Ελληνική Εταιρεία
Γενικής Ιατρικής

Περιφερειακό Τμήμα ΕΛΕΓΕΙΑ
Ανατολικής Μακεδονίας και Θράκης

Εργαστήριο Υγιεινής και Προστασίας Περιβάλλοντος
Τμήματος Ιατρικής Δημοκρίτειου Πανεπιστημίου Θράκης

Βασικές Αρχές Βιοστατιστικής

Εφαρμογές με χρήση του SPSS

Στυλιανός Κ. Τσίπος και ***Θ.Κ. Κωνσταντινίδης***

Φιλίππειο Φροντιστήριο Ερευνας

Επίπεδο ΙΙ

4-7 Μαρτίου 2010, Athena Palace, Ελιά Χαλκιδικής

Αλεξανδρούπολη, 2010

Πρόλογος

Από τη δημιουργία του Νεοελληνικού Κράτους τον 19ο αιώνα, οι γιατροί είχαν σημαίνουσα θέση στη διοίκηση αυτού του τόπου. Θυμίζω ονόματα, όπως του *Μαυροκορδάτου* και του *Κωλέττη.* Οι γιατροί αυτοί, σπουδασμένοι στην Ευρώπη, επέβαλαν τις αρχές της *Δυτικής Ιατρικής* στην Ελλάδα μ' έναν τρόπο, θάλεγα ελάχιστα δημοκρατικό, που είχε διαχρονικά αμφίσημα αποτελέσματα. Βλέπουμε λοιπόν αυστηρές νομοθεσίες για «παράνομη άσκηση της Ιατρικής», που απήλλαξε μεν τον πληθυσμό από δεισιδαιμονίες και θρησκοληψίες – ή ίσως καλύτερα μαγείες – σε διάφορες «θεραπείες», αλλά εξαφάνισε ταυτόχρονα πολύτιμες εμπειρικές γνώσεις για θεραπευτικές ιδιότητες της πολυποίκιλης ελληνικής χλωρίδας. Με αντίστοιχο τρόπο επιβλήθηκε η *ορθολογική δυτικοευρωπαϊκή ιατρική* ως αναντίρρητη αυθεντία με απόλυτες αρχές, αναφορικά με τις κλινικές εκφάνσεις της, αφήνοντας τους άλλους τομείς της υγείας στην ημιαμφισβητούμενη περιοχή, που συνορεύει από την άλλη πλευρά με το «ύποπτο» και το «παράνομο».

Θεωρώ ότι οι πολλαπλές αμφισβητήσεις των εναλλακτικών μορφών της Ιατρικής στον τόπο μας αλλά και η γενικότερη στάση απαξίωσης των κλινικών γιατρών του τόπου μας στην αγωγή υγείας, την πρόληψη αλλά και την αποκατάσταση, έχουν τις ρίζες τους σ' αυτό το παρελθόν και σ' αυτήν την λογική. Ως γενικός γιατρός, συνάντησα αυτή τη νοοτροπία πολλές φορές σε εξαίρετους κατά τα άλλα συναδέλφους, που δε μπορούσαν π.χ. να συμπεριλάβουν τα πολλά και βαριά τροχαία ατυχήματα στα προβλήματα υγείας της περιοχής ευθύνης του Κέντρου Υγείας που υπηρετούσαν.

Το 1986 η Ελληνική Πολιτεία καθιέρωσε τις «νέες» ειδικότητες: της *Κοινωνικής Ιατρικής,* της *Γενικής Ιατρικής* και της *Ιατρικής της Εργασίας.* Για τους ίδιους λόγους που ανέφερα παραπάνω, οι ειδικότητες αυτές, καθόλου δεν καλωσορίσθηκαν από τις λοιπές παραδοσιακές κλινικές ειδικότητες του *Ελληνικού Ιατρικού Πανθέου.* Ακόμη και σήμερα, έπειτα από ένα τέταρτο του αιώνα, πολλοί σημερινοί γιατροί, σαν εκείνους τους παλιούς «φραγκομαθημένους», μας κοιτούν με υποψία και δυσκολεύονται πολύ να μας αναγνωρίσουν ως ισότιμους συναδέλφους τους. Αγνοώντας τον Ιπποκράτη, αυτοανακηρύχθηκαν «επιστήμονες», αφέθηκαν – άν δεν τον προκάλεσαν οι ίδιοι – στον ιατρικό πληθωρισμό και, μπροστά στα επαγγελματικά τους αδιέξοδα, επιτίθενται συχνά στις ειδικότητές μας, κραδαίνοντας την «επιστημοσύνη» τους σαν όπλο, το ίδιο όπλο που χρησιμοποίησαν εναντίον των *Βικογιατρών* έναν αιώνα πριν.

Η *Γενική Ιατρική* αγκάλιασε την *Ιατρική της Εργασίας* από τα πρώτα της βήματα και θα συνεχίσει να συνεργάζεται αρμονικά μαζί της εις το διηνεκές. Εχουν και οι δύο αποστολή και στόχο την υγεία των λαών και όχι απλά την θεραπεία των ασθενειών. Απέναντι λοιπόν στις επιστημονίζουσες θεωρίες και πρακτικές, εμείς, με πλήρη συνείδηση και αποδοχή του γεγονότος ότι είμαστε τεχνίτες της υγείας, έχουμε υποχρέωση να τονίζουμε καθημερινά

προς κάθε κατεύθυνση ότι έχουμε τη γνώση και την εμπειρία να σταθούμε άξιοι της αποστολής και του λειτουργήματός μας.

Το βιβλίο αυτό, που ο φίλος και συνάδελφος Πρόεδρος της *Ελληνικής Εταιρείας Ιατρικής της Εργασίας και Περιβάλλοντος,* καθηγητής κ. *Κωνσταντινίδης* μου έκανε την τιμή να ζητήσει να προλογίσω, παρέχει μια διπλή απόδειξη, στην κατεύθυνση τόσο της γνώσης που προανέφερα, όσο και της αρμονικής συνεργασίας, αφού ουσιαστικά θα χρησιμοποιηθεί στην διδασκαλία των γενικών γιατρών στην Βιοστατιστική.

Εκ μέρους της *Γενικής Ιατρικής* ευχαριστώ τον κ. *Κωνσταντινίδη* και τον κ. *Τσίπο* για την προσπάθειά τους αυτήν, αλλά και προσωπικά για την τιμή που μου έκαναν.

Μποδοσάκης - Πρόδρομος Ρ. Μερκούρης
Πρόεδρος ΕΛΕΓΕΙΑ

Περιεχόμενα

1ο ΚΕΦΑΛΑΙΟ

Εισαγωγή στη Στατιστική

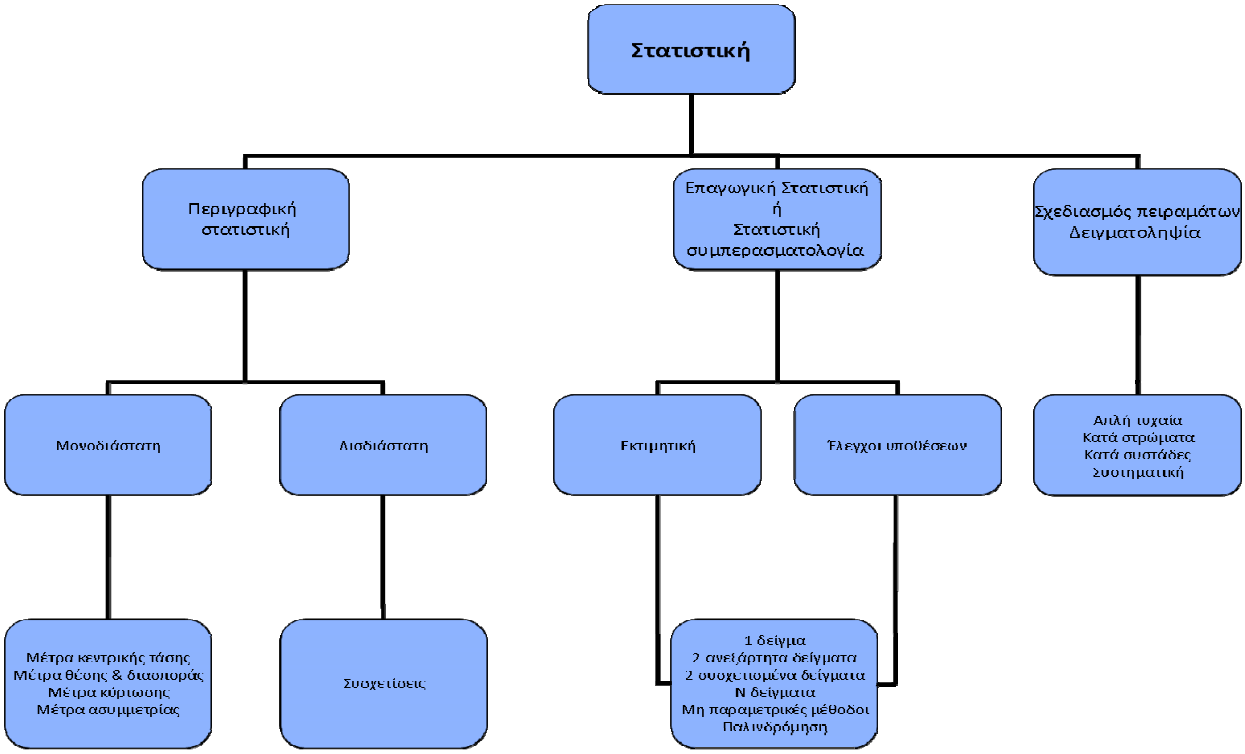
Εισαγωγή

Η στατιστική, όπως προκύπτει και από την ετυμολογία της λέξης (status=κράτος) συνδεόταν αρχικά με την λειτουργία του κράτους και για αυτό το λόγο περιοριζόταν σε πληροφορίες αναγκαίες για φορολογικούς ή στρατιωτικούς σκοπούς. Αργότερα όμως, επικάλυψε κάθε είδους δεδομένα που έχουν αριθμητική μορφή.

Ο συνηθέστερος και γνωστότερος ορισμός της "Στατιστικής" δόθηκε από τον πατέρα της σύγχρονης Στατιστικής Sir R. A. Fisher (1890-1962):

Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- το σχεδιασμό της διαδικασίας συλλογής δεδομένων (σχεδιασμός πειραμάτων-δειγματοληψία)
- τη συνοπτική και αποτελεσματική παρουσίασή τους (περιγραφική στατιστική)
- την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων, για όλο το σύνολο ή την ικανότητα μιας διαδικασίας, κάτω από συνθήκες αβεβαιότητας (επαγωγική στατιστική ή στατιστική συμπερασματολογία).



1.1 Βασικά στοιχεία της στατιστικής ανάλυσης

Τα βασικότερα στοιχεία της στατιστικής ανάλυσης είναι:

Πληθυσμός (Population): είναι ένα σύνολο στοιχείων που μας ενδιαφέρει να μελετήσουμε ως προς ένα ή περισσότερα χαρακτηριστικά του.

Μεταβλητή (Variable): είναι μία καλά ορισμένη μετρήσιμη έκφραση ενός χαρακτηριστικού του πληθυσμού ή της ικανότητας μιας διαδικασίας που μας ενδιαφέρει να εξετάσουμε και που παίρνει περισσότερες από μία διαφορετικές τιμές.

Τιμές της μεταβλητής: είναι οι δυνατές τιμές που μπορεί να πάρει μία μεταβλητή.

Δείγμα (Sample): είναι ένα υποσύνολο ενός πληθυσμού ή παρατηρηθέντων αποτελεσμάτων μιας διαδικασίας για μια χρονική περίοδο.

Παράμετρος (Parameter): είναι μία αριθμητική ποσότητα που συνοψίζει κάποιο χαρακτηριστικό του πληθυσμού ή της ικανότητας μιας διαδικασίας.

Στατιστική συνάρτηση (Statistic): είναι μία αριθμητική ποσότητα που συνοψίζει κάποιο χαρακτηριστικό του δείγματος και που μπορεί να χρησιμοποιηθεί για την εκτίμηση μίας άγνωστης αντίστοιχης παραμέτρου του πληθυσμού.

Εμπιστοσύνη (Confidence): είναι η πιθανοφάνεια ότι η στατιστική συμπερασματολογία στην οποία καταλήξαμε είναι σωστή ή ότι έχει κάποιο λάθος, το οποίο όμως δεν υπερβαίνει κάποια προκαθορισμένη ποσότητα.

Διάστημα εμπιστοσύνης (Confidence Interval): η εκτίμηση ενός διαστήματος πιθανών τιμών, με τη χρήση του δείγματος, για μία άγνωστη παράμετρο του πληθυσμού.

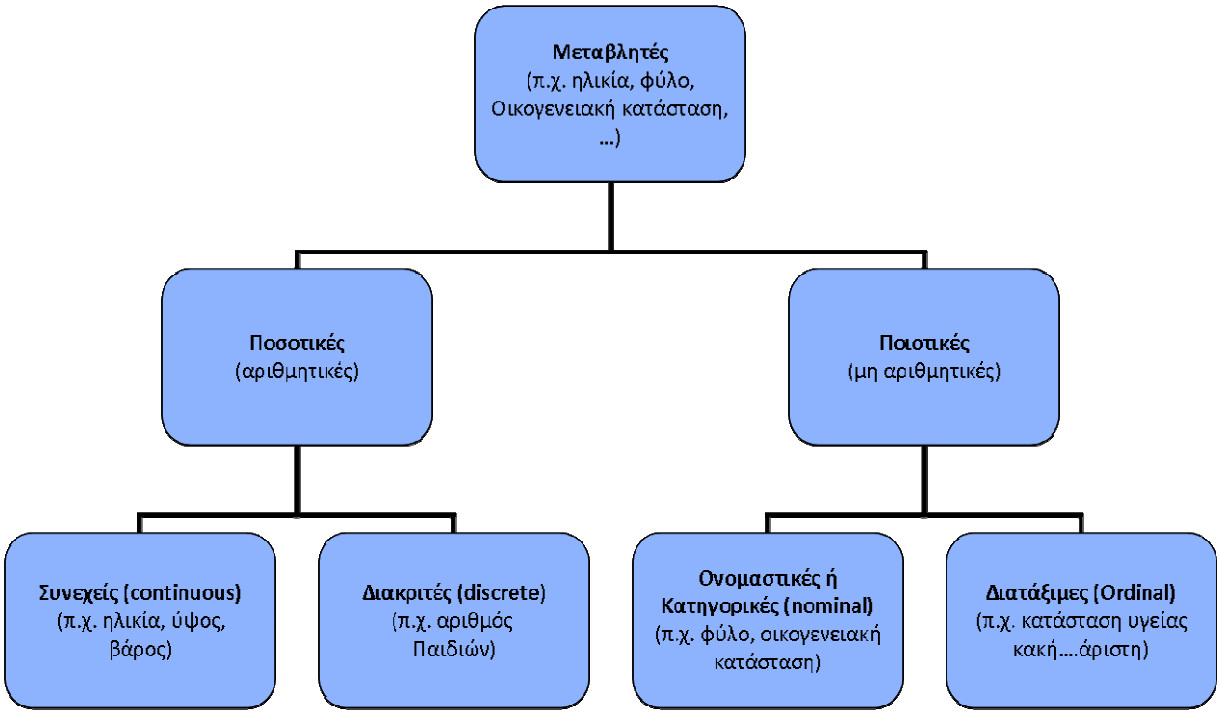
1.2 Η έννοια της μεταβλητής

Η λέξη «μεταβλητή» χρησιμοποιείται ως ουσιαστικοποιημένο επίθετο για να περιγράψει φαινόμενα που από την φύση τους παίρνουν διαφορετικές τιμές και δεν είναι σταθερά. Όταν λέμε «μεταβλητή» εννοούμε μεταβλητή ποσότητα, ποσότητα που μεταβάλλεται- όπως το ύψος των παιδιών το οποίο μεταβάλλεται από παιδί σε παιδί.

Οι μεταβλητές μπορεί να είναι ποσοτικές (αριθμητικές) ή ποιοτικές (μη αριθμητικές).

Όταν μια ποσοτική μεταβλητή δύναται να λάβει οποιαδήποτε τιμή εντός ενός διαστήματος πραγματικών αριθμών (α, β), ονομάζεται συνεχής (continuous). Διαφορετικά ονομάζεται ασυνεχής ή διακριτή (discrete). Για παράδειγμα το ύψος μιας ομάδας παιδιών είναι συνεχής μεταβλητή ενώ το αποτέλεσμα της ρίψης ενός ζαριού (1,2,…,6) είναι διακριτή.

Όταν η προσδιορισμένη σχέση μεταξύ των διαφορετικών τιμών της ποιοτικής μεταβλητής είναι απλά η ύπαρξη διαφοράς τότε ονομάζεται ονομαστική (nominal). Για παράδειγμα το χρώμα των ματιών, το φύλο κλπ. Όταν η προσδιορισμένη σχέση μεταξύ των διαφορετικών τιμών εμπεριέχει και την έννοια της "καλύτερης", "προτιμότερης", "μεγαλύτερης" δηλαδή της διαβάθμισης-ιεράρχησης μεταξύ τους ονομάζεται διάταξης (ordinal). Για παράδειγμα ένας ασθενής μπορεί να είναι φυσιολογικός < υπέρβαρος < παχύσαρκος κλπ.



