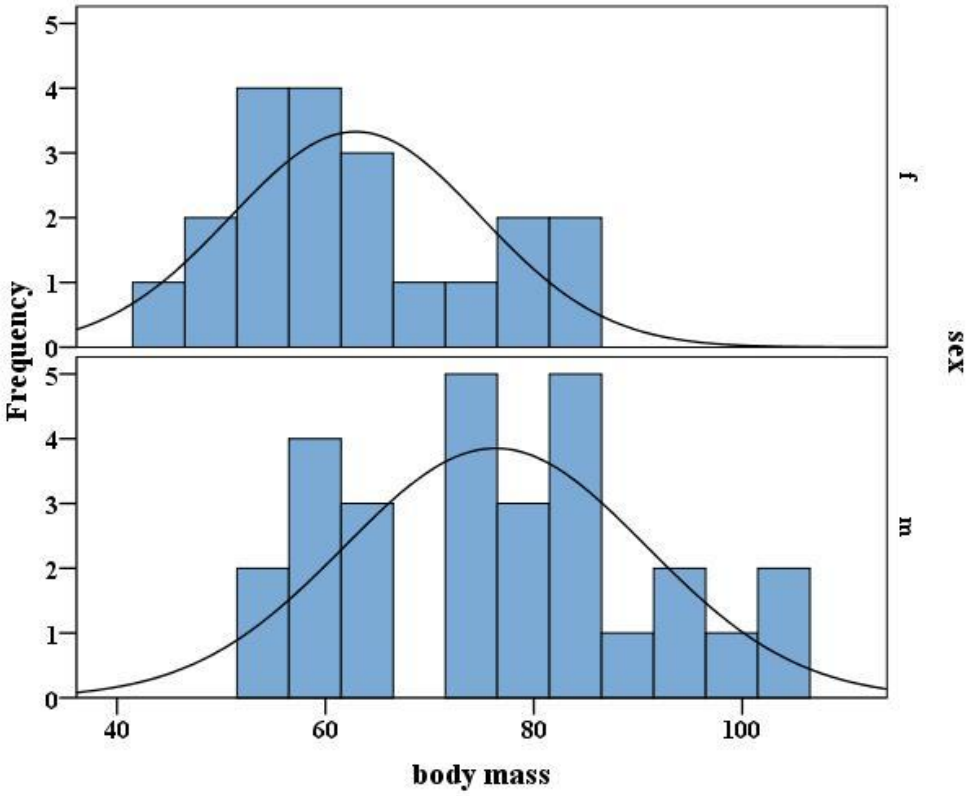
Σχήμα 2.11. Το ιστόγραμμα της μεταβλητής body mass

Τα ιστογράμματα γενικά μας δείχνουν εποπτικά πώς κατανέμονται οι τιμές ενός δείγματος γύρω από τη μέση τιμή ή την κορυφή. Όμως όταν το δείγμα είναι σχετικά μικρό, η εικόνα μπορεί να είναι πλασματική. Έτσι στο Σχήμα 2.11 παρατηρούμε ότι τα δεδομένα φαίνεται να ακολουθούν δύο κατανομές. Αυτό μπορεί να οφείλεται στο γεγονός ότι το δείγμα περιέχει τις τιμές του βάρους (εκτιμώμενου) τόσο των ανδρών όσο και των γυναικών. Μπορεί όμως η εικόνα του Σχήματος 2.11 να είναι και παραπλανητική επειδή το δείγμα είναι σχετικά μικρό.

Για να εξετάσουμε την πρώτη περίπτωση ξανακάνουμε το ιστόγραμμα, όμως αυτή τη φορά στο πλαίσιο Histogram εισάγουμε τη μεταβλητή sex στο πεδίο Panel by Rows (Σχήμα 2.12). Με αυτή την επιλογή θα γίνουν δύο ιστογράμματα, ένα για τους άνδρες και ένα για τις γυναίκες (Σχήμα 2.13).



Σχήμα 2.12. Το παράθυρο διαλόγου *Histogram*

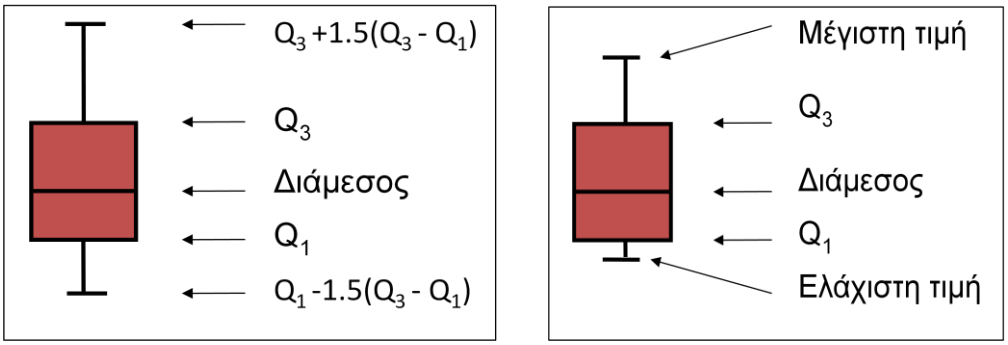


Σχήμα 2.13. Ιστογράμματα της μεταβλητής body mass ανάλογα με το φύλλο

Παρατηρούμε ότι πράγματι υπάρχουν δύο διαφορετικές κατανομές, όμως και πάλι λόγω του μικρού μεγέθους του δείγματος η εικόνα των ιστογραμμάτων δεν είναι καλή.

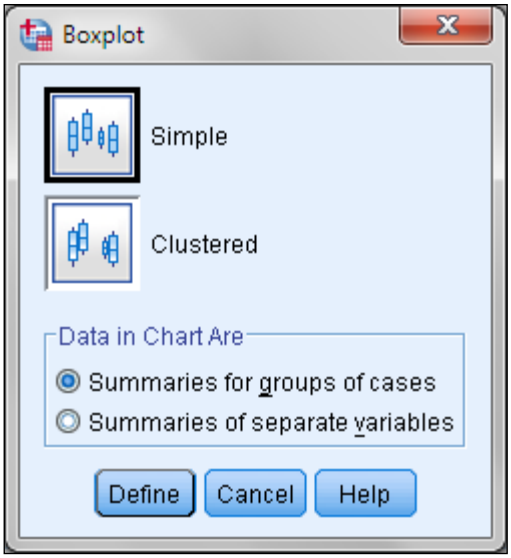
δ) Θηκόγραμμα (boxplot)

Το θηκόγραμμα απαρτίζεται από ένα ορθογώνιο με δύο κεραίες, μία στην κάτω βάση του ορθογωνίου και μία στην επάνω βάση του (Σχήμα 2.14). Η κάτω βάση του ορθογωνίου βρίσκεται στο Q_1 (πρώτο τεταρτημόριο) και η επάνω στο Q_3 (τρίτο τεταρτημόριο). Η διάμεσος αναπαρίσταται με ένα ευθύγραμμο οριζόντιο τμήμα στο εσωτερικό του ορθογωνίου. Οι κεραίες εκτείνονται μέχρι τις οριακές τιμές που μπορεί να είναι: α) η μέγιστη και η ελάχιστη τιμή του δείγματος, β) η μεγαλύτερη τιμή του δείγματος που είναι μικρότερη ή ίση από $Q_3 + 1.5(Q_3 - Q_1)$ και η μικρότερη τιμή του δείγματος που είναι μεγαλύτερη ή ίση από $Q_1 – 1.5(Q_3 - Q_1)$, γ) η μεγαλύτερη τιμή του δείγματος που είναι μικρότερη ή ίση από $Q_3 + 3(Q_3 - Q_1)$ και η μικρότερη τιμή του δείγματος που είναι μεγαλύτερη ή ίση από $Q_1 - 3(Q_3 - Q_1)$.

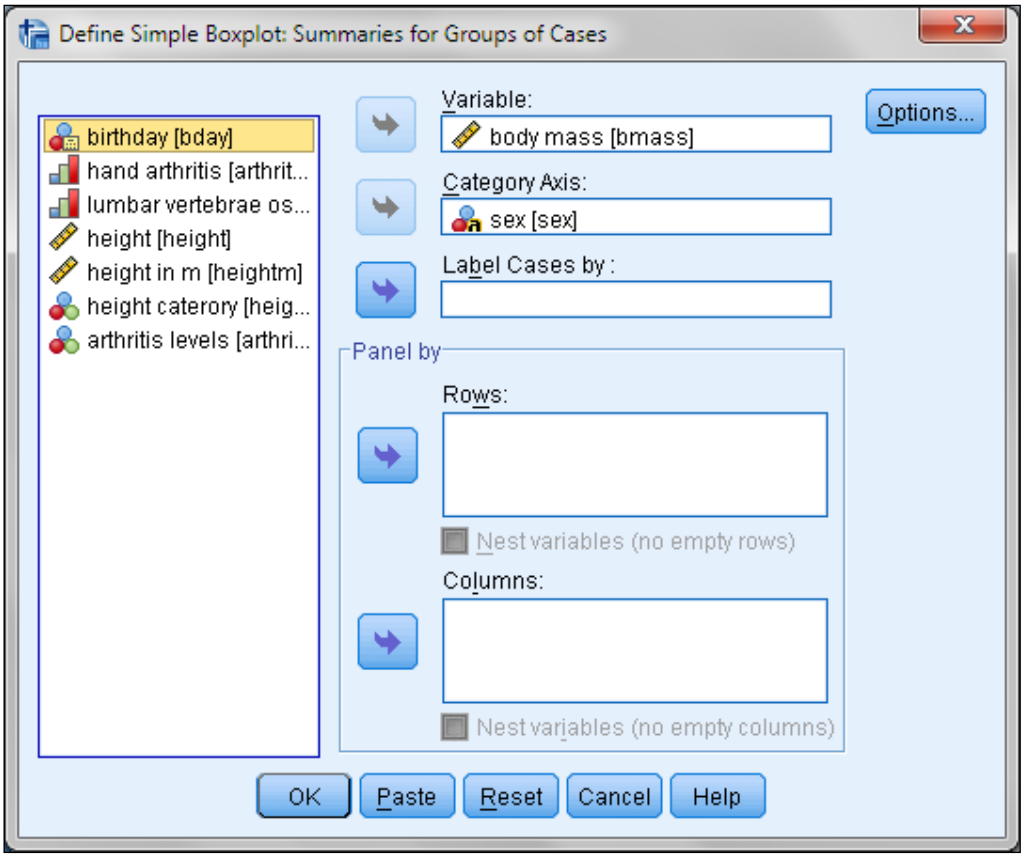


Σχήμα 2.14. *Θηκογράμματα*

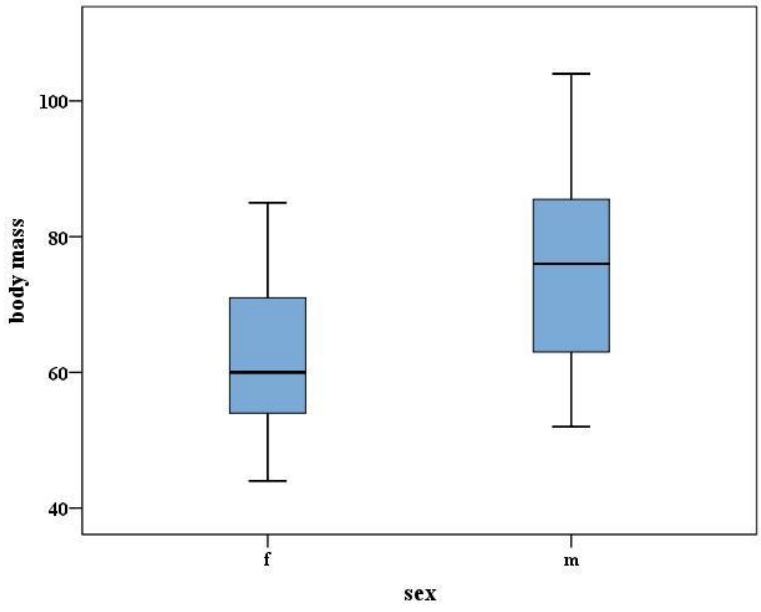
Τονίζεται και πάλι ότι τα θηκογράμματα αφορούν scale μεταβλητές. Παρόλα αυτά, για λόγους συνέχειας, θα σχηματίσουμε τα θηκογράμματα της μεταβλητής body mass με βάση το φύλλο. Για να τα κατασκευάσουμε ακολουθούμε τη διαδικασία *Graphs* → *Legacy Dialogs* → *Boxplot.* Στο παράθυρο διαλόγου *Boxplot* επιλέγουμε Simple, Summaries for groups of cases και κάνουμε κλικ στο Define (Σχήμα 2.15). Ακολούθως συμπληρώνουμε το παράθυρο διαλόγου *Define Simple Boxplot: Summaries for Groups of Cases* όπως στο Σχήμα 2.16 και πατάμε OK. Θα πάρουμε τα θηκογράμματα του Σχήματος 2.17.



Σχήμα 2.15. Το παράθυρο διαλόγου *Boxplot*



Σχήμα 2.16. Το παράθυρο διαλόγου *Define Simple Boxplot: Summaries for Groups of Cases*



Σχήμα 2.17. Θηκογράμματα της μεταβλητής body mass κατά φύλο

Παρατηρούμε τη διαφοροποίηση των θηκογραμμάτων με βάση το φύλλο. Όπως αναμένεται, οι γυναίκες έχουν σημαντικά μικρότερο σωματικό βάρος.

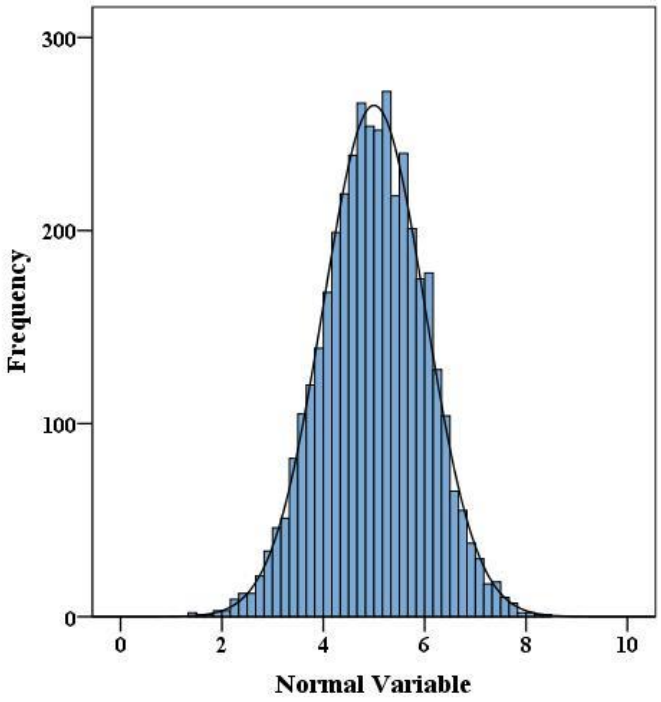
Τα θηκογράμματα είναι ιδιαίτερα χρήσιμα στη σύγκριση δειγμάτων, όταν αυτά δεν ακολουθούν την κανονική κατανομή. Το θέμα αυτό αναλύεται στα επόμενα κεφάλαια.

3. Η ΕΝΝΟΙΑ ΤΗΣ ΚΑΤΑΝΟΜΗΣ -ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

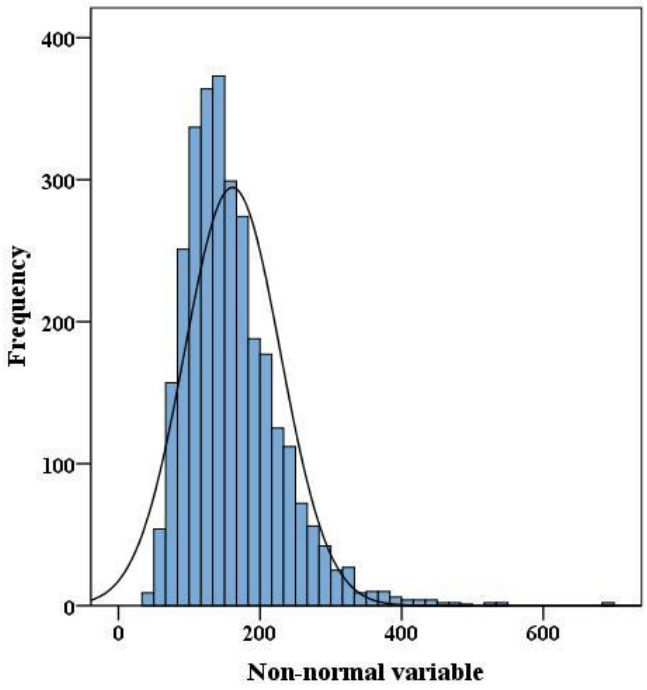
3.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΝΝΟΙΑ ΤΗΣ ΣΥΝΑΡΤΗΣΗΣ ΚΑΤΑΝΟΜΗΣ

Όπως αναφέρθηκε, τα ιστογράμματα δείχνουν εποπτικά πώς κατανέμονται οι τιμές ενός δείγματος. Αυτό όμως ισχύει μόνο όταν το δείγμα έχει πάρα πολλές τιμές, ίσως περισσότερες από 1000. Επιπλέον, τα ιστογράμματα μας εισάγουν με εποπτικό τρόπο στην πολύ βασική έννοια της κατανομής και της συνάρτησης κατανομής.

Έστω τα ιστογράμματα των Σχημάτων 3.1 και 3.2. Παρατηρούμε ότι η κατανομή των τιμών στα δύο δείγματα είναι διαφορετική. Στο πρώτο δείγμα η κατανομή είναι συμμετρική ως προς την μέση τιμή, ενώ στο δεύτερο είναι ασύμμετρη. Σε αυτή την περίπτωση λέμε ότι τα ιστογράμματα προέρχονται από διαφορετικές **κατανομές**.



Σχήμα 3.1. Ιστόγραμμα δείγματος 3000 τιμών που ακολουθούν την κανονική κατανομή με μ=5 και σ=1



Σχήμα 3.2. Ιστόγραμμα δείγματος 3000 τιμών που αποκλίνουν από την κανονική κατανομή

Η συστηματική μελέτη των ιστογραμμάτων έδειξε ότι δείγματα τιμών που προέρχονται από μετρήσεις που γίνονται σε ένα συγκεκριμένο σύστημα κάτω από σταθερές συνθήκες έχουν ιστογράμματα συμμετρικά, όπως του Σχήματος 3.1. Κάθε δείγμα που έχει αυτή την ιδιότητα ονομάζεται **κανονικό** δείγμα ή δείγμα με τιμές που ακολουθούν την **κανονική κατανομή**.

Αν οι συνθήκες δειγματοληψίας διαφέρουν από τις παραπάνω, τότε το δείγμα αρχίζει να αποκλίνει από την κανονικότητα με αποτέλεσμα το ιστόγραμμά του να γίνεται ασύμμετρο και να μην περιγράφεται από τη συνάρτηση της κανονικής κατανομής, όπως το ιστόγραμμα του Σχήματος 3.2. Σε κάθε περίπτωση όμως θα υπάρχει μια συνάρτηση η οποία τα περιγράφει το ιστόγραμμα. Η συνάρτηση αυτή ονομάζεται γενικά **συνάρτηση πυκνότητας πιθανότητας της κατανομής**.

3.2 ΒΑΣΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Υπάρχει μια πολύ μεγάλη ποικιλία κατανομών που ισχύουν σε διάφορα δείγματα. Από αυτές οι πιο χρήσιμες είναι:

- Διωνυμική κατανομή

- Κατανομή Poisson
- Κανονική κατανομή
- Τυπικά κανονική κατανομή
- Κατανομή Student ή t
- Κατανομή χι-τετράγωνο
- Κατανομή Fisher ή F

Λόγω της πληθώρας των κατανομών, όταν έχουμε ένα δείγμα δεν γνωρίζουμε πάντα την κατανομή που ακολουθούν οι τιμές του και συνεπώς δεν γνωρίζουμε ποια είναι η συνάρτηση πυκνότητας πιθανότητας με εξαίρεση ορισμένες μόνο περιπτώσεις, όπως για παράδειγμα:

- Όταν εκτελούμε ένα πείραμα τύχης που έχει δύο μόνο δυνατά αποτελέσματα με πιθανότητες p και q=1-p, αντίστοιχα, τότε το δείγμα που προκύπτει ακολουθεί τη διωνυμική κατανομή.
- Τα σπάνια γεγονότα ακολουθούν την κατανομή Poisson.
- Τα πειραματικά δεδομένα, κυρίως μετρήσεις που γίνονται κάτω από σταθερές και ελεγχόμενες συνθήκες, ακολουθούν την κανονική κατανομή.

3.3 ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ

Πολλές στατιστικές αναλύσεις απαιτούν την κανονικότητα των τιμών των δειγμάτων που αναλύονται. Έτσι, ο έλεγχος της κανονικότητας πρέπει να είναι ο πρώτος και ίσως ο βασικότερος έλεγχος για μια σωστή στατιστικά ανάλυση των δεδομένων ενός πειράματος. Οι βασικοί έλεγχοι είναι τα κριτήρια Kolmogorov-Smirnov και Shapiro-Wilk. Το SPSS όχι μόνο υπολογίζει τα κριτήρια αυτά αλλά υπολογίζει και την πιθανότητα να κάνουμε λάθος αν δεχτούμε ότι τα δεδομένα του δείγματος δεν ακολουθούν την κανονική κατανομή. Η πιθανότητα αυτή στο SPSS συμβολίζεται με Sig. Συνήθως όταν το Sig. έχει τιμές μεγαλύτερες από 0.05 δεχόμαστε ότι ισχύει η κανονική κατανομή για τις τιμές του δείγματος. Η σημασία του Sig. σε στατιστικούς ελέγχους εξετάζεται πιο διεξοδικά στο επόμενο κεφάλαιο.

Για τον έλεγχο της κανονικότητας με το SPSS ακολουθούμε την πορεία: *Analyze* → *Descriptive Statistics* → *Explore* και στο παράθυρο διαλόγου εισάγουμε τη μεταβλητή που μελετάμε στο πλαίσιο *Dependent List*. Το πρόγραμμα *Explore*

εκτός από την εφαρμογή των κριτηρίων Kolmogorov-Smirnov και Shapiro-Wilk προσφέρει αρκετές επιλογές και συγκεκριμένα:

Από το *Statistics* έχουμε τις δυνατότητες:

- Descriptives: Υπολογίζει τα κυριότερα στατιστικά μέτρα.
- Outliers: Υπολογίζει τις 5 μεγαλύτερες και 5 μικρότερες τιμές.

*Από το Plots μ*πορούμε να κατασκευάσουμε τα διαγράμματα: Boxplots, Histograms και Normality plots with tests.

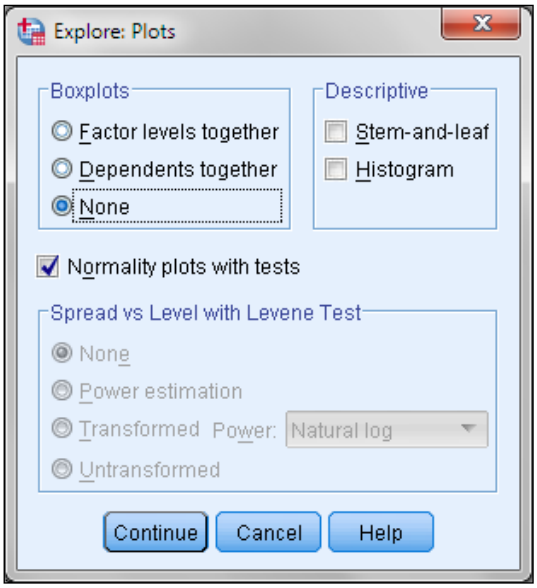
Από το *Options* μπορούμε να χειριστούμε τις απούσες τιμές με βάση τις επιλογές:

- Exclude cases listwise: Στους υπολογισμούς χρησιμοποιούνται μόνο οι περιπτώσεις που είναι ταυτόχρονα έγκυρες σε όλες τις μεταβλητές που υπάρχουν στις Dependent List και Factor List.
- Exclude cases pairwise: Στους υπολογισμούς χρησιμοποιούνται όλες οι περιπτώσεις που είναι έγκυρες για κάθε μεταβλητή που υπάρχει στην *Dependent List*.

Παράδειγμα

Να ελεγχθεί η κανονικότητα των τιμών της μεταβλητής height του αρχείου osteological data.sav.

◆ Ανοίγουμε το αρχείο osteological data.sav και ακολουθούμε τη διαδικασία *Analyze → Descriptive Statistics → Explore*. Στο παράθυρο διαλόγου εισάγουμε τη μεταβλητή height στο πλαίσιο *Dependent List* και κάνουμε κλικ στο κουμπί *Plots*. Στο πλαίσιο διαλόγου που εμφανίζεται κάνουμε κλικ στην επιλογή *None* στο πάνελ των *Boxplots,* απενεργοποιούμε την επιλογή *Stem-and-leaf* στο πάνελ Descriptive και επιλέγουμε μόνο το *Normality plots with tests* (Σχήμα 3.3).



Σχήμα 3.3. Παράθυρο διαλόγου Explore: Plots

Από τα αποτελέσματα που παίρνουμε ενδιαφέρον έχει ο πίνακας *Tests of Normality* (Πίνακας 3.1) και το διάγραμμα Q-Q που δίνεται στο Σχήμα 3.4. Στον πίνακα 3.1 η ποσότητα Sig. (significance) είναι η πιθανότητα να κάνουμε λάθος αν αποδεχθούμε ότι τα δεδομένα του δείγματος δεν ακολουθούν την κανονική κατανομή. Γενικά, όπως αναφέρθηκε, όταν η Sig. είναι μεγαλύτερη από 0.05 δεχόμαστε ότι η κατανομή είναι κανονική. Άρα στο δείγμα που εξετάζουμε οι τιμές ακολουθούν την κανονική κατανομή δεδομένου ότι ισχύει Sig.(Kolmogorov-Smirnov) =0.2 και Sig.(Shapiro-Wilk) = 0.095.

Πίνακας 3.1. Αποτελέσματα ελέγχου κανονικότητας

Tests of Normality

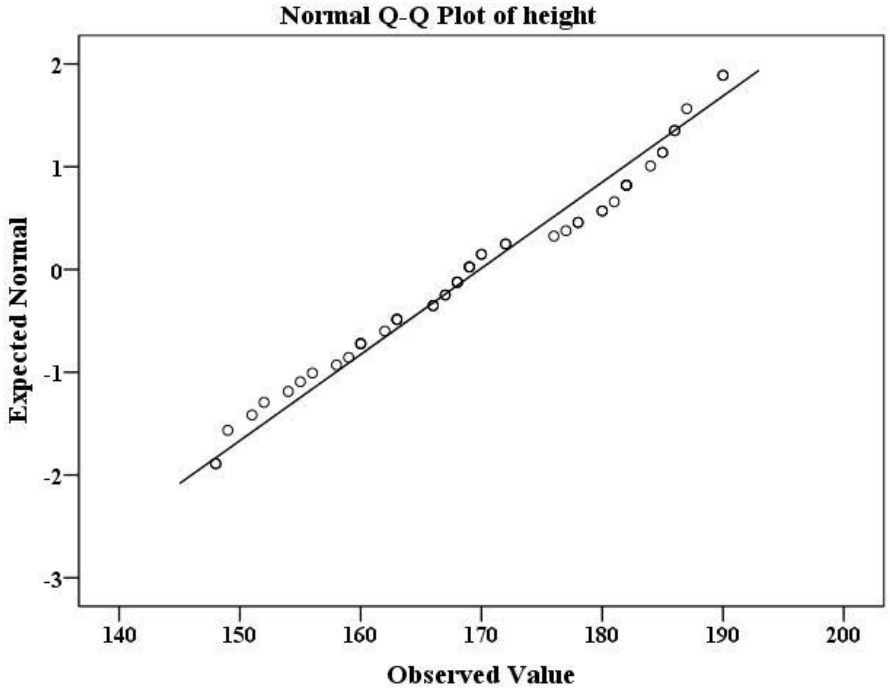
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
height	,102	50	,200*	,961	50	,095

a. Lilliefors Significance Correction

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

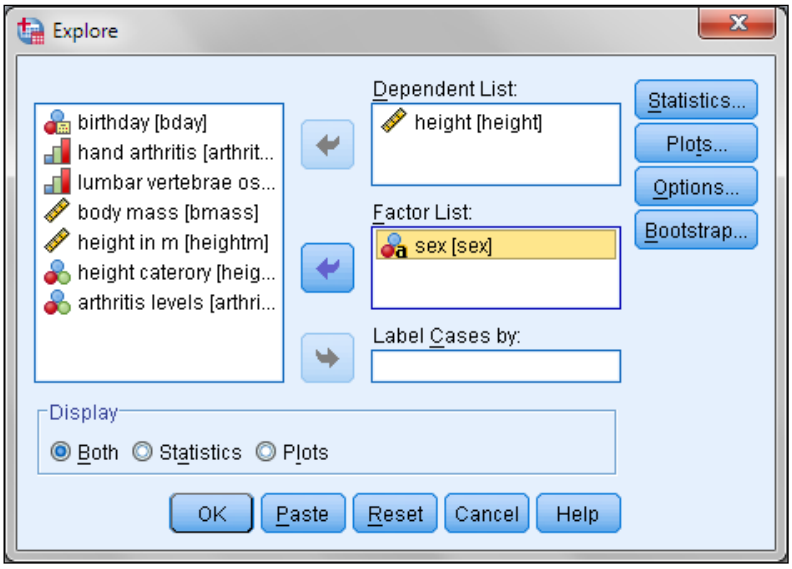
Το ίδιο συμπέρασμα προκύπτει και από το διάγραμμα *Normal Q-Q (quartile-quartile) plot* του Σχήματος 3.4. Για να είναι κανονικό το δείγμα θα πρέπει στο

διάγραμμα αυτό όλα τα σημεία είναι πάνω στην ευθεία. Παρατηρούμε ότι αυτό συμβαίνει για τη μεταβλητή που εξετάζουμε. Πάντως κατά κανόνα όταν έχουμε λίγα σημεία είναι δυνατόν το *Q-Q* διάγραμμα να μην αποδώσει την πραγματικότητα. Γι αυτό και κυρίως στηριζόμαστε στα αποτελέσματα του πίνακα *Tests of Normality*.



Σχήμα 3.4. Διάγραμμα Q-Q για τον έλεγχο της κανονικής κατανομής

Η μεταβλητή height αποτελείται από τις εκτιμώμενες τιμές ύψους των ανδρών και των γυναικών του δείγματος. Αν θέλουμε να εξετάσουμε την κανονικότητα των τιμών κάθε φύλου χωριστά εργαζόμαστε ως εξής. Ακολουθούμε τη διαδικασία *Analyze* → *Descriptive Statistics* → *Explore* και συμπληρώνουμε το παράθυρο διαλόγου που ανοίγει όπως στο Σχήμα 3.5. Παίρνουμε τα αποτελέσματα του Πίνακα 3.2, από τον οποίο προκύπτει ότι και οι τιμές ύψους του κάθε φύλου χωριστά ακολουθούν την κανονική κατανομή.



Σχήμα 3.5. Παράθυρο διαλόγου Explore

Πίνακας 3.2. Αποτελέσματα ελέγχου κανονικότητας

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	sex	Statistic	df	Sig.	Statistic	df	Sig.
height	f	,110	21	,200*	,941	21	,231
	m	,121	29	,200*	,945	29	,139

a. Lilliefors Significance Correction

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

3.4 ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Όταν δημιουργούμε ένα δείγμα δεν μας ενδιαφέρει άμεσα ούτε η μέση τιμή $\bar{x}$ ούτε η τυπική απόκλιση s στο δείγμα των μετρήσεών μας. Εκείνο που μας ενδιαφέρει είναι να προσδιορίσουμε ή έστω να εκτιμήσουμε την **πραγματική τιμή** μ της μεταβλητής x που μελετάμε και της τυπικής απόκλισης σ, δηλαδή τη μέση τιμή μ και την τυπική απόκλιση σ του πληθυσμού από τον οποίον προέρχεται το δείγμα. Σ' αυτές τις περιπτώσεις χρησιμοποιούμε **τα διαστήματα εμπιστοσύνης**.

Συγκεκριμένα ονομάζεται **P% διάστημα εμπιστοσύνης** (confidence interval) μιας παραμέτρου θ του πληθυσμού, το διάστημα (δ_1, δ_2) μέσα στο οποίο αναμένεται να υπάρχει η θ με πιθανότητα P%. Συνήθως η πιθανότητα P%

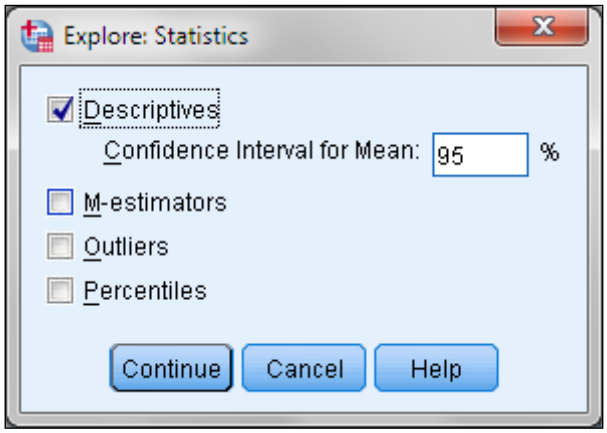
γράφεται ως: P = 100(1-α). Σ' αυτή την περίπτωση το α εκφράζει τον **κίνδυνο σφάλματος**, δηλαδή την πιθανότητα η παράμετρος θ του πληθυσμού να βρίσκεται έξω από το διάστημα εμπιστοσύνης.

Από τα διαστήματα εμπιστοσύνης εκείνο που χρησιμοποιείται ευρύτατα είναι το 95% διάστημα εμπιστοσύνης της μέσης τιμής. Δηλαδή το διάστημα (δ_1, δ_2) μέσα στο οποίο αναμένεται να υπάρχει η μέση τιμή του πληθυσμού, δηλαδή η πραγματική μέση τιμή, με πιθανότητα 95%. Το διάστημα αυτό υπολογίζεται μόνο σε δείγματα που ακολουθούν την κανονική κατανομή, όπως περιγράφεται στο παρακάτω παράδειγμα.

Παράδειγμα

Να υπολογιστεί το 95% διάστημα εμπιστοσύνης για τη μέση τιμή των τιμών της μεταβλητής height του αρχείου osteological data.sav.

◆ Ανοίγουμε το αρχείο osteological data.sav και ακολουθούμε τη διαδικασία *Analyze* → *Descriptive Statistics* → *Explore*. Συμπληρώνουμε το παράθυρο διαλόγου όπως στο Σχήμα 3.5 και πατάμε στο *Statistics*. Στο παράθυρο διαλόγου που ανοίγει επιλέγουμε *Descriptives*, αν δεν είναι επιλεγμένο, και ορίζουμε τα όρια του διαστήματος εμπιστοσύνης, 95% (Σχήμα 3.6). Πατάμε *Continue* και *OK*. Τα διαστήματα εμπιστοσύνης παρουσιάζονται στον πίνακα Descriptives (Πίνακας 3.3).



Σχήμα 3.6. *Επιλογή διαστήματος εμπιστοσύνης*

Πίνακας 3.3. Αποτελέσματα περιγραφικών μέτρων και διαστημάτων εμπιστοσύνης

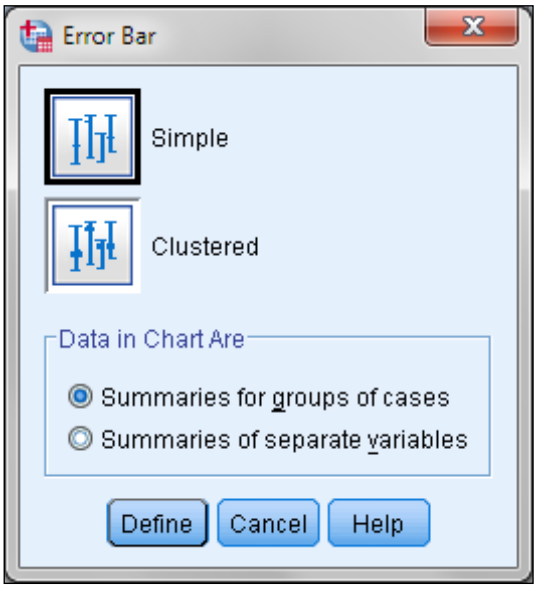
Descriptives

	sex			Statistic	Std. Error
height	f	Mean		161,52	2,246
		95% Confidence Interval for Mean	Lower Bound	156,84	
			Upper Bound	166,21	
		5% Trimmed Mean		160,98	
		Median		162,00	
		Variance		105,962	
		Std. Deviation		10,294	
		Minimum		148	
		Maximum		185	
		Range		37	
		Interquartile Range		15	
		Skewness		,707	,501
		Kurtosis		,208	,972
	m	Mean		175,90	1,698
		95% Confidence Interval for Mean	Lower Bound	172,42	
			Upper Bound	179,37	
		5% Trimmed Mean		176,03	
		Median		178,00	
		Variance		83,596	
		Std. Deviation		9,143	
		Minimum		159	
		Maximum		190	
		Range		31	
		Interquartile Range		15	
		Skewness		-,274	,434
		Kurtosis		-,989	,845

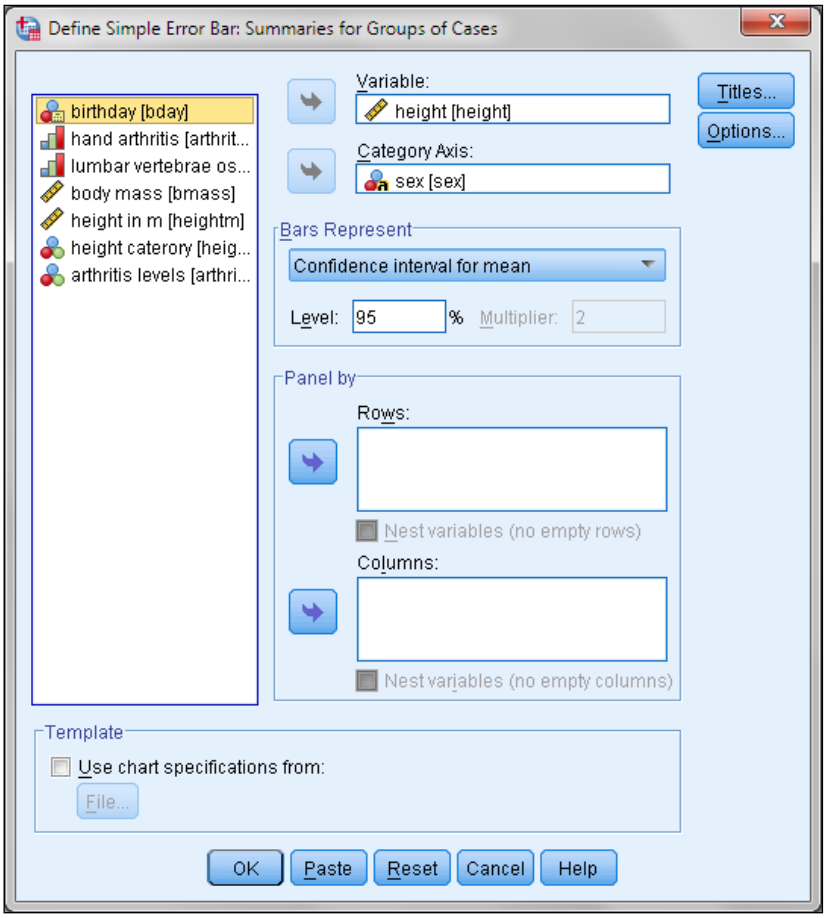
Παρατηρούμε ότι με πιθανότητα 95% η πραγματική μέση τιμή για το ύψος των γυναικών είναι στο διάστημα μεταξύ 156,84 και 166,21 cm, ενώ των ανδρών είναι μεταξύ 172,42 και 179,37 cm.

3.5 ΔΙΑΓΡΑΜΜΑΤΑ ΔΙΑΣΤΗΜΑΤΩΝ

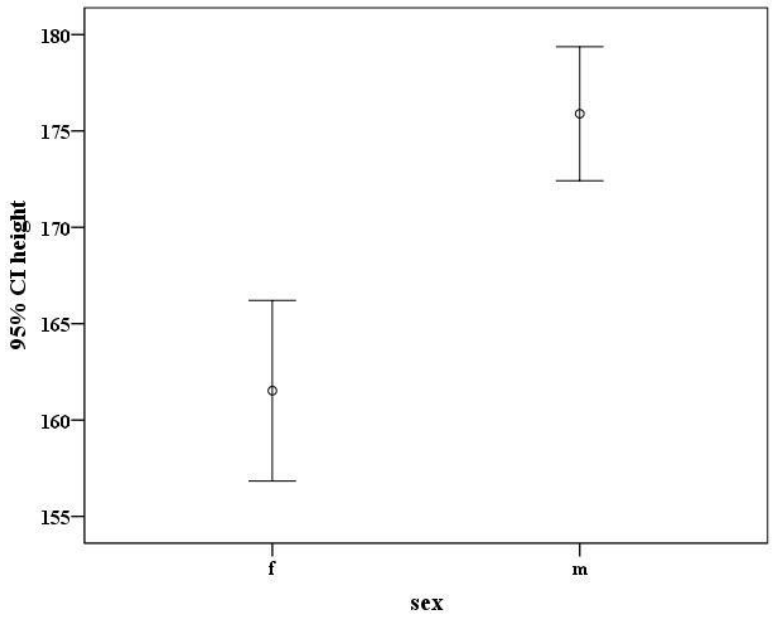
Με τα διαγράμματα διαστημάτων βλέπουμε εποπτικά τα διαστήματα εμπιστοσύνης. Για να σχηματίσουμε ένα τέτοιο διάγραμμα πηγαίνουμε *Graphs* → *Legacy Dialogs* → *Error Bar*. Έστω ότι θέλουμε να δημιουργήσουμε το διάγραμμα διαστημάτων εμπιστοσύνης του προηγούμενου παραδείγματος. Στο παράθυρο διαλόγου που ανοίγει κάνουμε κλικ στο *Simple*, επιλέγουμε *Summaries for groups of cases* (Σχήμα 3.7) και πατάμε στο *Define*. Ακολούθως συμπληρώνουμε το παράθυρο διαλόγου *Define simple Error Bar: Summaries of Groups of Cases*, όπως στο Σχήμα 3.8. Το διάγραμμα που παίρνουμε δίνεται στο Σχήμα 3.9 και μας δίνει εποπτικά την ίδια πληροφορία με τα διαστήματα εμπιστοσύνης που υπολογίσαμε στην προηγούμενη ενότητα.



Σχήμα 3.7. Το παράθυρο διαλόγου *Error Bar*



Σχήμα 3.8. Πλαίσιο διαλόγου Define simple Error Bar: Summaries of Groups of Cases



Σχήμα 3.9. Διάγραμμα με 95% διαστήματα εμπιστοσύνης της μεταβλητής height

4. ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ

4.1 ΓΕΝΙΚΑ

Στην πράξη καλούμαστε συχνά να πάρουμε αποφάσεις σχετικά με την πιθανότητα να συμβεί ή να μη συμβεί ένα γεγονός. Οι αποφάσεις αυτές λέγονται **στατιστικές αποφάσεις**. Για να λάβουμε στατιστικές αποφάσεις είναι απαραίτητο να κάνουμε υποθέσεις. Μια πολύ βασική υπόθεση, που ονομάζεται **μηδενική υπόθεση** (null hypothesis) και συμβολίζεται με H_0, δέχεται ότι οι διαφορές σε δύο ή περισσότερα δείγματα οφείλονται μόνο σε τυχαία σφάλματα, δηλαδή δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων. Μια εναλλακτική υπόθεση της H_0 συμβολίζεται με H_1.

Αν με βάση τα στατιστικά δεδομένα απορρίψουμε μια υπόθεση που είναι αληθινή, τότε λέμε ότι κάνουμε ένα **σφάλμα τύπου Ι**. Αντίθετα αν δεχθούμε μια λανθασμένη υπόθεση, τότε κάνουμε ένα **σφάλμα τύπου ΙΙ**. Δυστυχώς όταν προσπαθούμε να περιορίσουμε ένα σφάλμα τύπου Ι αυξάνουμε την πιθανότητα να κάνουμε ένα σφάλμα τύπου ΙΙ. Η μόνη περίπτωση να ελαττώσουμε την πιθανότητα να κάνουμε σφάλμα τύπου Ι και ΙΙ είναι να αυξήσουμε το μέγεθος των δειγμάτων που μελετάμε.

Ονομάζουμε **επίπεδο ή στάθμη σημαντικότητας** (significant level) τη μέγιστη πιθανότητα με την οποία δεχόμαστε να κάνουμε σφάλμα τύπου Ι όταν εξετάζουμε μια στατιστική υπόθεση. Η πιθανότητα αυτή συμβολίζεται με α και οι τιμές που συνήθως χρησιμοποιούμε είναι α = 0.05 ή α = 0.01. Αυτό σημαίνει ότι η πιθανότητα να απορρίψουμε μια σωστή υπόθεση είναι μικρότερη από 5% όταν α = 0.05 και μικρότερη από 1% όταν α = 0.01.

Θα πρέπει να τονιστεί ότι οι στατιστικοί έλεγχοι μας επιτρέπουν να παίρνουμε αποφάσεις στηριζόμενοι σε πιθανότητες, δεν αποδεικνύουν όμως την απόλυτη ισχύ ή όχι μιας υπόθεσης. Επίσης οι στατιστικοί έλεγχοι απαντούν με τρόπο θετικό μόνο στην απόρριψη της μηδενικής υπόθεσης. Έτσι όταν απορρίπτουμε τη H_0 υπάρχει μια πιθανότητα μικρότερη από α% να είναι ορθή. Αν όμως τα στοιχεία του

δείγματος είναι τέτοια, ώστε να δεχτούμε τη H_0 στο επίπεδο σημαντικότητας α, τότε δεν μπορούμε να εκτιμήσουμε τον κίνδυνο να έχουμε κάνει λάθος. Επίσης, όταν απορρίπτουμε τη H_0 σε επίπεδο σημαντικότητας α%, είναι σφάλμα να συμπεράνουμε ότι η H_0 είναι ορθή με πιθανότητα 1-α/100.

Το SPSS σε κάθε έλεγχο σημαντικότητας υπολογίζει την **p-value, δηλαδή την πιθανότητα να κάνουμε λάθος απορρίπτοντας τη μηδενική υπόθεση**. Συνεπώς αν έχουμε επιλέξει το επίπεδο σημαντικότητας α (0.05 ή 0.01), ισχύει:

- Αν $p < \alpha$ τότε η H_0 απορρίπτεται
- Αν $p > \alpha$ τότε η H_0 δεν απορρίπτεται

Στο SPSS οι p τιμές βρίσκονται στη στήλη των αποτελεσμάτων με τίτλο Sig. (Significance).

Όλοι οι έλεγχοι στατιστικών υποθέσεων προϋποθέτουν ότι τα δεδομένα του δείγματος ακολουθούν την κανονική κατανομή. Συνεπώς ο πρώτος έλεγχος που πρέπει να γίνεται είναι ο έλεγχος της κανονικότητας, όπως περιγράφηκε παραπάνω. Στα παραδείγματα αυτού του κεφαλαίου θα υποθέσουμε ότι ο έλεγχος αυτός έχει γίνει και έδειξε ότι όλα τα δείγματα ή ακολουθούν την κανονική κατανομή ή δεν παρουσιάζουν σημαντικές αποκλίσεις από αυτή.

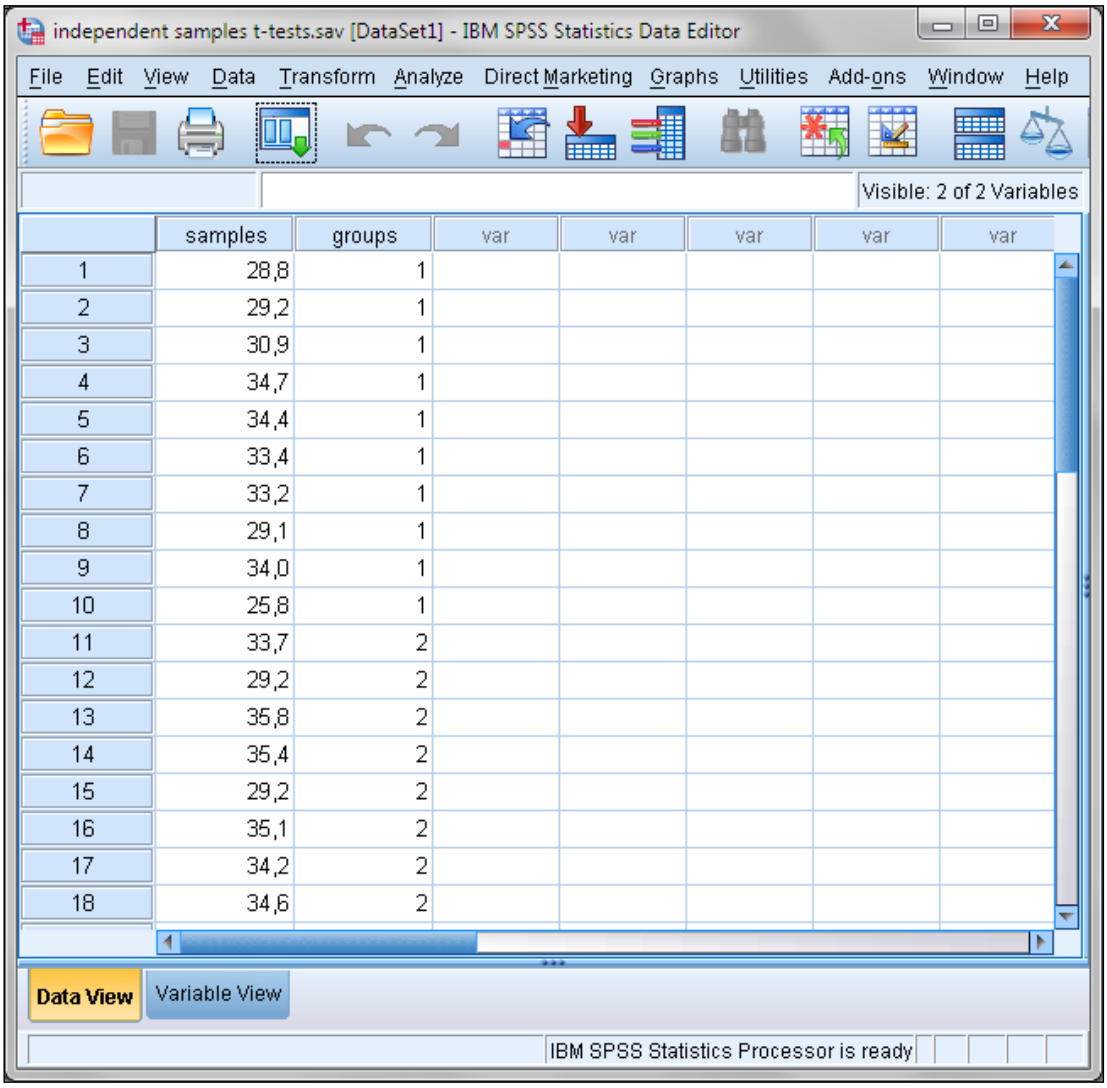
4.2 ΔΙΑΦΟΡΕΣ ΜΕΣΩΝ ΤΙΜΩΝ ΔΕΙΓΜΑΤΩΝ
(Independent samples t-tests)

Σε πολλές περιπτώσεις, θέλουμε να συγκρίνουμε δύο ανεξάρτητα δείγματα και να δούμε εάν υπάρχει στατιστικά σημαντική διαφορά μεταξύ τους. Για παράδειγμα, μπορεί να θέλουμε να συγκρίνουμε το ύψος των μαθητών μιας τάξης με το ύψος των μαθητών μιας άλλης τάξης. Για το σκοπό αυτό συγκρίνουμε τις μέσες τιμές των δειγμάτων.

Παράδειγμα

Ο προσδιορισμός του πλάτους της λεκάνης 10 ενήλικων ανδρών και 8 γυναικών έδωσαν τα αποτελέσματα των δειγμάτων 1 και 2, αντίστοιχα, στο Σχήμα 4.1 (σε cm). Να εξετασθεί αν οι τιμές των δύο δειγμάτων παρουσιάζουν στατιστικά σημαντική απόκλιση σε επίπεδο σημαντικότητας 0.05.

◆ Για να μπορέσουμε να αναλύσουμε τα δύο αυτά δείγματα στο SPSS τα μεταφέρουμε σε **μια** στήλη, που την ονομάζουμε έστω samples. Στη διπλανή στήλη, που την ονομάζουμε groups, χρησιμοποιούμε τους αριθμούς 1 και 2 για να διακρίνουμε τα δύο δείγματα, όπως φαίνεται στο Σχήμα 4.1.



Σχήμα 4.1. Δεδομένα παραδείγματος

Για να χρησιμοποιήσουμε το SPSS στον παραπάνω έλεγχο, ακολουθούμε τη διαδικασία *Analyze → Compare Means → Independent-Samples T Test*, στο παράθυρο που ανοίγει μεταφέρουμε τη μεταβλητή Samples στο πλαίσιο *Test Variable(s)*, τη μεταβλητή Groups στο *Grouping Variable* και κάνουμε κλικ στο *Define Groups*. Στο νέο παράθυρο εισάγουμε την τιμή 1 στο *Group 1* και την τιμή

2 στο *Group 2*. Με κλικ στο *Continue* και στο *OK* παίρνουμε τους Πίνακες 4.1 και 4.2.

Πίνακας 4.1. Γενικά περιγραφικά στατιστικά στοιχεία για τα δείγματα

Group Statistics

	groups	N	Mean	Std. Deviation	Std. Error Mean
samples	1	10	31,350	3,0252	,9566
	2	8	33,400	2,6753	,9459

Πίνακας 4.2. Τμήμα αποτελεσμάτων στατιστικού ελέγχου

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means				
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Equal variances assumed	,567	,462	-1,502	16	,153	-2,0500	1,3648
Equal variances not assumed			-1,524	15,793	,147	-2,0500	1,3453

Παρατηρούμε ότι ο στατιστικός έλεγχος διασπορών με το κριτήριο Levene δίνει την τιμή $p = 0.462 > 0.05$ που δείχνει ότι η H_0 δεν απορρίπτεται. Συνεπώς δεν υπάρχει στατιστικά σημαντική διαφορά στις διασπορές των δύο δειγμάτων. Με βάση αυτό το αποτέλεσμα στον παραπάνω πίνακα ισχύει η επάνω γραμμή (Equal variances assumed). Αν υπήρχε στατιστικά σημαντική διαφορά στις διασπορές, τότε θα εξετάζαμε τα αποτελέσματα της κάτω γραμμής (Equal variances not assumed) στον παραπάνω πίνακα.

Από τα αποτελέσματα της επάνω γραμμής παίρνουμε για τη μηδενική υπόθεση την τιμή $p = 0.153 > 0.05$ που δείχνει ότι η H_0 δεν απορρίπτεται.

Επομένως σε επίπεδο σημαντικότητας 0,05 η μηδενική υπόθεση ισχύει: Δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δύο δειγμάτων.

4.3 ΣΥΓΚΡΙΣΗ ΖΕΥΓΩΝ ΔΕΙΓΜΑΤΩΝ
(Paired samples t-tests)

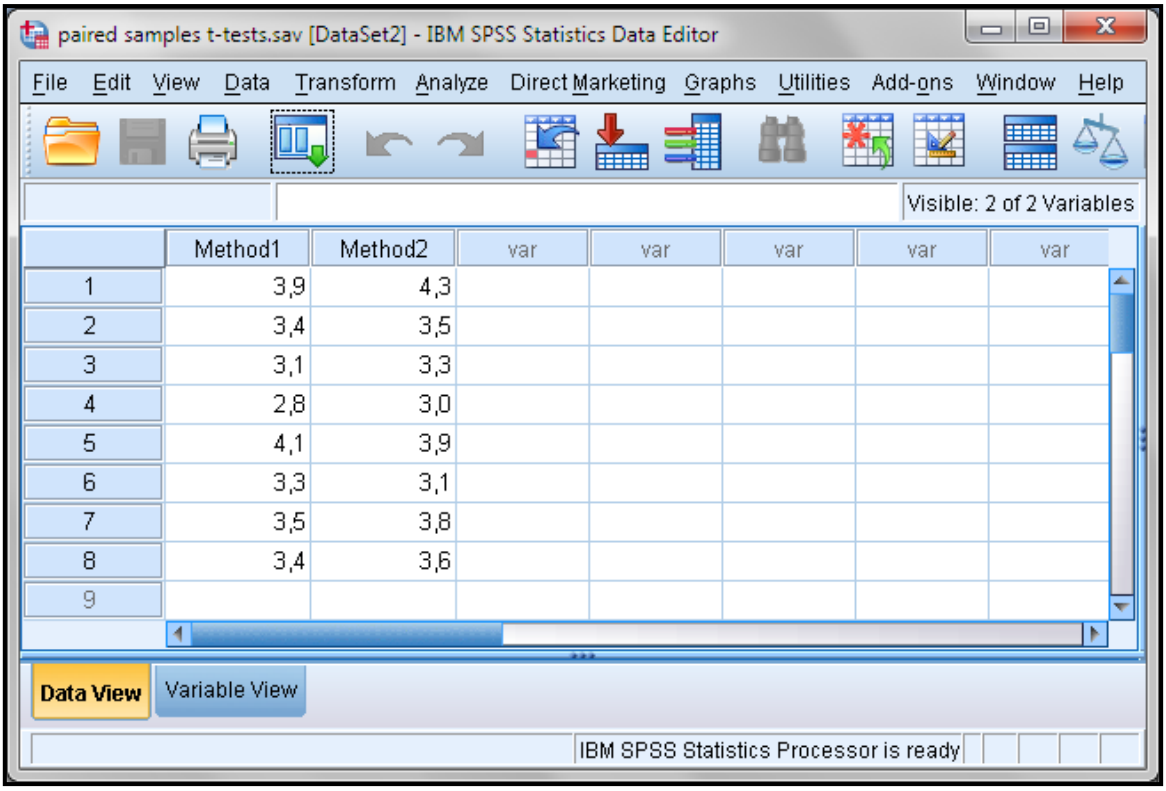
Δύο δείγματα σχηματίζουν ένα ζεύγος αν υπάρχει ένα προς ένα αντιστοιχία μεταξύ των τιμών των δειγμάτων. Για παράδειγμα, στον πίνακα του Σχήματος 4.2 χρονολογούνται 8 ταφικά αντικείμενα με δύο διαφορετικές μεθόδους. Οι τιμές αυτές σχηματίζουν ένα ζεύγος δειγμάτων.

Σ' αυτές τις περιπτώσεις το κύριο ερώτημα που εγείρεται είναι αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μετρήσεων των δύο μεθόδων. Αν δεν υπάρχει στατιστικά σημαντική διαφορά, θα πρέπει ο μέσος όρος των διαφορών μεταξύ των αντίστοιχων τιμών των δύο δειγμάτων να είναι μηδέν.

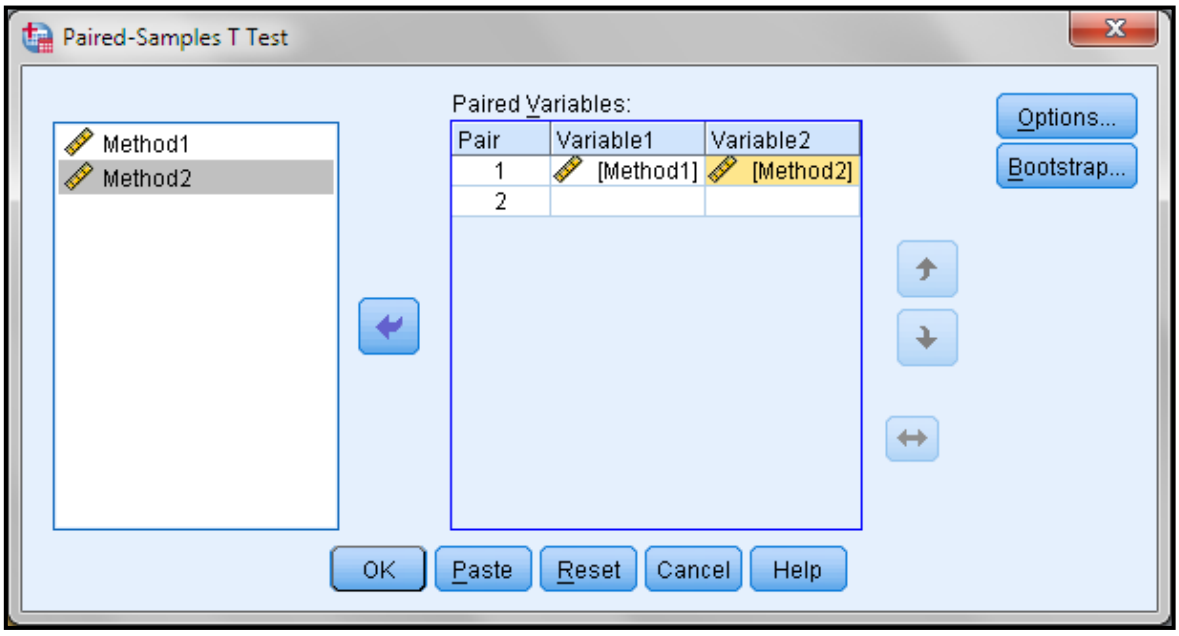
Παράδειγμα

Σ' έναν τάφο χρονολογούμε 8 διαφορετικά αντικείμενα με δύο τεχνικές χρονολόγησης. Τα αποτελέσματα που πήραμε σε χιλιάδες χρόνια δίνονται στο Σχήμα 4.2. Να ελεγχθεί αν οι δύο μέθοδοι δίνουν στατιστικά τα ίδια αποτελέσματα.