Υπάρχει η δυνατότητα να γίνει ποιοτική μια ποσοτική μεταβλητή μέσω ομαδοποιήσεων, είναι όμως δυσκολότερο να συμβεί το αντίστροφο. Σε μια τέτοια περίπτωση, κάποιες τιμές κοντινές μεταξύ τους εντάσσονται σε μια κατηγορία η οποία προσδιορίζεται ποιοτικά. Το ίδιο γίνεται ύστερα και για όλες τις διαθέσιμες τιμές μέχρι αυτές να ενταχθούν σε κάποια κατηγορία. Πχ άτομα μικρότερα των 18 ετών εντάσσονται στην κατηγορία «Α», μεταξύ 18-30 στην κατηγορία «Β» κοκ.

Στη στατιστική συμπερασματολογία μία μεταβλητή που θα αναλυθεί μπορεί να είναι ανεξάρτητη (independent) ή εξαρτημένη (dependent). Εξαρτημένη είναι η μεταβλητή που εξετάζει-μελετάει ο ερευνητής ενώ ανεξάρτητη η μεταβλητή που ελέγχει και η οποία επηρεάζει κάποια άλλη (αίτιο). Για παράδειγμα, στο ερευνητικό ερώτημα εάν η έντονη χρήση του Η/Υ έχει δυσμενείς επιπτώσεις στην σχολική επίδοση των μαθητών, ανεξάρτητη μεταβλητή είναι η έντονη χρήση Η/Υ ενώ εξαρτημένη η σχολική επίδοση των μαθητών.

1.3 Συλλογή δεδομένων

Η συλλογή δεδομένων μπορεί να γίνει με:

- Απογραφή, δηλαδή άντληση των πληροφοριών που χρειαζόμαστε για κάποιο πληθυσμό από όλα τα άτομα που τον αποτελούν
- Δειγματοληψία, δηλαδή άντληση των πληροφοριών που χρειαζόμαστε για κάποιο πληθυσμό από ένα υποσύνολο του
- Άμεση παρατήρηση
- Πειράματα

Οι αρχές και οι μέθοδοι για τη συλλογή και ανάλυση δεδομένων από πεπερασμένους πληθυσμούς είναι το αντικείμενο της Δειγματοληψίας (Sampling). Υπάρχει αρκετή βιβλιογραφία σχετικά με τον τρόπο που θα επιλέξουμε έναν δείγμα από έναν πληθυσμό και φυσικά αρκετοί κίνδυνοι που συνδέονται με την επιλογή λάθος δείγματος. Οι πιο σημαντικές μέθοδοι δειγματοληψίας είναι:

- η απλή τυχαία δειγματοληψία
- η κατά στρώματα (τυχαία) δειγματοληψία
- η κατά συστάδες (τυχαία) δειγματοληψία
- η συστηματική (τυχαία) δειγματοληψία.

Η πιο συνήθης περίπτωση της δειγματοληψίας αφορά τη «στρωματοποιημένη». Αυτό σημαίνει ότι πρέπει να προσδιορίσουμε ορισμένα χαρακτηριστικά (φτιάχνοντας έτσι αμοιβαία αποκλειόμενες ομάδες του πληθυσμού δηλαδή στρώματα) που θεωρούμε ότι επηρεάζουν το φαινόμενο που εξετάζουμε και να επιλέξουμε ένα δείγμα από κάθε στρώμα, με απλή τυχαία δειγματοληψία, με αναλογίες ίδιες με αυτές του πληθυσμού. Πχ αν θεωρούμε ότι το κάπνισμα προκαλεί καρκίνο του πνεύμονα και γνωρίζουμε από την Ελληνική Αντικαπνιστική Εταιρεία ότι το 43% των Ελλήνων καπνίζει, τότε και το δικό μας δείγμα θα πρέπει να έχει 43% καπνιστές. Γνωρίζοντας επίσης την αναλογία αντρών γυναικών δημιουργούμε αντίστοιχη αναλογία και στο δείγμα. Γνωρίζοντας την ηλικιακή κατανομή των γυναικών, δημιουργούμε αντίστοιχη κατανομή στο δείγμα επίσης. Η διαδικασία συνεχίζεται έως ότου καλυφθούν όλα τα χαρακτηριστικά που εξαρχής υποθέσαμε ότι δύναται να είναι παράγοντες επηρεασμού. Όλα αυτά τα χαρακτηριστικά ονομάζονται «παράμετροι». Μέσω της έρευνας και της βιβλιογραφίας είναι δυνατό να

βρεθούν νέοι παράμετροι και να προστεθούν για μια ενδεχόμενη εξέταση στο μέλλον. Βάσει των ανωτέρω, θα μπορούσε κάποιος να φτιάξει ένα δείγμα που να είναι «μικρογραφία» του γενικότερου πληθυσμού που θέλει να μελετήσει και πάνω σε αυτήν να εφαρμόσει Βιοστατιστική. Αν η «μικρογραφία» αυτή δεν είναι αντιπροσωπευτική, δεν έχει δηλαδή ποσοστά αντίστοιχα με αυτά του πληθυσμού, τότε τα αποτελέσματα μπορεί να είναι παραπλανητικά. Λανθασμένο είναι επίσης το επιχείρημα ότι η αύξηση του δείγματος βελτιώνει την αξιοπιστία. Η αύξηση του δείγματος δεν βοηθά όταν βασίζεται σε παρατηρήσεις που δεν έχουν αντίστοιχη βαρύτητα στο γενικό πληθυσμό. Αντίθετα επηρεάζει τα αποτελέσματα και προκαλεί αυτό που ονομάζουμε «μεροληψία δειγματοληψίας». Είναι όμως χρήσιμη όταν αυξάνουμε το δείγμα με αντιπροσωπευτικό τρόπο γιατί μειώνει το βαθμό αβεβαιότητας και το εύρος εντός του οποίου κινούνται οι αληθινές τιμές των μεγεθών (αύξηση ισχύος ελέγχου). Όταν η διαδικασία αυτή τελειώσει, προσδιορίζουμε με τυχαίο τρόπο (συνήθως μέσω μια μηχανικής διαδικασίας με ηλεκτρονικό υπολογιστή χωρίς να παρεμβαίνει ανθρώπινη κρίση) τα υποκείμενα της μελέτης.

Γενικά, μπορούμε να πούμε ότι η οργάνωση της συλλογής και επεξεργασίας των σχετικών δεδομένων και πληροφοριών γίνεται κατά τρόπο που για δεδομένη ακρίβεια να επιτυγχάνεται το χαμηλότερο δυνατό κόστος ή, αντιστρόφως, να εξασφαλίζεται η μέγιστη δυνατή ακρίβεια την οποίαν επιτρέπουν τα μέσα που διαθέτουμε.

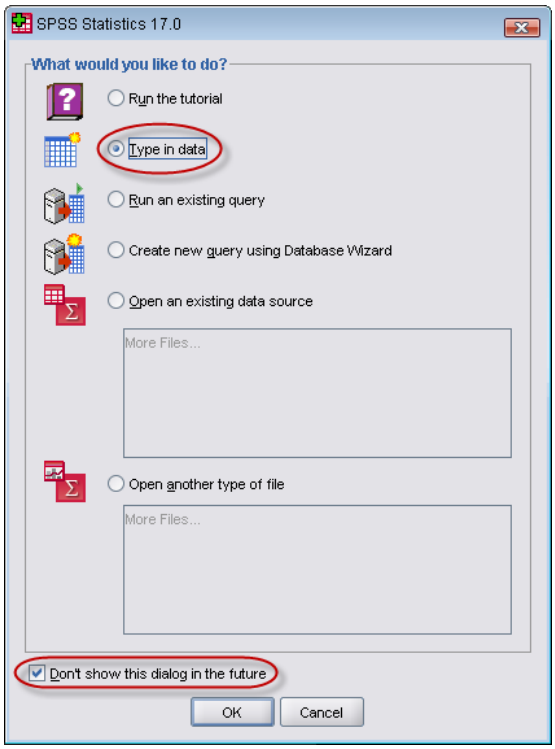
2ο ΚΕΦΑΛΑΙΟ
Εισαγωγή στο SPSS

Εισαγωγή

Σε αυτό το κεφάλαιο που ακολουθεί προσφέρεται μία σύντομη περιήγηση στο στατιστικό πακέτο SPSS 17.0 για το λειτουργικό σύστημα Windows. Το SPSS είναι μία εξελιγμένη εφαρμογή η οποία μπορεί να εκτελέσει σχεδόν οποιαδήποτε στατιστική ανάλυση και επεξεργασία των δεδομένων του σε ένα χρηστικό παραθυρικό περιβάλλον (ωστόσο, είναι δυνατή η στατιστική επεξεργασία των δεδομένων και μέσα από ένα προγραμματιστικό περιβάλλον). Εάν χρησιμοποιήσετε διαφορετική έκδοση του SPSS οι οθόνες που παρουσιάζονται εδώ, όπως και οι εντολές, μπορεί να διαφέρουν. Για πιο διεξοδική περιγραφή του SPSS καθώς και των στατιστικών μεθόδων που χρησιμοποιούνται, θα πρέπει να ανατρέξετε σε πιο αναλυτικά εγχειρίδια (συμπεριλαμβανομένου και του Συστήματος Βοηθείας [Help] του SPSS) καθώς και σε βιβλία Στατιστικής που κυκλοφορούν στο εμπόριο και παρέχουν αναλυτικό εννοιολογικό και θεωρητικό υπόβαθρο των στατιστικών μεθόδων.

2.1 Εκκίνηση - Περιήγηση στο SPSS

Για να ξεκινήσει η εφαρμογή επιλέγεται από το μενού [Έναρξη → Όλα τα προγράμματα → SPSS] το εικονίδιο του [SPSS Statistics 17.0]. Η εφαρμογή θα ξεκινήσει σχετικά σύντομα (πρόκειται για μια εφαρμογή αρκετά απαιτητική σε πόρους του συστήματος) και θα εμφανιστεί η παρακάτω οθόνη (βλ. Εικόνα 2.1). Εδώ δίνεται η δυνατότητα να πληκτρολογηθούν δεδομένα (σ' ένα κενό φύλλο δεδομένων), να εισαχθούν δεδομένα από μία βάση δεδομένων και να ανοιχθεί ένα ήδη υπάρχον φύλλο δεδομένων του SPSS ή άλλης εφαρμογής (πχ. MS Excel). Επιλογή του [Type in data] και τσεκάρισμα του [Don't show this dialog in the future].

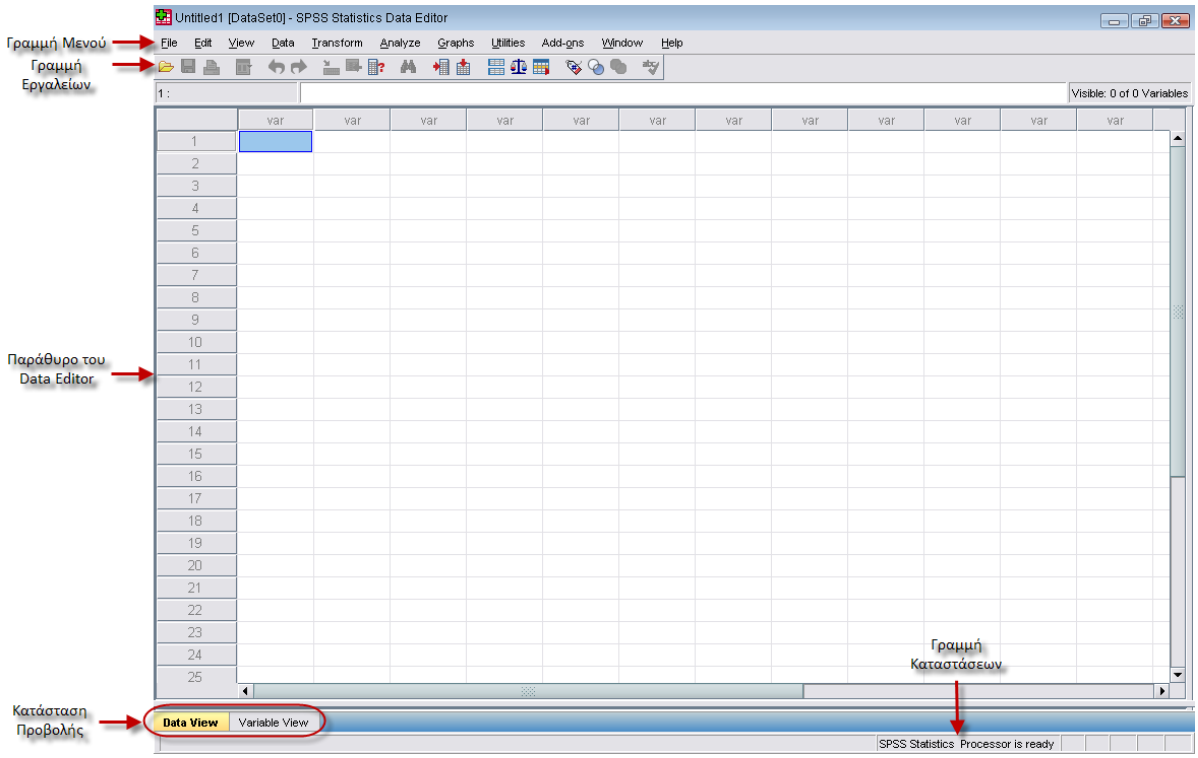


Εικόνα 2.1: Πρώτη οθόνη του SPSS μετά την εγκατάσταση του

Στη συνέχεια, μετά το πάτημα του κουμπιού [OK], εμφανίζεται το παρακάτω παράθυρο (βλ. Εικόνα 2.2), ένα άδειο φύλλο δεδομένων (dataset), το οποίο ονομάζεται Data Editor. Το κελί (cell) με το εντονότερο περίγραμμα και το γαλάζιο χρώμα ονομάζεται ενεργό κελί και προστίθεται σ' αυτό ότι πληκτρολογηθεί. Η μετακίνηση μεταξύ των κελιών πραγματοποιείται με το ποντίκι ή με τα βέλη (→,↓, ←,↑) του πληκτρολογίου.

Σε κάθε παράθυρο του Data Editor εμφανίζονται:

- Η γραμμή μενού (Menu bar), που αποτελούν τον τρόπο επικοινωνίας του χρήστη με το λογισμικό, ορισμένα απ' αυτά (File, Edit, View, Window, Help) είναι αντίστοιχα με αυτά που υπάρχουν σε άλλες εφαρμογές των windows και τα υπόλοιπα πράττουν μια σειρά από διάφορες λειτουργίες του λογισμικού.
- Η γραμμή εργαλείων (Toolbar), ορισμένα εικονίδια συντόμευσης για άμεση πρόσβαση σε λειτουργίες που γίνονται συχνότερα. Τοποθετώντας για λίγο το δείκτη του ποντικιού πάνω από τη γραμμή εργαλείων εμφανίζεται μια σύντομη επεξήγηση της λειτουργίας τους.
- Η γραμμή καταστάσεων (Status bar).
- Ο διαθέσιμος χώρος για να κάνετε εισαγωγή δεδομένων.