◆ Μεταφέρουμε τα δεδομένα σε φύλλο εργασίας του SPSS, όπως φαίνεται στο Σχήμα 4.2 όπου οι μεταβλητές έχουν τα ονόματα Method1 και Method2. Ακολουθούμε τη διαδικασία *Analyze* → *Compare Means* → *Paired-Samples T Test*. Στο παράθυρο που ανοίγει κάνουμε κλικ στις δύο μεταβλητές και με κλικ στο βέλος ▶ τις μεταφέρουμε στο πλαίσιο *Paired Variables* (Σχήμα 4.3). Με κλικ στο *OK* παίρνουμε τους πίνακες αποτελεσμάτων. Ο πίνακας που κυρίως μας ενδιαφέρει είναι ο *Paired Samples Tests* (Πίνακας 4.3). Παρατηρούμε ότι p = 0.150 > 0.05 που δείχνει ότι και εδώ η H_0 δεν απορρίπτεται. Συνεπώς δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δύο δειγμάτων και άρα οι δύο μέθοδοι δίνουν στατιστικά ίδια αποτελέσματα.



Σχήμα 4.2. Οργάνωση δεδομένων για σύγκριση ζευγών δειγμάτων



Σχήμα 4.3. Πλαίσιο διαλόγου Paired-samples T test

Πίνακας 4.3. Αποτελέσματα στατιστικού ελέγχου

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Method1 - Method2	-,1250	,2188	,0773	-,3079	,0579	-1,616	7	,150

4.4 ΕΛΕΓΧΟΣ ΔΙΑΣΠΟΡΩΝ (ANOVA)

4.4.1 ΜΟΝΟΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ (One-way ANOVA)

Προηγουμένως εξετάσαμε τον στατιστικό έλεγχο της διαφοράς μέσων τιμών για δύο δείγματα. Σε πολλές περιπτώσεις θέλουμε να εξετάσουμε αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών τριών ή περισσοτέρων δειγμάτων. Αυτό επιτυγχάνεται με τη μέθοδο που ονομάζεται **ανάλυση διασποράς** (*Analysis of Variance - ANOVA*). Αν και υπάρχουν πολλές παραλλαγές της μεθόδου, μπορούμε να διακρίνουμε δύο περιπτώσεις: Τη μονο-παραγοντική ανάλυση διασποράς (*One-way ANOVA*) και τη δι-παραγοντική ανάλυση διασποράς (*Two-way ANOVA*). Η δεύτερη έχει επίσης δύο υποπεριπτώσεις: την ανάλυση χωρίς αλληλεπιδράσεις ή με αλληλεπιδράσεις. Οι δύο αυτές υποπεριπτώσεις ονομάζονται και ανάλυση χωρίς επαναλήψεις ή με επαναλήψεις.

Γενικά στη μονο-παραγοντική ανάλυση διασποράς έχουμε n δείγματα (cases) με m δεδομένα (variables) το καθένα. Στον Πίνακα 4.4 δίνεται μια γενική διευθέτηση δειγμάτων για μονο-παραγοντική ανάλυση διασποράς.

Πίνακας 4.4. Διευθέτηση δειγμάτων για μονο-παραγοντική ανάλυση διασποράς

Δείγμα 1	x_{11}	x_{12}	. . .	x_{1m}
Δείγμα 2	x_{21}	x_{22}	. . .	x_{2m}
. . .			. . .	
Δείγμα n	x_{n1}	x_{n2}	. . .	x_{nm}

Για να είναι επιτρεπτή η εφαρμογή της μεθόδου πρέπει να πληρούνται οι εξής προϋποθέσεις:

Α. Δεν πρέπει να υπάρχουν στατιστικά σημαντικές διαφορές στις διασπορές των δειγμάτων. Δηλαδή θα πρέπει να υπάρχει **ομοιογένεια της διασποράς** (*Homogeneity of variance*). Τα περισσότερα στατιστικά προγράμματα, όπως και το SPSS χρησιμοποιούν το κριτήριο ή τον έλεγχο Levene. Αν δεν υπάρχει ομοιογένεια της διασποράς, μπορούμε να χρησιμοποιήσουμε μη παραμετρική ANOVA, όπως εξετάζεται σε επόμενο κεφάλαιο.

Β. Τα δείγματα πρέπει να ακολουθούν την κανονική κατανομή. Μικρές αποκλίσεις από την κανονική κατανομή δεν επηρεάζουν τα αποτελέσματα της μεθόδου. Αν όμως υπάρχουν σημαντικές αποκλίσεις τότε εφαρμόζουμε μη παραμετρική ANOVA.

Για να εφαρμόσουμε την απλή One-way ANOVA στο SPSS όλα τα δείγματα τοποθετούνται σε μια στήλη, ενώ σε μια άλλη στήλη, γειτονική ή μη γειτονική, χρησιμοποιούμε τους αριθμούς 1, 2, 3, … για να διακρίνουμε τα δείγματα. Ακολούθως από το *Analyze* → *Compare Means* → *One-Way ANOVA* ανοίγουμε το παράθυρο *One-Way ANOVA* και μεταφέρουμε τη μεταβλητή των δειγμάτων στο πλαίσιο *Dependent List*, τη μεταβλητή με τους αριθμούς 1, 2, 3, … στο *Factor* και από το *Options* επιλέγουμε το *Homogeneity of variance test* για να ελέγξουμε την ομοιογένεια της διασποράς. Ολοκληρώνουμε με κλικ στο *Continue* και στο *OK*.

Παράδειγμα

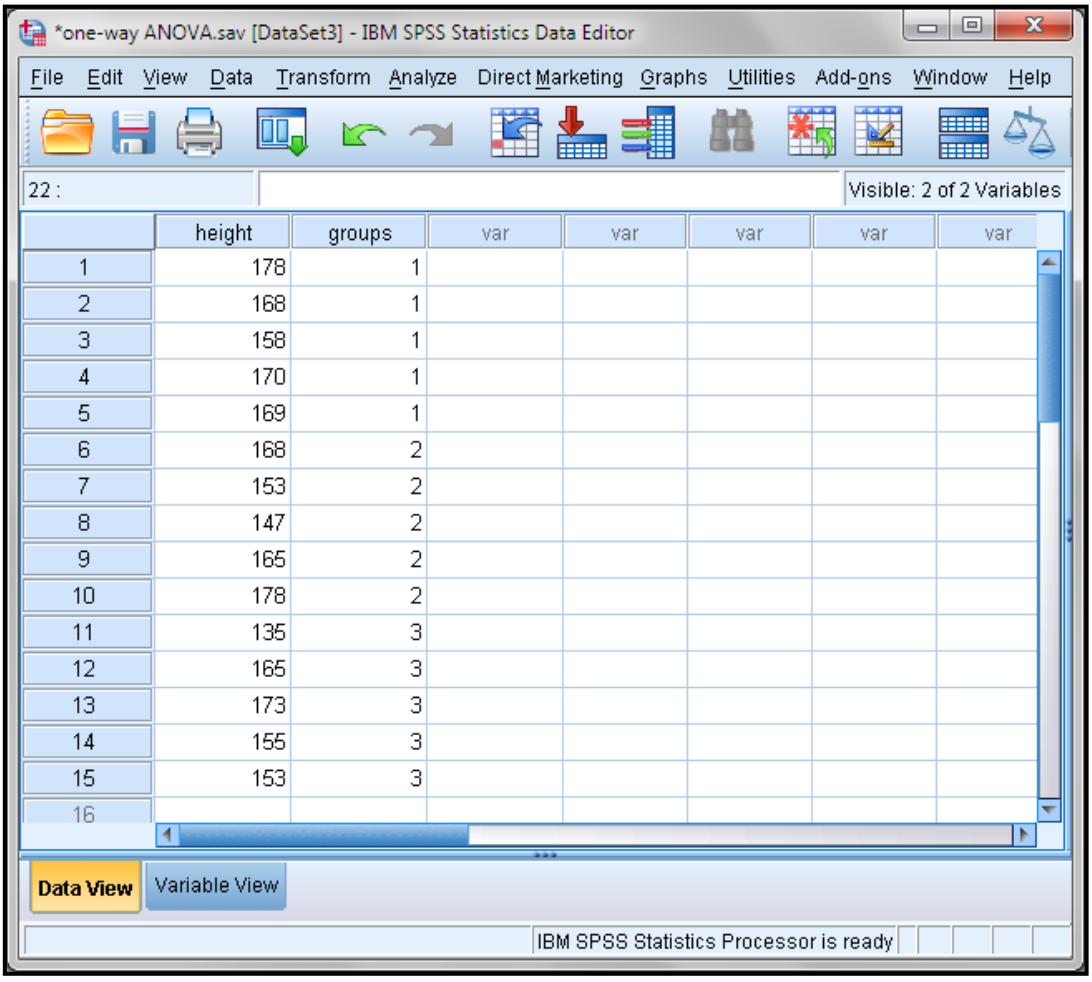
Σε τρεις διαφορετικές τοποθεσίες προσδιορίστηκε το ύψος πέντε ενήλικων ανδρών. Τα αποτελέσματα δίνονται στον Πίνακα 4.5. Να εξετασθεί αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των υψών σε επίπεδο σημαντικότητας α = 0,05.

◆ Μεταφέρουμε τα παραπάνω δεδομένα σε φύλλο εργασίας του SPSS, όπως φαίνεται στο Σχήμα 4.4, στήλες Height και Groups. Για να εφαρμόσουμε τώρα τη μέθοδο Ανάλυση Διασποράς ακολουθούμε την πορεία που προαναφέραμε. Δηλαδή, από το *Analyze* → *Compare Means* → *One-Way ANOVA* ανοίγουμε το παράθυρο *One-Way ANOVA* και μεταφέρουμε τη μεταβλητή Height στο πλαίσιο *Dependent List*, τη μεταβλητή Groups στο *Factor* και από το *Options* επιλέγουμε

το *Homogeneity of variance test*. Με κλικ στο *Continue* και στο *OK* παίρνουμε τα αποτελέσματα των Πινάκων 4.6 και 4.7.

Πίνακας 4.5. Ύψος σε cm ενήλικων ανδρών όπως προσδιορίστηκε από ταφικά δεδομένα τριών διαφορετικών τοποθεσιών

Τοποθεσία	Ύψος σε cm				
Α	178	158	148	170	139
Β	168	153	147	165	178
Γ	135	175	173	155	153



Σχήμα 4.4. Δεδομένα παραδείγματος

Πίνακας 4.6. Αποτελέσματα ελέγχου ομοιογένειας της διασποράς

Test of Homogeneity of Variances

height

Levene Statistic	df1	df2	Sig.
1.157	2	12	.347

Πίνακας 4.7. Αποτελέσματα σύγκρισης μεταξύ ομάδων

ANOVA

height

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	384.533	2	192.267	1.415	.281
Within Groups	1630.800	12	135.900		
Total	2015.333	14			

Από τον πίνακα Test of Homogeneity of Variances παρατηρούμε ότι ο στατιστικός έλεγχος διασπορών με το κριτήριο Levene δίνει την τιμή $p = 0.347 > 0.05$ που δείχνει ότι η H_0 δεν απορρίπτεται. Συνεπώς δεν υπάρχει στατιστικά σημαντική διαφορά στις διασπορές των δειγμάτων και άρα η ANOVA μπορεί να εφαρμοστεί. Παρατηρούμε επίσης ότι για την p-value της ANOVA ισχύει $p = 0.281 > 0.05$. Συνεπώς σε επίπεδο σημαντικότητας $\alpha = 0.05$ δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των υψών των ανδρών των τριών περιοχών.

Παράδειγμα

Ας υποθέσουμε ότι στη τοποθεσία Α τα ύψη ήταν 195, 180, 170, 185 και 190. Τι συμπεράσματα προκύπτουν σ' αυτή την περίπτωση;

◆ Αν αναλύσουμε τα αποτελέσματα όπως παραπάνω, παίρνουμε τον Πίνακα 4.8 από τον οποίο φαίνεται ότι σε επίπεδο σημαντικότητας $\alpha = 0.05$ (αλλά και σε επίπεδο σημαντικότητας $\alpha = 0.01$) υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των υψών των τριών τοποθεσιών ($p = 0.009 < 0.01$).

Πίνακας 4.8. Αποτελέσματα σύγκρισης μεταξύ ομάδων

ANOVA

height2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2140.133	2	1070.067	7.143	.009
Within Groups	1797.600	12	149.800		
Total	3937.733	14			

Για να δούμε μεταξύ ποιών δειγμάτων εντοπίζονται οι διαφορές, εργαζόμαστε ως εξής: Στο παράθυρο *One-Way ANOVA* κάνουμε κλικ στο *Post Hoc* και επιλέγουμε το κριτήριο *Tukey*, εφόσον έχουμε ομοιογένεια της διασποράς. Τώρα στα αποτελέσματα θα πάρουμε και τον Πίνακα 4.9. Από τον πίνακα αυτό εύκολα προκύπτει ότι έχουμε στατιστικά σημαντικές διαφορές μεταξύ των groups 1 και 2 ($p = 0.038 < 0.05$) και των 1 και 3 ($p = 0.01 < 0.05$). Αντίθετα, δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των groups 2 και 3. Σημειώνεται ότι ισχύει 1 = Α, 2 = Β και 3 = Γ.

Πίνακας 4.9. Αποτελέσματα επιμέρους συγκρίσεων

Multiple Comparisons

height2

Tukey HSD

		Mean			95% Confidence Interval	
(I) groups	(J) groups	Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	21.80000*	7.74080	.038	1.1486	42.4514
	3	27.80000*	7.74080	.010	7.1486	48.4514
2	1	-21.80000*	7.74080	.038	-42.4514	-1.1486
	3	6.00000	7.74080	.725	-14.6514	26.6514
3	1	-27.80000*	7.74080	.010	-48.4514	-7.1486
	2	-6.00000	7.74080	.725	-26.6514	14.6514

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

4.4.2 ΔΙΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ (Two-way ANOVA)

Όπως έχουμε ήδη αναφέρει, η δι-παραγοντική ανάλυση διασποράς (*Two-way ANOVA*) διακρίνεται σε δύο υποπεριπτώσεις: την ανάλυση χωρίς αλληλεπιδράσεις και την ανάλυση με αλληλεπιδράσεις.

4.4.2.1 ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ ΧΩΡΙΣ ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ

Στην δι-παραγοντική ανάλυση διασποράς οι τιμές του δείγματος επηρεάζονται από δύο παράγοντες, σε αντίθεση με την μονο-παραγοντική ανάλυση διασποράς όπου έχουμε μόνο έναν παράγοντα.

Παράδειγμα

Έστω ότι μελετάμε την περίμετρο κρανίων διαφορετικής χρονολογίας (Περίοδος I, II, III και IV) που βρέθηκαν σε τρεις διαφορετικές τοποθεσίες: (A) νησιά, (B) πεδιάδες, και (C) ορεινά. Σε κάθε τοποθεσία και χρονική περίοδο προσδιορίστηκε ο μέσος όρος της περιμέτρου των κρανίων και τα αποτελέσματα που ελήφθησαν δίνονται στον Πίνακα 4.10. Να εξετασθεί κατά πόσο είναι στατιστικά σημαντική η επίδραση της χρονικής περιόδου και της τοποθεσίας στην περίμετρο του κρανίου.

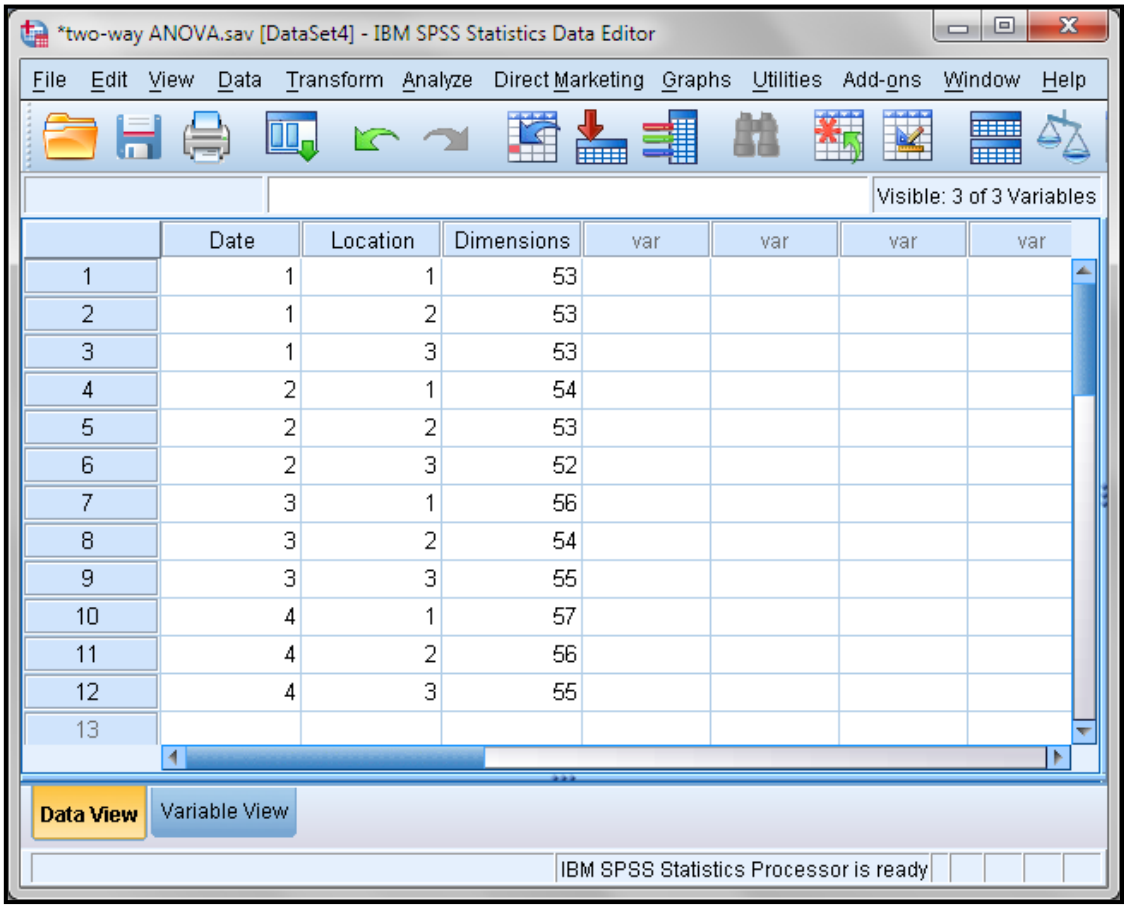
Πίνακας 4.10. Δεδομένα παραδείγματος

	A	B	C
I	53	53	53
II	54	53	52
III	56	54	55
IV	57	56	55

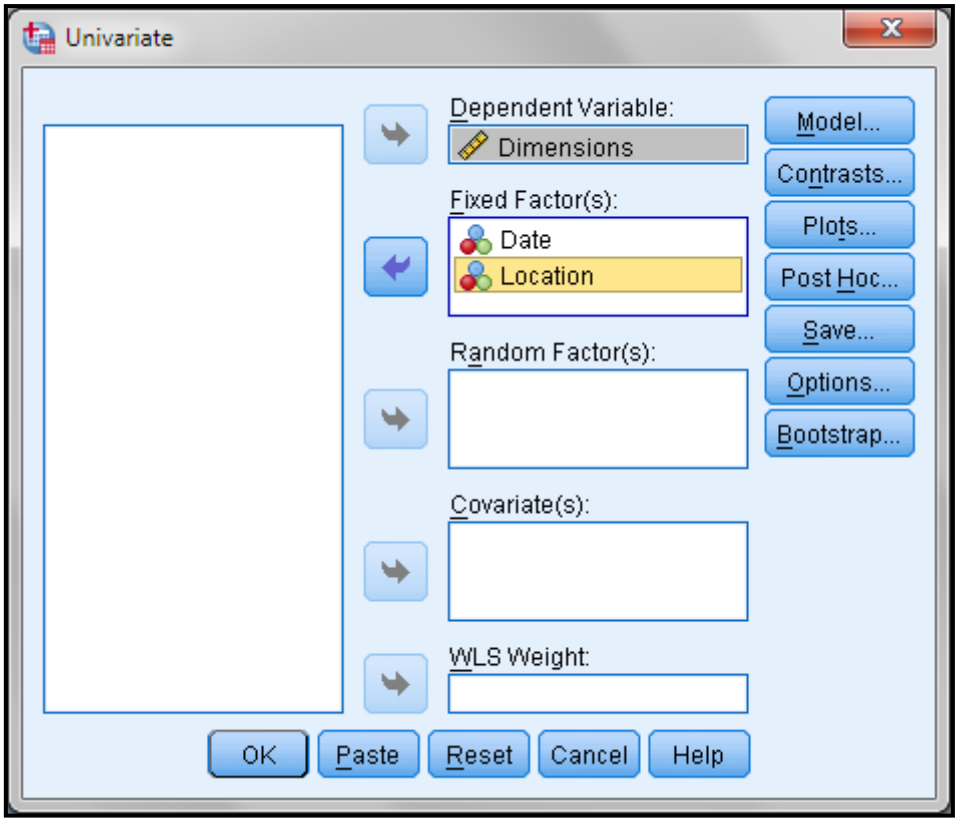
◆ Μεταφέρουμε τα δεδομένα του Πίνακα 4.10 σε ένα φύλλο εργασίας του SPSS, όπως φαίνεται στο Σχήμα 4.5. Στο σχήμα αυτό η μεταβλητή Dimensions εκφράζει τις τιμές της περιμέτρου των κρανίων, η Date τις χρονικές περιόδους και η Location την τοποθεσία. Προφανώς οι τιμές 1, 2, 3, 4 της Date αντιστοιχούν στις περιόδους I, II, III, IV και οι 1, 2, 3 της Location στις τοποθεσίες A, B και C.

Για να εφαρμόσουμε τώρα τη μέθοδο της δι-παραγοντικής Ανάλυσης Διασποράς ακολουθούμε την πορεία *Analyze* → *General Linear Model* → *Univariate*. Στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές Dimensions, Date, Location στα πλαίσια *Dependent Variable* και *Fixed Factor(s)* όπως φαίνεται στο Σχήμα 4.6. Κάνουμε κλικ στο *Model* και στο παράθυρο διαλόγου που ανοίγει κάνουμε τις ακόλουθες ενέργειες: Επιλέγουμε *Custom*, μεταφέρουμε τις

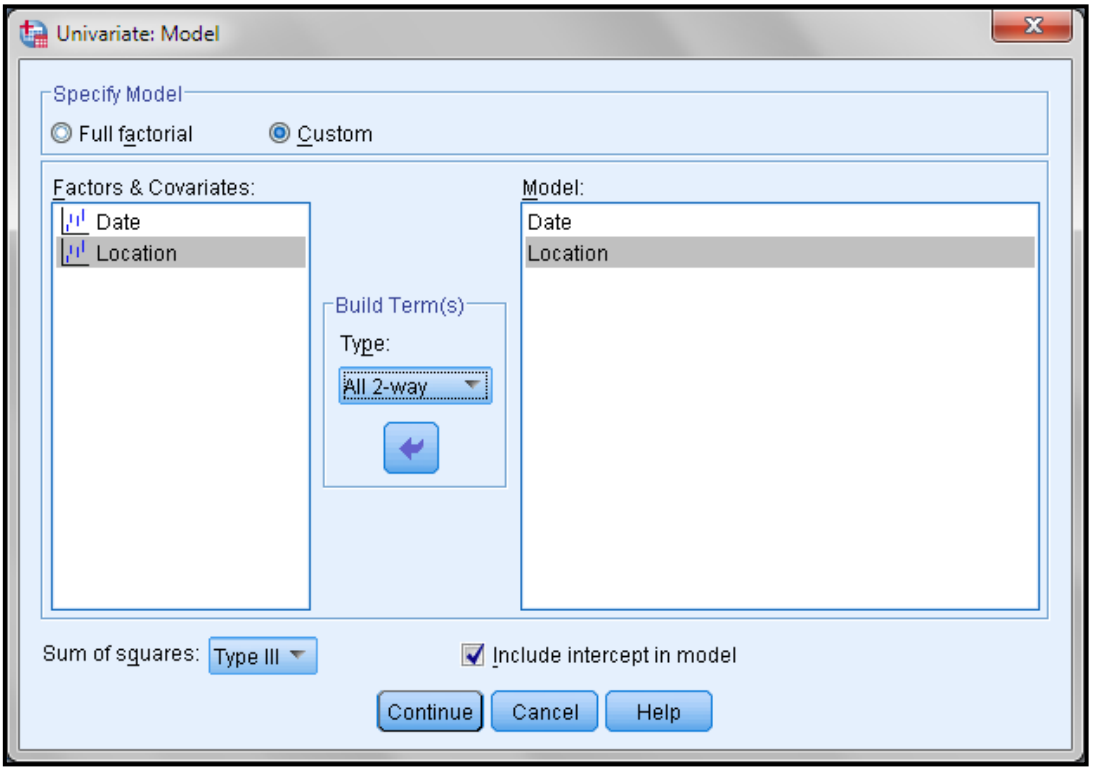
μεταβλητές Date και Location στο πλαίσιο *Model* και επιλέγουμε *All 2-way*, όπως φαίνεται στο Σχήμα 4.7. Συνεχίζουμε με *Continue* και στο παράθυρο *Univariate* κάνουμε κλικ στο *Options*, όπου επιλέγουμε το *Homogeneity tests*. Με κλικ στο *Continue* και στο *OK* παίρνουμε τα αποτελέσματα του Πίνακα 4.11.



Σχήμα 4.5. Δεδομένα παραδείγματος στο SPSS



Σχήμα 4.6. Το παράθυρο *Univariate*



Σχήμα 4.7. Το παράθυρο *Univariate: Model*

Πίνακας 4.11. Αποτελέσματα συγκρίσεων

Tests of Between-Subjects Effects

Dependent Variable:Dimensions

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	23,750[a]	5	4,750	11,400	,005
Intercept	35316,750	1	35316,750	84760,200	,000
Date	20,250	3	6,750	16,200	,003
Location	3,500	2	1,750	4,200	,072
Error	2,500	6	,417		
Total	35343,000	12			
Corrected Total	26,250	11			

a. R Squared = ,905 (Adjusted R Squared = ,825)

Η τιμή p = 0.003 < 0.05 δείχνει ότι σε επίπεδο σημαντικότητας α = 0.05 η επίδραση του παράγοντα *χρονική περίοδος* είναι στατιστικά σημαντική. Αντίθετα, η επίδραση της τοποθεσίας είναι έστω και οριακά στατιστικά ασήμαντη (p = 0.072 > 0.05).

Επειδή παρατηρήσαμε ότι η επίδραση του παράγοντα *χρονική περίοδος* είναι στατιστικά σημαντική, μπορούμε με το SPSS να εξετάσουμε μεταξύ ποιών περιόδων υπάρχουν οι μεγαλύτερες διαφορές. Όπως και στην απλή *One-Way ANOVA*, αυτό γίνεται με κλικ στο *Post Hoc* στο παράθυρο *Univariate*. Στο νέο παράθυρο που ανοίγει μεταφέρουμε τη μεταβλητή Date στο πλαίσιο *Post Hoc Tests for* και επιλέγουμε το κριτήριο *Tukey*, εφόσον έχουμε ομοιογένεια της διασποράς. Τώρα στα αποτελέσματα θα πάρουμε και τον Πίνακα 4.12, από τον οποίο προκύπτει ότι στατιστικά σημαντικές διαφορές υπάρχουν μεταξύ των περιόδων I – III, I – IV, II – III και II - IV. Ποιο σημαντικές είναι οι διαφορές I – IV και II - IV.

Πίνακας 4.12. Αποτελέσματα επιμέρους συγκρίσεων

Multiple Comparisons

Dimensions

Tukey HSD

(I) Date	(J) Date	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
I	II	,00	,527	1,000	-1,82	1,82
	III	-2,00*	,527	,034	-3,82	-,18
	IV	-3,00*	,527	,005	-4,82	-1,18
II	I	,00	,527	1,000	-1,82	1,82
	III	-2,00*	,527	,034	-3,82	-,18
	IV	-3,00*	,527	,005	-4,82	-1,18
III	I	2,00*	,527	,034	,18	3,82
	II	2,00*	,527	,034	,18	3,82
	IV	-1,00	,527	,321	-2,82	,82
IV	I	3,00*	,527	,005	1,18	4,82
	II	3,00*	,527	,005	1,18	4,82
	III	1,00	,527	,321	-,82	2,82

Based on observed means.

The error term is Mean Square(Error) = .417.

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

4.4.2.2 ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ ΜΕ ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ

Σε πολλές περιπτώσεις μπορεί να υπάρχει σημαντική αλληλεπίδραση μεταξύ των παραγόντων με αποτέλεσμα η τελική τους επίδραση στα δεδομένα να είναι ή πολύ μεγαλύτερη ή πολύ μικρότερη από αυτή που θα αναμέναμε. Σ' αυτή την περίπτωση πρέπει να χρησιμοποιήσουμε **Ανάλυση Διασποράς με αλληλεπιδράσεις**. Για να εφαρμοστεί αυτή η ανάλυση απαιτείται να υπάρχουν περισσότερες από μια τιμές σε κάθε τιμή του πρώτου ή του δεύτερου παράγοντα.

Παράδειγμα

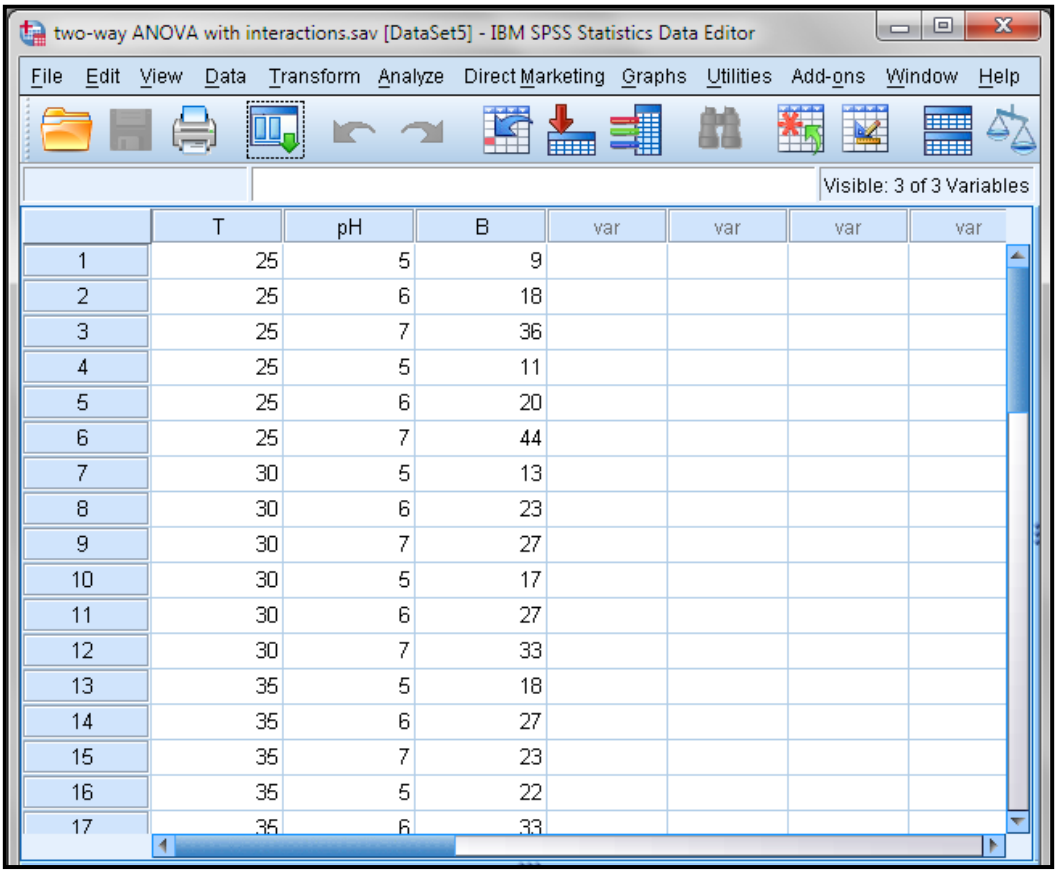
Σ' ένα πείραμα μελέτης της επίδρασης της θερμοκρασίας και του pH στην ανάπτυξη ενός βακτηρίου σε 24 φιάλες ελήφθησαν τα αποτελέσματα που δίνονται

στον Πίνακα 4.13. Να εξετασθεί η επίδραση της θερμοκρασίας και του pH στην ανάπτυξη του βακτηρίου.

Πίνακας 4.13. Δεδομένα παραδείγματος

T / ^{0}C	pH = 5	pH = 6	pH = 7
25	9	18	36
25	11	20	44
30	13	23	27
30	17	27	33
35	18	27	23
35	22	33	27
40	22	20	7
40	28	24	13

◆ Μεταφέρουμε τα δεδομένα σε ένα φύλλο εργασίας του SPSS, όπως φαίνεται στο Σχήμα 4.8, όπου η μεταβλητή B εκφράζει την πυκνότητα, η T τη θερμοκρασία και η pH την οξύτητα του διαλύματος.



	T	pH	B	var	var	var	var
1	25	5	9				
2	25	6	18				
3	25	7	36				
4	25	5	11				
5	25	6	20				
6	25	7	44				
7	30	5	13				
8	30	6	23				
9	30	7	27				
10	30	5	17				
11	30	6	27				
12	30	7	33				
13	35	5	18				
14	35	6	27				
15	35	7	23				
16	35	5	22				
17	35	6	33				

Σχήμα 4.8. Τμήμα δεδομένων του παραδείγματος στο SPSS

Για να αναλύσουμε τα δεδομένα ακολουθούμε την πορεία *Analyze* → *General Linear Model* → *Univariate*. Στο παράθυρο που ανοίγει μεταφέρουμε τη μεταβλητή Β στο πλαίσιο *Dependent Variable* και τις Τ και pH στο *Fixed Factor(s)*. Στο *Options* επιλέγουμε το *Homogeneity tests*, και στο *Model* επιλέγουμε *Full Factorial*. Τα αποτελέσματα που παίρνουμε δίνονται στον Πίνακα 4.14.

Τα συμπεράσματα που προκύπτουν από τον πίνακα αυτόν είναι τα ακόλουθα:

(i) Δεν υπάρχει στατιστικά σημαντική επίδραση της θερμοκρασίας στην ανάπτυξη του βακτηρίου ($p = 0.065 > 0.05$).

(ιι) Η επίδραση του pH είναι στατιστικά σημαντική ($p = 0.001 < 0.05$).

(iii) Υπάρχει σημαντική αλληλεπίδραση μεταξύ θερμοκρασίας και pH ($p = 0.000 < 0.05$). Αυτό σημαίνει ότι η ανταπόκριση του βακτηρίου στο pH εξαρτάται από τη θερμοκρασία και αντίστροφα.

Πίνακας 4.14. Αποτελέσματα ANOVA

Tests of Between-Subjects Effects

Dependent Variable:B

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1649.833[a]	11	149.985	12.161	.000
Intercept	12240.167	1	12240.167	992.446	.000
T	116.500	3	38.833	3.149	.065
pH	330.333	2	165.167	13.392	.001
T * pH	1203.000	6	200.500	16.257	.000
Error	148.000	12	12.333		
Total	14038.000	24			
Corrected Total	1797.833	23			

a. R Squared = .918 (Adjusted R Squared = .842)

5. ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΔΟΚΙΜΕΣ

5.1 ΓΕΝΙΚΑ

Όπως αναφέρθηκε, για να εφαρμοστούν οι έλεγχοι του προηγούμενου κεφαλαίου είναι απαραίτητο τα δεδομένα να ακολουθούν την κανονική κατανομή. Οι έλεγχοι που προϋποθέτουν την κανονική κατανομή ονομάζονται **παραμετρικοί έλεγχοι**. Όταν μια μεταβλητή δεν ακολουθεί την κανονική κατανομή είμαστε υποχρεωμένοι να εφαρμόσουμε **μη παραμετρικές δοκιμές**. Ως μη-παραμετρικές στατιστικές μέθοδοι ορίζονται οι μέθοδοι στις οποίες δεν υπάρχουν παραδοχές ως προς τη μορφή των πληθυσμιακών κατανομών των χρησιμοποιούμενων δεδομένων. Το μειονέκτημα αυτών των μεθόδων είναι ότι οι πληροφορίες που παίρνουμε είναι λιγότερες από τις αντίστοιχες των παραμετρικών μεθόδων. Επίσης, βασικοί παραμετρικοί έλεγχοι, όπως *Post Hoc* έλεγχοι και δι-παραγοντική ανάλυση διασποράς με επαναλήψεις δεν μπορούν να γίνουν μη παραμετρικά.

5.2 ΣΥΓΚΡΙΣΗ ΔΥΟ ΑΝΕΞΑΡΤΗΤΩΝ ΔΕΙΓΜΑΤΩΝ

Ο έλεγχος αυτός χρησιμοποιείται όταν έχουμε δύο ή περισσότερα δείγματα και θέλουμε να διαπιστώσουμε αν προέρχονται από τον ίδιο ή όχι πληθυσμό. Είναι αντίστοιχος του παραμετρικού *Independent samples t-test*.

Παράδειγμα

Να εξεταστούν τα δεδομένα του αρχείου independent samples t-tests.sav με μη παραμετρικό έλεγχο.

◆ Ανοίγουμε το αρχείο independent samples t-tests.sav και ακολουθούμε τη διαδικασία *Analyze → Nonparametric Tests → Legacy Dialogs → 2 Independent Samples*. Μεταφέρουμε τη μεταβλητή Samples στο πλαίσιο *Test Variable List*, τη μεταβλητή *Groups* στο *Grouping Variable* και κάνουμε κλικ στο *Define Groups*. Στο νέο παράθυρο εισάγουμε την τιμή 1 στο *Group 1* και την τιμή 2 στο *Group* 2. Με κλικ στο *Continue* και στο *OK* παίρνουμε τον Πίνακα 5.1.

Παρατηρούμε ότι p = 0.062 (ή 0.068) > 0.05 που δείχνει ότι η H_0 δεν απορρίπτεται. Εδώ η μηδενική υπόθεση H_0 είναι ότι τα δύο δείγματα προέρχονται

από τον ίδιο πληθυσμό. Συνεπώς, με βάση το αποτέλεσμα αυτό προκύπτει ότι σε επίπεδο σημαντικότητας 0.05 δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δύο δειγμάτων.

Πίνακας 5.1. Αποτελέσματα σύγκρισης

Test Statistics[b]

	samples
Mann-Whitney U	19.000
Wilcoxon W	74.000
Z	-1.870
Asymp. Sig. (2-tailed)	.062
Exact Sig. [2*(1-tailed Sig.)]	.068[a]

a. Not corrected for ties.

b. Grouping Variable: groups

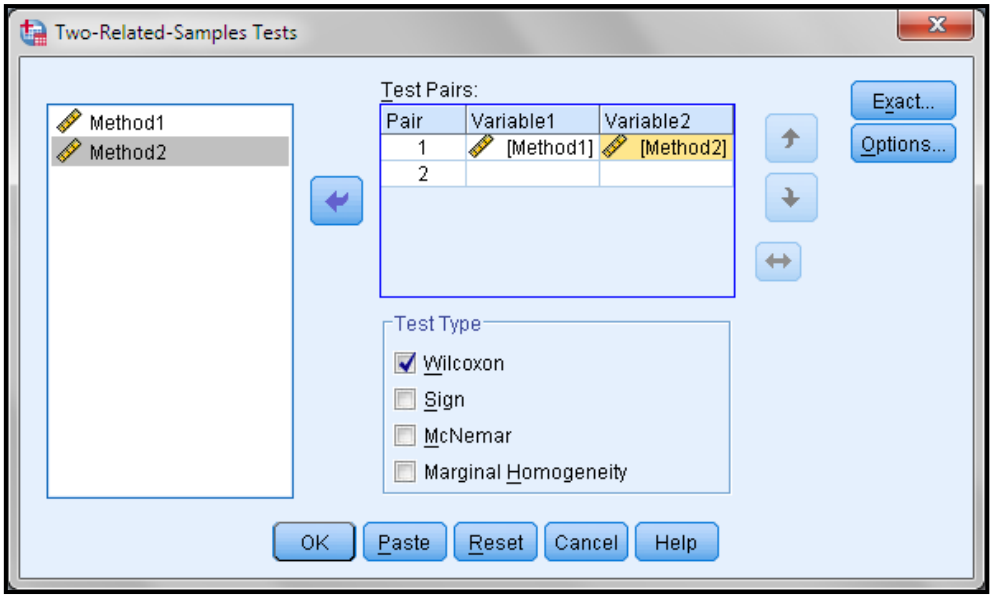
5.3 ΣΥΓΚΡΙΣΗ ΖΕΥΓΩΝ ΔΕΙΓΜΑΤΩΝ

Πρόκειται για αντίστοιχη διαδικασία με το παραμετρικό Paired samples t-test.

Παράδειγμα

Να εξεταστούν τα δεδομένα του αρχείου paired samples t-tests.sav με μη παραμετρικό έλεγχο.

◆ Ανοίγουμε το αρχείο paired samples t-tests.sav και ακολουθούμε τη διαδικασία *Analyze* → *Nonparametric Tests* → *Legacy Dialogs* → *2 Related Samples*. Στο παράθυρο που ανοίγει κάνουμε κλικ στις μεταβλητές Method1 και Method2 και με κλικ στο βέλος ▶ τις μεταφέρουμε στο πλαίσιο *Test Pair(s)* (Σχήμα 5.1). Επιλέγουμε το *Wilcoxon* (αν δεν είναι επιλεγμένο) στο *Test Type* και με κλικ στο *OK* παίρνουμε τον Πίνακα 5.2. Παρατηρούμε ότι p = 0.151 > 0.05 που δείχνει ότι και εδώ η H_0 δεν απορρίπτεται. Συνεπώς δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δύο δειγμάτων και συνεπώς οι δύο μέθοδοι δίνουν στατιστικά ίδια αποτελέσματα.



Σχήμα 5.1. Πλαίσιο διαλόγου Two-Related-Samples Tests

Πίνακας 5.2. Αποτελέσματα σύγκρισης

Test Statistics[b]

	Method2 - Method1
Z	-1,436[a]
Asymp. Sig. (2-tailed)	,151

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

5.4 ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

5.4.1 ΜΟΝΟ-ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ (ΚΡΙΤΗΡΙΟ KRUSKAL-WALLIS)

Πρόκειται για αντίστοιχη διαδικασία προς την παραμετρική One-way ANOVA.

Παράδειγμα

Να εξεταστούν τα δεδομένα του αρχείου one-way ANOVA.sav με μη παραμετρικό έλεγχο.

◆ Ανοίγουμε το αρχείο one-way ANOVA.sav και ακολουθούμε την πορεία *Analyze → Nonparametric Tests→ Legacy Dialogs → K Independent Samples,* ανοίγουμε το παράθυρο *Tests for Several Independent Samples* και μεταφέρουμε τη μεταβλητή Height στο πλαίσιο *Test Variable List* και τη μεταβλητή *Groups* στο *Grouping Variable*. Κάνουμε κλικ στο *Define Groups* και στο νέο παράθυρο εισάγουμε την τιμή 1 στο *Minimum* και την τιμή 3 στο *Maximum*. Με κλικ στο *Continue* και στο *OK* παίρνουμε τον Πίνακα 5.3.

Παρατηρούμε ότι ισχύει p = 0.247 > 0.05. Συνεπώς σε επίπεδο σημαντικότητας α = 0.05 δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των υψών των ανδρών από τις τρεις τοποθεσίες.

Πίνακας 5.3. Αποτελέσματα συγκρίσεων

Test Statistics[a,b]

	height
Chi-Square	2,800
df	2
Asymp. Sig.	,247

a. Kruskal Wallis Test
b. Grouping Variable: groups

Αν εξετάσουμε τη μεταβλητή height2 με μη παραμετρικό έλεγχο παίρνουμε τον παρακάτω πίνακα αποτελεσμάτων από τον οποίο φαίνεται ότι σε επίπεδο σημαντικότητας α = 0.05 (αλλά **όχι** και σε επίπεδο σημαντικότητας α = 0.01) υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των υψών των ανδρών.

Πίνακας 5.4. Αποτελέσματα συγκρίσεων

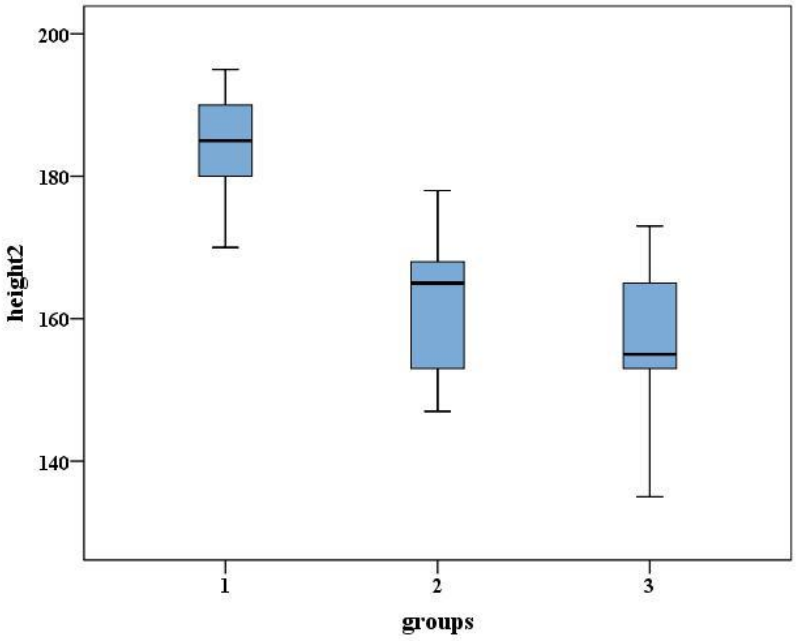
Test Statistics[a,b]

	height2
Chi-Square	8,089
df	2
Asymp. Sig.	,018

a. Kruskal Wallis Test
b. Grouping Variable: groups

Παρατήρηση 1. Σ’ αυτή την περίπτωση δεν μπορούμε να πάρουμε περισσότερες πληροφορίες για το πού βρίσκονται οι διαφορές. Για να ξεπεράσουμε αυτό το πρόβλημα κάνουμε τα θηκογράμματα των δειγμάτων (Σχήμα 5.2). Παρατηρούμε ότι η διαφοροποίηση βρίσκεται μεταξύ του πρώτου δείγματος και των υπολοίπων.

Παρατήρηση 2. Η παραπάνω διαδικασία ονομάζεται και τεστ Kruskal - Wallis.



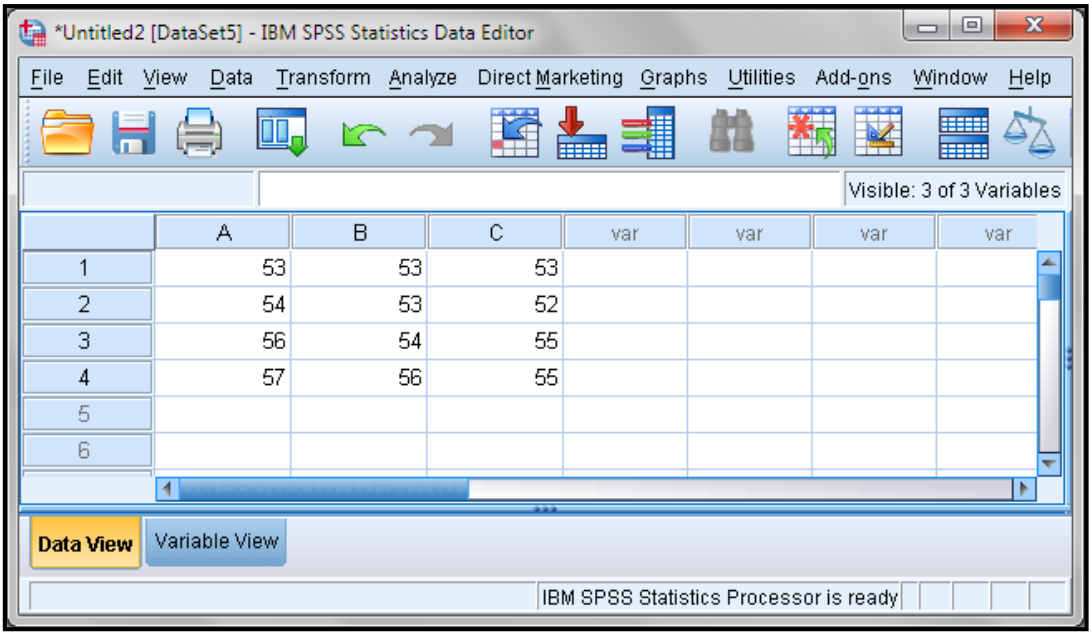
Σχήμα 5.2. Θηκογράμματα για να προσδιοριστούν οι στατιστικά σημαντικές διαφοροποιήσεις μεταξύ των δειγμάτων

5.4.2 ΔΙ-ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

Πρόκειται για αντίστοιχη ανάλυση της παραμετρικής Two-way ANOVA.

Παράδειγμα

Να εξεταστούν τα δεδομένα του αρχείου two-way ANOVA.sav με μη παραμετρικό έλεγχο.



Σχήμα 5.3. Δεδομένα παραδείγματος σε κατάλληλη διάταξη

◆ Ανοίγουμε ένα νέο φύλλο εργασίας και μεταφέρουμε τα δεδομένα του αρχείου two-way ANOVA.sav όπως φαίνεται στο Σχήμα 5.3. Ακολουθούμε τη διαδικασία *Analyze* → *Nonparametric Tests* → *Legacy Dialogs* → *K Related Samples* και στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές A, B, C στο πλαίσιο *Test Variables*. Επιλέγουμε το *Friedman* (αν δεν είναι επιλεγμένο) στο *Test Type* και με κλικ στο *OK* παίρνουμε τον Πίνακα 5.5. Παρατηρούμε ότι p = 0.097 > 0.05 που δείχνει ότι και εδώ η H_0 δεν απορρίπτεται. Συνεπώς δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων <u>όταν αυτά ορίζονται κατά στήλες</u>. Εφόσον κάθε στήλη αντιστοιχεί σε μια τοποθεσία, η επίδραση του παράγοντα τοποθεσίας είναι στατιστικά μη σημαντική.

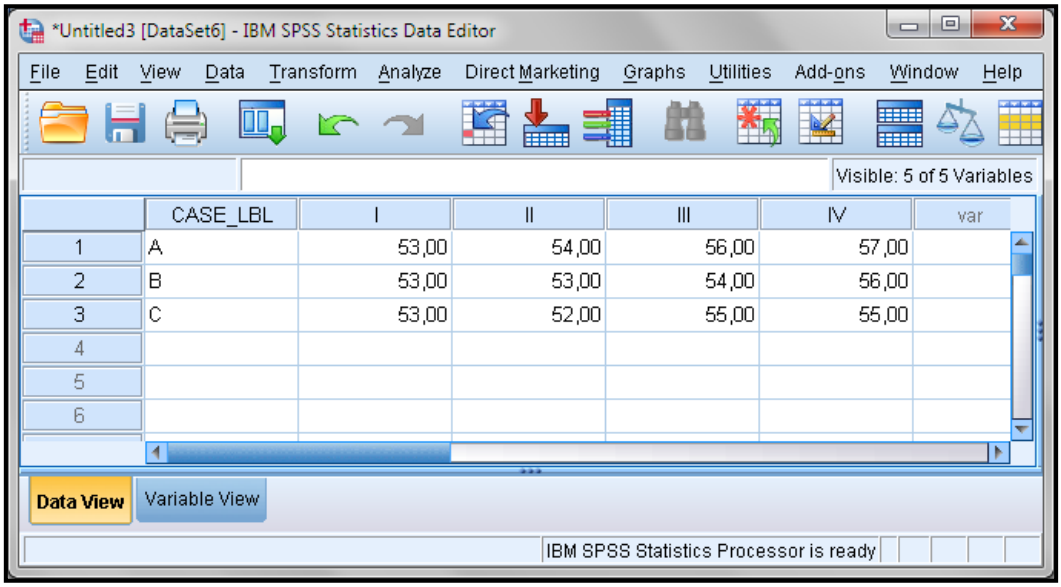
Πίνακας 5.5. Αποτελέσματα σύγκρισης

Test Statistics[a]

N	4
Chi-Square	4,667
df	2
Asymp. Sig.	,097

a. Friedman Test

Για να δούμε την επίδραση της χρονικής περιόδου κάνουμε τις γραμμές στήλες και τις στήλες γραμμές ως εξής: Από το *Data* → *Transpose* στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές A, B, C στο πλαίσιο *Variable(s)* και κάνουμε κλικ στο *OK*. Τα δεδομένα αναστρέφονται σ' ένα νέο φύλλο εργασίας που ανοίγει αυτόματα (Σχήμα 5.4). Στο νέο αυτό φύλλο επαναλαμβάνουμε τη διαδικασία *Analyze* → *Nonparametric Tests* → *K Related Samples* εισάγοντας τώρα τέσσερεις μεταβλητές στο πλαίσιο *Test Variables*. Τα αποτελέσματα που παίρνουμε δίνονται στον Πίνακα 5.6.



Σχήμα 5.4. Αναδιάταξη του αρχείου δεδομένων

Πίνακας 5.6. Αποτελέσματα σύγκρισης

Test Statistics[a]

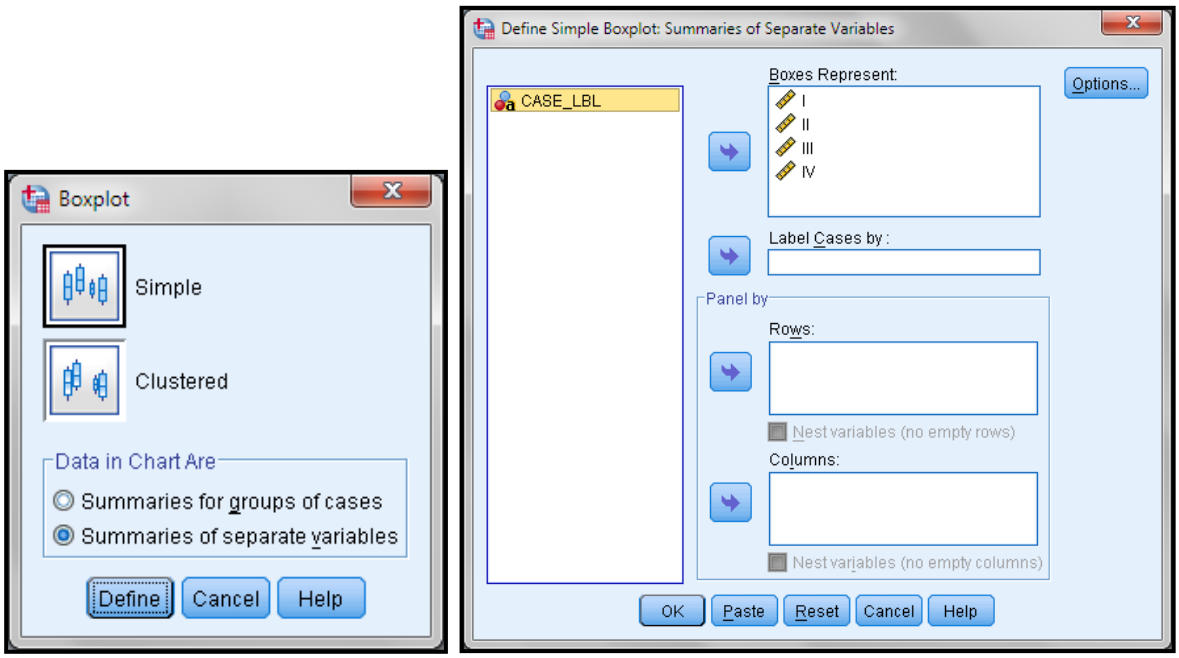
N	3
Chi-Square	8,143
df	3
Asymp. Sig.	,043

a. Friedman Test

Παρατηρούμε ότι τώρα η επίδραση της *χρονικής περιόδου* είναι στατιστικά σημαντική στο επίπεδο σημαντικότητας α = 0.05 (p = 0.043 < 0.05). Παρατηρούμε επίσης ότι μεταξύ των αποτελεσμάτων της μη παραμετρικής

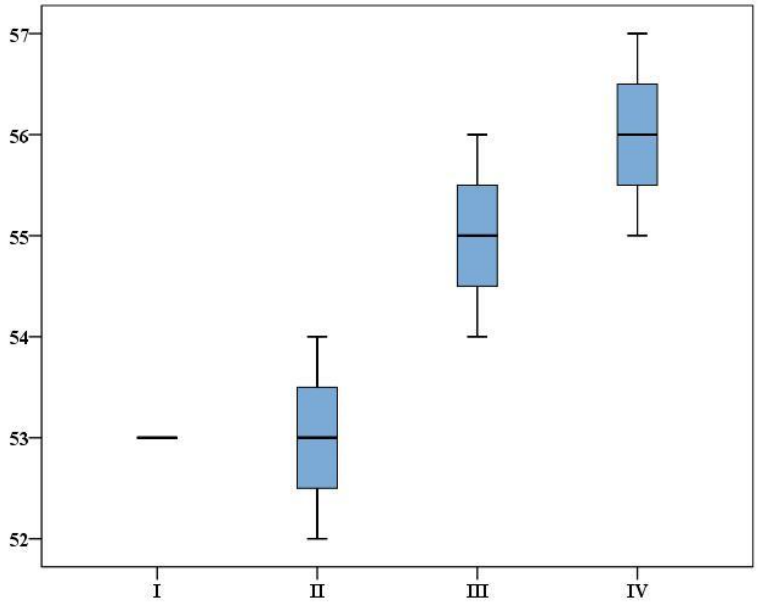
ανάλυσης και της ανάλυσης που έγινε στο προηγούμενο κεφάλαιο υπάρχουν μικρές διαφορές, που πρέπει να αναμένονται λόγω της διαφορετικής μεθοδολογίας των δύο μεθόδων.

Για να δούμε μεταξύ ποιών χρονικών περιόδων (I, II, III και IV) υπάρχουν οι στατιστικά σημαντικές διαφορές, κατασκευάζουμε τα θηκογράμματα αυτών των δειγμάτων. Έτσι, ακολουθούμε τη διαδικασία *Graphs → Legacy Dialogs → Boxplot.* Στο παράθυρο διαλόγου *Boxplot* επιλέγουμε Simple και Summaries of separate variables επειδή τα δείγματα βρίσκονται σε διαφορετικές στήλες (μεταβλητές) (Σχήμα 5.5). Ακολούθως στο παράθυρο διαλόγου *Define Simple Boxplot: Summaries of Separate Variables* μεταφέρουμε τις μεταβλητές I, II,II, IV στο πλαίσιο *Boxes Represent* και πατάμε *OK*. Θα πάρουμε τα θηκογράμματα του Σχήματος 5.6.



Σχήμα 5.5. Επιλογές για κατασκευή θηκογραμμάτων

Παρατηρούμε ότι οι στατιστικά σημαντικές διαφορές εμφανίζονται μεταξύ των χρονικών περιόδων (I και II) και (III και IV).



Σχήμα 5.6. Θηκογράμματα δειγμάτων σε διάφορες χρονικές περιόδους I, II, III, IV

6. ΕΛΕΓΧΟΙ ΣΕ ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ

Οι στατιστικοί έλεγχοι που εξετάσαμε στα προηγούμενα κεφάλαια αφορούν ποσοτικά δεδομένα. Πολλές φορές όμως είναι απαραίτητο να αναλύσουμε κατηγορικά δεδομένα. Τα κατηγορικά δεδομένα προκύπτουν όταν με βάση κάποιο ποιοτικό ή και ποσοτικό κριτήριο ταξινομούμε τα δεδομένα σε κατηγορίες.

6.1 ΠΙΝΑΚΕΣ ΔΙΑΣΤΑΥΡΩΣΗΣ (CROSS TABULATION)

Όταν έχουμε έναν πίνακα δεδομένων, όπως αυτός του Παραρτήματος Ι, είναι εύλογο να αναρωτηθούμε ποιά σχέση υπάρχει μεταξύ φύλου και αρθρίτιδας. Η σχέση αυτή, που είναι σχέση μεταξύ nominal ή ordinal μεταβλητών προσδιορίζεται με τη διαδικασία *crosstabs* (*cross tabulation*).

Παράδειγμα

Να προσδιοριστεί η σχέση φύλου και αρθρίτιδας στους οσφυϊκούς σπονδύλους στο αρχείο osteological data.sav.