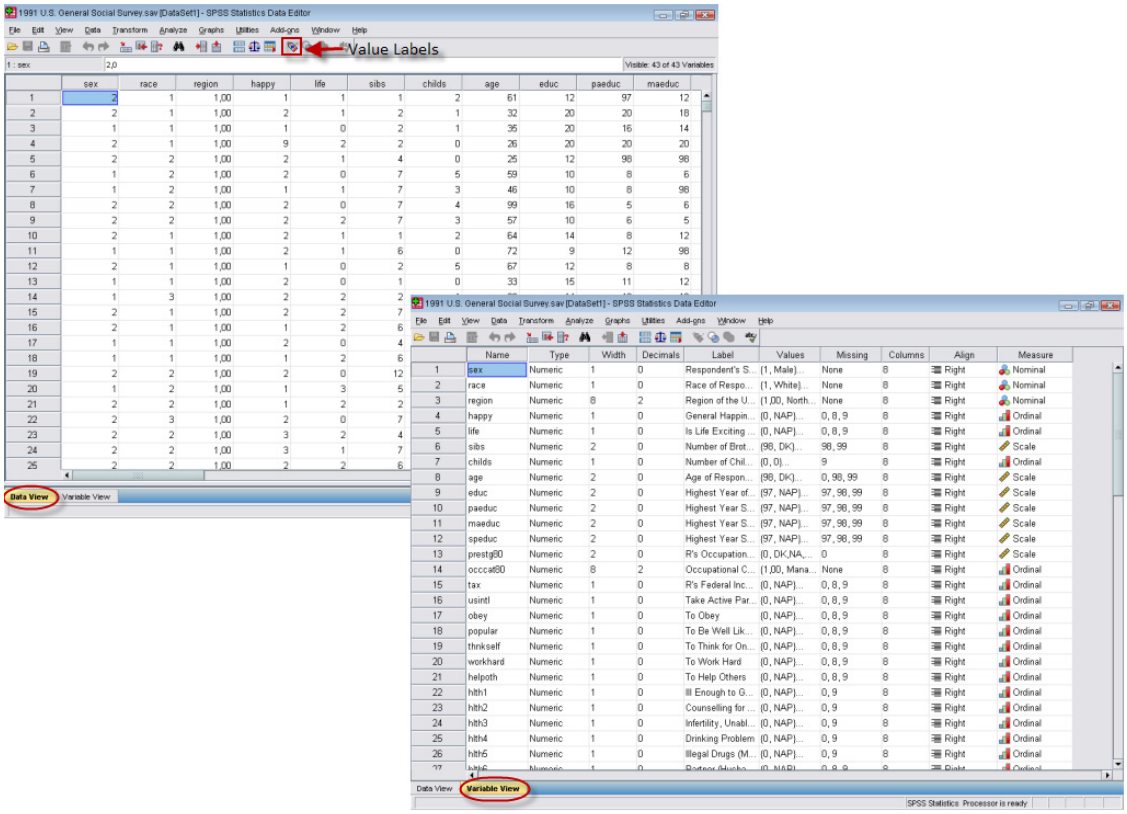


Εικόνα 2.2: Το παράθυρο του Data Editor

Στον Data Editor πραγματοποιείται η εισαγωγή και η διόρθωση των δεδομένων κατ' αντίστοιχο τρόπο μ' ένα λογιστικό φύλλο. Σ' αυτόν μπορεί να γίνει η εισαγωγή των δεδομένων ενός ήδη υπάρχοντος αρχείου ή να πληκτρολογηθούν νέα. Η μορφή των δεδομένων πρέπει να είναι αυτή ενός ορθογώνιου πίνακα όπου οι γραμμές είναι οι παρατηρήσεις (cases) (πχ. οι ασθενείς μίας έρευνας) και οι στήλες οι μεταβλητές (variables) (πχ. το φύλο, η ηλικία κλπ των ασθενών) ενός συνόλου δεδομένων.

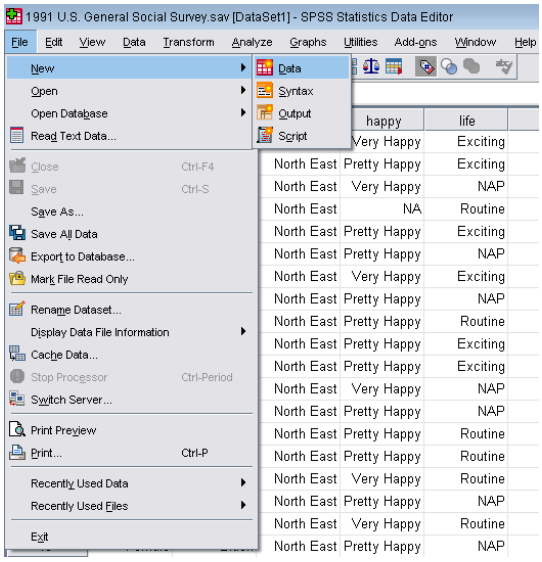
Τα δεδομένα ενός αρχείου στον Data Editor μπορούν να προβάλλονται (Κατάσταση Προβολής) με δύο διαφορετικούς τρόπους:

- **Data View** (προεπιλεγμένη): Εδώ εμφανίζονται οι αναλυτικές τιμές των δεδομένων μας είτε με την αριθμητική-αλφαριθμητική τιμή τους, είτε με τις λεκτικές επεξηγήσεις των κατηγοριών τους (value labels), κατ' αντίστοιχο τρόπο μ' ένα λογιστικό φύλλο.
- **Variable View**: Εδώ εμφανίζονται οι πληροφορίες για όλες τις μεταβλητές των δεδομένων μας. Οι πληροφορίες αυτές αφορούν το όνομα της μεταβλητής (Name), το είδος της (Type), τη λεκτική περιγραφή της (Label), τη λεκτική επεξήγηση των κατηγοριών της (Value), τον τρόπο χειρισμό των ελλειπουσών τιμών (Missing) και το επίπεδο μέτρησής της (Measure), (βλ. Εικόνα 2.3).



Εικόνα 2.3: Το παράθυρο του Data Editor με τις δύο καταστάσεις προβολών

Επιλέγοντας το μενού [File] στον Data Editor γίνεται η διαπίστωση ότι πολλές από τις επιλογές του είναι ήδη γνωστές από άλλες παραθυρικές εφαρμογές (βλ. Εικόνα 2.4).

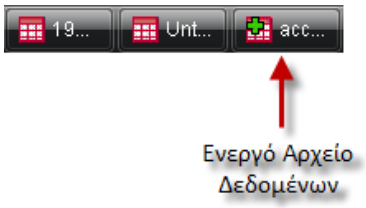


Εικόνα 2.4: Το μενού [File →New] του SPSS

Εδώ κυρίως δίνεται η δυνατότητα δημιουργίας ενός νέου αρχείου, το άνοιγμα κάποιου ήδη αποθηκευμένου αρχείου, η αποθήκευση και η εκτύπωση του αρχείου.

Επιλέγοντας το [New] δίνεται η δυνατότητα δημιουργίας ενός νέου αρχείου που μπορεί να είναι:

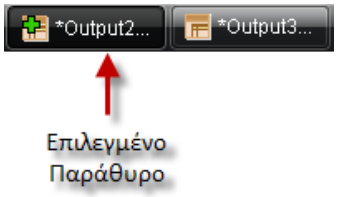
- [**Data**]: Ένα άδειο φύλλο δεδομένων (dataset), για πέρασμα νέων στοιχείων. Υπάρχει η δυνατότητα ταυτόχρονα να είναι ανοιχτά πολλά αρχεία δεδομένων (datasets), απ' αυτά μόνο το ένα είναι διαθέσιμο προς επεξεργασία και ονομάζεται ενεργό αρχείο δεδομένων (active dataset). Η αλλαγή του ενεργού αρχείου δεδομένων πραγματοποιείται κάνοντας κλικ στο επιθυμητό παράθυρο του Data Editor, οπότε στην γραμμή εργασιών των Windows το ενεργό αρχείο δεδομένων θα έχει επάνω του έναν πράσινο σταυρό (βλ. Εικόνα 2.5). Τα αρχεία δεδομένων (Data) του SPSS αποθηκεύονται με την επέκταση **sav**.



Εικόνα 2.5: Ενεργό αρχείο δεδομένων

- [**Syntax**]: Εδώ κάποιος μπορεί να γράψει κώδικα, της γλώσσας εντολών που χρησιμοποιεί το SPSS, αντί να χρησιμοποιήσει τα μενού επιλογών. Τα αρχεία Syntax του SPSS αποθηκεύονται με την επέκταση .**sps**.

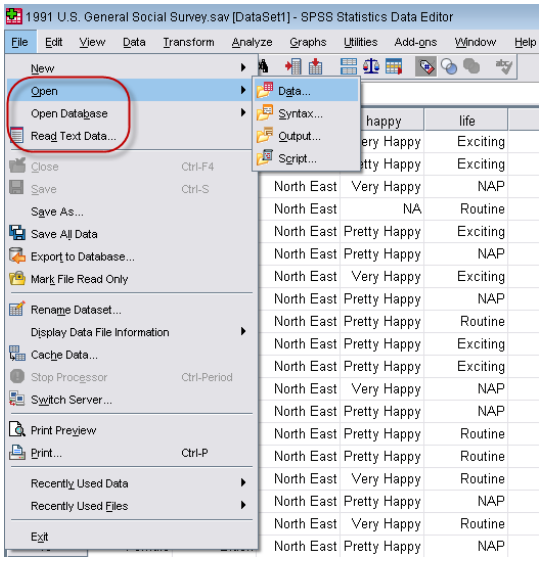
- [**Output**]: Κάθε φορά που εκτελείται μια εντολή στο SPSS, το αποτέλεσμά της κατευθύνεται σε ένα ξεχωριστό παράθυρο εξόδου, τον Output Viewer. Υπάρχει η δυνατότητα ταυτόχρονα να είναι ανοιχτά πολλά παράθυρα εξόδου [Output), απ' αυτά μόνο το ένα είναι επιλεγμένο και διαθέσιμο προς επεξεργασία (designated window).). Η αλλαγή του επιλεγμένου παράθυρου του Output Viewer πραγματοποιείται επιλέγοντας [Utilities → Designated Window] ή το πράσινο σταυρό στη γραμμή εργαλείων του, οπότε στην γραμμή εργασιών των Windows το επιλεγμένο παράθυρο θα έχει επάνω έναν πράσινο σταυρό (βλ. Εικόνα 2.6). Τα αρχεία Output του SPSS αποθηκεύονται με την επέκταση .**spv**.



Εικόνα 2.6: Επιλεγμένο παράθυρο Output Viewer

- [**Script**]: Το παράθυρο αυτό δίνει στο χρήστη τη δυνατότητα να συντάξει πλήρη προγράμματα σε μια γλώσσα που μοιάζει αρκετά με τη VISUAL BASIC. Τα προγράμματα αυτά έχουν πρόσβαση στις λειτουργίες του SPSS και είναι δυνατόν ο χρήστης να δημιουργήσει δικές του διαδικασίες - που δεν αποτελούν μέρος του SPSS - αξιοποιώντας τις λειτουργίες του SPSS. Τα αρχεία Script του SPSS αποθηκεύονται με την επέκταση .**sbs**.

Επίσης, στο μενού [File] υπάρχουν τρεις επιλογές για την ανάγνωση δεδομένων, από υπάρχοντα αρχεία (βλ. Εικόνα 2.7).



Εικόνα 2.7: Επιλογές ανάγνωσης δεδομένων από το μενού [File] του SPSS

- [**Open**]: Άνοιγμα ενός αρχείου με επέκταση σε .sav (πχ. patients.sav) αλλά και άλλων αρχείων δεδομένων όπως excel, lotus, sas, stata κλπ (προεπιλογή είναι τα αρχεία με επέκταση σε .sav).

- [**Open Database**]: Άνοιγμα μίας βάσης δεδομένων (πχ. Excel, DBase, Access κλπ) και εισαγωγής δεδομένων.

- `[**Read Text Data**]: Άνοιγμα ενός αρχείου με επέκταση σε .txt (αρχείο κειμένου-ASCII file) αλλά και άλλα αρχεία δεδομένων όπως excel, lotus, sas, stata κλπ (προεπιλογή είναι τα αρχεία με επέκταση σε .txt/.dat).

3ο ΚΕΦΑΛΑΙΟ
Εισαγωγή Δεδομένων στο SPSS

Εισαγωγή

Το SPSS δίνει τη δυνατότητα για εισαγωγή δεδομένων με τους παρακάτω τρόπους:

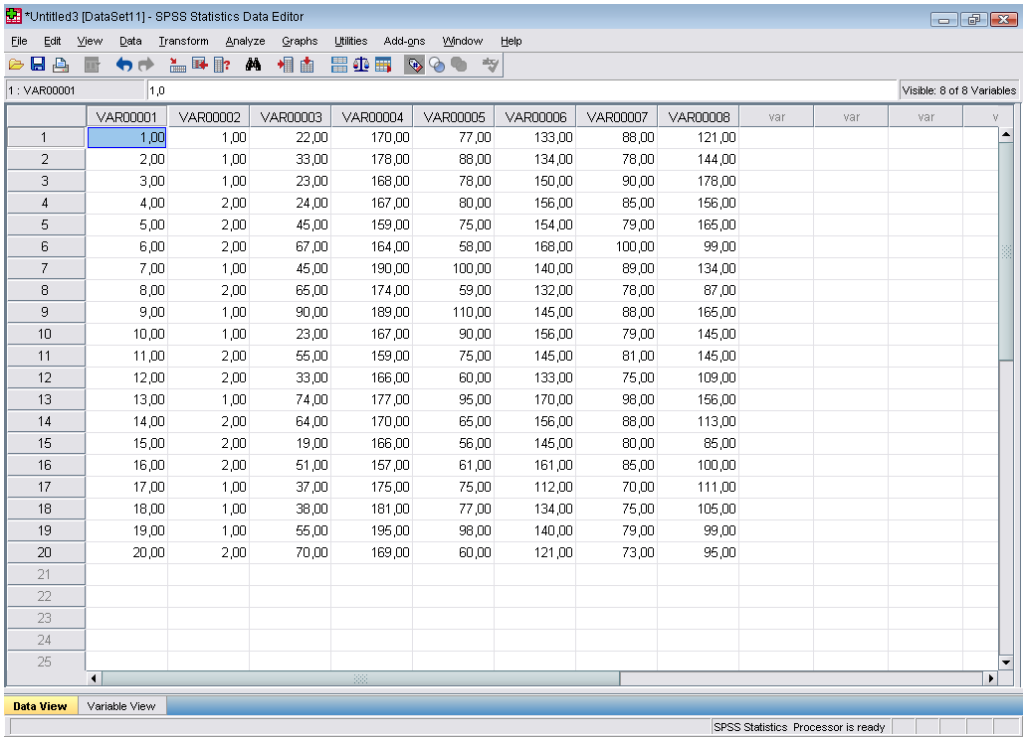
- Πληκτρολογώντας δεδομένα απευθείας στο Data View.
- Με άνοιγμα ενός απλού αρχείου (πχ. αρχείου κειμένου).
- Με άνοιγμα ενός αρχείου από Βάση Δεδομένων.
- Με μετακίνηση και αντιγραφή δεδομένων.

3.1 Πληκτρολόγηση δεδομένων

Επιλέχθηκαν τυχαία 20 ασθενείς που νοσηλεύονταν σε παθολογική κλινική μεγάλου νοσοκομείου και καταγράφηκαν το φύλο, η ηλικία, το ύψος, το βάρος, η συστολική και η διαστολική πίεση καθώς και το επίπεδο της LDL χοληστερίνης τους, όπως φαίνεται στον παρακάτω πίνακα.

ID	Sex	Age	Height	Weight	Sys_Pr	Dias_Pr	Chol
1	Άνδρας	22	170	77	133	88	121
2	Άνδρας	33	178	88	134	78	144
3	Άνδρας	23	168	78	150	90	178
4	Γυναίκα	24	167	80	156	85	156
5	Γυναίκα	45	159	75	154	79	165
6	Γυναίκα	67	164	58	168	100	99
7	Άνδρας	45	190	100	140	89	134
8	Γυναίκα	65	174	59	132	78	87
9	Άνδρας	90	189	110	145	88	165
10	Άνδρας	23	167	90	156	79	145
11	Γυναίκα	55	159	75	145	81	145
12	Γυναίκα	33	166	60	133	75	109
13	Άνδρας	74	177	95	170	98	156
14	Γυναίκα	64	170	65	156	88	113
15	Γυναίκα	19	166	56	145	80	85
16	Γυναίκα	51	157	61	161	85	100
17	Άνδρας	37	175	75	112	70	111
18	Άνδρας	38	181	77	134	75	105
19	Άνδρας	55	195	98	140	79	99
20	Γυναίκα	70	169	60	121	73	95

Η εισαγωγή των δεδομένων στον Data View πραγματοποιείται έχοντας υπόψη ότι οι γραμμές του πίνακα αντιστοιχούν σε περιπτώσεις (στο παράδειγμα οι ασθενείς) και οι στήλες σε μεταβλητές (στο παράδειγμα το φύλο, το βάρος κλπ). Έτσι, η συνήθης πρακτική είναι να πραγματοποιείται η εισαγωγή των δεδομένων μίας περίπτωσης για όλες τις μεταβλητές. Ο Data View θα έχει τη μορφή της παρακάτω εικόνας (βλ. Εικόνα 3.1).