◆ Ανοίγουμε το αρχείο osteological data.sav και ακολουθούμε τη διαδικασία *Analyze* → *Descriptive Statistics* → *Crosstabs*. Εισάγουμε στο πλαίσιο *Row(s)* τη μεταβλητή sex και στο *Column(s)* τη μεταβλητή osteophytosis. Κάνουμε κλικ στο *Cells* και επιλέγουμε τα *Observed, Expected, Row, Column* και *Total*. Ολοκληρώνουμε με κλικ στο *Continue* και στο *OK*. Ο κύριος πίνακας αποτελεσμάτων είναι ο Πίνακας 6.1. Στον πίνακα αυτό *Count* είναι οι περιπτώσεις που υπάρχουν στον αρχικό πίνακα δεδομένων και *Expected Count* οι περιπτώσεις που αναμένονται αν η κατανομή ήταν τυχαία. Για παράδειγμα, στα δεδομένα υπάρχουν 4 γυναίκες με eburnation, ενώ η αναμενόμενη τιμή αν το φύλο δεν έπαιζε κανένα ρόλο είναι μόνο 4.6. Γενικά παρατηρούμε ότι στα δεδομένα του παραδείγματος που εξετάζουμε το φύλο δεν καθορίζει σημαντικά το επίπεδο πάθησης.

Ακολουθώντας την ίδια διαδικασία θα μπορούσαμε να εξετάσουμε την επίδραση του φύλου στην αρθρίτιδα στα χέρια.

Πίνακας 6.1. Αποτελέσματα σχέσης φύλου και αρθρίτιδας στους οσφυϊκούς σπονδύλους

sex * lumbar vertebrae osteophytosis Crosstabulation

			lumbar vertebrae osteophytosis			
			lipping	pitting	eburnation	Total
sex	f	Count	11	6	4	21
		Expected Count	10,9	5,5	4,6	21,0
		% within sex	52,4%	28,6%	19,0%	100,0%
		% within lumbar vertebrae osteophytosis	42,3%	46,2%	36,4%	42,0%
		% of Total	22,0%	12,0%	8,0%	42,0%
	m	Count	15	7	7	29
		Expected Count	15,1	7,5	6,4	29,0
		% within sex	51,7%	24,1%	24,1%	100,0%
		% within lumbar vertebrae osteophytosis	57,7%	53,8%	63,6%	58,0%
		% of Total	30,0%	14,0%	14,0%	58,0%
Total		Count	26	13	11	50
		Expected Count	26,0	13,0	11,0	50,0
		% within sex	52,0%	26,0%	22,0%	100,0%
		% within lumbar vertebrae osteophytosis	100,0%	100,0%	100,0%	100,0%
		% of Total	52,0%	26,0%	22,0%	100,0%

6.2 ΤΟ ΚΡΙΤΗΡΙΟ χ^2

Ένα πιο αυστηρό κριτήριο για το αν μια μεταβλητή Nominal ή Ordinal επιδρά σε μια άλλη είναι το κριτήριο χ^2 (*chi square test*). Για να διενεργήσουμε αυτόν τον έλεγχο, στο παράθυρο διαλόγου *Crosstabs* που ανοίγει από *Analyze* → *Descriptive Statistics* → *Crosstabs*, κάνουμε κλικ στο *Statistics* και επιλέγουμε *Chi-square*.

Στον έλεγχο αυτό η μηδενική υπόθεση (H_0) είναι ότι οι μεταβλητές είναι ανεξάρτητες μεταξύ τους και η p-value δίνεται στη στήλη Assymp. Sig.

Για τον έλεγχο της σχέσης μεταξύ φύλου και επιπέδου αρθρίτιδας στους οσφυϊκούς σπονδύλους παίρνουμε τον Πίνακα 6.2. Παρατηρούμε ότι p = 0.888 > 0.05 και συνεπώς δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση. Αυτό σημαίνει ότι δεν φαίνεται να υπάρχει στατιστικά σημαντική επίδραση μεταξύ φύλου και επιπέδου αρθρίτιδας στους οσφυϊκούς σπονδύλους.

Πίνακας 6.2. Αποτελέσματα του χ^2 test για τη σχέση φύλου και επιπέδου αρθρίτιδας στους οσφυϊκούς σπονδύλους

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	,237[a]	2	,888
Likelihood Ratio	,238	2	,888
N of Valid Cases	50		

a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,62.

Στο ίδιο συμπέρασμα καταλήγουμε αν εξετάσουμε την επίδραση φύλου και αρθρίτιδας στα χέρια (Πίνακας 6.3).

Πίνακας 6.3. Αποτελέσματα του χ^2 test για την επίδραση φύλου και επιπέδου αρθρίτιδας στα χέρια

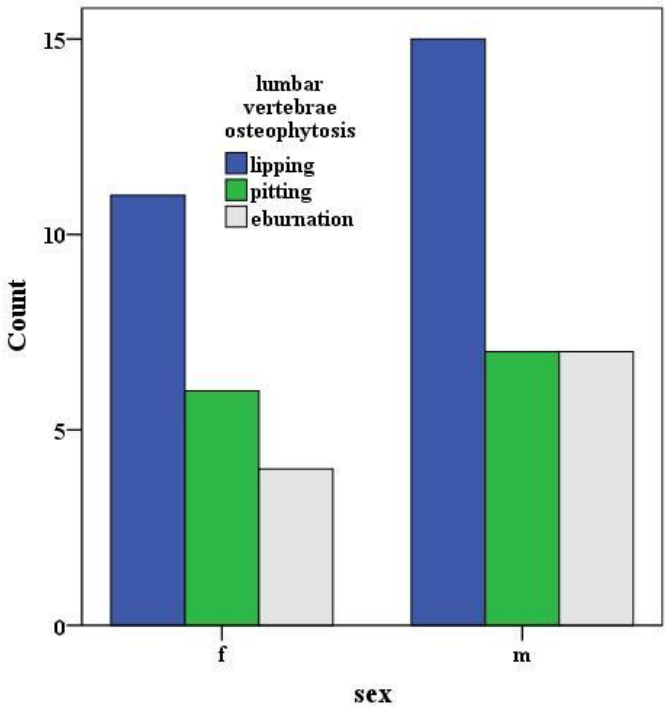
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,381[a]	5	,794
Likelihood Ratio	2,418	5	,789
N of Valid Cases	50		

a. 8 cells (66,7%) have expected count less than 5. The minimum expected count is ,84.

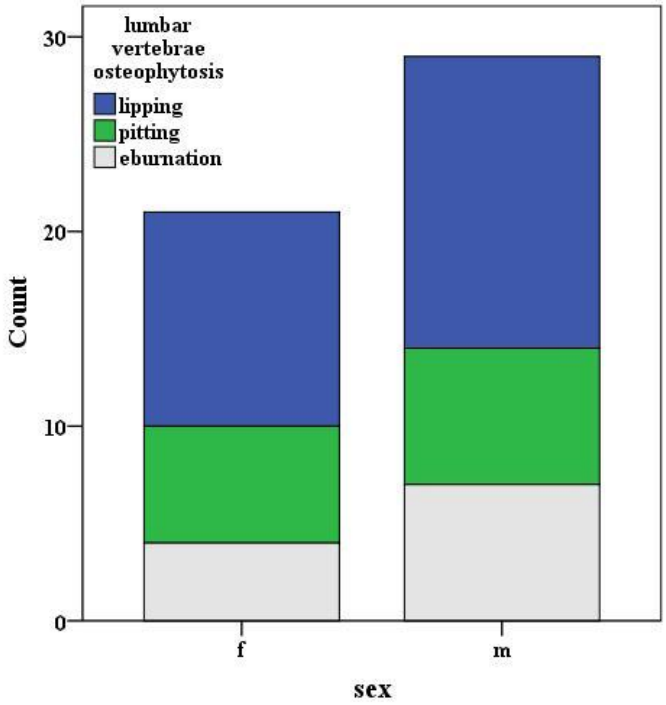
6.3 ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι γραφικές παραστάσεις που σχετίζονται με πίνακες διασταύρωσης. Στο παράδειγμα για τη σχέση μεταξύ φύλου και επιπέδου αρθρίτιδας στους οσφυϊκούς σπονδύλους η γραφική παράσταση κατασκευάζεται αν επιλέξουμε το *Display clustered bar chart* στο παράθυρο διαλόγου *Crosstabs* που ανοίγει μέσω της διαδικασίας *Analyze* → *Descriptive Statistics* → *Crosstabs*. Αυτή δίνεται στο Σχήμα 6.1 και μας δείχνει εποπτικά τη συμμετοχή των δύο φύλων στα τρία επίπεδα πάθησης.



Σχήμα 6.1. Ραβδόγραμμα σχέσης φύλου και επιπέδου αρθρίτιδας στους οσφυϊκούς σπονδύλους

Εναλλακτικά το γράφημα αυτό γίνεται και από το *Graphs* → *Legacy Dialogs* → *Bar*. Στο παράθυρο διαλόγου επιλέγουμε *Clustered* και *Summaries for groups of cases* και κάνουμε κλικ στο *Define*. Στο νέο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τη μεταβλητή sex στο πλαίσιο *Category Axis* και τη μεταβλητή osteophytosis στο πλαίσιο *Define clusters by*. Με κλικ στο *OK* παίρνουμε πάλι το Σχήμα 6.1. Μια εναλλακτική μορφή αυτού του σχήματος προκύπτει αν επιλέξουμε *Stacked* αντί για *Clustered* (Σχήμα 6.2) που ουσιαστικά δίνει τις ίδιες πληροφορίες με αυτές του Σχήματος 6.1.



Σχήμα 6.2. Σωρευμένο ραβδόγραμμα σχέσης φύλου και επιπέδου αρθρίτιδας στους οσφυϊκούς σπονδύλους

6.4 ΑΝΑΛΥΣΗ LOGLINEAR

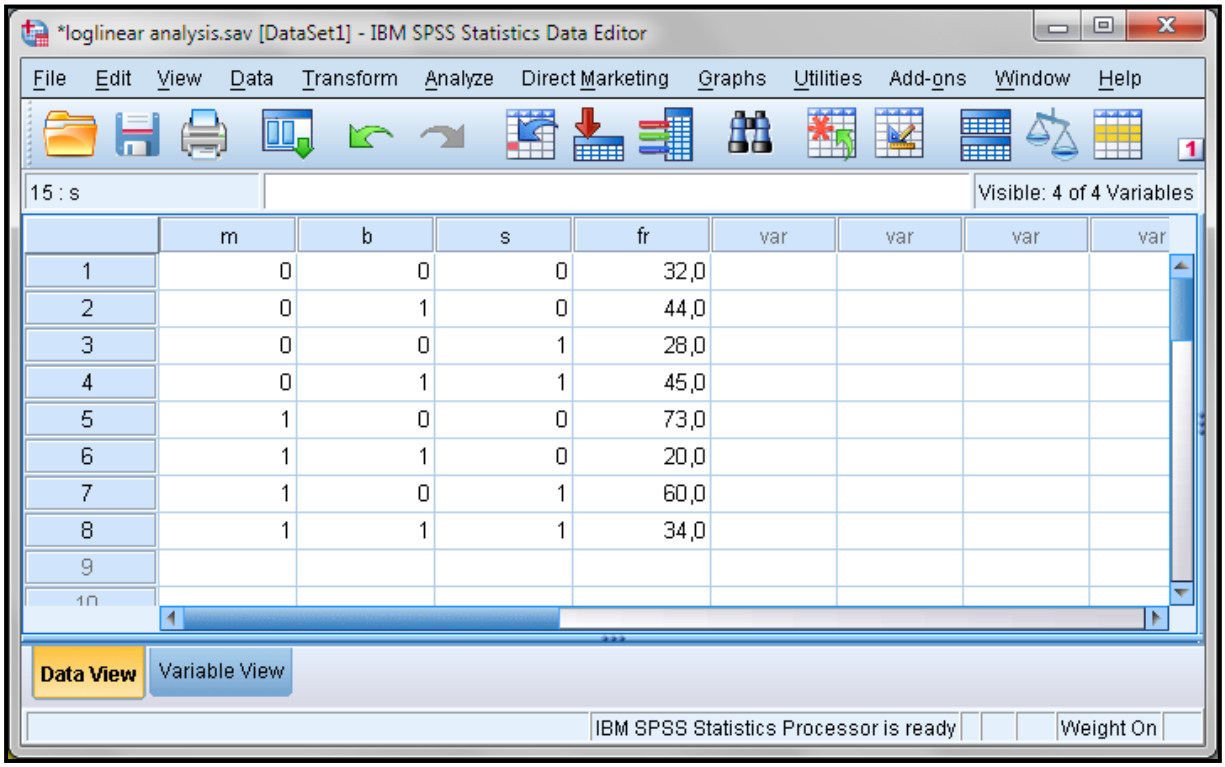
Η Ανάλυση Loglinear χρησιμοποιείται για να μελετήσουμε τη συσχέτιση μεταξύ τριών ή περισσότερων κατηγορικών μεταβλητών. Το πρόβλημα αυτό θα το εξετάσουμε μέσα από το παρακάτω παράδειγμα.

Παράδειγμα

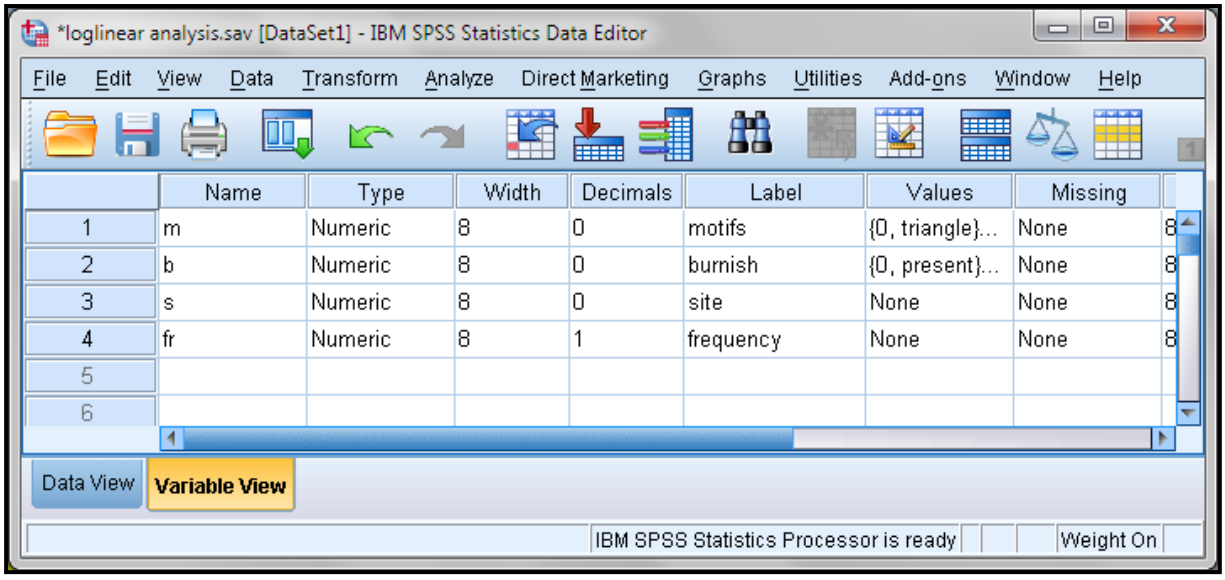
Έστω ότι θέλουμε να μελετήσουμε τη συσχέτιση μεταξύ των διακοσμητικών μοτίβων (τρίγωνα-τετράγωνα) και της παρουσίας επιφανειακού βερνικιού (παρόν-απόν) σε κεραμικά αγγεία από δύο αρχαιολογικές θέσεις (0 – 1). Συγκεκριμένα ενδιαφερόμαστε να εξετάσουμε αν υπάρχει διαφοροποίηση στα αγγεία αυτών των θέσεων. Τα δεδομένα δίνονται στα Σχήματα 6.3 και 6.4.

◆ Στην ανάλυση Loglinear προσπαθούμε να προσδιορίσουμε το καλύτερο μοντέλο που προβλέπει τις συχνότητες του αρχικού πίνακα. Το απλούστερο

μοντέλο είναι των **ανεξάρτητων μεταβλητών (independence model)**. Στο παράδειγμα που εξετάζουμε η συχνότητα εξαρτάται από τις κατηγορικές μεταβλητές m, b και s. Σε αυτή την περίπτωση το ανεξάρτητο μοντέλο δίνεται από τη σχέση:



Σχήμα 6.3. Διάταξη δεδομένων για loglinear analysis



Σχήμα 6.4. Ορισμός μεταβλητών Σχήματος 6.3

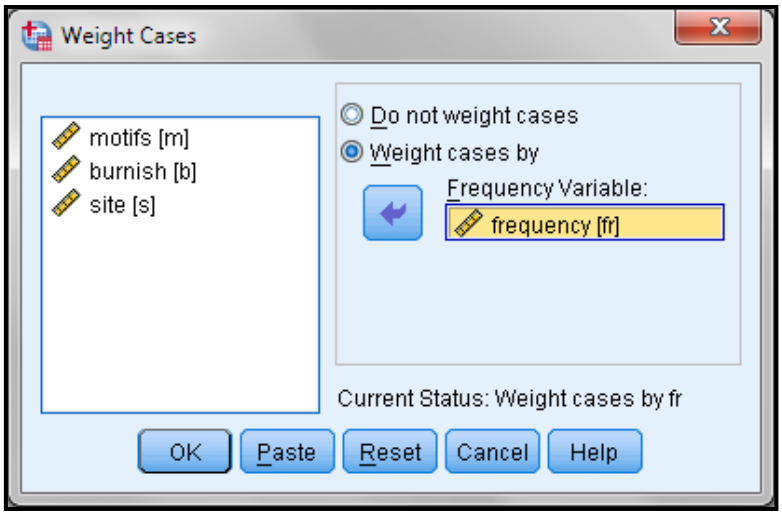
$$\ln(fr_{ijk}) = c_0 + c_1 m_i + c_2 b_j + c_3 s_k$$

όπου c_0, c_1, c_2, c_3 είναι προσαρμόσιμοι παράμετροι που υπολογίζονται με το SPSS και m_i, b_j, s_k είναι οι κατηγορικές που μεταβλητές στο παράδειγμα που εξετάζουμε και παίρνουν τις τιμές 0 και 1. Το ανεξάρτητο μοντέλο σπάνια περιγράφει ικανοποιητικά τα δεδομένα. Για το λόγο αυτό, συνήθως ξεκινάμε από το **κορεσμένο μοντέλο (saturated model)**

$$\ln(fr_{ijk}) = c_0 + c_1 m_i + c_2 b_j + c_3 s_k + c_4 mb_{ij} + c_5 ms_{ik} + c_6 bs_{jk} + c_7 mbs_{ijk}$$

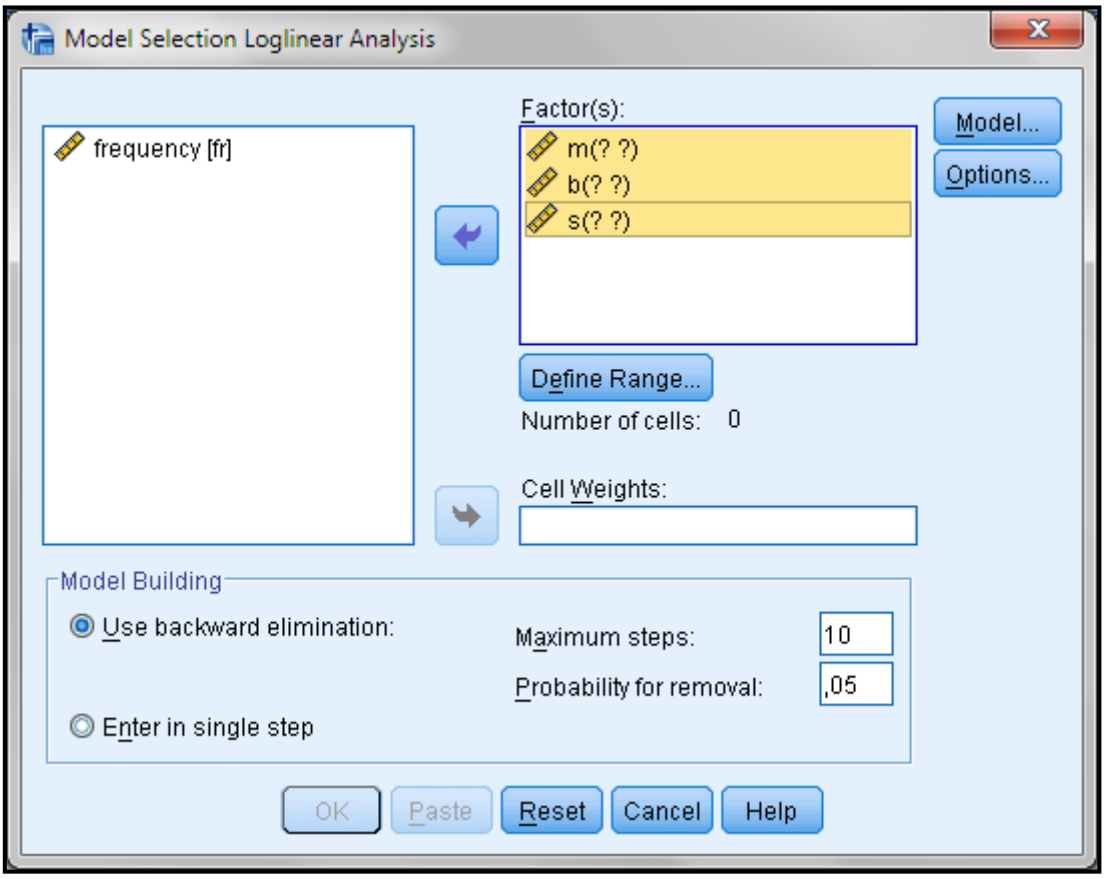
και απαλείφουμε διαδοχικά έναν-έναν τους στατιστικά μη σημαντικούς όρους μέχρι να καταλήξουμε στο μοντέλο που προβλέπει τις συχνότητες του αρχικού πίνακα, Σχήμα 6.3. Στο κορεσμένο μοντέλο οι μεταβλητές mb_{ij}, ms_{ik}, bs_{jk}, mbs_{ijk} ονομάζονται αλληλεπιδράσεις (interactions) και εκφράζουν την αλληλεπίδραση των μεταβλητών m_i - b_j, m_i - s_k, b_j - s_k και m_i - b_j - s_k, αντίστοιχα.

Για να εφαρμόσουμε ανάλυση Loglinear στο SPSS, από τη γραμμή εργαλείων κάνουμε κλικ στην επιλογή *Data* → *Weight cases*. Ακολούθως ενεργοποιούμε την επιλογή *Weight cases by* και μεταφέρουμε τη μεταβλητή frequency στο πλαίσιο διαλόγου Frequency Variable (Σχήμα 6.5). Με τον τρόπο αυτό το πρόγραμμα καταλαβαίνει ότι η μεταβλητή frequency αντιστοιχεί σε συχνότητες.

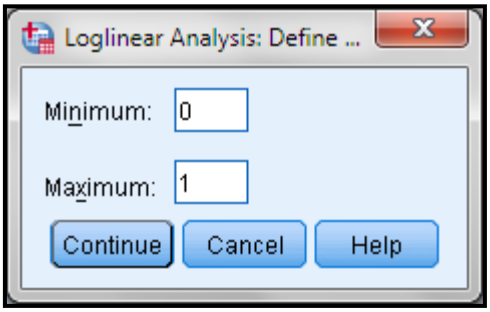


Σχήμα 6.5. Παράθυρο διαλόγου Weight Cases

Στη συνέχεια ακολουθούμε την πορεία: *Analyze* → *Loglinear* → *Model Selection*. Επιλέγουμε τις μεταβλητές των οποίων την αλληλεπίδραση θέλουμε να εξετάσουμε και τις μεταφέρουμε στο πλαίσιο *Factors*, όπως φαίνεται στο Σχήμα 6.6. Κάνουμε κλικ στο κουμπί *Define Range* και για κάθε μεταβλητή ορίζουμε τη μέγιστη και την ελάχιστη τιμή που λαμβάνει (Σχήμα 6.7). Στο παράδειγμα που μελετάμε για όλες τις μεταβλητές έχουμε ορίσει ως τιμές κωδικοποίησης τις τιμές 0 και 1.



Σχήμα 6.6. Πλαίσιο διαλόγου Model Selection Loglinear Analysis



Σχήμα 6.7 Πλαίσιο ορισμού τιμών κωδικοποίησης των μεταβλητών

Στη συνέχεια κάνουμε κλικ στο κουμπί *Options* και ενεργοποιούμε την επιλογή *Association table* ώστε στα αποτελέσματα να λάβουμε έναν πίνακα συσχέτισης (χ^2) μεταξύ όλων των μεταβλητών. Με κλικ στο *Continue* και *OK* παίρνουμε πολλούς πίνακες, από τους οποίους ενδιαφέρον παρουσιάζουν οι παρακάτω:

Πίνακας 6.4. Ο πίνακας Goodness-of-Fit Tests

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	.000	0	.
Pearson	.000	0	.

Υπάρχουν δύο πίνακες Goodness-of-Fit tests. Από αυτούς ο πρώτος είναι ο Πίνακας 6.4 και αναφέρεται στο κορεσμένο μοντέλο. Στον πίνακα αυτόν η τιμή Chi-Square είναι 0 και αυτό σημαίνει πως το κορεσμένο μοντέλο περιγράφει απόλυτα καλά τα δεδομένα των συχνοτήτων. Αυτό φαίνεται και από τον προηγούμενο πίνακα, Cell counts and Residuals, όπου παρατηρούμε ότι οι αρχικές (Observed) συχνότητες και οι προβλεπόμενες (Expected) ταυτίζονται.

Ο Πίνακας K-Way and Higher-Order Effects μας δίνει τις μεταβλητές που μπορούν να απομακρυνθούν από το μοντέλο χωρίς να επηρεάσουν σημαντικά τα αποτελέσματα. Σε αυτό τον πίνακα εστιάζουμε στο πλαίσιο K-way Effects. Παρατηρούμε ότι η επίδραση των μεμονωμένων μεταβλητών γενικά (K=1) στο μοντέλο είναι στατιστικά σημαντική (Sig.=0.002). Επίσης, οι διμερείς αλληλεπιδράσεις μεταξύ των μεταβλητών (motifs-burnish, motifs-site, burnish-site) επιδρούν σημαντικά στο μοντέλο (Sig.=0). Αντίθετα, η τριμερής αλληλεπίδραση των μεταβλητών (motifs-site-burnish) έχει μία στατιστικά μη σημαντική επίδραση (Sig.=0.225) και συνεπώς η τελευταία αυτή μεταβλητή μπορεί να απαλειφθεί.

Πίνακας 6.5. Πίνακας K-Way and Higher-Order Effects

K-Way and Higher-Order Effects

			Likelihood Ratio		Pearson		
	K	df	Chi-Square	Sig.	Chi-Square	Sig.	Number of Iterations
K-way and Higher Order Effects[a]	1	7	49,655	,000	51,000	,000	0
	2	4	37,868	,000	36,450	,000	2
	3	1	1,476	,224	1,473	,225	4
K-way Effects[b]	1	3	11,787	,008	14,550	,002	0
	2	3	36,393	,000	34,977	,000	0
	3	1	1,476	,224	1,473	,225	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Πίνακας 6.6. Ο Πίνακας Partial associations

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
m*b	1	33,275	,000	2
m*s	1	,671	,413	2
b*s	1	3,681	,055	2
m	1	4,307	,038	2
b	1	7,468	,006	2
s	1	,012	,913	2

Ο επόμενος πίνακας (Partial associations) μας δίνει μια πιο λεπτομερή εικόνα για τη συνεισφορά των μεταβλητών αφού αφαιρεθεί η επίδραση motifs-site-burnish. Παρατηρούμε ότι τώρα η μεταβλητή site (s) και οι αλληλεπιδράσεις της, motifs-site (m*s), burnish-site (b*s), δεν είναι στατιστικά σημαντικές. Συνεπώς μπορούμε να καταλήξουμε στο συμπέρασμα ότι δεν υπάρχει διαφοροποίηση στα αγγεία των αρχαιολογικών θέσεων που εξετάζονται.

7. ΠΑΛΙΝΔΡΟΜΗΣΗ-ΣΥΣΧΕΤΙΣΗ

7.1 ΠΑΛΙΝΔΡΟΜΗΣΗ

Σ' ένα μεγάλο αριθμό προβλημάτων έχουμε πειραματικά δεδομένα της γενικής μορφής (x, y), και απαιτείται να προσδιορίσουμε την εξίσωση που τα περιγράφει. Η διαδικασία εύρεσης της εξίσωσης αυτής ονομάζεται **παλινδρόμηση** (regression) και είναι ιδιαίτερα χρήσιμη επειδή επιτρέπει να αντικαθίσταται ένας πίνακας δεδομένων από μια απλή εξίσωση. Το κριτήριο, το οποίο ορίζει τον καλύτερο τρόπο περιγραφής των πειραματικών δεδομένων ονομάζεται κριτήριο ***των ελαχίστων τετραγώνων*** και ορίζει ως καλύτερη καμπύλη εκείνη που περνά μέσα από τα σημεία (x_i, y_i) και για την οποία το άθροισμα των τετραγώνων των **υπολοίπων** είναι ελάχιστο. Το υπόλοιπο (residual) είναι η διαφορά μεταξύ πειραματικής και θεωρητικής τιμής y σε μια ορισμένη τιμή x.

Παράδειγμα

Στον Πίνακα 7.1 δίνεται η μεταβολή του μήκους του βραχίονα νηπίων σε mm με το χρόνο σε εβδομάδες. Να γίνει η γραφική παράσταση weeks - mm και να εκτιμηθεί η ηλικία δύο νηπίων με μήκος βραχίονα 50 και 55 mm, αντίστοιχα.