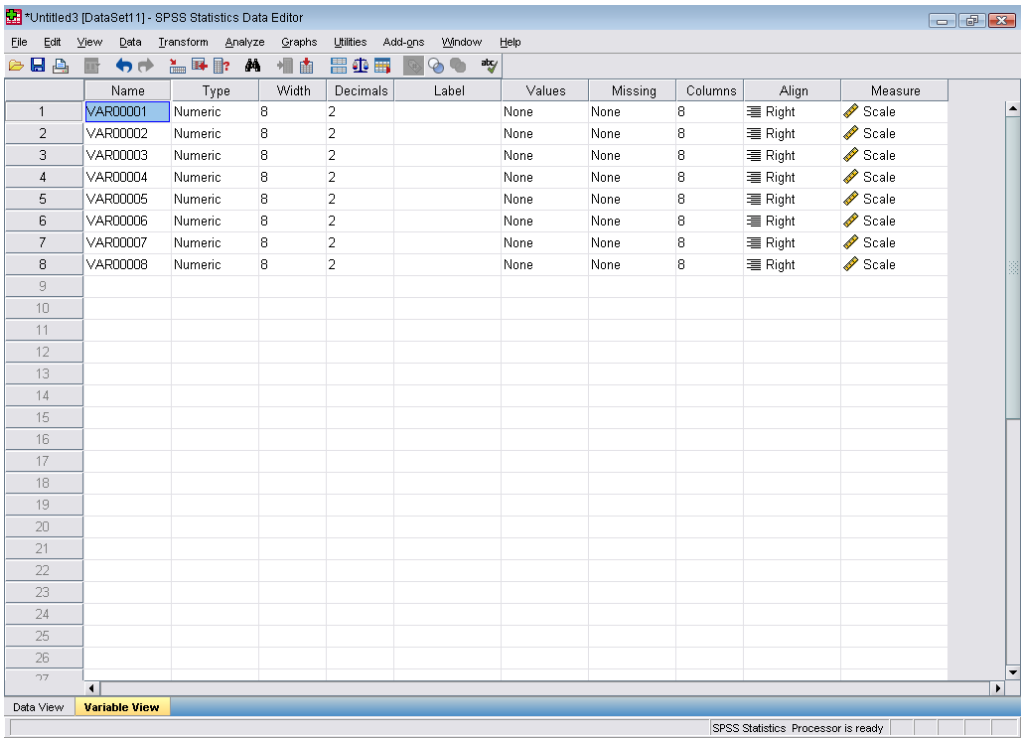


	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008	var	var	var	v
1	1,00	1,00	22,00	170,00	77,00	133,00	88,00	121,00				
2	2,00	1,00	33,00	178,00	88,00	134,00	78,00	144,00				
3	3,00	1,00	23,00	168,00	78,00	150,00	90,00	178,00				
4	4,00	2,00	24,00	167,00	80,00	156,00	85,00	156,00				
5	5,00	2,00	45,00	159,00	75,00	154,00	79,00	165,00				
6	6,00	2,00	67,00	164,00	58,00	168,00	100,00	99,00				
7	7,00	1,00	45,00	190,00	100,00	140,00	89,00	134,00				
8	8,00	2,00	65,00	174,00	59,00	132,00	78,00	87,00				
9	9,00	1,00	90,00	189,00	110,00	145,00	88,00	165,00				
10	10,00	1,00	23,00	167,00	90,00	156,00	79,00	145,00				
11	11,00	2,00	55,00	159,00	75,00	145,00	81,00	145,00				
12	12,00	2,00	33,00	166,00	60,00	133,00	75,00	109,00				
13	13,00	1,00	74,00	177,00	95,00	170,00	98,00	156,00				
14	14,00	2,00	64,00	170,00	65,00	156,00	88,00	113,00				
15	15,00	2,00	19,00	166,00	56,00	145,00	80,00	85,00				
16	16,00	2,00	51,00	157,00	61,00	161,00	85,00	100,00				
17	17,00	1,00	37,00	175,00	75,00	112,00	70,00	111,00				
18	18,00	1,00	38,00	181,00	77,00	134,00	75,00	105,00				
19	19,00	1,00	55,00	195,00	98,00	140,00	79,00	99,00				
20	20,00	2,00	70,00	169,00	60,00	121,00	73,00	95,00				
21												
22												
23												
24												
25												

Εικόνα 3.1: Ο Data View μετά την εισαγωγή των δεδομένων

Πραγματοποιείται η κωδικοποίηση του φύλου και ορίζεται για τους άνδρες ο κωδικός '1' και για τις γυναίκες ο κωδικός '2'. Η κωδικοποίηση είναι μία τεχνική που χρησιμοποιείται για μεταβλητές που περιέχουν αλφαριθμητικά δεδομένα και μπορούν εύκολα να κατηγοριοποιηθούν (πχ. το φύλο σε άνδρες – γυναίκες, η οικογενειακή κατάσταση σε έγγαμος – άγαμος – διαζευγμένος – χήρος, το μορφωτικό επίπεδο σε απόφοιτος πρωτοβάθμιας – δευτεροβάθμιας – τριτοβάθμιας εκπαίδευσης κλπ).

Στο παράθυρο Variable View καθορίζονται τα χαρακτηριστικά κάθε μεταβλητής. Εδώ η παρουσίαση των μεταβλητών είναι σε αντιμετάθεση, οι οποίες πλέον εμφανίζονται στις σειρές του πίνακα αντί στις στήλες, ενώ στις στήλες καθορίζονται τα χαρακτηριστικά κάθε μεταβλητής (βλ. Εικόνα 3.2).



Εικόνα 3.2: Το Variable View μετά την εισαγωγή των δεδομένων

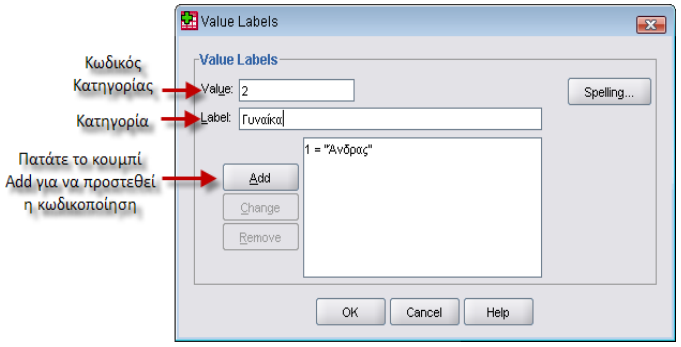
Πιο συγκεκριμένα:

- [**Name**]: Εδώ καθορίζεται το όνομα της μεταβλητής. Προτείνεται η χρήση ονομάτων με λατινικούς χαρακτήρες, όχι ελληνικούς, και μικρό σχετικά μέγεθος, όχι πάνω από 8 χαρακτήρες, προκειμένου να διατηρηθεί η συμβατότητα του αρχείου σας με παλαιότερες εκδόσεις του SPSS. Ο ορισμός πιο αναλυτικού ονόματος με ελληνικούς χαρακτήρες πραγματοποιείται με την επιλογή [Label] (βλ. παρακάτω). Στην περίπτωση ορισμού του ίδιου ονόματος σε δυο μεταβλητές θα εμφανιστεί ένα προειδοποιητικό μήνυμα ότι κάτι τέτοιο δεν επιτρέπεται. Τα ονόματα των μεταβλητών πρέπει να ακολουθούν τους παρακάτω κανόνες:
    - Να αρχίζουν με γράμμα του ελληνικού ή λατινικού αλφαβήτου και οι υπόλοιποι χαρακτήρες μπορεί να είναι γράμματα, αριθμοί, μια τελεία ή τα σύμβολα @, #, _, $.
    - Δεν μπορούν να τελειώνουν σε τελεία.
    - Δεν μπορεί να είναι το μέγεθός της πάνω από 64 χαρακτήρες.
    - Δεν μπορεί να περιλαμβάνει κενά ή ειδικούς χαρακτήρες όπως !, ?, ¨, *.
    - Δεν μπορούν να χρησιμοποιηθούν λέξεις κλειδιά που χρησιμοποιεί το SPSS όπως ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO , WITH.

- Δεν μπορούν διαφορετικές μεταβλητές να έχουν το ίδιο όνομα (στο ίδιο αρχείο). Το SPSS στην ονομασία μεταβλητών δεν ξεχωρίζει μικρά με κεφαλαία γράμματα οπότε λέξεις όπως VAR00001, VaR00001 και vAR00001 τις θεωρεί ίδιες.
- Καλό είναι να μη τελειώνει σε κάτω παύλα (_) ώστε να μη συγχέονται οι μεταβλητές που παράγει αυτόματα το SPSS με τις δικές σας.

- [**Type**]: Εδώ επιλέγεται το είδος των δεδομένων για κάθε μεταβλητή. Οι μεταβλητή μπορεί να είναι αριθμητική (numeric), αλφαριθμητική (string) ή ημερομηνία (date). Εξ' ορισμού οι νέες μεταβλητές έχουν ως type το [numeric], δηλαδή θεωρείται ότι το είδος των δεδομένων είναι αριθμητικό. Τα [dot], [comma] και [scientific notation], [dollar] και [custom currency] αφορούν στον τρόπο παρουσίασης των αριθμών στα κελιά. Οι ημερομηνίες εσωτερικά στο SPSS αποθηκεύονται σε δευτερόλεπτα που έχουν παρέλθει από τις 14 Οκτωβρίου 1582.
- [**Width**]: Εδώ επιλέγεται ο συνολικός αριθμός χαρακτήρων της μεταβλητής. Οι αλφαριθμητικές μεταβλητές αποθηκεύονται εσωτερικά στο SPSS με τόσους χαρακτήρες όσους έχετε ορίσει. Εάν για παράδειγμα καταχωρηθεί η τιμή "Ναι" και το Width είναι ορισμένο σε 10 χαρακτήρες τότε μετά το "Ναι" θα ακολουθήσουν 7 κενοί χαρακτήρες (spaces) δηλαδή εσωτερικά το SPSS το αποθηκεύει σαν "Ναι " ενώ στον Data View θα παρουσιάζεται σαν "Ναι".
- [**Decimals**]: Εδώ επιλέγεται ο μέγιστος αριθμός δεκαδικών που θα εμφανίζονται στον Data View. Εάν οριστεί μία αριθμητική μεταβλητή να έχει 2 δεκαδικά και καταχωρηθούν 5 δεκαδικά στον Data View τότε το SPSS θα αποθηκεύσει τα δεδομένα που καταχωρήθηκαν (δηλαδή 5) αλλά θα εμφανίζονται στον Data View στρογγυλοποιημένα με 2 δεκαδικά. Μπορούν να καταχωρηθούν μέχρι και 16 δεκαδικά.
- [**Label**]: Εδώ ορίζεται η περιγραφή της μεταβλητής με μέγιστο μήκος 256 χαρακτήρες. Είναι σημαντικό να προσεχθεί η ορθογραφία των περιγραφών, πεζά - κεφαλαία γράμματα κ.λπ., ώστε να βελτιωθεί η αναγνωσιμότητα των αποτελεσμάτων, καθώς το [Label] (για όσες μεταβλητές έχει οριστεί) χρησιμοποιείται ως επεξήγηση στον Output Viewer (πίνακες, γραφήματα) καθώς και στον Data View (αφήνοντας το δείκτη του ποντικιού επάνω στο όνομα της μεταβλητής).
- [**Values**]: Πολύ συχνά στο SPSS χρειάζεται να πραγματοποιηθεί κωδικοποίηση των κατηγορικών-διατεταγμένων δεδομένων σε αριθμητική μορφή. Για παράδειγμα, το "Άνδρας" και "Γυναίκα" μπορεί να κωδικοποιηθεί ως "1" και "2" αντίστοιχα. Για να

πραγματοποιηθεί πραγματοποιείται επιλογή της μεταβλητής [Sex] και στη στήλη [Values] πάτημα του κουμπιού με τις 3 τελείες οπότε και ανοίγει το πλαίσιο διαλόγου που ακολουθεί (βλ. Εικόνα 3.3).



Εικόνα 3.3: Πλαίσιο διαλόγου κωδικοποίησης μεταβλητών

- [**Missing**]: Εδώ ορίζεται ποιες συγκεκριμένες τιμές το SPSS θα εκλαμβάνει ως ελλείπουσες (user-missing values), κατά την ανάλυση των δεδομένων, σε περιπτώσεις που είναι απαραίτητη η διάκριση της έλλειψης τιμών ανάλογα με το λόγο που αυτές προέκυψαν. Εάν δεν οριστούν ελλείπουσες τιμές, σε αριθμητικές μεταβλητές που είναι κενές, κατά την ανάλυση αυτές θα χαρακτηρίζονται ως system missing values και στο Data View θα εμφανίζονται με το σύμβολο του δεκαδικού σημείου. Εάν δεν οριστούν ελλείπουσες τιμές, σε αλφαριθμητικές μεταβλητές που είναι κενές, κατά την ανάλυση αυτές δεν θα χαρακτηρίζονται ως system missing values, εκτός και εάν οριστούν, και στο Data View θα εμφανίζονται με το σύμβολο του κενού (space). Μια αλφαριθμητική τιμή, η οποία είναι κενή, δεν θα οριστεί ως user-missing value εάν στο πλαίσιο διαλόγου [Missing Values] στο πεδίο [Discrete missing values] πληκτρολογηθεί ένα κενό (space). Μπορούν να οριστούν μέχρι και 3 διακριτές ελλείπουσες τιμές ή ένα εύρος και 1 ελλείπουσα διακριτή τιμή (βλ. Πίνακα 3.1).

Ηλικιακή Ομάδα Ασθενών

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<=30	4	20,0	21,1	21,1
	Μεταξύ 31 και 60	7	35,0	36,8	57,9
	>=61	3	15,0	15,8	73,7
	Δεν απάντησαν	3	15,0	15,8	89,5
	Δεν Γνώριζαν	2	10,0	10,5	100,0
	Total	19	95,0	100,0	
Missing	System	1	5,0		
Total		20	100,0		

User-missing values

System missing values

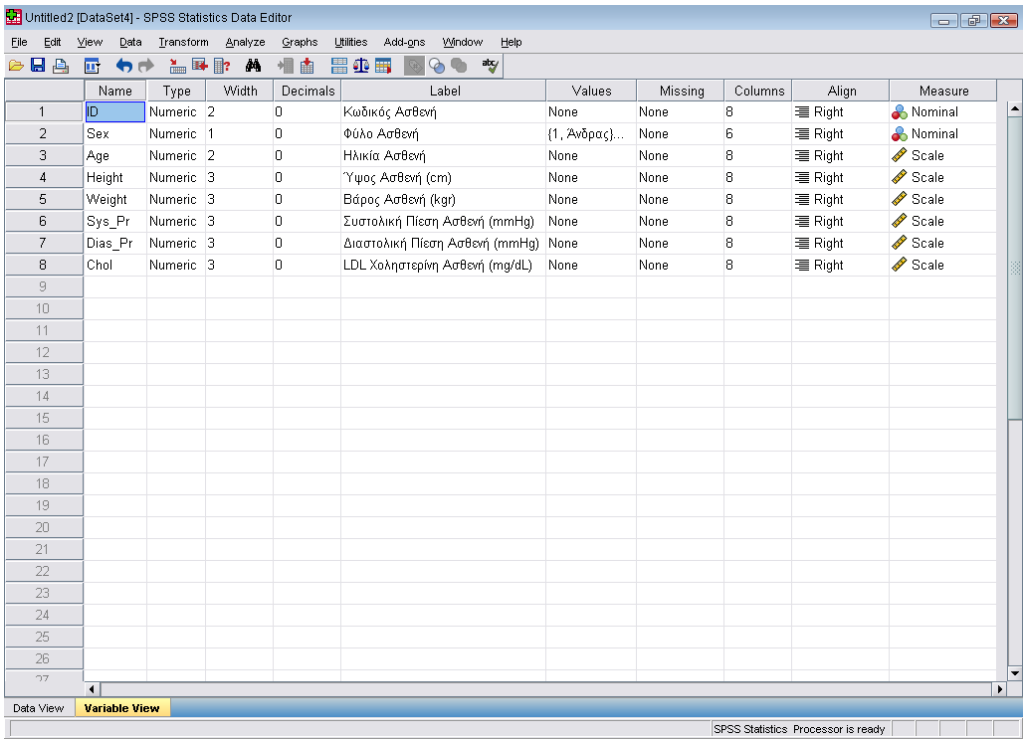
Πίνακας 3.1: Χειρισμός Missing Values

- [**Columns**]: Με αυτή τη ρύθμιση τροποποιείται το πλάτος κάθε στήλης στο Data view. Η διαδικασία αυτή μπορεί να γίνει και από το Data view "τραβώντας" τα όρια κάθε στήλης μέχρι το επιθυμητό σημείο.

- [**Align**]: Με τη ρύθμιση αυτή τροποποιείται τη στοίχιση των δεδομένων, κάθε στήλης, στο Data view.
- [**Measure**]: Αυτή η ρύθμιση αντικατοπτρίζει την ταξινόμηση της μεταβλητής σε μία από τις γενικές κατηγορίες κλιμάκων μέτρησης. Στην πράξη δεν επηρεάζει το αποτέλεσμα των διαδικασιών που θα εκτελεστούν στο SPSS παρά μόνο στη διαδικασία Custom Tables. Υπάρχουν τρεις επιλογές:
    - [**Scale**]: Μεταβλητές Κλίμακας Ίσων Διαστημάτων (ή αναλογικές) που αντιστοιχούν σε ποσοτικά μεγέθη (πχ. αποτελέσματα διάφορων μικροβιολογικών εξετάσεων, βάρος, ύψος κλπ)
    - [**Ordinal**]: Μεταβλητές διατεταγμένες (ή τακτικές) που έχει νόημα η Ιεραρχία και αντιστοιχούν σε αλφαριθμητικές τιμές ή σε αριθμητικές τιμές που είναι στην ουσία η κωδικοποίηση των ιεραρχημένων κατηγοριών (πχ. οι απαντήσεις στην ερώτηση το πόσο κατανόησαν το μάθημα του SPSS στο τέλος της ημέρας: καθόλου, λίγο, αρκετά, πολύ, τέλεια).
    - [**Nominal**]: Μεταβλητές Ονομαστικές (ή κατηγορικές) που δεν είναι μετρήσιμα μεγέθη αλλά κατηγοριοποιούν τα στοιχεία ενός συνόλου σε ομάδες (διακεκριμένες κατηγορίες) και μπορεί να είναι είτε αριθμητικές είτε αλφαριθμητικές μεταβλητές (πχ. 1=Άνδρας, 2=Γυναίκα).