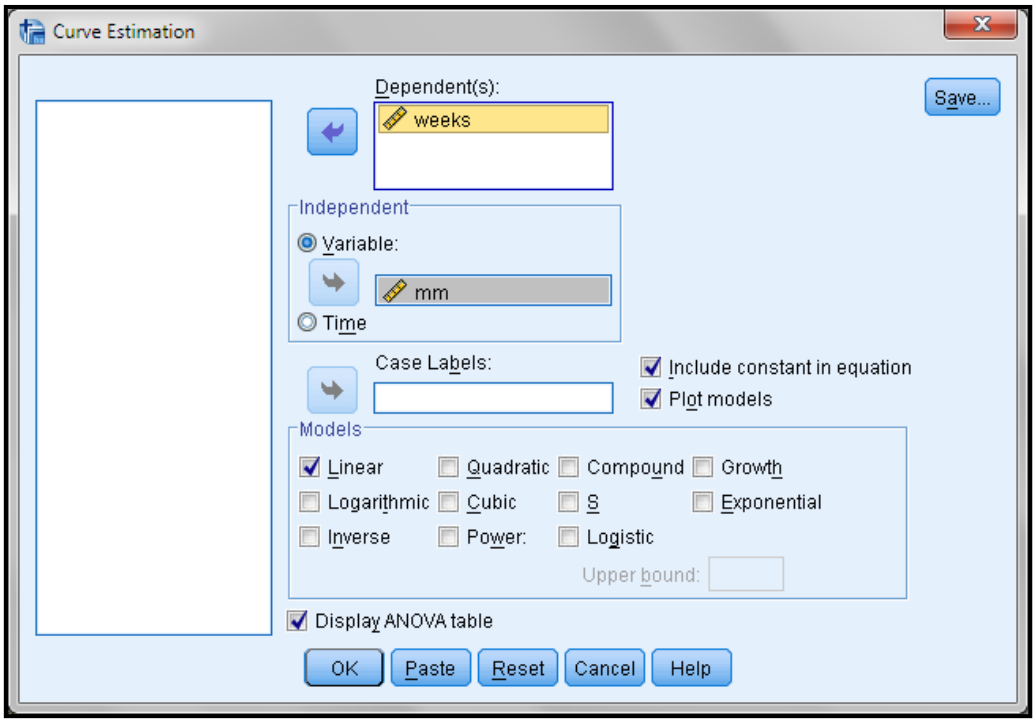
Πίνακας 7.1. Μεταβολή του μήκους του βραχίονα νηπίων σε mm με το χρόνο σε εβδομάδες

mm	weeks	mm	weeks
42	28	65	37
45	27	65	38
58	32	68	40
59	34	70	40
59	35	70	40
61	35	72	41
64	36	75	45

◆ Μεταφέρουμε τα δεδομένα σ' ένα φύλλο εργασίας του SPSS και αποφασίζουμε ποια μεταβλητή θα είναι ανεξάρτητη και ποια εξαρτημένη. Ο

γενικός κανόνας που ισχύει είναι ότι ως ανεξάρτητη μεταβλητή επιλέγεται αυτή που τη μεταβάλλουμε κατά βούληση και εξαρτημένη αυτή που την προσδιορίζουμε πειραματικά ως συνέπεια των μεταβολών της ανεξάρτητης. Στην περίπτωσή μας, εφόσον ενδιαφερόμαστε να προσδιορίσουμε την ηλικία των νηπίων ως συνάρτηση του μήκους του βραχίονα, ως ανεξάρτητη μεταβλητή θα οριστεί το μήκος του βραχίονα και εξαρτημένη η ηλικία.

Μετά την επιλογή των μεταβλητών, από το *Analyze* → *Regression* → *Curve Estimation* εισάγουμε τη μεταβλητή weeks στο πλαίσιο *Variable(s)* και τη μεταβλητή mm στο *Independent Variable*, όπως φαίνεται στο Σχήμα 7.1. Επίσης, επιλέγουμε *Display ANOVA table, Linear, Plot models* και *Include constant in equation*. Η τελευταία επιλογή γίνεται πάντα εκτός κι αν έχουμε στοιχεία ότι όταν x = 0 τότε και y = 0. Με κλικ στο *OK* παίρνουμε αρκετούς πίνακες και τη γραφική παράσταση του Σχήματος 7.2. Από τους πίνακες ενδιαφέρον έχει ο Πίνακας 7.2, που δίνεται παρακάτω.

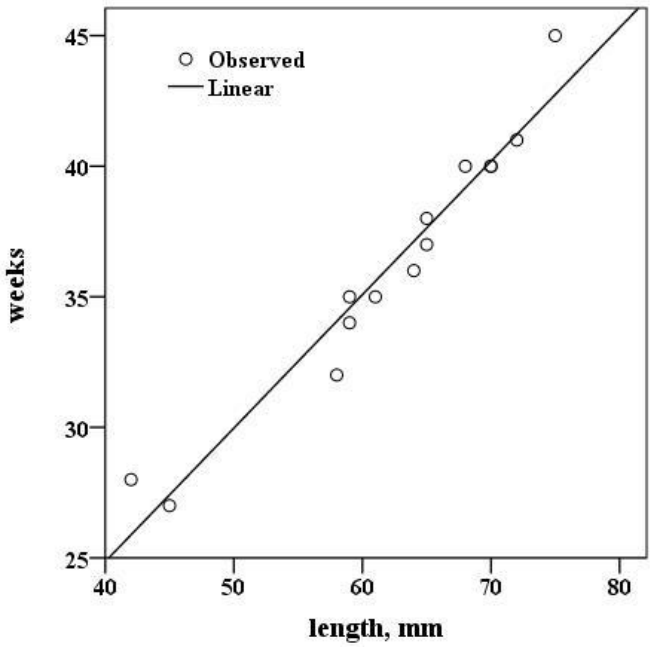


Σχήμα 7.1. Πλαίσιο διαλόγου Curve estimation

Πίνακας 7.2. Συντελεστές προσαρμογής

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
mm	,511	,035	,973	14,541	,000
(Constant)	4,420	2,215		1,996	,069



Σχήμα 7.2. Η γραφική παράσταση weeks – mm με την ευθεία των ελαχίστων τετραγώνων.

Από τον πίνακα προκύπτει ότι η εξίσωση της ευθείας (y = a + bx) είναι η:

$$y = 4.42 + 0.511x$$

Η τυπική απόκλιση της σταθεράς a είναι 2.215 και της κλίσης b είναι 0.035, δηλαδή έχουμε a = 4.42 ± 2.215 και b = 0.511 ± 0.035. Η τελευταία στήλη μας ενημερώνει αν μια σταθερά των ελαχίστων τετραγώνων, a ή b, είναι στατιστικά σημαντική. **Πρέπει η τιμή Sig. να είναι μικρότερη από 0.05**. Παρατηρούμε ότι η σταθερά a μπορεί να θεωρηθεί ως στατιστικά μη σημαντική. Δηλαδή θα μπορούσαμε στο παράθυρο του Σχήματος 7.1 να μην επιλέγαμε Include constant

in equation. Γενικά αν μια σταθερά είναι στατιστικά μη σημαντική μπορεί να απαλειφθεί από τη μελέτη, εκτός κι αν υπάρχουν ισχυροί λόγοι να παραμείνει.

Για να προβλέψουμε τώρα την ηλικία των νηπίων με μήκος βραχίονα 50 και 55 mm, αντίστοιχα, απλά κάνουμε τις πράξεις:

$$4{,}42 + 0{,}511*50 = 29{,}97 \approx 30 \text{ βδομάδες}$$

$$4{,}42 + 0{,}511*55 = 32{,}525 \approx 32{,}5 \text{ βδομάδες}$$

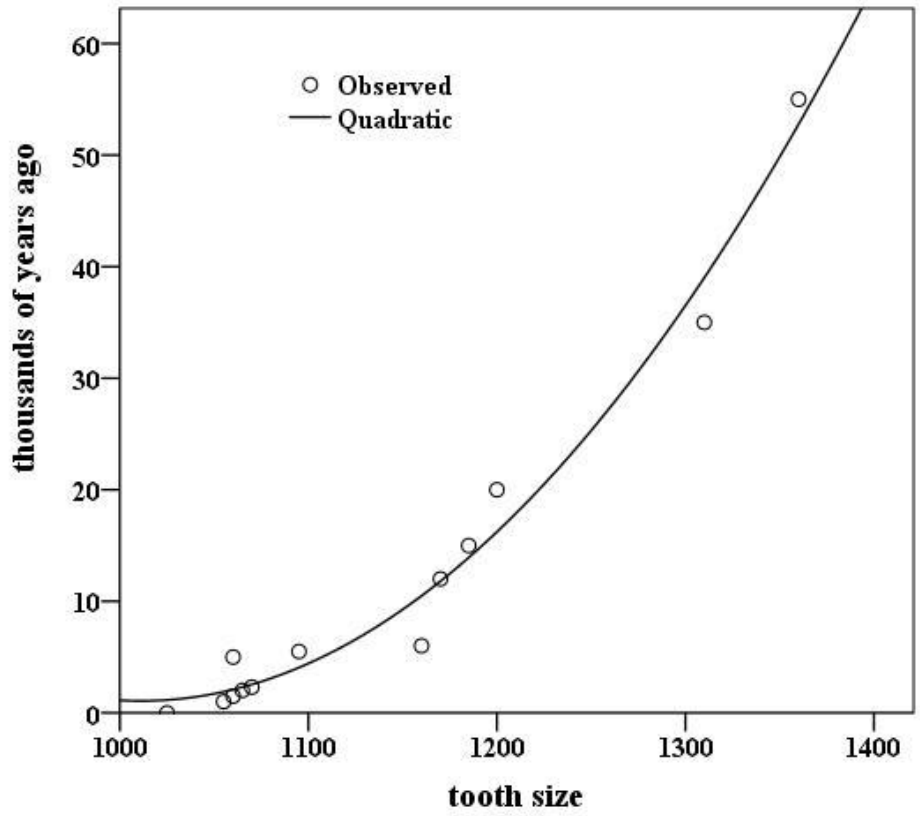
Παράδειγμα

Στον Πίνακα 7.3 δίνεται η μεταβολή των διαστάσεων των δοντιών με το πέρασμα των χιλιετιών. Να γίνει η γραφική παράσταση των τιμών του πίνακα και να εκτιμηθεί η χρονολογία των δειγμάτων 1150 και 1250 mm^2.

Πίνακας 7.3. Μεταβολή των δοντιών με το πέρασμα χιλιετιών

Thousands years ago	Tooth-size (mm^2)	Thousands years ago	Tooth-size (mm^2)
0	1025	6	1160
1	1055	12	1170
1.5	1060	15	1185
2	1065	20	1200
2.3	1070	35	1310
5	1060	55	1360
5.5	1095		

◆ Θα πρέπει καταρχήν να επιλέξουμε την ανεξάρτητη μεταβλητή. Επειδή μας ζητείται να κάνουμε εκτιμήσεις της χρονολογίας των δειγμάτων 1150 και 1250 mm^2 θα χρησιμοποιήσουμε ως ανεξάρτητη μεταβλητή το tooth-size και ως εξαρτημένη το χρόνο, years. Στη συνέχεια, ακολουθούμε ακριβώς την ίδια πορεία με αυτή στο προηγούμενο παράδειγμα, με μόνη διαφορά ότι επιλέγουμε το *Quadratic* και απενεργοποιούμε το *Linear* στο παράθυρο *Curve Estimation*. Με κλικ στο *OK* παίρνουμε το Σχήμα 7.3 και τον Πίνακα 7.4.



Σχήμα 7.3. Η γραφική παράσταση years – tooth size με την καμπύλη των ελαχίστων τετραγώνων

Πίνακας 7.4. Συντελεστές καμπύλης

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
tooth size	-,859	,184	-5,520	-4,677	,001
tooth size ** 2	,000	,000	6,479	5,490	,000
(Constant)	435,137	108,157		4,023	,002

Οι συντελεστές της καμπύλης $y = a + bx + cx^2$ στον πίνακα Coefficients είναι ως εξής: a = (Constant) = 435.137, b = toothsize = -0.859 και c = toothsize**2 = 0.000. (Στο SPSS το σύμβολο ** σημαίνει δύναμη). Εδώ όμως

πρέπει να προσέξουμε. Η τιμή toothsize**2 = 0.000 όταν μάλιστα είναι στατιστικά σημαντική (Sig.=0,000) σημαίνει ότι δεν είναι 0 αλλά υπάρχουν ψηφία που δεν εμφανίζονται. Για το σκοπό αυτό επιλέγουμε τον πίνακα και τον μεταφέρουμε σε ένα φύλλο του Excel. Τότε αν κάνουμε κλικ επάνω στην τιμή 0,000 της μεταβλητής toothsize**2 διαπιστώνουμε ότι αυτή είναι 0,0004246. Επίσης θα πρέπει να αυξήσουμε την ακρίβεια και της μεταβλητής toothsize = -0,858616, εφόσον θα την χρησιμοποιήσουμε σε υπολογισμούς παρακάτω.

Για να προσδιορίσουμε τη χρονολογία των δειγμάτων 1150 και 1250 mm^2, τοποθετούμε τις τιμές αυτές στη συνάρτηση που έχει προκύψει και παίρνουμε ότι οι ηλικίες των νηπίων είναι:

$$435,137 - 0,858616*1150 + 0,0004246*1150^2 = 9,26 \text{ βδομάδες}$$

$$435,137 - 0,858616*1250 + 0,0004246*1250^2 = 25,3 \text{ βδομάδες}$$

Το πόσο σημαντική είναι η ακρίβεια των μεταβλητών στις παραπάνω πράξεις φαίνεται από το γεγονός ότι αν χρησιμοποιήσουμε -0,859 αντί για 0,858616, παίρνουμε

$$435,137 - 0,859*1250 + 0,0004246*1250^2 = 24,83 \text{ βδομάδες}$$

ενώ αν θέσουμε 0 αντί για 0,0004246 θα πάρουμε

$$435,137 - 0,859*1250 = -638,6 \text{ βδομάδες}$$

Δηλαδή ένα εντελώς παράλογο αποτέλεσμα.

Παράδειγμα

Η ποσότητα y του νερού που εξατμίζεται από το έδαφος εξαρτάται από τη μέγιστη (T_1) και την ελάχιστη (T_2) θερμοκρασία του εδάφους και τη μέγιστη (T_3) και την ελάχιστη (T_4) θερμοκρασία του αέρα σύμφωνα με τα δεδομένα του Πίνακα 26. Να προσδιοριστεί το γραμμικό μοντέλο, δηλαδή η συνάρτηση

$$y = a_0 + a_1T_1 + a_2T_2 + a_3T_3 + a_4T_4$$

Πίνακας 7.5. Δεδομένα εξάρτησης της ποσότητας x του νερού που εξατμίζεται από το έδαφος από τις θερμοκρασίες T_1, T_2, T_3, T_4.

y	T_1	T_2	T_3	T_4
30	28	18	29	15
34	28	18	30	16
33	26	18	28	17
26	27	19	28	18
41	28	20	31	20
10	23	18	25	19
12	22	18	25	20
20	23	19	28	20
31	28	20	31	21
38	30	22	32	24
43	31	22	32	24
47	32	23	34	24
45	31	22	34	23
45	31	22	33	21
22	27	20	30	20
5	15	20	28	20
30	28	15	30	18
29	28	21	30	20
23	25	21	31	21

◆ Μεταφέρουμε τα δεδομένα στο SPSS σε στήλες ανάλογες με του Πίνακα 7.5. Ακολουθούμε την πορεία: *Analyze* → *Regression* → *Linear* και εισάγουμε τη μεταβλητή y στο πλαίσιο *Dependent* και τις μεταβλητές T_1, T_2, T_3 και T_4 στο *Independent(s)*. Από το Options επιλέγουμε το Include constant in equation και στο Method επιλέγουμε τη μέθοδο που θα χρησιμοποιηθεί για τον υπολογισμό των σταθερών της συνάρτησης. Όταν επιλέγουμε Enter το πρόγραμμα υπολογίζει όλες τις σταθερές, στην περίπτωση που εξετάζουμε τις σταθερές a_0, a_1, a_2, a_3, a_4. Αν επιλέξουμε Backward το πρόγραμμα αρχικά υπολογίζει όλες τις σταθερές και μετά αρχίζει να αφαιρεί μία-μία τις στατιστικά μη σημαντικές. Με την επιλογή *Forward* το πρόγραμμα πρώτα εισάγει τον σταθερό όρο και μετά τη σταθερά που αντιστοιχεί στη μεταβλητή που έχει τη μεγαλύτερη συσχέτιση με την εξαρτημένη

μεταβλητή. Εξετάζεται αν είναι στατιστικά σημαντική και μετά το πρόγραμμα εισάγει την επόμενη μεταβλητή με την καλύτερη συσχέτιση με την εξαρτημένη μεταβλητή κ.ο.κ. Τέλος, η επιλογή *Stepwise* είναι συνδυασμός των μεθόδων *Backward* και *Forward*. Γενικά οι μέθοδοι *Stepwise*, *Forward* και *Backward* χρησιμοποιούνται για να πάρουμε μόνο τους στατιστικά σημαντικούς όρους, ενώ η Enter όλες τις σταθερές. Δυστυχώς και οι τρεις μέθοδοι δε δίνουν πάντα το ίδιο αποτέλεσμα, οπότε καλούμαστε να επιλέξουμε εμείς τη μέθοδο με άλλα κριτήρια. Ένα από αυτά είναι η φυσική σημασία των όρων του συμμετέχουν στο μοντέλο.

Αν στο παράδειγμα που εξετάζουμε επιλέξουμε το *Enter,* παίρνουμε τον Πίνακα 7.6, ενώ με Backward τον Πίνακα 7.7. Είναι χαρακτηριστικό ότι ο πίνακας της μεθόδου *Backward* περιέχει όλα τα βήματα μέχρι το τελικό αποτέλεσμα.

Παρατηρούμε ότι ο φυσικός νόμος μπορεί να εκφραστεί ως:

$$y = -75.494 + 1.933T_1 + 1.776T_3$$

Πίνακας 7.6. Αποτελέσματα με τη μέθοδο Enter

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-75.412	12.218		-6.172	.000
	T1	1.882	.386	.621	4.876	.000
	T2	.212	.959	.035	.221	.829
	T3	1.990	.774	.418	2.572	.022
	T4	-.465	.659	-.098	-.705	.492

a. Dependent Variable: x

Πίνακας 7.7. Αποτελέσματα με τη μέθοδο Backward

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-75.412	12.218		-6.172	.000
	T1	1.882	.386	.621	4.876	.000
	T2	.212	.959	.035	.221	.829
	T3	1.990	.774	.418	2.572	.022
	T4	-.465	.659	-.098	-.705	.492
2	(Constant)	-75.037	11.710		-6.408	.000
	T1	1.870	.370	.617	5.058	.000
	T3	2.063	.677	.433	3.047	.008
	T4	-.367	.473	-.077	-.776	.450
3	(Constant)	-75.494	11.549		-6.537	.000
	T1	1.933	.356	.638	5.425	.000
	T3	1.776	.560	.373	3.172	.006

a. Dependent Variable: x

7.2 ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

Ένα θέμα που σχετίζεται έμμεσα με την παλινδρόμηση και τα ελάχιστα τετράγωνα είναι το πρόβλημα της συσχέτισης (correlation) δύο μεταβλητών. Είναι χρήσιμο σε αρκετές περιπτώσεις να γνωρίζουμε αν δύο τυχαίες μεταβλητές σχετίζονται ή όχι. Αν δηλαδή η μεταβολή της μιας μεταβάλει και την άλλη.

7.2.1 ΣΥΝΤΕΛΕΣΤΕΣ PEARSON ΚΑΙ SPEARMAN

Για να ελέγξουμε αν δύο μεταβλητές, x και y, σχετίζονται, υπολογίζουμε συνήθως τον **συντελεστή Pearson**, **r**. Ο συντελεστής r παίρνει τιμές στο διάστημα από −1 έως 1. Αρνητικές τιμές του r σημαίνουν ότι όταν η μεταβλητή x αυξάνει, η y ελαττώνεται και το αντίστροφο. r = 0 σημαίνει παντελή έλλειψη **συσχέτισης** και r θετικό σημαίνει ότι όταν η μια μεταβλητή αυξάνει, αυξάνει και η άλλη.

Θα πρέπει πάντως να τονιστεί ότι ο συντελεστής Pearson χρησιμοποιείται μόνο όταν τα δεδομένα ακολουθούν την κανονική κατανομή. Αν δεν ακολουθούν την κανονική κατανομή, υπολογίζουμε τον **συντελεστή Spearman, ρ**, που επίσης παίρνει τιμές στο διάστημα από −1 έως 1, αλλά ανήκει στις μη παραμετρικές μεθόδους.

Παράδειγμα

Να εξετασθεί αν υπάρχει συσχέτιση μεταξύ των μεταβλητών height και body mass του αρχείου osteological data.sav.

◆ Ανοίγουμε το αρχείο osteological data.sav και ελέγχουμε την κανονικότητα των δειγμάτων. Θα πρέπει εδώ να τονίσουμε ότι ήδη έχουμε εξετάσει τη μεταβλητή height και διαπιστώσαμε ότι ακολουθεί την κανονική κατανομή. Το ίδιο ισχύει και για τη μεταβλητή body mass. Επειδή αυτός ο έλεγχος είναι ιδιαίτερα απλός μπορούμε να τον επαναλάβουμε και εδώ. Πηγαίνουμε *Analyze* → *Descriptive Statistics* → *Explore*, στο παράθυρο διαλόγου εισάγουμε και τις δύο μεταβλητές, height και body mass, στο πλαίσιο *Dependent List* και κάνουμε κλικ στο κουμπί *Plots*. Στο πλαίσιο διαλόγου που εμφανίζεται κάνουμε κλικ στην επιλογή *None* στο πάνελ των *Boxplots,* απενεργοποιούμε την επιλογή *Stem-and-leaf* στο πάνελ Descriptive και επιλέγουμε μόνο το *Normality plots with tests.* Από τον πίνακα των αποτελεσμάτων, Πίνακας 7.8, παρατηρούμε ότι πράγματι οι μεταβλητές ακολουθούν την κανονική κατανομή.

Πίνακας 7.8. Αποτελέσματα ελέγχου κανονικότητας

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
height	,098	48	,200*	,963	48	,134
body mass	,116	48	,111	,965	48	,167

a. Lilliefors Significance Correction

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

Επομένως μπορούμε να χρησιμοποιήσουμε τον συντελεστή Pearson και προφανώς τον συντελεστή Spearman που δεν υπόκειται σε περιορισμούς. Στη συνέχεια ακολουθούμε τη διαδικασία *Analyze* → *Correlate* → *Bivariate*. Στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές height και body mass στο πλαίσιο *Variables* και επιλέγουμε στο *Correlation Coefficients* τα κριτήρια *Pearson* και Spearman. Επίσης επιλέγουμε (αν δεν είναι default) και το *Flag significant correlations*. Με την επιλογή αυτή το πρόγραμμα θα μας ενημερώνει και για το

επίπεδο σημαντικότητας των αποτελεσμάτων. Με κλικ στο *OK* παίρνουμε τον Πίνακα 7.9. Βλέπουμε ότι υπάρχει θετική συσχέτιση των μεταβλητών και μάλιστα υψηλή συσχέτιση (r = 0,863 και ρ = 0,878).

Πίνακας 7.9. Αποτελέσματα συσχέτισης με τον συντελεστή Pearson (επάνω) και Spearman (κάτω)

Correlations

		height	body mass
height	Pearson Correlation	1	,863**
	Sig. (2-tailed)		,000
	N	50	48
body mass	Pearson Correlation	,863**	1
	Sig. (2-tailed)	,000	
	N	48	48

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

Correlations

			height	body mass
Spearman's rho	height	Correlation Coefficient	1,000	,878**
		Sig. (2-tailed)	.	,000
		N	50	48
	body mass	Correlation Coefficient	,878**	1,000
		Sig. (2-tailed)	,000	.
		N	48	48

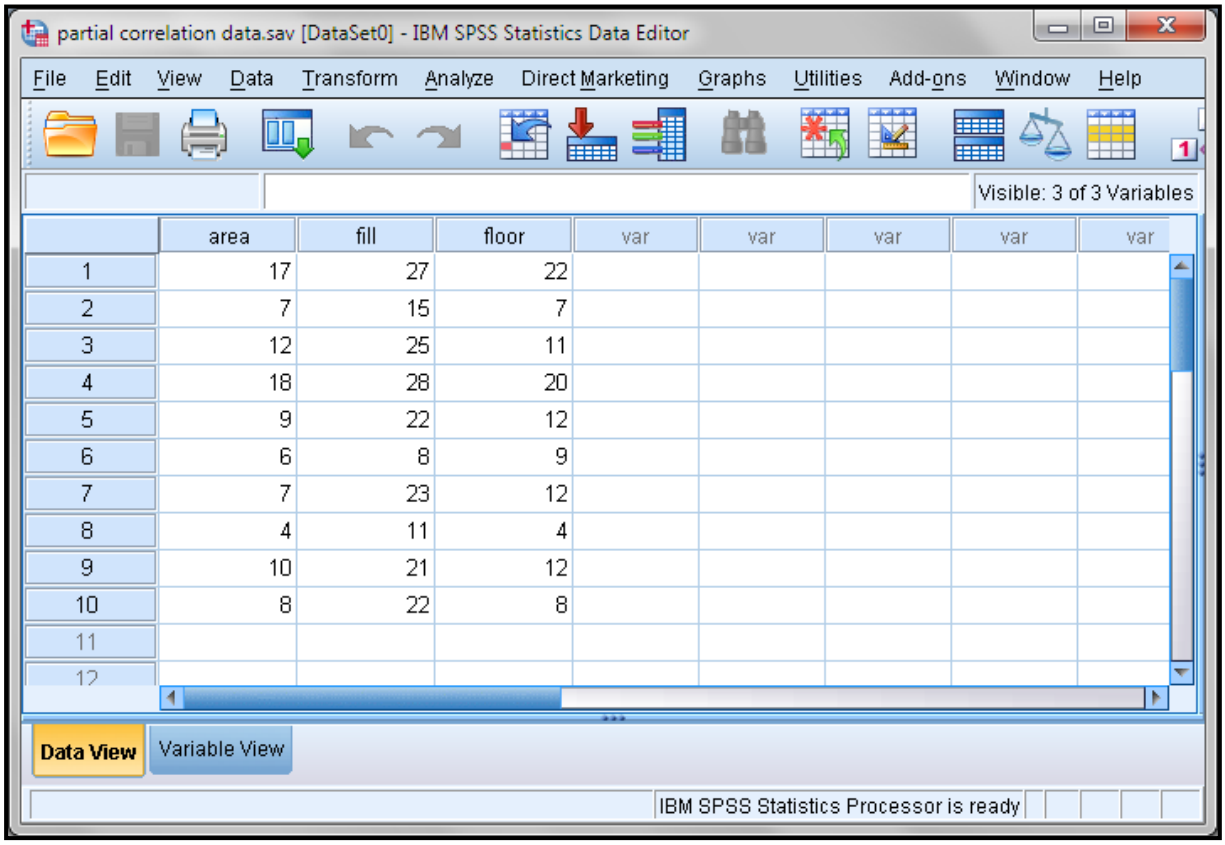
**. Correlation is significant at the 0.01 level (2-tailed).

7.2.2. ΜΕΡΙΚΗ ΣΥΣΧΕΤΙΣΗ (Partial correlation)

Η μερική συσχέτιση (Partial correlation) χρησιμοποιείται προκειμένου να εξετάσουμε την αλληλεπίδραση δύο μεταβλητών, ενώ συγχρόνως ελέγχουμε την επίδραση μίας τρίτης μεταβλητής. Πιο συγκεκριμένα, εξετάζουμε αν δύο μεταβλητές συσχετίζονται όταν η επίδραση μιας τρίτης μεταβλητής θεωρείται σταθερή. Οι μεταβλητές μπορεί να είναι συνεχείς, κατηγορικές (dichotomous), όπως για παράδειγμα το φύλο (άντρας-γυναίκα), ή συνδυασμός συνεχών και κατηγορικών.

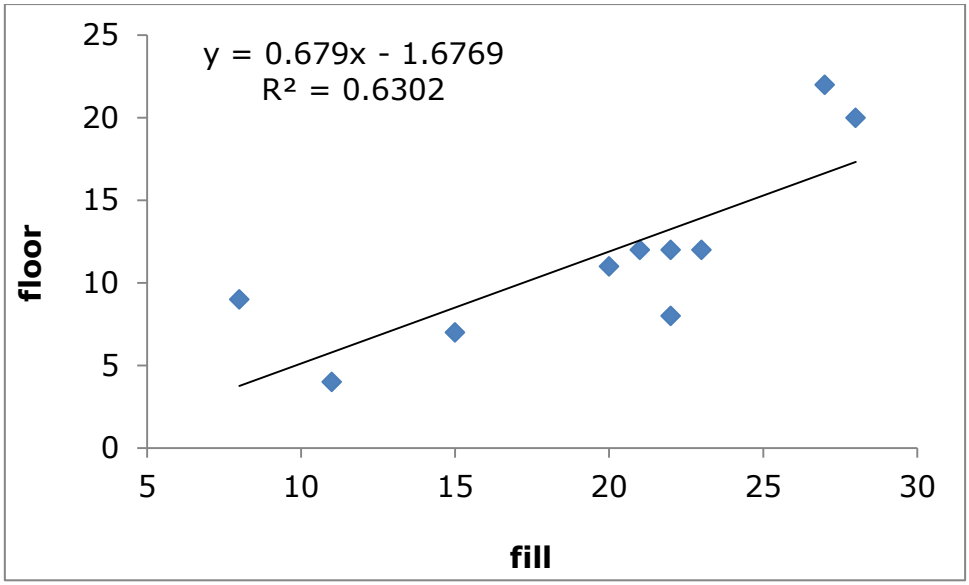
Παράδειγμα

Σε συγκρότημα δωματίων Ρωμαϊκής εποχής καταγράφηκαν μεταξύ των άλλων το εμβαδόν των δωματίων (σε m^2) και ο αριθμός των διαφορετικών τύπων λίθων που χρησιμοποιήθηκαν στα δάπεδα και σε γεμίσματα. Τα δεδομένα της καταγραφής δίνονται στο Σχήμα 7.4, όπου η μεταβλητή area παρέχει το εμβαδόν κάθε δωματίου, η μεταβλητή fill εκφράζει το πλήθος των διαφορετικών τύπων λίθων στα γεμίσματα και η floor το πλήθος των λίθων στα δάπεδα. Να εξετασθεί η συσχέτιση μεταξύ των μεταβλητών fill και floor.