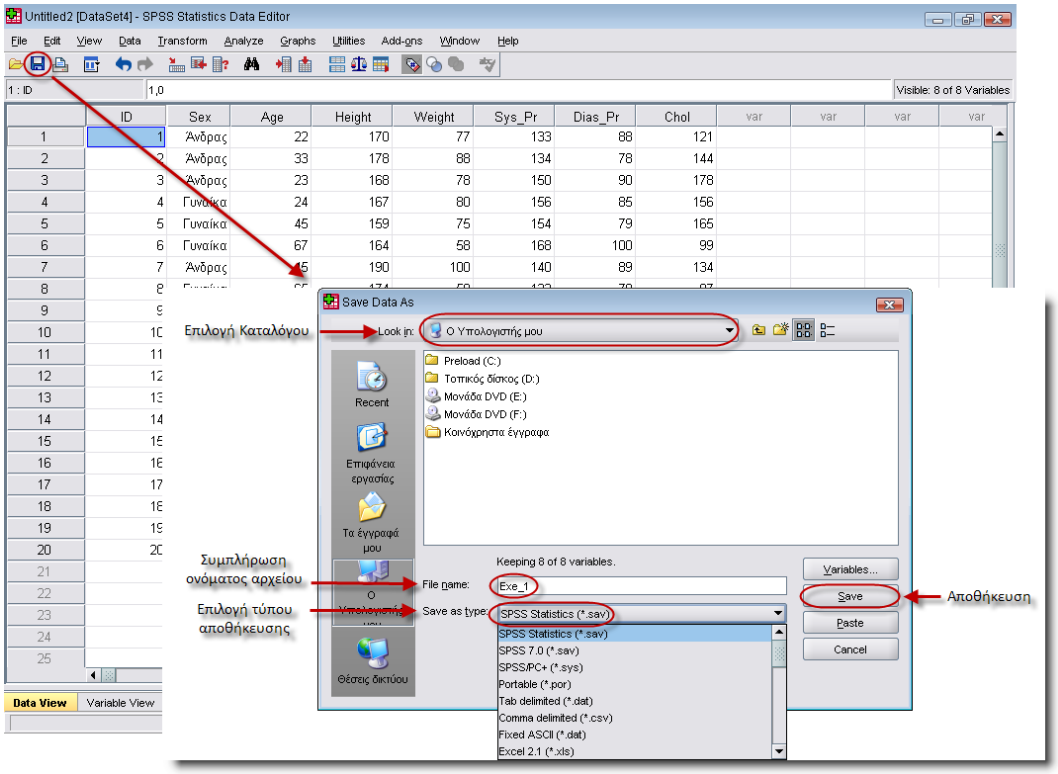
Αφού καθοριστούν τα χαρακτηριστικά κάθε μεταβλητής το Variable View θα έχει την παρακάτω εμφάνιση (βλ. Εικόνα 3.4).

Μπορούν να εφαρμοστούν οι ορισμένες ιδιότητες μίας μεταβλητής και σε άλλες καινούριες. Αυτό πραγματοποιείται επιλέγοντας τις επιθυμητές ιδιότητες (που μπορεί να είναι γραμμή, γραμμές, κελιά ή κελί) με το ποντίκι, στη συνέχεια πάτημα στο [Edit → Copy] και τέλος να επιλεχθεί με το ποντίκι το επιθυμητό σημείο επικόλλησης και πάτημα στο [Edit → Paste].



Εικόνα 3.4: Το Variable View μετά τον καθορισμό των ιδιοτήτων των μεταβλητών

Πατώντας το κουμπί της αποθήκευσης στη γραμμή εργαλείων αναδύεται το πλαίσιο διαλόγου [Save Data As] (βλ. Εικόνα 3.5).



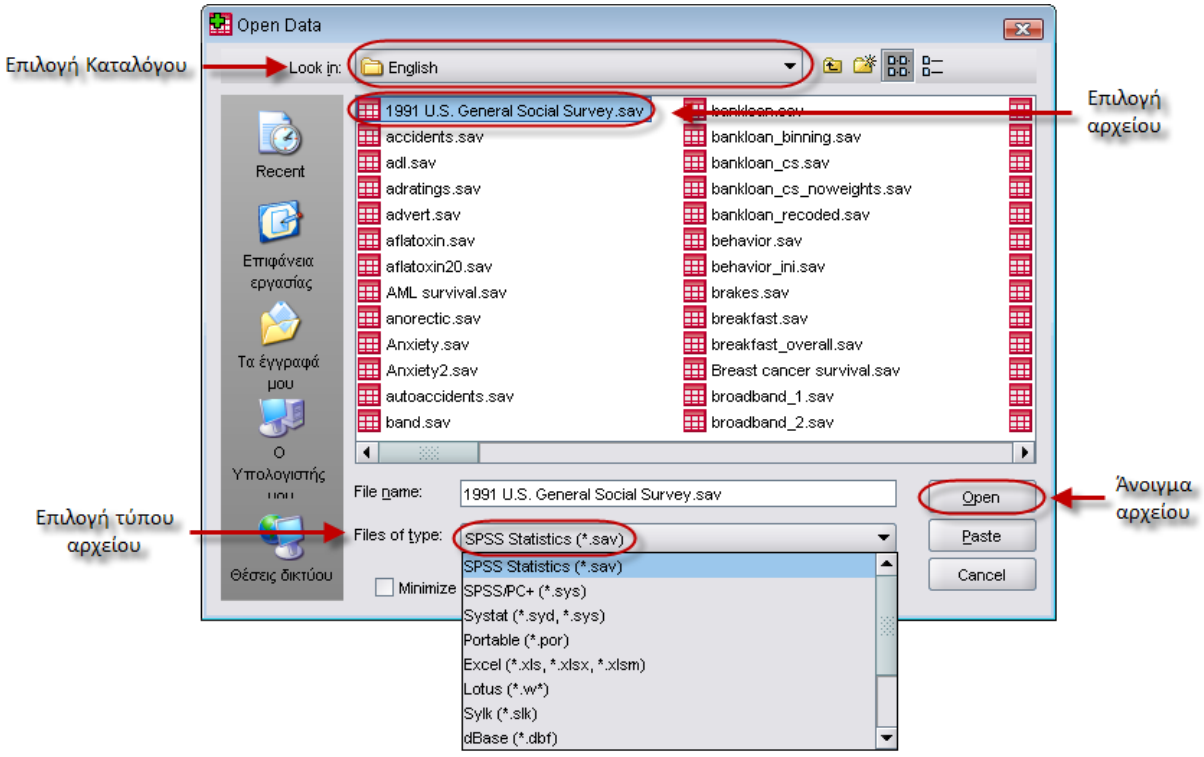
Εικόνα 3.5: Η διαδικασία αποθήκευσης αρχείου δεδομένων SPSS

Επιλέγεται ο επιθυμητός κατάλογος που θα αποθηκευτεί το αρχείο, συμπληρώνεται το όνομα του αρχείου, επιλέγεται ο τύπος (μορφή) με τον οποίο θα αποθηκευτεί (υπάρχει μία πληθώρα τύπων εκτός ως αρχείο SPSS όπως ως Excel, ως Stata, ως SAS κλπ) και τέλος πάτημα στο κουμπί [Save] (αποθήκευση).

3.2 Άνοιγμα απλού αρχείου δεδομένων

Το SPSS έχει τη δυνατότητα να ανοίγει εκτός των δικών του αρχείων και άλλους τύπους αρχείων όπως Excel, Stata, SAS, ASCII κλπ. Η διαδικασία που ακολουθείται έχει ως εξής (βλ. Εικόνα 3.6):

- Επιλογή από τη γραμμή μενού του [File → Open → Data].
- Από το πλαίσιο διαλόγου [Open Data] που αναδύθηκε επιλογή του καταλόγου που περιέχει το αρχείο.
- Επιλογή του τύπου του αρχείου που θα ανοιχτεί.
- Επιλογή του αρχείου κάνοντας κλικ με το ποντίκι επάνω του.
- Τέλος, πάτημα του κουμπιού [Open].

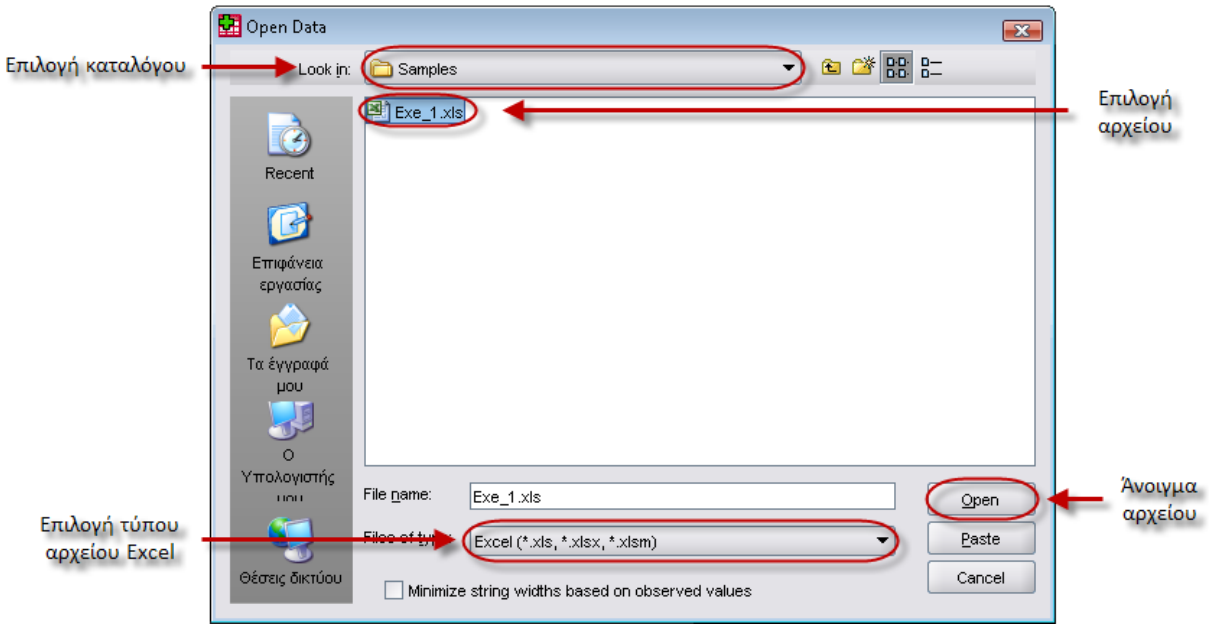


Εικόνα 3.6: Η διαδικασία ανοίγματος αρχείου δεδομένων SPSS

3.2.1 Άνοιγμα αρχείου δεδομένων Excel

Η διαδικασία που ακολουθείται έχει ως εξής (βλ. Εικόνα 3.7):

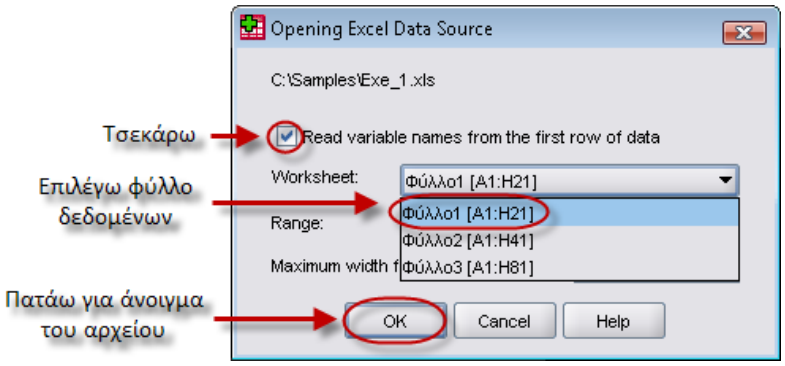
- Επιλογή από τη γραμμή μενού του [File → Open → Data].
- Από το πλαίσιο διαλόγου [Open Data] που αναδύθηκε επιλογή του καταλόγου που περιέχει το αρχείο Excel.
- Επιλογή στον τύπο αρχείου το [Excel (*.xlx, *.xlsx, *.xlsm)].
- Επιλογή του αρχείου κάνοντας κλικ με το ποντίκι επάνω του.
- Τέλος, πάτημα του κουμπιού [Open].



Εικόνα 3.7: Επιλογή αρχείου δεδομένων Excel για άνοιγμα

- Στο πλαίσιο διαλόγου [Opening Excel Data Source] που αναδύθηκε (βλ. Εικόνα 3.8):
    - Τσεκάρισμα του [Read variable names from the first line of data], εάν τα δεδομένα έχουν επικεφαλίδες, έτσι ώστε να οριστούν ως τα ονόματα των μεταβλητών.
    - Επιλογή του επιθυμητού [Worksheet] (φύλλο εργασίας) που περιέχει τα δεδομένα.
    - Δυνατότητα ορισμού μίας περιοχής κελιών στο πεδίο [Range] έτσι ώστε το SPSS να διαβάσει ένα τμήμα των δεδομένων.

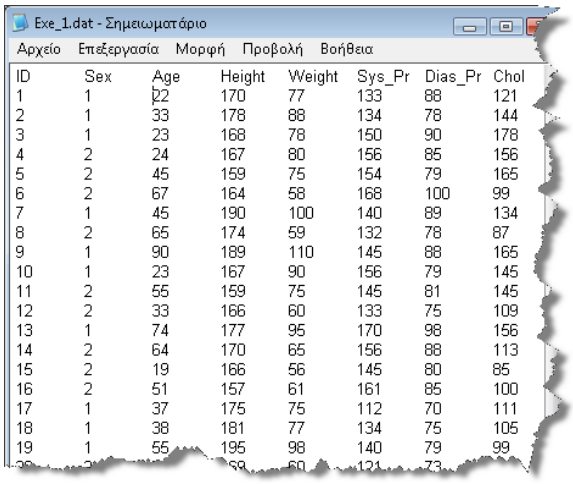
- Πάτημα του κουμπιού [OK].



Εικόνα 3.8: Άνοιγμα φύλλου δεδομένων Excel

3.2.2 Άνοιγμα αρχείου κειμένου (*.txt, *.dat)

Στα αρχεία αυτού του τύπου οι μεταβλητές διαχωρίζονται μεταξύ τους με tabs, κόμματα, κενά (spaces) ή με σταθερό πλάτος (fixed width format). Στη συνέχεια θα παρουσιαστεί ο τρόπος ανοίγματος και διαβάσματος δεδομένων από ένα αρχείο κειμένου με τις μεταβλητές να διακρίνονται μεταξύ τους με tab (η διαδικασία είναι ίδια και με τα άλλα διαχωριστικά) (βλ. Εικόνα 3.9).


Exe_1.dat - Σημειωματάριο

Αρχείο Επεξεργασία Μορφή Προβολή Βοήθεια

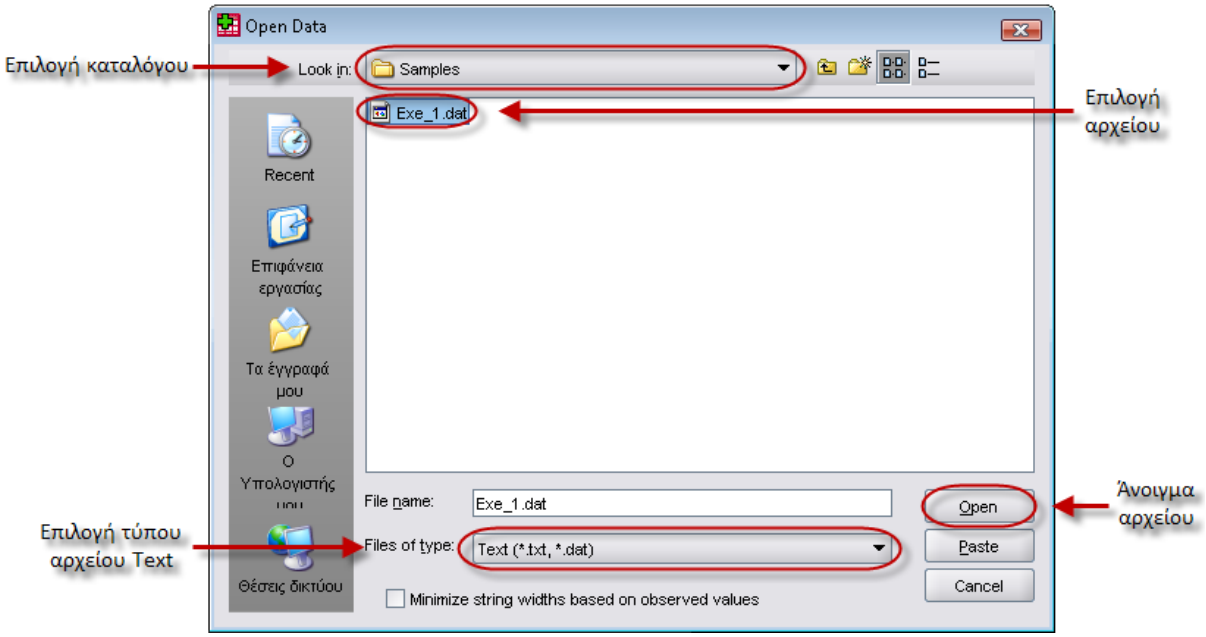
ID	Sex	Age	Height	Weight	Sys_Pr	Dias_Pr	Chol
1	1	22	170	77	133	88	121
2	1	33	178	88	134	78	144
3	1	23	168	78	150	90	178
4	2	24	167	80	156	85	156
5	2	45	159	75	154	79	165
6	2	67	164	58	168	100	99
7	1	45	190	100	140	89	134
8	2	65	174	59	132	78	87
9	1	90	189	110	145	88	165
10	1	23	167	90	156	79	145
11	2	55	159	75	145	81	145
12	2	33	166	60	133	75	109
13	1	74	177	95	170	98	156
14	2	64	170	65	156	88	113
15	2	19	166	56	145	80	85
16	2	51	157	61	161	85	100
17	1	37	175	75	112	70	111
18	1	38	181	77	134	75	105
19	1	55	195	98	140	79	99

Εικόνα 3.9: Αρχείο δεδομένων Text με διαχωριστικό μεταβλητών το tab

Η διαδικασία έχει ως εξής (βλ. Εικόνα 3.10):

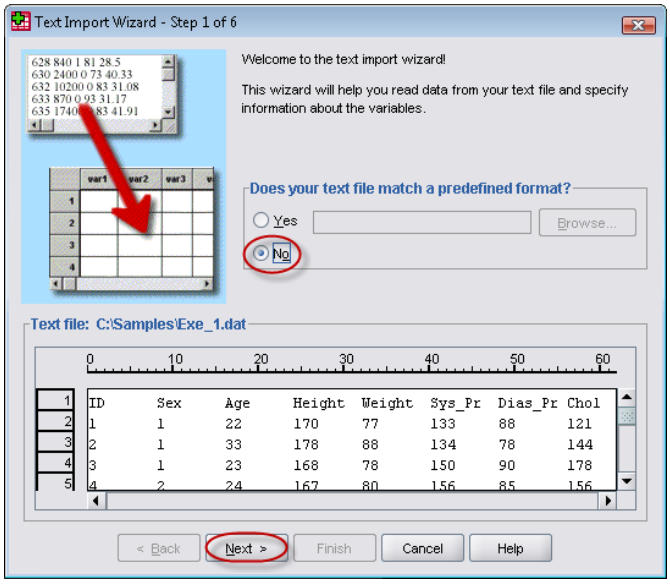
- Επιλογή από τη γραμμή μενού του [File → Read Text Data…].
- Από το πλαίσιο διαλόγου [Open Data] που αναδύθηκε επιλογή του καταλόγου που περιέχει το αρχείο κειμένου.

- Στον τύπο αρχείου επιλογή του [Text (*.txt, *.dat)], εάν δεν είναι επιλεγμένο.
- Επιλογή του αρχείου κάνοντας κλικ με το ποντίκι επάνω του.
- Τέλος, πάτημα του κουμπιού [Open] όπου και ανοίγει ο οδηγός [Text Import Wizard], ο οποίος έχει 6 βήματα.



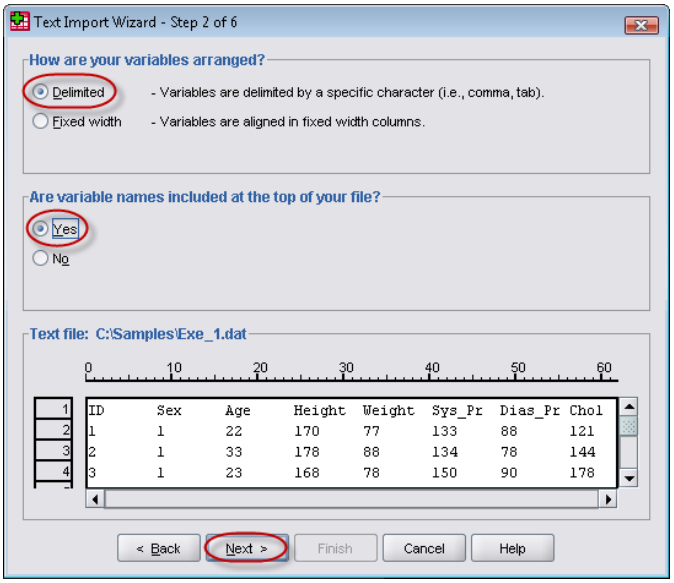
Εικόνα 3.10: Επιλογή αρχείου δεδομένων Text για άνοιγμα

- [Βήμα 1]: Ορισμός εάν θα χρησιμοποιηθεί, για το αρχείο που θα ανοιχτεί, μία προκαθορισμένη διαμόρφωση. Επιλογή του [No] και πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.11).



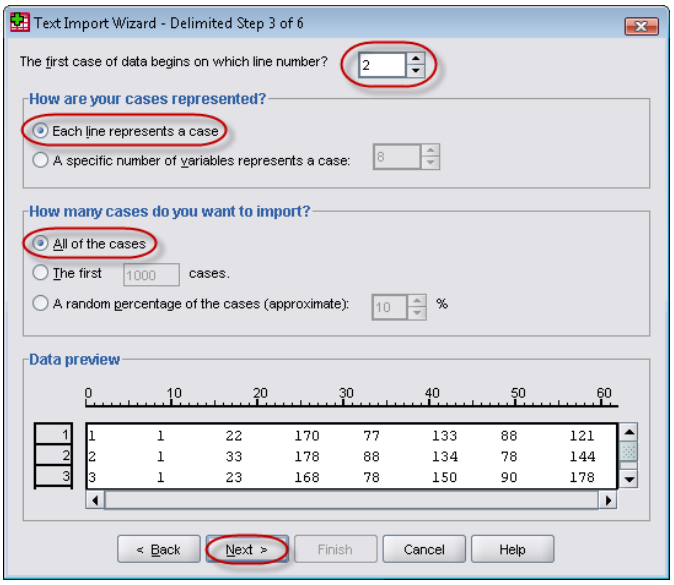
Εικόνα 3.11: 1ο βήμα του [Text Import Wizard]

- [Βήμα 2]: Ορισμός για το πως είναι διαχωρισμένες μεταξύ τους οι μεταβλητές καθώς και εάν υπάρχουν τα ονόματα των μεταβλητών στην αρχή (πρώτη γραμμή) του αρχείου κειμένου που θα ανοιχτεί. Επιλογή των [Delimited] και [Yes] και στη συνέχεια πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.12).



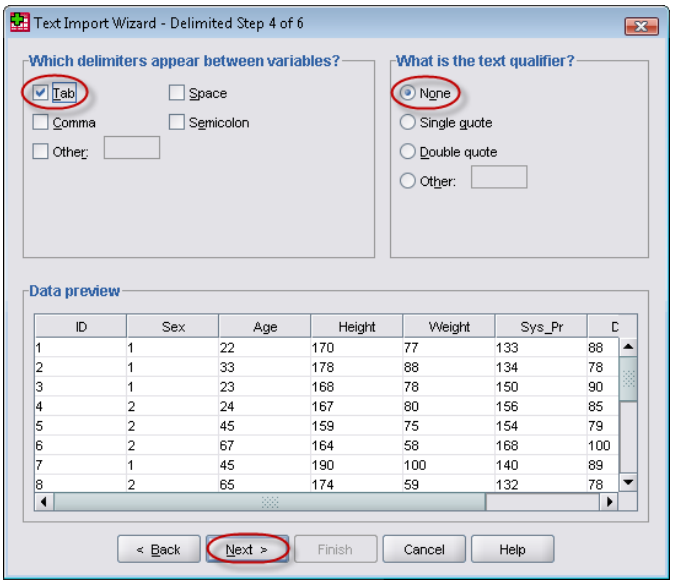
Εικόνα 3.12: 2ο βήμα του [Text Import Wizard]

- [Βήμα 3]: Ορισμός ποια είναι η πρώτη γραμμή των δεδομένων, τον τρόπο εμφάνισης των περιπτώσεων (cases) καθώς και του επιθυμητού αριθμού των δεδομένων που θα εισαχθούν στο SPSS. Επιλογή [2] (γιατί η πρώτη γραμμή περιλαμβάνει τα ονόματα των μεταβλητών), [Each line represents a case], [All of the cases] και πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.13).



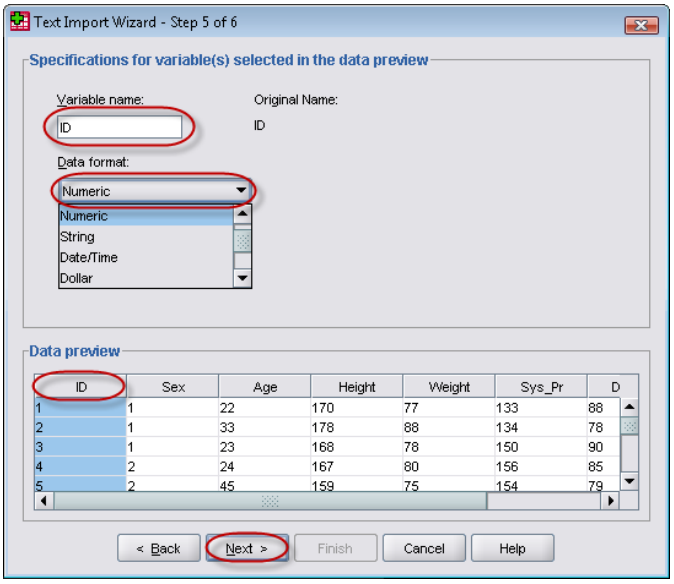
Εικόνα 3.13: 3ο βήμα του [Text Import Wizard]

- [Βήμα 4]: Ορισμός του διαχωριστικού παράγοντα μεταξύ των μεταβλητών καθώς και τον προσδιοριστικό παράγοντα των δεδομένων κειμένου. Επιλογή του [tab] (επειδή συνήθως δεν είναι γνωστό σε ένα αρχείο δεδομένων κειμένου ποιος είναι ο διαχωριστικός παράγοντας, μπορεί να επιλεχτεί ή να οριστεί κάποιος άλλος ανάλογα με την επιθυμητή εμφάνιση των δεδομένων στο παράθυρο [Data preview], επιλογή του [None] και πατάμε [Next] (βλ. Εικόνα 3.14).



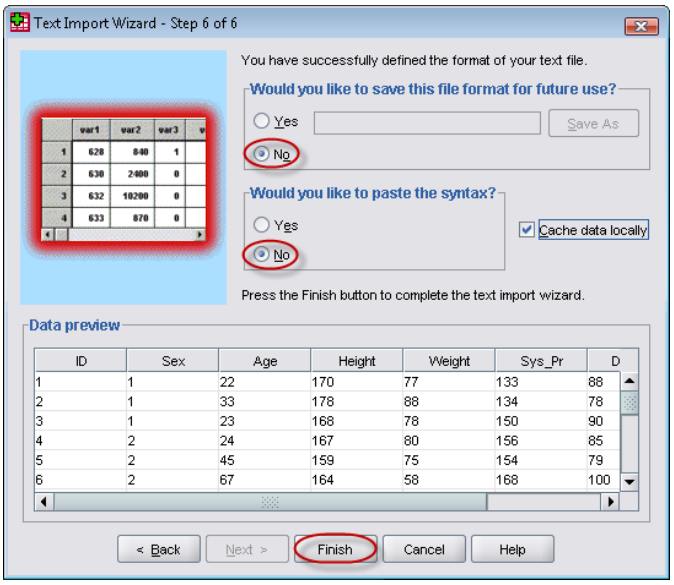
Εικόνα 3.14: 4ο βήμα του [Text Import Wizard]

- [Βήμα 5]: Ορισμός του ονόματος κάθε μεταβλητής καθώς και του τύπου των δεδομένων της (κάθε μεταβλητή επιλέγεται κάνοντας κλικ επάνω στο όνομα της μεταβλητής στο παράθυρο [Data preview]). Επιλογή, μία-μία, όλων των μεταβλητών και ορισμός τους ως [Numeric] και πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.15).



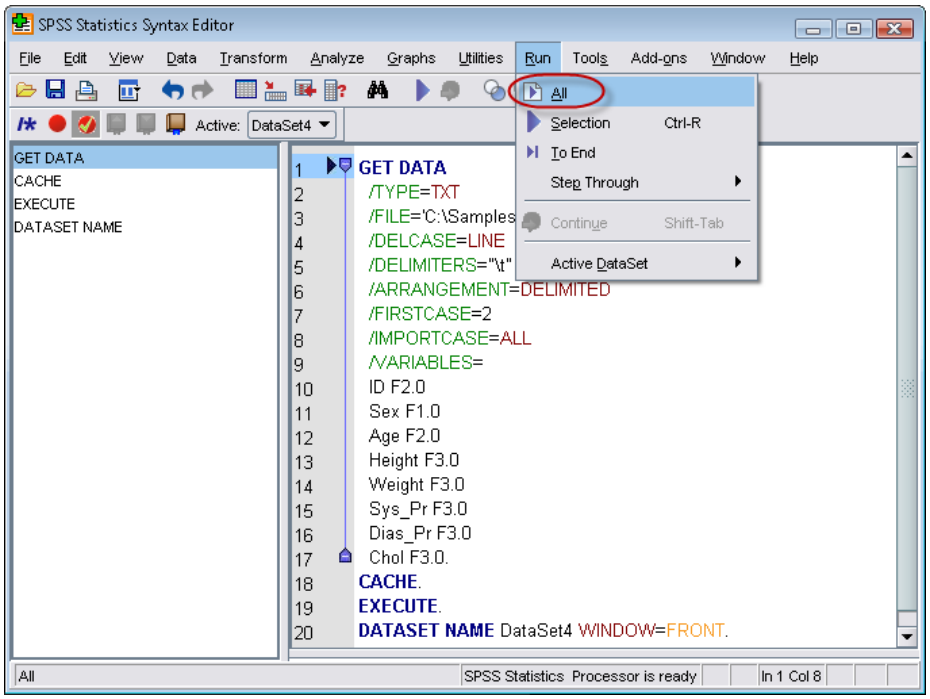
Εικόνα 3.15: 5ο βήμα του [Text Import Wizard]

- [Βήμα 6]: Ορισμός εάν θα αποθηκευτεί η διαμόρφωση του αρχείου για μελλοντική χρήση καθώς και εάν θα επικολληθεί όλη τη διαδικασία της ανάγνωσης του αρχείου στο παράθυρο του Syntax Editor (βλ. Εικόνα 3.17). Επιλογή [No], [No] και πάτημα του κουμπιού [Finish] (βλ. Εικόνα 3.16).



Εικόνα 3.16: 6ο βήμα του [Text Import Wizard]

Μπορεί να ξαναεκτελεστεί η ίδια διαδικασία χωρίς να χρειαστεί να γίνουν ένα-ένα όλα τα βήματα από την αρχή εάν στο παράθυρο του Syntax Editor επιλεχθεί από τη γραμμή μενού το [Run → All].



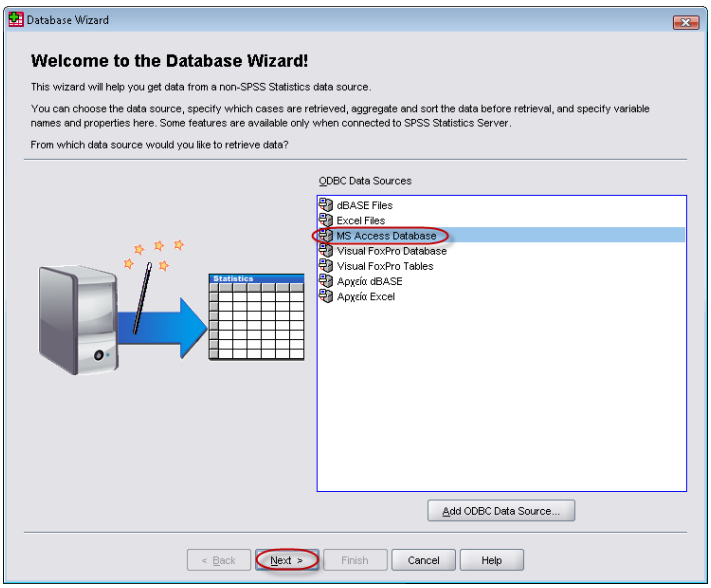
Εικόνα 3.17: Το παράθυρο τουSyntax Editor και με τα 6 βήματα του [Text Import Wizard]

3.3 Άνοιγμα αρχείου από Βάση Δεδομένων

Τέτοιου είδους αρχεία είναι συνήθως βάσεις δεδομένων της Microsoft Access, χωρίς να αποκλείεται και το Microsoft Excel το οποίο μπορεί να ανοιχτεί απευθείας από το [File → Open → Data]. Άλλου είδους βάσεις δεδομένων μπορεί να είναι η Oracle, ο SQL Server, η MySQL, η dBASE κλπ. Στη συνέχεια θα παρουσιαστεί ο τρόπος ανοίγματος και διαβάσματος δεδομένων από μία βάση δεδομένων της Microsoft Access.