Σχήμα 7.4. Δεδομένα παραδείγματος για μερική συσχέτιση

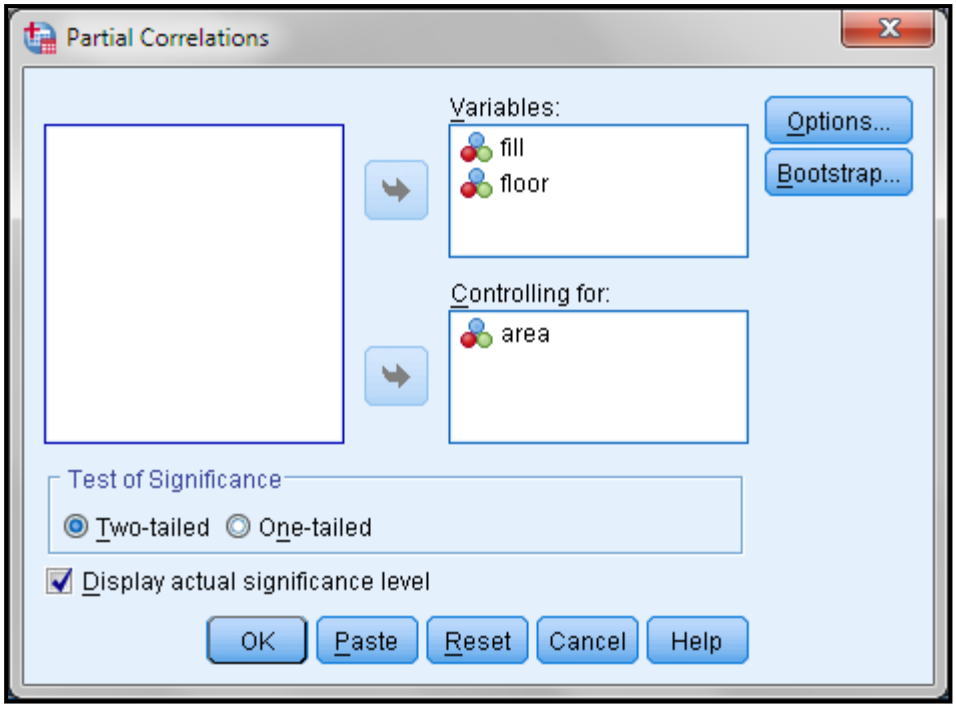
◆ Αν κάνουμε τη γραφική παράσταση fill – floor, παρατηρούμε ότι υπάρχει μια θετική συσχέτιση ανάμεσα σε αυτές τις δύο μεταβλητές (Σχήμα 7.5). Όμως η συνολική επιφάνεια κάθε δωματίου αναμένεται να παίζει ρόλο στη συσχέτιση αυτή, δεδομένου ότι δωμάτια με μεγάλη επιφάνεια θα έχουν μεγαλύτερο πλήθος διαφορετικών τύπων λίθων. Για να δούμε την επίδραση της επιφάνειας των δωματίων (area) στην συσχέτιση των μεταβλητών fill και floor εργαζόμαστε ως εξής.



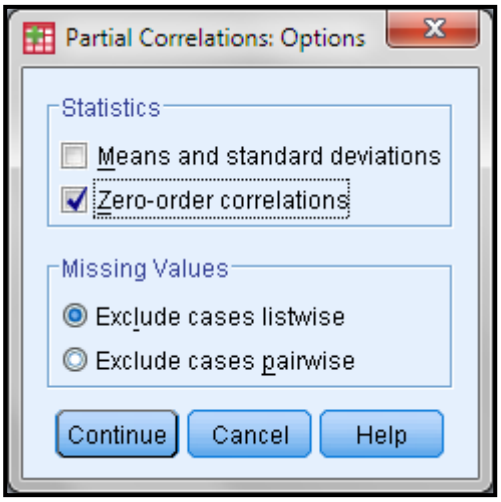
Σχήμα 7.5. Διάγραμμα μεταβολής της μεταβλητής floor με τη μεταβλητή fill

Αφού διατάξουμε τα δεδομένα στο φύλο του SPSS όπως φαίνεται στο Σχήμα 7.4, ακολουθούμε την πορεία: *Analyze* → *Correlate* → *Partial*. Στο παράθυρο διαλόγου που εμφανίζεται μεταφέρουμε τις μεταβλητές fill και floor στο πλαίσιο *Variables* και τη μεταβλητή *area* στο πλαίσιο *Controlling for*, όπως φαίνεται στο Σχήμα 7.6. Στη συνέχεια κάνουμε κλικ στο κουμπί *Options* και ενεργοποιούμε την επιλογή *Zero-order correlations*, όπως φαίνεται στο Σχήμα 7.7. Η επιλογή αυτή θα μας δώσει στα αποτελέσματα τον συντελεστή συσχέτισης όλων των μεταβλητών χωρίς έλεγχο της επίδρασης της τρίτης μεταβλητής. Δηλαδή, στον πίνακα αποτελεσμάτων θα παρουσιάζονται, εκτός από την επίδραση της μεταβλητής area στον συντελεστή συσχέτισης των μεταβλητών fill και area, και ο συντελεστής Pearson για όλες τις απλές συσχετίσεις των μεταβλητών fill – floor, fill - area και floor - area.

Με κλικ στο *OK* παίρνουμε τα αποτελέσματα του Πίνακα 7.10, ο οποίος χωρίζεται σε δύο τμήματα. Στο επάνω τμήμα έχουμε τα αποτελέσματα των απλών συσχετίσεων όλων των μεταβλητών ανά δύο και στο κάτω την επίδραση της μεταβλητής area στη συσχέτιση των μεταβλητών fill και floor.



Σχήμα 7.6. Παράθυρο διαλόγου για Partial correlations



Σχήμα 7.7. Παράθυρο διαλόγου Options

Τα αποτελέσματα δείχνουν σημαντική θετική συσχέτιση ανάμεσα στις μεταβλητές fill και floor (r=0.758 και p=0.011) καθώς επίσης και ανάμεσα στις fill - area (r=0.809 και p=0.005) και floor - area (r=0.929 και p=0). Η συσχέτιση ανάμεσα στις μεταβλητές fill και floor όταν η επίδραση της επιφάνειας των δωματίων δεν ελέγχεται δίνεται εποπτικά στο Σχήμα 7.5.

Ωστόσο, όταν η επίδραση της επιφάνειας ελέγχεται, δηλαδή όταν εξετάζουμε τη συσχέτιση των μεταβλητών fill και floor κρατώντας σταθερή τη

μεταβλητή area, τότε ο συντελεστής συσχέτισης ουσιαστικά μηδενίζεται (r=0.032) και επιπλέον η τιμή αυτή παύει να είναι στατιστικά σημαντική (p=0.934). Δηλαδή όταν λαμβάνουμε υπόψη το εμβαδόν των δωματίων, παρατηρούμε ότι οι μεταβλητές fill και floor παύουν να συσχετίζονται.

Πίνακας 7.10. Αποτελέσματα των zero-order και partial correlations

Correlations

Control Variables			fill	floor	area
-none-[a]	fill	Correlation	1,000	,758	,809
		Significance (2-tailed)		,011	,005
		df	0	8	8
	floor	Correlation	,758	1,000	,929
		Significance (2-tailed)	,011		,000
		df	8	0	8
	area	Correlation	,809	,929	1,000
		Significance (2-tailed)	,005	,000	
		df	8	8	0
area	fill	Correlation	1,000	,032	
		Significance (2-tailed)		,934	
		df	0	7	
	floor	Correlation	,032	1,000	
		Significance (2-tailed)	,934		
		df	7	0	

a. Cells contain zero-order (Pearson) correlations.

8. ΑΝΑΛΥΣΗ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ

8.1 ΓΕΝΙΚΑ

Συχνά συσσωρεύουμε πληθώρα δεδομένων και θέλουμε να ερευνήσουμε αν υπάρχουν ομάδες δειγμάτων με παρόμοιες ιδιότητες, και ποιες είναι αυτές. Για παράδειγμα, θέλουμε να δούμε κατά πόσο υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στα κεραμικά αγγεία από τις θέσεις Παλιάμπελα και Μακρύγιαλος χρησιμοποιώντας ως μεταβλητές συγχρόνως το ύψος των αγγείων, το πλάτος, τη διάμετρο του στομίου, τη διάμετρο της βάσης και άλλες διαστάσεις. Στο ερώτημα αυτό απάντηση προσπαθεί να δώσει η **Ανάλυση Πολλών Μεταβλητών** (**Multivariate Analysis**).

Από τις αναλύσεις πολλών μεταβλητών εδώ θα εξετάσουμε τις μεθόδους: α) Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis - PCA), β) Ανάλυση σε Ομάδες (Cluster Analysis – CA), γ) Διαχωριστική Ανάλυση (Discriminant Analysis - DA) και δ) Ανάλυση Διασποράς Πολλών Μεταβλητών (Multivariate Analysis of Variance – MANOVA).

Για την εφαρμογή των μεθόδων PCA και CA δεν απαιτείται καμία παραδοχή σχετικά με τη μορφή των πληθυσμιακών κατανομών των δεδομένων. Αντίθετα, η εφαρμογή των μεθόδων DA και MANOVA προϋποθέτει τουλάχιστον την κανονικότητα των δειγμάτων.

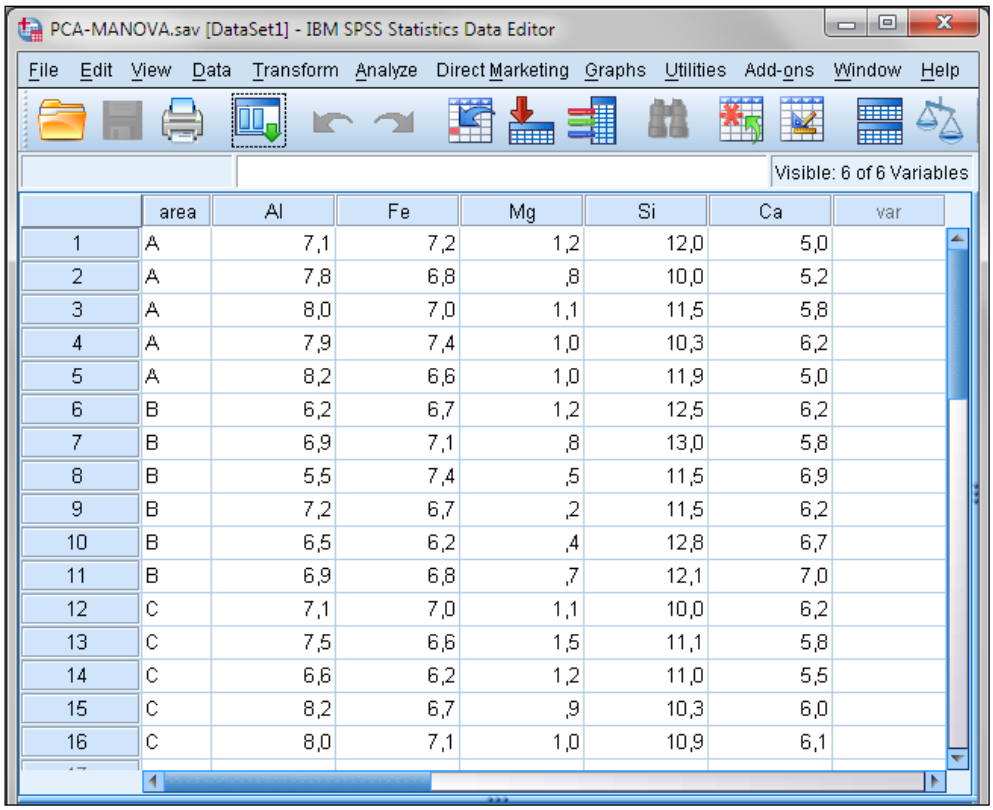
8.2 ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ - PRINCIPAL COMPONENT ANALYSIS (PCA)

Για να μπορέσουμε να εξετάσουμε αν σ’ έναν πίνακα δεδομένων υπάρχουν ομάδες ομοειδών δεδομένων (clusters) θα πρέπει να **ελαττώσουμε τις διαστάσεις του πίνακα** σε δύο ή τρεις. Ποιοτικά για να το πετύχουμε αυτό φέρνουμε έναν άξονα (μια γραμμή) μέσα από τα σημεία της γραφικής παράστασης και κατά μήκος της μεγαλύτερης διασποράς των σημείων και

προβάλουμε τα σημεία αυτά πάνω στον άξονα. Ο άξονας ονομάζεται PC1 ή πρώτη κύρια συνιστώσα. Ακολούθως φέρνουμε ένα δεύτερο άξονα, τον PC2, που είναι κάθετος στον PC1 και τον περιστρέφουμε, πάντα κάθετα στον PC1, έτσι ώστε και αυτός να είναι κατά μήκος της μεγαλύτερης διασποράς των σημείων ως προς τη διεύθυνσή του. Οι δύο αυτοί άξονες ορίζουν ένα επίπεδο. Στο επίπεδο αυτό προβάλουμε όλα τα σημεία. Συνεχίζουμε με τον ίδιο τρόπο μέχρι να καταλήξουμε με αρκετά PCs ώστε να εξηγηθεί όλη η διασπορά του δείγματος.

Παράδειγμα

Στις 6 πρώτες στήλες του πίνακα του Σχήματος 8.1 δίνονται τα αποτελέσματα της χημικής ανάλυσης ειδωλίων ίδιας χρονολογίας που βρέθηκαν σε τρεις διαφορετικές περιοχές Α, Β, C. Να εξαχθούν συμπεράσματα σχετικά με την προέλευση των ειδωλίων.



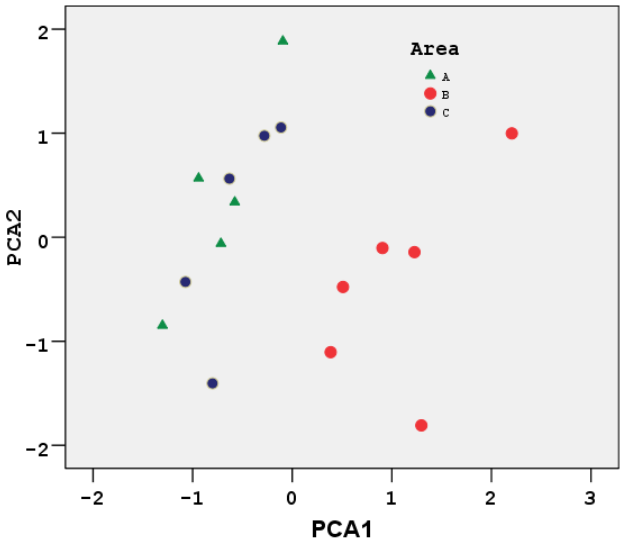
	area	Al	Fe	Mg	Si	Ca	var
1	A	7,1	7,2	1,2	12,0	5,0	
2	A	7,8	6,8	,8	10,0	5,2	
3	A	8,0	7,0	1,1	11,5	5,8	
4	A	7,9	7,4	1,0	10,3	6,2	
5	A	8,2	6,6	1,0	11,9	5,0	
6	B	6,2	6,7	1,2	12,5	6,2	
7	B	6,9	7,1	,8	13,0	5,8	
8	B	5,5	7,4	,5	11,5	6,9	
9	B	7,2	6,7	,2	11,5	6,2	
10	B	6,5	6,2	,4	12,8	6,7	
11	B	6,9	6,8	,7	12,1	7,0	
12	C	7,1	7,0	1,1	10,0	6,2	
13	C	7,5	6,6	1,5	11,1	5,8	
14	C	6,6	6,2	1,2	11,0	5,5	
15	C	8,2	6,7	,9	10,3	6,0	
16	C	8,0	7,1	1,0	10,9	6,1	

Σχήμα 8.1. Αποτελέσματα της χημικής σύστασης σε Al, Fe, Mg, Si, Ca ειδωλίων από τις περιοχές Α, Β, C σε φύλλο εργασίας του SPSS

◆ Ακολουθούμε την πορεία *Analyze* → *Dimension Reduction* → *Factor* και στο παράθυρο που ανοίγει εισάγουμε τις μεταβλητές Al, Fe, Mg, Si, Ca στο πλαίσιο

Variables. Στο *Rotation* επιλέγουμε ως μέθοδο περιστροφής την *Varimax* και στο *Extraction* επιλέγουμε ως μέθοδο την *Principal Components*, το *Correlation Matrix* και το *Eigenvalues over 1* για να πάρουμε μόνο τους άξονες που έχουν τιμές μεγαλύτερες από 1 και που συνεπώς είναι οι πιο σημαντικοί. Τέλος, από το *Scores* επιλέγουμε το *Save as variables*. Με αυτή την επιλογή οι τιμές των PC1, PC2 αποθηκεύονται στο φύλλο εργασίας με τίτλους FAC1_1, FAC2_1. Με κλικ στο *OK* δημιουργούνται αυτόματα στον SPSS Data Editor οι στήλες FAC1_1, FAC2_1 οι οποίες περιλαμβάνουν τις τιμές των αξόνων PC1 και PC2.

Για να δούμε γραφικά τα αποτελέσματα ακολουθούμε τη διαδικασία *Graphs* → *Legacy Dialogs* → *Scatter/Dot* και στο πρώτο παράθυρο που ανοίγει επιλέγουμε *Simple Scatter* και συνεχίζουμε με κλικ στο *Define*. Στο νέο παράθυρο μεταφέρουμε τη μεταβλητή REGR factor score 1 στο πλαίσιο *X Axis*, τη μεταβλητή REGR factor score 2 στο *Y Axis* και τη μεταβλητή Area στο *Set Markers by*. Με αυτόν τον τρόπο η κάθε περιοχή, A, B, C, θα έχει διαφορετικό σύμβολο. Με κλικ στο *OK* παίρνουμε (μετά από κατάλληλη μορφοποίηση) το Σχήμα 8.2.



Σχήμα 8.2. Διάγραμμα αποτελεσμάτων (PC1 vs. PC2)

Παρατηρούμε ότι τα σημεία της περιοχής Β σχηματίζουν μια ξεχωριστή ομάδα (cluster), ενώ τα σημεία των περιοχών Α και C μαζί μια άλλη ομάδα. Πιθανές ερμηνείες είναι ότι η πηγή αργίλου που χρησιμοποιούσαν οι κάτοικοι της περιοχής Β ήταν διαφορετική από αυτή των περιοχών Α και C και επιπλέον οι κάτοικοι της Β δεν είχαν ανταλλαγές με τους κατοίκους των περιοχών Α και C,

τουλάχιστον ως προς τα ειδώλια. Σ’ ότι αφορά τους κατοίκους των περιοχών Α και C ή είχαν κοινή πηγή αργίλου ή ανταλλαγές μεταξύ τους.

Παρατήρηση 1. Ανάλογα με το πρόβλημα είναι δυνατόν να δημιουργηθούν περισσότερες από δύο στήλες στο φύλλο εργασίας, FAC1_1, FAC2_1, FAC3_1, …

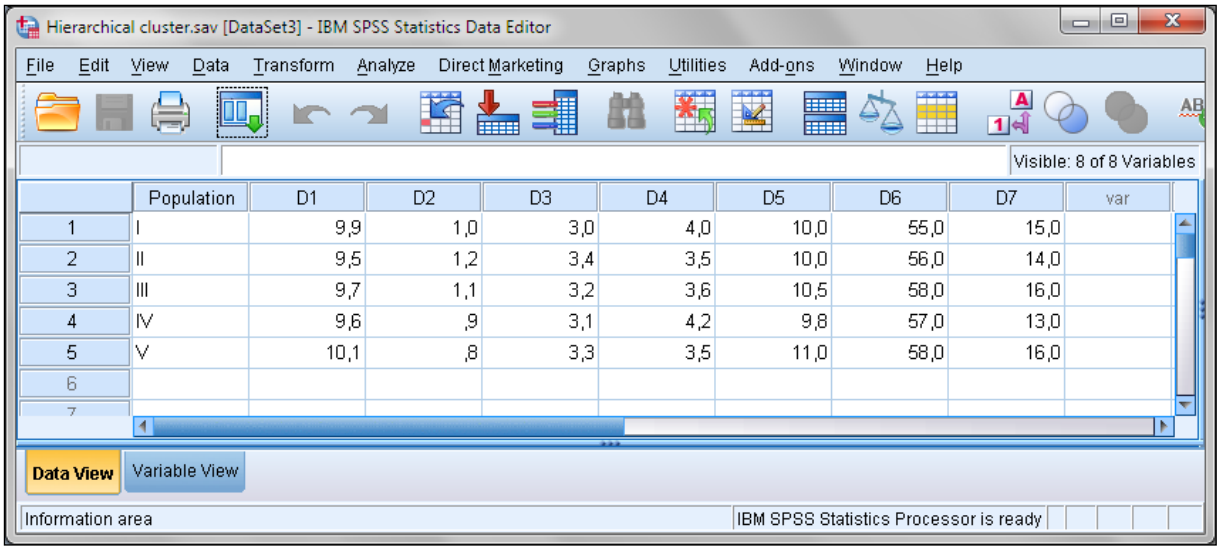
Παρατήρηση 2. Αν στο διάγραμμα αποτελεσμάτων δεν ξεχωρίσουν ομάδες (clusters) δοκιμάζουμε διαφορετικές μεθόδους περιστροφής. Δηλαδή ξανα-εφαρμόζουμε τη μέθοδο και από το *Rotation* επιλέγουμε ή δεν επιλέγουμε καμία μέθοδο περιστροφής (None) ή δοκιμάζουμε τις άλλες μεθόδους, Quartimax, Equamax, Promax.

8.3 ΑΝΑΛΥΣΗ ΣΕ ΟΜΑΔΕΣ - CLUSTER ANALYSIS (CA)

Η ανάλυση σε ομάδες περιλαμβάνει μεθόδους που διαχωρίζουν τα δείγματα σε **ομάδες (clusters)** με παρόμοιες ιδιότητες. Η δημιουργία των ομάδων μπορεί να γίνεται με τρόπο διαδοχικό ενώνοντας στην ομάδα ένα δείγμα κάθε φορά ή με μη διαδοχικό τρόπο ελέγχοντας πολλά δείγματα ταυτόχρονα. Οι μέθοδοι που ανήκουν στην πρώτη κατηγορία ονομάζονται **Ιεραρχικές**, ενώ αυτές της δεύτερης κατηγορίας ονομάζονται **Μη ιεραρχικές**.

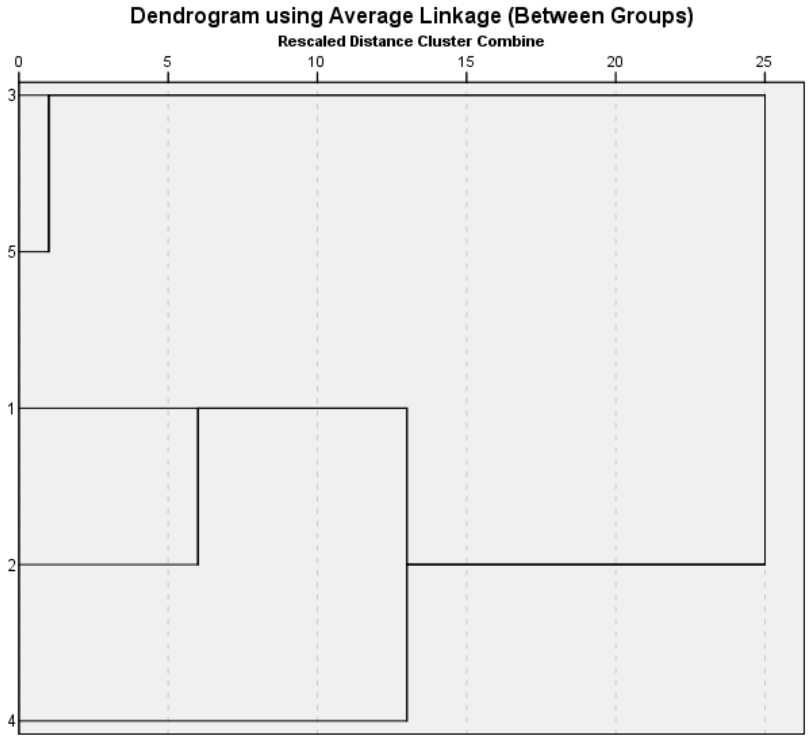
Παράδειγμα

Στο Σχήμα 8.3 δίνονται οι τιμές 7 δεικτών που χαρακτηρίζουν το σχήμα των κρανίων πέντε προϊστορικών πληθυσμών. Με βάση αυτόν τον πίνακα να ελεγχθούν πιθανές συγγένειες μεταξύ των πληθυσμών.



Σχήμα 8.3 Πίνακας δεδομένων

◆ Ακολουθούμε την πορεία *Analyze* → *Classify* → *Hierarchical Cluster* και στο παράθυρο που ανοίγει μεταφέρουμε όλες τις μεταβλητές D1 – D7 στο πλαίσιο *Variable(s)*. Με κλικ στο *Plots* επιλέγουμε το *Dendogram* και ολοκληρώνουμε με κλικ στο *Continue* και στο *OK*. Το δενδρόγραμμα που παίρνουμε δίνεται στο Σχήμα 8.4.

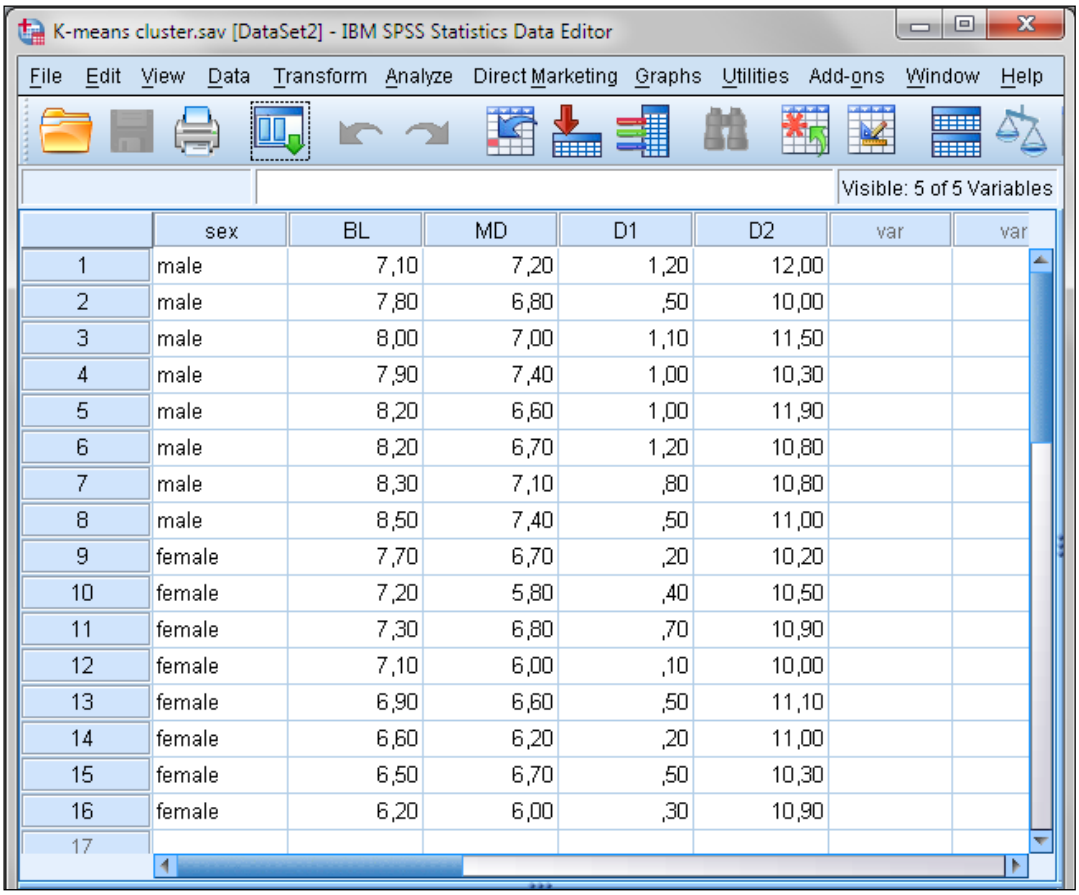


Σχήμα 8.4. Δενδρόγραμμα πληθυσμών

Από το δενδρόγραμμα παρατηρούμε ότι οι πληθυσμοί με βάση τα κρανιακά δεδομένα μπορούν να χωριστούν σε δύο ομάδες: Οι πληθυσμοί III και V έχουν στενή συγγένεια, ενώ οι I, II και IV σχηματίζουν μια δεύτερη ομάδα. Στην ομάδα αυτή οι I με τους II φαίνεται να σχηματίζουν μια υποομάδα.