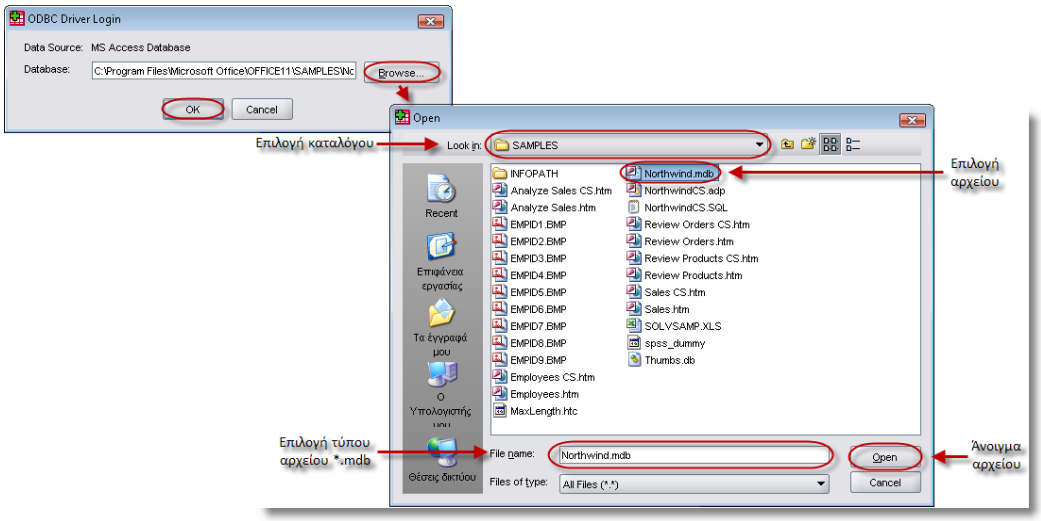
Η διαδικασία έχει ως παρακάτω:

- Επιλογή από τη γραμμή μενού του [File → Open Database → New Query…], όπου και ανοίγει ο οδηγός [Database Wizard], ο οποίος έχει 6 βήματα.
- [Βήμα 1]: Επιλογή από το [ODBC Data Sources] το [MS Access Database] και πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.18).



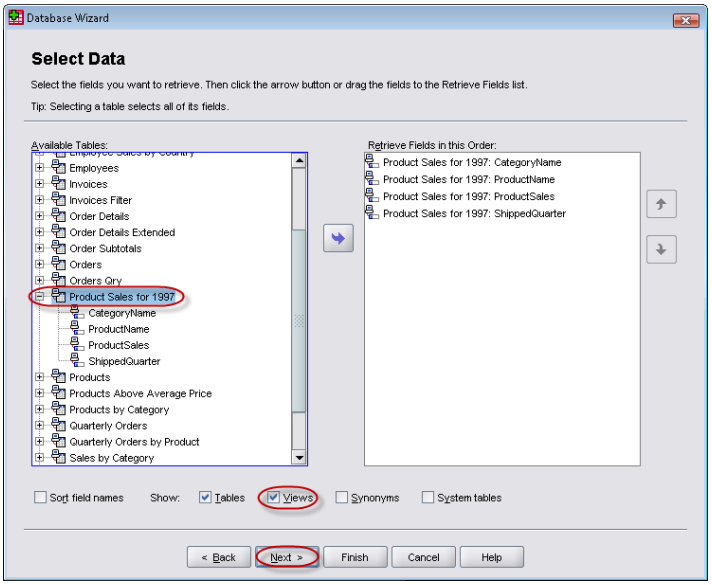
Εικόνα 3.18: 1^{ο} βήμα του [Database Wizard]

- [Βήμα 2]: Στο πλαίσιο διαλόγου [ODBC Driver Login] πάτημα του κουμπιού [Browse…] και στο πλαίσιο διαλόγου που αναδύεται επιλογή του κατάλογου που είναι εγκατεστημένη η Microsoft Access και επιλογή-άνοιγμα του αρχείου [Northwind.mdb] που βρίσκεται στο φάκελο [SAMPLES]. Στη συνέχεια, αφού κλείσει το πλαίσιο διαλόγου [Open] πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.19).



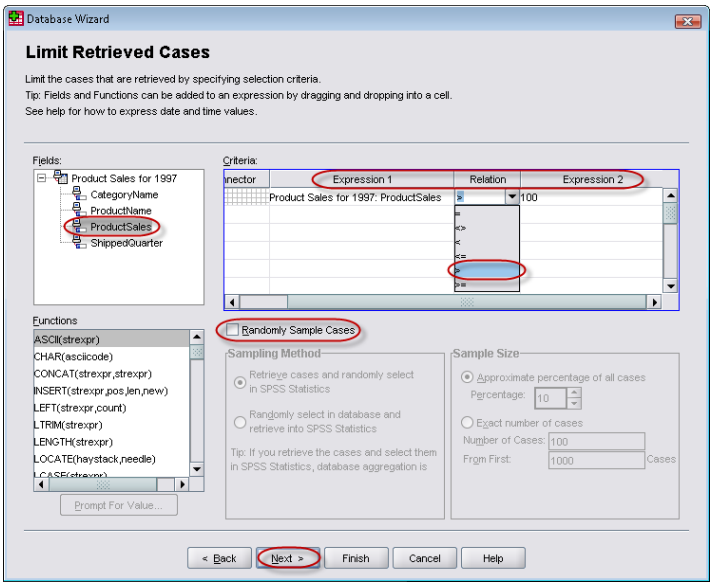
Εικόνα 3.19: 2^{ο} βήμα του [Database Wizard]

- [Βήμα 3]: Σ' αυτό το βήμα δίνεται η δυνατότητα να επιλεχτεί κάποιος ή κάποιοι από τους διαθέσιμους πίνακες ή ένας συνδυασμός αυτών από το [Available Tables]. Για να εμφανιστούν και τα ερωτήματα (queries) που ήδη υπάρχουν στη βάση δεδομένων τσεκάρισμα του [Views]. Τα ονόματα που είναι δεξιά του [⊞] στο [Available Tables] αντιστοιχούν στα ονόματα των πινάκων που είναι διαθέσιμοι στη βάση δεδομένων και κάνοντας κλικ επάνω τους αναπτύσσεται η λίστα με τα πεδία τους. Κάνοντας διπλό κλικ επάνω στο όνομα του πίνακα ή κάνοντας διπλό κλικ στο όνομα κάθε πεδίου να μεταφέρεται στο [Retrieve Fields in this Order] (ή επιλέγοντας τα και πατώντας το [▶]). Στο παράδειγμα τσεκάρουμε το [Views], κάνουμε διπλό κλικ στο [Product Sales for 1997] και πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.20).



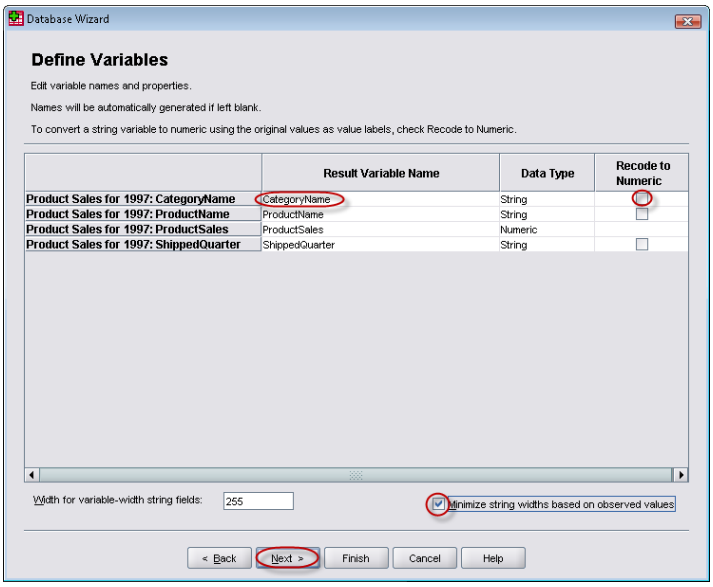
Εικόνα 3.20: 3^{ο} βήμα του [Database Wizard]

- [Βήμα 4]: Σ' αυτό το βήμα δίνεται η δυνατότητα να περιοριστούν οι περιπτώσεις (cases) των δεδομένων που θα εισαχθούν στο SPSS. Για παράδειγμα η εισαγωγή μόνο των περιπτώσεων που το πεδίο [ProductSales] είναι πάνω από 100 το συγκεκριμένο πεδίο το σέρνουμε και το αφήνουμε (drag and drop) στην πρώτη γραμμή του πεδίου [Expression 1], στο [Relation] επιλογή του συμβόλου [>] και στο [Expression 2] πληκτρολόγηση του 100. Στο παράδειγμα δεν μπαίνει κάποιο κριτήριο, πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.21).



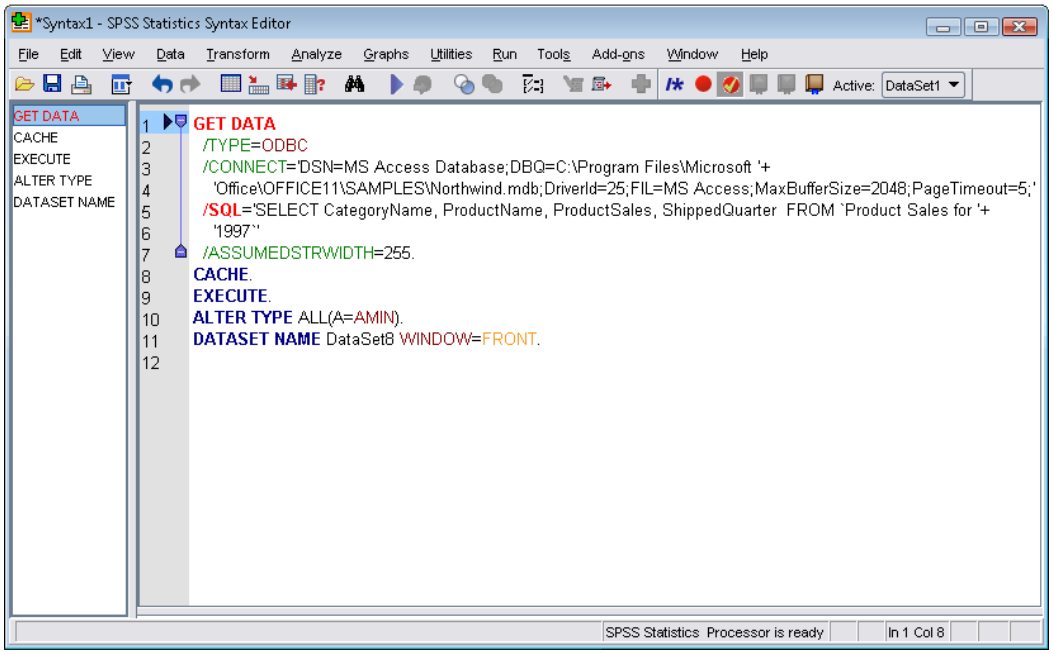
Εικόνα 3.21: 4ο βήμα του [Database Wizard]

- [Βήμα 5]: Σ' αυτό το βήμα δίνεται η δυνατότητα να οριστούν ξανά, εάν χρειάζεται, οι ιδιότητες των μεταβλητών. Κάνοντας διπλό κλικ επάνω στο όνομα της μεταβλητής μπορεί να μετονομαστεί, καθώς και μία μεταβλητή που δεν είναι αλφαριθμητική (string) να επανα-κωδικοποιηθεί σε αριθμητική (numeric), τσεκάροντας το [Recode to Numeric]. Μπορεί, επίσης, να περιοριστεί το πλάτος (width) των αλφαριθμητικών (string) μεταβλητών αυτόματα τσεκάροντας το [Minimize string widths based on observed values], βασισμένο στο μήκος της "πλατύτερης" παρατηρούμενης τιμής της μεταβλητής. Στο παράδειγμα δεν αλλάζουμε τίποτα, τσεκάρισμα του [Minimize string widths based on observed values] και πάτημα του κουμπιού [Next] (βλ. Εικόνα 3.22).



Εικόνα 3.22: 5^{ο} βήμα του [Database Wizard]

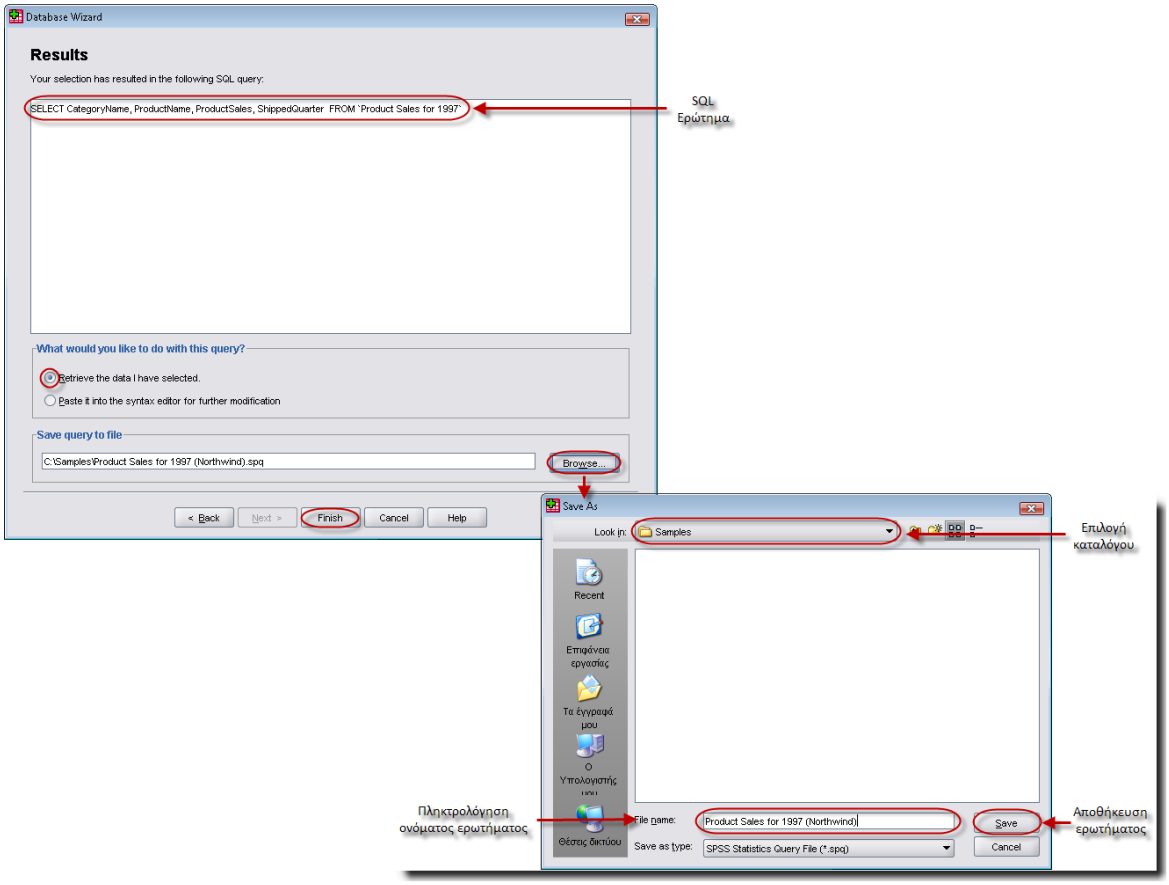
- [Βήμα 6]: Σ' αυτό το βήμα δίνεται η δυνατότητα για:
    - Άμεση εισαγωγή των δεδομένων στο SPSS, βάσει του SQL ερωτήματος, επιλέγοντας το [Retrieve the data I have selected] ή
    - Επικόλληση όλης της διαδικασίας στον Syntax Editor για περεταίρω επεξεργασία επιλέγοντας το [Paste it into the syntax editor for further modification] (βλ. Εικόνα 3.23).



Εικόνα 3.23: Το παράθυρο τουSyntax Editor και με τα 6 βήματα του [Database Wizard]

- Αποθήκευση του ερωτήματος (query), επιλέγοντας [Browse] και κατάλογο-όνομα αποθήκευσης, για να:
  - Εκτελεστεί ξανά εάν χρειαστεί, επιλέγοντας και ανοίγοντας το από το μενού [File → Open Database → Run Query…], ή
  - Επεξεργασία επιλέγοντας και ανοίγοντας το από το μενού [File → Open Database → Edit Query…].