Παράδειγμα

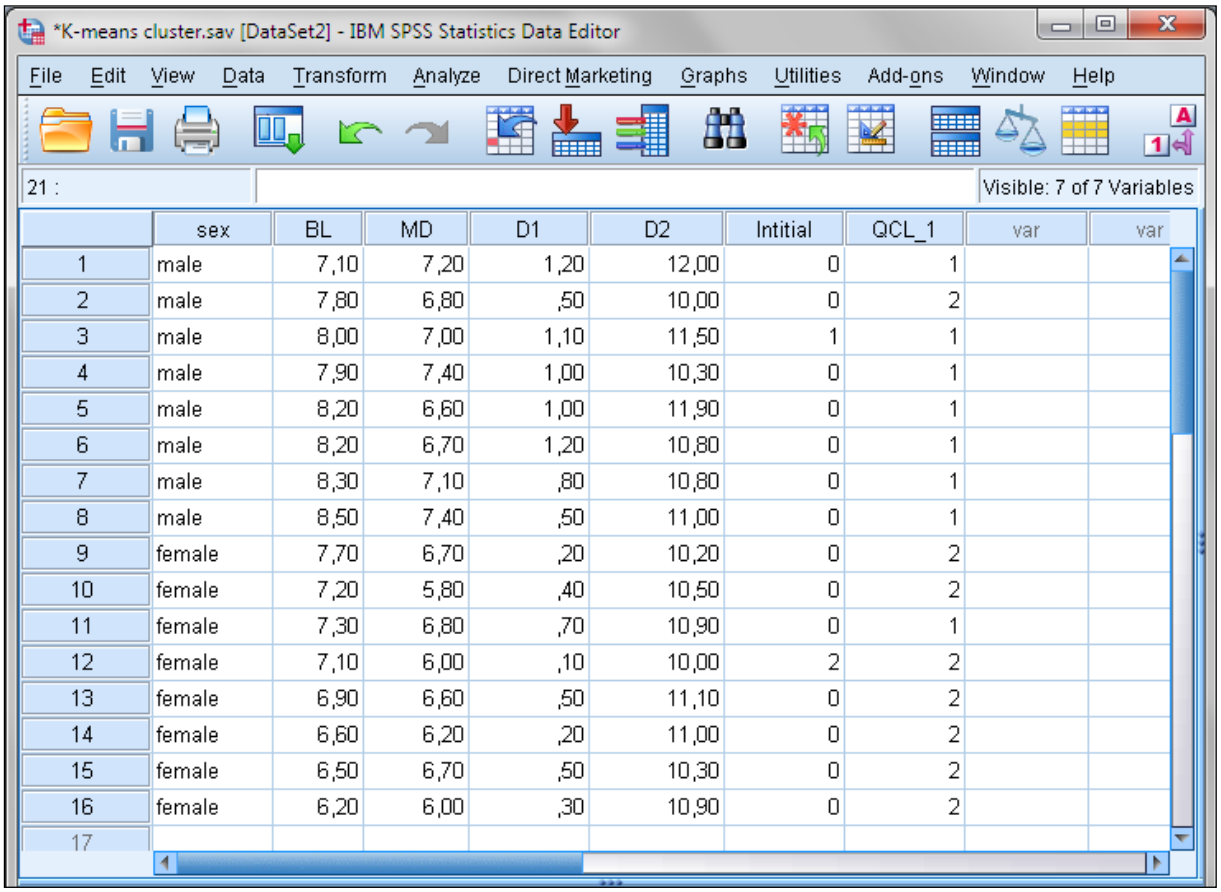
Στο Σχήμα 8.5 δίνονται τέσσερεις χαρακτηριστικοί δείκτες των δοντιών ενηλίκων ανδρών και γυναικών. Θεωρήστε ότι η πρώτη στήλη μας είναι άγνωστη, δηλαδή δεν γνωρίζουμε ποια δείγματα είναι ανδρών και ποια γυναικών. Γνωρίζουμε όμως ότι από τα δείγματα αυτά η περίπτωση 3 είναι χαρακτηριστική των ανδρών και η 12 των γυναικών. Με βάση αυτή την πληροφορία **να εκτιμηθεί ποια δείγματα είναι ανδρικά και ποια γυναικεία**.



	sex	BL	MD	D1	D2	var	var
1	male	7,10	7,20	1,20	12,00		
2	male	7,80	6,80	,50	10,00		
3	male	8,00	7,00	1,10	11,50		
4	male	7,90	7,40	1,00	10,30		
5	male	8,20	6,60	1,00	11,90		
6	male	8,20	6,70	1,20	10,80		
7	male	8,30	7,10	,80	10,80		
8	male	8,50	7,40	,50	11,00		
9	female	7,70	6,70	,20	10,20		
10	female	7,20	5,80	,40	10,50		
11	female	7,30	6,80	,70	10,90		
12	female	7,10	6,00	,10	10,00		
13	female	6,90	6,60	,50	11,10		
14	female	6,60	6,20	,20	11,00		
15	female	6,50	6,70	,50	10,30		
16	female	6,20	6,00	,30	10,90		
17							

Σχήμα 8.5. Τιμές δεικτών (BL, MD, D1, D2) δοντιών ενήλικων ανδρών και γυναικών σε φύλλο εργασίας του SPSS

♦ Στο SPSS το πρόβλημα αυτό λύνεται με τη μέθοδο *K-Means Cluster*. Πρώτα όμως κάνουμε τις εξής ενέργειες: Στην έκτη στήλη γράφουμε τον τίτλο Initial και τη συμπληρώνουμε με 0. Στη γραμμή 3 το μηδέν το κάνουμε 1 και στη 12 το 0 γίνεται 2, εφόσον αυτές οι περιπτώσεις είναι χαρακτηριστικές των ανδρών και γυναικών, αντίστοιχα. Στη συνέχεια ακολουθούμε την πορεία *Analyze* → *Classify* → *K-Means Cluster* και στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές BL, MD, D1 και D2 στο πλαίσιο *Variables* ενώ τη μεταβλητή Initial στο *Label Cases by*. Κάνουμε κλικ στο *Save* και επιλέγουμε *Cluster membership*. Ολοκληρώνουμε με κλικ στο *Continue* και στο *OK*. Τα αποτελέσματα της μεθόδου δίνονται σε μια νέα στήλη που προστίθεται στο αρχικό φύλλο εργασίας με τίτλο QCL_1. Στη στήλη αυτή με 1 δηλώνονται τα ανδρικά δείγματα και με 2 τα γυναικεία, εφόσον αυτούς τους αριθμούς χρησιμοποιήσαμε στη στήλη Initial για να ξεχωρίσουμε τα ανδρικά από τα γυναικεία δείγματα (Σχήμα 8.6).



	sex	BL	MD	D1	D2	Intitial	QCL_1	var	var
1	male	7,10	7,20	1,20	12,00	0	1		
2	male	7,80	6,80	,50	10,00	0	2		
3	male	8,00	7,00	1,10	11,50	1	1		
4	male	7,90	7,40	1,00	10,30	0	1		
5	male	8,20	6,60	1,00	11,90	0	1		
6	male	8,20	6,70	1,20	10,80	0	1		
7	male	8,30	7,10	,80	10,80	0	1		
8	male	8,50	7,40	,50	11,00	0	1		
9	female	7,70	6,70	,20	10,20	0	2		
10	female	7,20	5,80	,40	10,50	0	2		
11	female	7,30	6,80	,70	10,90	0	1		
12	female	7,10	6,00	,10	10,00	2	2		
13	female	6,90	6,60	,50	11,10	0	2		
14	female	6,60	6,20	,20	11,00	0	2		
15	female	6,50	6,70	,50	10,30	0	2		
16	female	6,20	6,00	,30	10,90	0	2		
17									

Σχήμα 8.6. Φύλλο δεδομένων με προσθήκη της μεταβλητής Initial και αποτελέσματα κατάταξης

8.4. ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ - DISCRIMINANT ANALYSIS

Η *Διαχωριστική Ανάλυση (discriminant analysis)* είναι μια στατιστική μέθοδος που μας επιτρέπει να βρούμε σε ποια κατηγορία ανήκουν ένα ή περισσότερα δείγματα με την προϋπόθεση ότι υπάρχουν και είναι γνωστές δύο ή περισσότερες κατηγορίες στις οποίες μπορούν να ανήκουν αυτά. Απαραίτητη προϋπόθεση για την εφαρμογή της μεθόδου είναι τα δεδομένα να ακολουθούν την κανονική κατανομή.

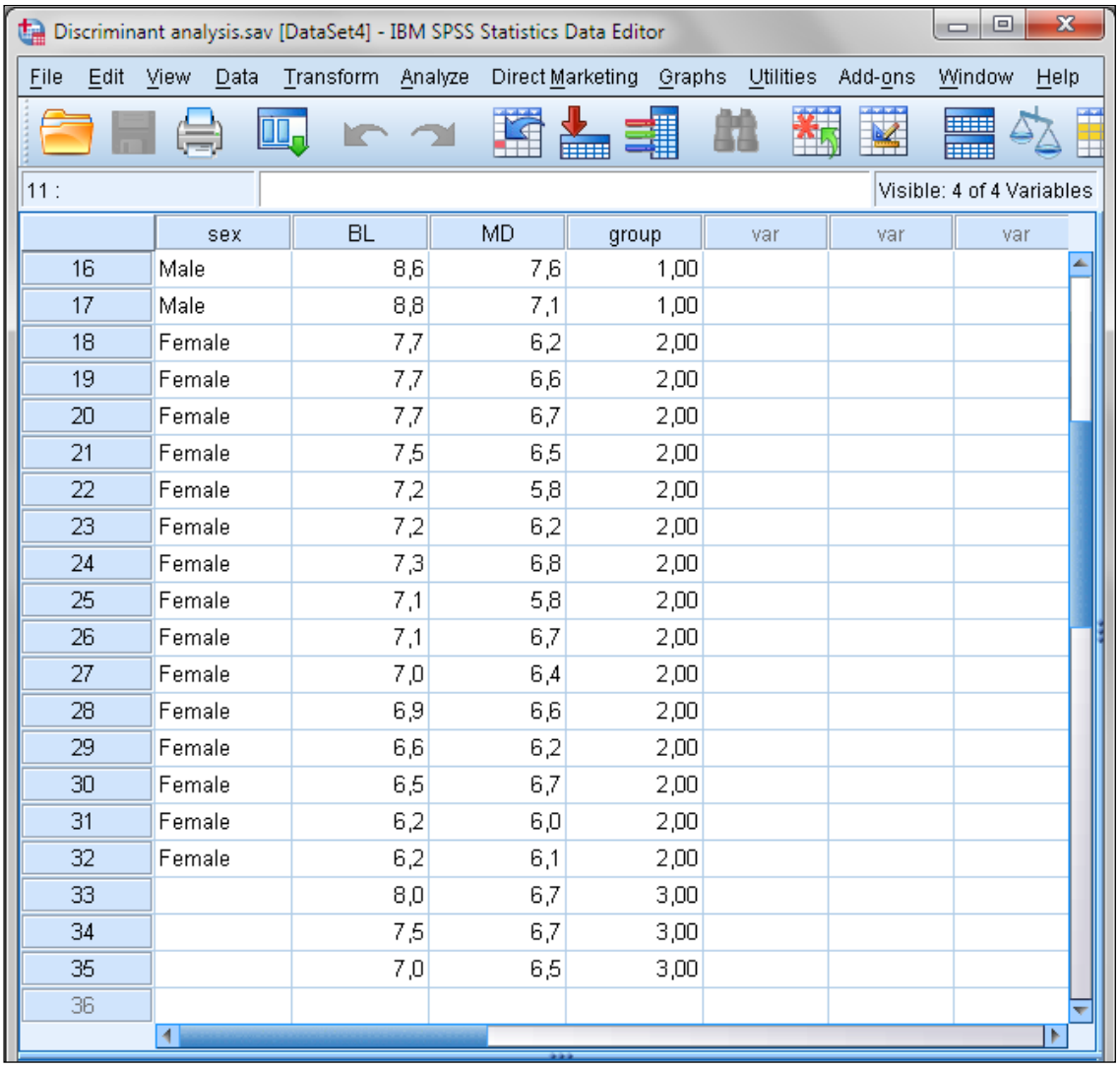
Παράδειγμα

Στον Πίνακα 8.1 δίνονται οι δείκτες BL και MD των δοντιών ενήλικων ανδρών και γυναικών. Να προσδιορίσετε αν τα δείγματα (BL, MD) = (8, 6.7), (7.5, 6.7), (7, 6.5) ανήκουν σε άνδρα ή σε γυναίκα.

Πίνακας 8.1. Τιμές των δεικτών BL και MD των δοντιών ενήλικων ανδρών και γυναικών

Sex	BL	MD	Sex	BL	MD
Male	7.9	6.4	Female	7.7	6.2
Male	7.4	6.6	Female	7.7	6.6
Male	7.2	6.7	Female	7.7	6.7
Male	7.1	7.2	Female	7.5	6.5
Male	7.8	6.8	Female	7.2	5.8
Male	8.1	6.9	Female	7.2	6.2
Male	8.1	7	Female	7.3	6.8
Male	8	7	Female	7.1	5.8
Male	8	7.1	Female	7.1	6.7
Male	7.9	7.4	Female	7	6.4
Male	8.2	6.6	Female	6.9	6.6
Male	8.2	6.7	Female	6.6	6.2
Male	8.3	7.1	Female	6.5	6.7
Male	8.5	7.3	Female	6.2	6
Male	8.5	7.4	Female	6.2	6.1
Male	8.6	7.6			
Male	8.8	7.1			

◆ Για να προχωρήσουμε στην εξέταση των δειγμάτων πρέπει να εφαρμόσουμε *discriminant analysis*. Για το σκοπό αυτό μεταφέρουμε τα δεδομένα στο SPSS σε τρεις στήλες, όπως φαίνεται στο Σχήμα 8.7. **Προσοχή**, τα άγνωστα δείγματα τοποθετούνται στο τέλος των στηλών BL και MD. Ακολούθως δημιουργούμε μια νέα στήλη με όνομα Group, της οποίας η μεταβλητή παίρνει τις τιμές 1 όταν αντιστοιχεί σε male, 2 σε female και 3 στα άγνωστα δείγματα.



	sex	BL	MD	group	var	var	var
16	Male	8,6	7,6	1,00			
17	Male	8,8	7,1	1,00			
18	Female	7,7	6,2	2,00			
19	Female	7,7	6,6	2,00			
20	Female	7,7	6,7	2,00			
21	Female	7,5	6,5	2,00			
22	Female	7,2	5,8	2,00			
23	Female	7,2	6,2	2,00			
24	Female	7,3	6,8	2,00			
25	Female	7,1	5,8	2,00			
26	Female	7,1	6,7	2,00			
27	Female	7,0	6,4	2,00			
28	Female	6,9	6,6	2,00			
29	Female	6,6	6,2	2,00			
30	Female	6,5	6,7	2,00			
31	Female	6,2	6,0	2,00			
32	Female	6,2	6,1	2,00			
33		8,0	6,7	3,00			
34		7,5	6,7	3,00			
35		7,0	6,5	3,00			
36							

Σχήμα 8.7. Τμήμα του πίνακα δεδομένων στο SPSS

Στο SPSS ακολουθούμε τώρα την πορεία *Analyze* → *Classify* → *Discriminant* και στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές BL, MD στο πλαίσιο *Independent* και τη μεταβλητή *Group* στο *Grouping Variable*. Με κλικ στο *Define*

Range εισάγουμε στο *Minimum* την τιμή 1 και στο *Maximum* την τιμή 2 (όχι την 3). Κάνουμε κλικ στο *Continue* και στο *Save* επιλέγουμε *Predicted group membership* και *Probabilities of group membership*. Επίσης, στο *Classify* επιλέγουμε το *Summary table* και ολοκληρώνουμε με κλικ στο *Continue* και στο *OK*.

Από τους πίνακες που παίρνουμε ενδιαφέρον παρουσιάζει ο Classification Results (Πίνακας 8.2). Επίσης, το πρόγραμμα στο φύλλο εργασίας δημιουργεί τρεις νέες στήλες με τίτλους Dis_1, Dis1_1 και Dis2_1 (Σχήμα 8.8).

Πίνακας 8.2. Αποτελέσματα ανάλυσης

Classification Results[a]

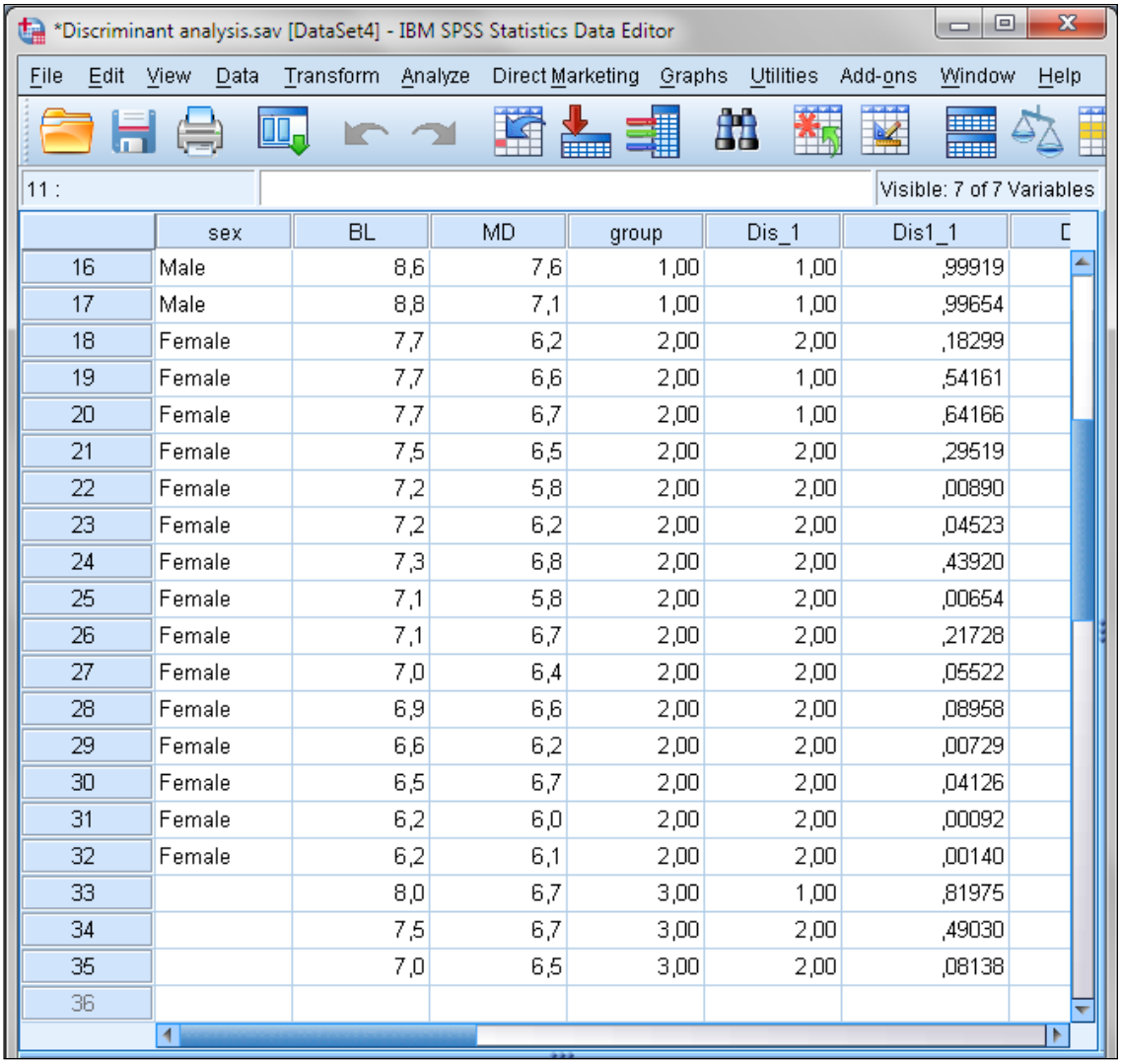
			Predicted Group Membership		
		group	1,00	2,00	Total
Original	Count	1,00	14	3	17
		2,00	2	13	15
		Ungrouped cases	1	2	3
	%	1,00	82,4	17,6	100,0
		2,00	13,3	86,7	100,0
		Ungrouped cases	33,3	66,7	100,0

a. 84.4% of original grouped cases correctly classified.

Στον Πίνακα 8.2 αξιολογείται αν πράγματι τα αρχικά δεδομένα σχηματίζουν δύο διακριτές κατηγορίες. Παρατηρούμε ότι από τους 17 άνδρες το πρόγραμμα ξεχωρίζει τους 14 και από τις 15 γυναίκες τις 13. Μπορούμε επομένως να πούμε ότι υπάρχει ένας ικανοποιητικός διαχωρισμός των δύο κατηγοριών.

Από τις στήλες, η Dis_1 μας δίνει την πρόβλεψη του προγράμματος για κάθε ζεύγος (BL, MD), ενώ στις επόμενες στήλες είναι η εκτιμώμενη πιθανότητα μια περίπτωση (Case) να είναι άνδρας (στήλη Dis1_1) ή γυναίκα (στήλη Dis2_1). Για τα άγνωστα δείγματα έχουμε τα ακόλουθα: Το πρώτο δείγμα είναι με πιθανότητα 0.82 = 82% άνδρας, ενώ για το δεύτερο υπάρχει πλήρης αβεβαιότητα δεδομένου ότι είναι άνδρας με πιθανότητα 49% και γυναίκα με πιθανότητα 51%. Τέλος, το

τρίτο δείγμα ανήκει σε γυναίκα με πιθανότητα 91.9% ≈ 92% (Σχήμα 8.8).



	sex	BL	MD	group	Dis_1	Dis1_1	D
16	Male	8,6	7,6	1,00	1,00	,99919	
17	Male	8,8	7,1	1,00	1,00	,99654	
18	Female	7,7	6,2	2,00	2,00	,18299	
19	Female	7,7	6,6	2,00	1,00	,54161	
20	Female	7,7	6,7	2,00	1,00	,64166	
21	Female	7,5	6,5	2,00	2,00	,29519	
22	Female	7,2	5,8	2,00	2,00	,00890	
23	Female	7,2	6,2	2,00	2,00	,04523	
24	Female	7,3	6,8	2,00	2,00	,43920	
25	Female	7,1	5,8	2,00	2,00	,00654	
26	Female	7,1	6,7	2,00	2,00	,21728	
27	Female	7,0	6,4	2,00	2,00	,05522	
28	Female	6,9	6,6	2,00	2,00	,08958	
29	Female	6,6	6,2	2,00	2,00	,00729	
30	Female	6,5	6,7	2,00	2,00	,04126	
31	Female	6,2	6,0	2,00	2,00	,00092	
32	Female	6,2	6,1	2,00	2,00	,00140	
33		8,0	6,7	3,00	1,00	,81975	
34		7,5	6,7	3,00	2,00	,49030	
35		7,0	6,5	3,00	2,00	,08138	
36							

Σχήμα 8.8. Πρόβλεψη φύλου άγνωστων δειγμάτων

8.5 ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ – MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Με τη μονοπαραγοντική ανάλυση διασποράς εξετάζουμε αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών τριών ή περισσοτέρων δειγμάτων. Η ανάλυση διασποράς πολλών μεταβλητών (MANOVA) επεκτείνει αυτή τη δυνατότητα και εξετάζει την ύπαρξη στατιστικά σημαντικών διαφορών μεταξύ ομάδων δειγμάτων. Ως επέκταση της μονοπαραγοντικής ανάλυσης διασποράς, η εφαρμογή της MANOVA προϋποθέτει την ομοιογένεια της διασποράς και την

κανονικότητα των δειγμάτων που πρέπει πάντα να ελέγχονται, όπως και στην περίπτωση της ANOVA.

Ως παράδειγμα εφαρμογής θα εξετάσουμε αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των ομάδων A, B, C των δειγμάτων του Σχήματος 8.1. Από τη μελέτη του παραδείγματος αυτού με τη μέθοδο PCA έχουμε διαπιστώσει ότι τα δείγματα της ομάδας B διαφοροποιούνται από αυτά των ομάδων A και C που σχηματίζουν μια ενιαία ομάδα. Έτσι έχει ενδιαφέρων να δούμε αν αυτό το συμπέρασμα επιβεβαιώνεται με την MANOVA.

Για να εφαρμόσουμε τη MANOVA στο SPSS διευθετούμε τα δεδομένα όπως στο Σχήμα 8.1 και πηγαίνουμε *Analyze → General Linear Model → Multivariate.* Στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές Al, Fe, Mg, Si, Ca στο πλαίσιο *Dependent Variables* και τη μεταβλητή *area* στο *Fixed Factor(s)*. Από το *Options* επιλέγουμε να γίνει έλεγχος της ομοιογένειας της διασποράς κάνοντας κλικ στο *Homogeneity tests* και από το *Model* επιλέγουμε το *Full Factorial* και τσεκάρουμε το *Include intercept in the model*. Από το *Post Hoc* μπορούμε να επιλέξουμε πολλαπλούς ελέγχους, αλλά αυτοί περιλαμβάνουν και ελέγχους μεταξύ των μεταβλητών και των ομάδων, οδηγώντας σε έναν μάλλον πολύπλοκο πίνακα αποτελεσμάτων.

Όταν ολοκληρώσουμε τις επιλογές και κάνουμε κλικ στο *OK,* παίρνουμε αρκετούς πίνακες αποτελεσμάτων, από τους οποίους οι σημαντικότεροι είναι ο Πίνακας 8.3, πίνακας ελέγχου της ομοιογένειας της διασποράς με το κριτήριο *Levene,* και ο Πίνακας 8.4 που δείχνει αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών.

Πίνακας 8.3. Έλεγχος ομοιογένειας διασποράς

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

	F	df1	df2	Sig.
Al	,714	2	13	,508
Fe	,071	2	13	,932
Mg	1,656	2	13	,229
Si	3,203	2	13	,074
Ca	2,622	2	13	,110

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

a. Design: Intercept + area

Πίνακας 8.4. Αποτελέσματα MANOVA

Multivariate Tests[c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	,999	3228,486[a]	5,000	9,000	,000
	Wilks' Lambda	,001	3228,486[a]	5,000	9,000	,000
	Hotelling's Trace	1793,603	3228,486[a]	5,000	9,000	,000
	Roy's Largest Root	1793,603	3228,486[a]	5,000	9,000	,000
area	Pillai's Trace	1,377	4,425	10,000	20,000	,002
	Wilks' Lambda	,049	6,370[a]	10,000	18,000	,000
	Hotelling's Trace	10,825	8,660	10,000	16,000	,000
	Roy's Largest Root	9,942	19,885[b]	5,000	10,000	,000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + area

Παρατηρούμε στον πρώτο πίνακα ότι Sig. > 0.05 για όλες τις μεταβλητές και συνεπώς υπάρχει ομοιογένεια της διασποράς. Άρα η πρώτη προϋπόθεση για εφαρμογή της μεθόδου ισχύει. Η δεύτερη προϋπόθεση, ο έλεγχος της κανονικότητας των δειγμάτων γίνεται με τα κριτήρια Kolmogorov-Smirnov και Shapiro-Wilk από *Analyze* → *Descriptive Statistics* → *Explore* και δείχνει ότι πληρείται και η προϋπόθεση αυτή. Επομένως μπορούμε να εφαρμόσουμε τη μέθοδο και συνεπώς τα αποτελέσματα του Πίνακα 8.4 είναι έγκυρα. Στον πίνακα αυτό πηγαίνουμε στο πάνελ area όπου παρατηρούμε ότι όλοι οι έλεγχοι που χρησιμοποιεί το SPSS δείχνουν Sig. < 0.05, δηλαδή υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των ομάδων.

Για να δούμε μεταξύ ποιών ομάδων υπάρχουν αυτές οι διαφορές, εφαρμόζουμε τη μέθοδο στα δείγματα των ομάδων A-B, B-C και A-C, δηλαδή αφαιρούμε μία ομάδα δειγμάτων και εφαρμόζουμε MANOVA στα υπόλοιπα δείγματα. Οι βασικοί πίνακες αποτελεσμάτων που παίρνουμε δίνονται στους Πίνακες 8.5-8.7. Παρατηρούμε ότι, σε πλήρη συμφωνία με τα αποτελέσματα της PCA και της CA, στατιστικά σημαντικές διαφορές υπάρχουν μόνο μεταξύ της ομάδας B και των υπολοίπων ομάδων.

Πίνακας 8.5. Πίνακας αποτελεσμάτων MANOVA για τις ομάδες Α-Β

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	1.000	2466.486[a]	5.000	5.000	.000
	Wilks' Lambda	.000	2466.486[a]	5.000	5.000	.000
	Hotelling's Trace	2466.486	2466.486[a]	5.000	5.000	.000
	Roy's Largest Root	2466.486	2466.486[a]	5.000	5.000	.000
area	Pillai's Trace	.891	8.185[a]	5.000	5.000	.019
	Wilks' Lambda	.109	8.185[a]	5.000	5.000	.019
	Hotelling's Trace	8.185	8.185[a]	5.000	5.000	.019
	Roy's Largest Root	8.185	8.185[a]	5.000	5.000	.019

a. Exact statistic

b. Design: Intercept+area

Πίνακας 8.6. Πίνακας αποτελεσμάτων MANOVA για τις ομάδες Β-C

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	1.000	2543.306[a]	5.000	5.000	.000
	Wilks' Lambda	.000	2543.306[a]	5.000	5.000	.000
	Hotelling's Trace	2543.306	2543.306[a]	5.000	5.000	.000
	Roy's Largest Root	2543.306	2543.306[a]	5.000	5.000	.000
Area	Pillai's Trace	.931	13.465[a]	5.000	5.000	.006
	Wilks' Lambda	.069	13.465[a]	5.000	5.000	.006
	Hotelling's Trace	13.465	13.465[a]	5.000	5.000	.006
	Roy's Largest Root	13.465	13.465[a]	5.000	5.000	.006

a. Exact statistic

b. Design: Intercept+Area

Πίνακας 8.7. Πίνακας αποτελεσμάτων MANOVA για τις ομάδες Α-C

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.999	1064.206[a]	5.000	4.000	.000
	Wilks' Lambda	.001	1064.206[a]	5.000	4.000	.000
	Hotelling's Trace	1330.257	1064.206[a]	5.000	4.000	.000
	Roy's Largest Root	1330.257	1064.206[a]	5.000	4.000	.000
Area	Pillai's Trace	.730	2.166[a]	5.000	4.000	.237
	Wilks' Lambda	.270	2.166[a]	5.000	4.000	.237
	Hotelling's Trace	2.708	2.166[a]	5.000	4.000	.237
	Roy's Largest Root	2.708	2.166[a]	5.000	4.000	.237

a. Exact statistic

b. Design: Intercept+Area

ΠΑΡΑΡΤΗΜΑ

Απλοποιημένο δείγμα από οστεολογική συλλογή με άτομα γνωστού φύλου και ηλικίας

No	sex	bday	hand arthritis	lumbar vertebrae osteophytosis	height	body mass
1	m	5/5/1958	6	3	182	80
2	f	16/2/1951	4	3	168	59
3	m	22/3/1971	2	1	178	85
4	f	15/4/1975	1	1	163	65
5	f	9/8/1965	2	2	160	-
1	m	23/2/1952	4	2	168	75
7	m	6/8/1956	2	1	172	80
8	f	6/5/1976	1	1	154	50
9	f	23/11/1970	2	1	156	55
10	f	3/12/1970	1	1	163	69
11	f	7/2/1959	2	3	170	80
12	m	11/11/1966	1	1	181	92
13	f	15/1/1969	3	2	158	49
14	m	26/12/1949	2	1	176	83
15	m	9/8/1972	2	1	180	73
16	m	7/1/1964	1	1	182	87
17	f	8/2/1962	2	1	163	58
18	m	10/3/1956	2	3	167	60
19	m	10/8/1962	1	1	168	59
20	f	23/1/1970	1	2	148	53
21	f	19/12/1963	3	1	151	62
22	m	14/9/1970	1	1	190	103
23	f	11/3/1965	2	1	162	66
24	m	7/3/1978	1	1	186	-
25	f	11/7/1972	2	2	172	81
26	m	18/11/1976	2	1	166	62
27	f	9/3/1954	3	3	167	61
28	m	11/4/1973	4	2	160	53

29	m	28/7/1949	5	3	177	72
30	m	7/9/1971	3	2	169	64
31	f	4/2/1964	2	1	149	53
32	f	8/11/1954	6	3	166	73
33	m	8/11/1961	2	2	190	93
34	m	2/8/1949	5	3	187	84
35	m	20/8/1961	2	3	184	74
36	f	17/8/1973	1	1	155	44
37	m	9/10/1974	2	1	182	77
38	m	17/4/1968	1	1	186	83
39	m	12/6/1965	1	2	159	52
40	m	21/8/1973	2	3	169	60
41	f	18/3/1971	4	1	185	52
42	m	20/9/1963	5	1	170	65
43	m	18/11/1974	2	1	160	61
44	f	15/6/1963	1	1	152	58
45	m	8/8/1978	2	2	169	72
46	f	8/11/1970	2	2	148	55
47	m	8/4/1968	4	1	185	104
48	f	17/6/1967	5	2	182	54
49	m	16/4/1958	1	2	180	98
50	m	14/3/1960	1	3	178	86

sex = φύλο
bday = birthday, ημερομηνία γέννησης
hand arthritis:
1=slight osteophytic formation
2=moderate osteophytic formation
3=extensive osteophytic formation
4=pitting
5=eburnation
6=ankylosis
lumbar vertebrae osteophytosis:
1=lipping
2=pitting
3=eburnation
height = εκτιμώμενο ύψος σε cm
body mass = εκτιμώμενο βάρος σε kg