- Στο παράδειγμα επιλέγουμε το [Retrieve the data I have selected] και πάτημα του κουμπιού [Finish] (βλ. Εικόνα 3.24).



Εικόνα 3.24: 6^{ο} βήμα του [Database Wizard]

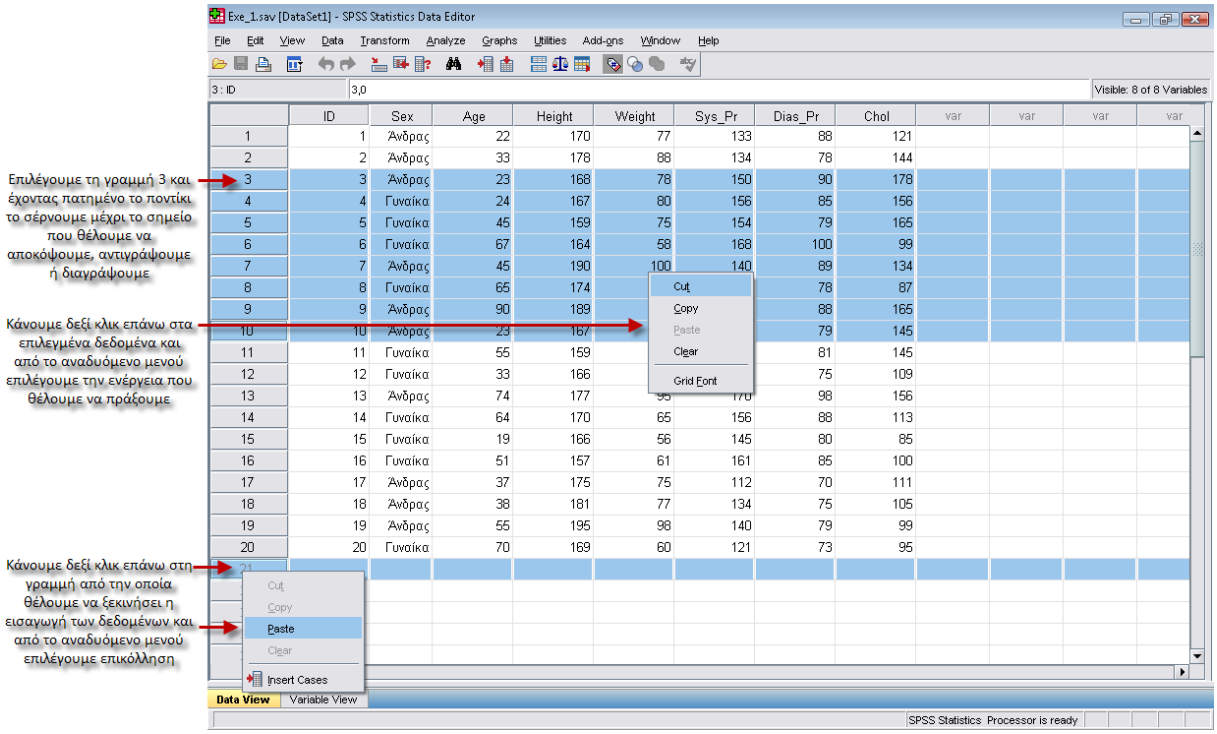
3.4 Μετακίνηση και αντιγραφή δεδομένων

Πολλές φορές κατά τη διάρκεια εισαγωγής των δεδομένων προκύπτει η ανάγκη για μετακίνηση, με αποκοπή και επικόλληση (cut & paste), ή για αντιγραφή, με αντιγραφή και επικόλληση (copy & paste), ορισμένων περιπτώσεων (cases) ή τιμών (values) σε ένα άλλο ή και στο ίδιο αρχείο δεδομένων. Επίσης, μπορεί να απαιτηθεί να αφαιρεθούν (clear) κάποια από τα δεδομένα. Η διαδικασία έχει ως παρακάτω (βλ. Εικόνα 3.25):

- Επιλογή μίας περίπτωσης (κάνοντας κλικ επάνω στον αριθμό της επιθυμητής γραμμής (περίπτωσης)) ή περισσότερων (κρατώντας πατημένο το ποντίκι και σέρνοντας το προς τα κάτω). Όλα τα κελιά της συγκεκριμένης/ων περίπτωσης/εων γίνονται γαλάζια.
- Δεξί κλικ επάνω στην επιλεγμένη περιοχή και στο αναδυόμενο μενού επιλογή της αποκοπής (Cut) ή της αντιγραφής (Copy) ή της διαγραφής (Clear) της περίπτωσης ή των περιπτώσεων.

- Δεξί κλικ στη γραμμή από την οποία θα ξεκινήσει, εάν δεν έχει επιλεχτεί διαγραφή, η επικόλληση της περίπτωσης ή των περιπτώσεων.
- Από το αναδυόμενο μενού επιλογή επικόλληση (paste).

Η διαδικασία είναι αντίστοιχη και για τιμές μεταβλητών δηλαδή την επιλογή μίας περιοχής δεδομένων και όχι μίας ή πολλών περιπτώσεων.



Εικόνα 3.25: Διαδικασία μετακίνησης και αντιγραφής δεδομένων

4ο ΚΕΦΑΛΑΙΟ
Περιγραφική Στατιστική

Εισαγωγή

Για να παρουσιαστούν τα διαθέσιμα δεδομένα με τον καλύτερο δυνατό τρόπο θα πρέπει να συνοψιστούν. Δύο είναι οι βασικές μέθοδοι σύνοψης αυτών. Η μια μέθοδος είναι η παρουσίαση αυτών με πίνακες και γραφικές παραστάσεις και η άλλη είναι η χρησιμοποίηση αριθμητικών περιγραφικών μέτρων που προέρχονται από αυτά.

Οι γραφικές μέθοδοι και οι συνεπτυγμένοι πίνακες είναι εξαιρετικά χρήσιμα στην παρουσίαση δεδομένων και προσφέρουν μία γρήγορη περιγραφή των δεδομένων αλλά έχουν περιορισμούς που μπορούν να αρθούν με τη χρησιμοποίηση αριθμητικών περιγραφικών μέτρων.

Παρουσίαση των τεχνικών που μπορούν να χρησιμοποιηθούν ανάλογα με τον τύπο της μεταβλητής και τη μέθοδο σύνοψης παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 1

Γενικός Οδηγός Επιλογής Γραφικής και Αριθμητικής Μεθόδου

		Μέθοδοι	
		Γραφικές-Πίνακες	Αριθμητικές
Μεταβλητή	Ποιοτική ή διακριτή ποσοτική	Πίνακες κατανομής συχνοτήτων Κυκλικά διαγράμματα Ραβδογράμματα	Ποσοστά (σχετικές συχνότητες)
	Συνεχής ποσοτική	Ιστογράμματα Διαγράμματα Μίσχου-Φύλλου Θηκογράμματα Γραμμογραφήματα Διαγράμματα διασποράς Διαγράμματα Ακιδωτά	Μέτρα θέσης ή κεντρικής τάσης Μέτρα σχετικής θέσης Μέτρα μεταβλητότητας ή διασποράς Μέτρα σχετικής μεταβλητότητας Μέτρα ασυμμετρίας Μέτρα κύρτωσης

Τα Διαγράμματα Μίσχου-Φύλλου και Θηκογράμματα, επειδή συνδυάζουν διάφορα μέτρα σύνοψης δεδομένων, εντάσσονται και στη μέθοδο ανιχνευτικής ανάλυσης (EDA).

4.1 Πίνακες συχνοτήτων (Frequency tables)

Έστω μία διακριτή ποσοτική ή ποιοτική μεταβλητή X με παρατηρήσεις $x_1, x_2, \ldots, x_n$, ενός δείγματος μεγέθους n. Ο φυσικός αριθμός που δείχνει πόσες φορές εμφανίζεται η τιμή x_i της εξεταζόμενης μεταβλητής X στο σύνολο των παρατηρήσεων ονομάζεται **συχνότητα** (frequency) n_i της παρατήρησης.

Αν διαιρέσουμε τη συχνότητα f_i με το μέγεθος n του δείγματος τότε προκύπτει η **σχετική συχνότητα** (relative frequency) f_i της τιμής x_i, δηλαδή:

$$f_i = \frac{n_i}{n}$$

η οποία μπορεί να εκφραστεί επί τοις εκατό, οπότε συμβολίζεται με $f_i\%$, δηλαδή $f_i\% = 100 f_i$.

Για τη σχετική συχνότητα ισχύουν οι ιδιότητες:

- $0 \le f_i \le 1$
- $f_1 + f_2 + \ldots + f_k = 1,\ k \le n$

Στις διακριτές ποσοτικές μεταβλητές, επιπλέον, καθώς και στις διατακτικές ποιοτικές μεταβλητές χρησιμοποιείται η **αθροιστική συχνότητα** (cumulative frequency) N_i και η **αθροιστική σχετική συχνότητα** (cumulative relative frequency) F_i, οι οποίες εκφράζουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i. Η αθροιστική σχετική συχνότητα F_i μπορεί να εκφραστεί επί τοις εκατό, οπότε συμβολίζεται με $F_i\%$, δηλαδή $F_i\% = 100 F_i$.

Ο πίνακας κατανομής συχνοτήτων ή απλά πίνακας συχνοτήτων είναι ένας συνοπτικός πίνακας που περιέχει συγκεντρωμένες σε στήλες τις ποσότητες x_i, n_i, f_i και F_i, $F_i\%$, όπου δύναται, ενός δείγματος.

Για παράδειγμα, σε μία έρευνα ζητείται, μεταξύ άλλων, να σημειωθεί και το μορφωτικό

επίπεδο των ασθενών οπότε, μετά την ανάλυση των δεδομένων, ο ερευνητής καταλήγει στον Πίνακα 2.

Πίνακας 2

Κατανομή συχνοτήτων για το μορφωτικό επίπεδο των ασθενών

i	Μορφωτικό επίπεδο x_i	Συχνότητα n_i	Σχετική συχνότητα f_i	Αθροιστική συχνότητα N_i	Αθροιστική σχετική συχνότητα F_i%
1	Καθόλου σχολείο	5	0,06	5	0,06
2	Δημοτικό	9	0,10	14	0,16
3	Λύκειο	33	0,38	33	0,38
4	Ιδ. Σχολή	19	0,22	66	0,76
5	ΑΕΙ/ΤΕΙ	15	0,17	81	0,93
6	Μεταπτυχιακό	6	0,07	87	1
Σύνολο:		87	1		

4.2 Μέτρα θέσης ή κεντρικής τάσης

Τα μέτρα κεντρικής τάσης δίνουν πληροφορίες για τη θέση του "κέντρου" των παρατηρήσεων. Με αυτά δίνεται η δυνατότητα να εξαχθούν άμεσα συμπεράσματα σχετικά με την συμπεριφορά των υποκείμενων τιμών, χωρίς ιδιαίτερες πολυπλοκότητες.
Τα βασικότερα μέτρα κεντρικής τάσης ή μέτρα θέσης, όπως αλλιώς ονομάζονται, είναι:

4.2.1 Μέση τιμή ή αριθμητικός μέσος (Mean)

Η μέση τιμή ενός συνόλου n παρατηρήσεων αποτελεί το σπουδαιότερο και χρησιμότερο μέτρο της Στατιστικής και ορίζεται ως το άθροισμα των παρατηρήσεων διά του πλήθους των παρατηρήσεων. Η μέση τιμή ενός πληθυσμού συμβολίζεται με μ ενώ του δείγματος με $\bar{x}$. Όταν σε ένα δείγμα μεγέθους n οι παρατηρήσεις μιας μεταβλητής X είναι $x_1, x_2, ..., x_n$, τότε η μέση τους δίνεται από τη σχέση:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Σε μια κατανομή συχνοτήτων, αν $y_1, y_2, ..., y_i$ είναι οι τιμές της μεταβλητής X με συχνότητες $\nu_1, \nu_2, ..., \nu_i$ αντίστοιχα, η μέση τιμή ορίζεται ισοδύναμα από τη σχέση:

$$\overline{x} = \frac{y_1\nu_1 + y_2\nu_2 + ... + y_n\nu_n}{\nu_1 + \nu_2 + ... + \nu_n} = \frac{\sum_{i=1}^{n} y_i\nu_i}{\sum_{i=1}^{n} \nu_i} = \frac{1}{\nu}\sum_{i=1}^{n} y_i\nu_i = \sum_{i=1}^{n} y_i f_i$$

Όπου, f_i οι σχετικές συχνότητες.

4.2.2 Επικρατούσα τιμή (Mode)

Η επικρατούσα τιμή (mode) συμβολίζεται με T_0. Εάν τα δεδομένα δεν είναι ομαδοποιημένα τότε ορίζεται ως η παρατήρηση με τη μεγαλύτερη συχνότητα εμφάνισης. Εάν τα δεδομένα έχουν μία μόνο επικρατούσα τιμή η κατανομή τους ονομάζεται μονοκόρυφη ενώ διαφορετικά εάν έχει δύο ονομάζεται δικόρυφη. Εάν τα δεδομένα είναι ομαδοποιημένα τότε ορίζεται:

$$T_0 = L_{T_0} + \delta\frac{\Delta_1}{\Delta_1 + \Delta_2}$$

όπου, L_{T_0} είναι το κάτω άκρο της επικρατούσας κλάσης, δ είναι το πλάτος της επικρατούσας κλάσης, Δ_1 η διαφορά της συχνότητας της προηγούμενης κλάσης από τη συχνότητα της επικρατούσας κλάσης και Δ_2 η διαφορά της συχνότητας της επόμενης κλάσης από τη συχνότητα της επικρατούσας κλάσης.

4.2.3 Διάμεσος (Median)

Η διάμεσος M ενός δείγματος n παρατηρήσεων χωρίζει τα δεδομένα στη μέση όταν αυτές έχουν διαταχθεί σε αύξουσα σειρά, δηλαδή το πολύ το 50% των παρατηρήσεων είναι

μικρότερες από αυτήν και το πολύ το 50% των παρατηρήσεων είναι μεγαλύτερες από αυτήν χωρίς να επηρεάζεται από ακραίες παρατηρήσεις.

Όταν το n είναι περιττός αριθμός τότε υπολογίζεται:

$$\mathrm{M} = \chi_{\frac{n+1}{2}}$$

Ενώ όταν είναι άρτιος:

$$\mathrm{M} = \frac{1}{2}(\chi_{\frac{n}{2}+1} + \chi_{\frac{n}{2}})$$

4.2.4 Σταθμικός μέσος (Weighted mean)

Εάν οι τιμές $x_1, x_2, \ldots, x_n$ ενός συνόλου δεδομένων έχουν διαφορετική βαρύτητα (έμφαση) και εκφράζεται με τους λεγόμενους συντελεστές στάθμισης (βαρύτητας), $\omega_1, \omega_2, \ldots, \omega_n$, τότε αντί του αριθμητικού μέσου χρησιμοποιούμε τον σταθμισμένο αριθμητικό μέσο ή σταθμικό μέσο (weighted mean). Ο σταθμικός μέσος υπολογίζεται:

$$\overline{x} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_\nu} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

4.3 Μέτρα σχετικής θέσης

Αρκετά συχνά τα μέτρα κεντρικής τάσης των δεδομένων δεν επιτυγχάνουν ικανοποιητική σύνοψη των πληροφοριών τους. Στις περιπτώσεις αυτές χρησιμοποιούμε τα μέτρα που αναφέρονται στη σχετική θέση των δεδομένων μεταξύ τους.

4.3.1 Τεταρτημόρια (Quartiles)

Τα τεταρτημόρια χωρίζουν τα δεδομένα σε 4 ίσα μέρη και συμβολίζονται με Q_1, Q_2 και Q_3. Για το Q_1 (κάτω τεταρτημόριο) έχουμε αριστερά το πολύ 25% των παρατηρήσεων και δεξιά

το πολύ 75%. Όμοια για το Q_3 (άνω τεταρτημόριο) έχουμε αριστερά το πολύ 75% των παρατηρήσεων και δεξιά το πολύ 25% των παρατηρήσεων. Προφανώς το Q_2 συμπίπτει και με τη διάμεσο, δηλαδή $Q_2 = M$. Για τον υπολογισμό τους εργαζόμαστε ως εξής:

- Διατάσσουμε τις παρατηρήσεις σε αύξουσα σειρά μεγέθους.
- Υπολογίζουμε τη διάμεσο. Η τιμή αυτή είναι το Q_2.
- Υπολογίζουμε τη διάμεσο του πρώτου μισού των διατεταγμένων παρατηρήσεων, δηλαδή των παρατηρήσεων που είναι αριστερά του Q_2. Η τιμή αυτή είναι το Q_1.
- Υπολογίζουμε τη διάμεσο του δεύτερου μισού των διατεταγμένων παρατηρήσεων, δηλαδή των παρατηρήσεων που είναι δεξιά του Q_2. Η τιμή αυτή είναι το Q_3.

4.3.2 Τυποποιημένες τιμές ή Z-τιμές (Standardized values)

Όταν θέλουμε να συγκρίνουμε αποδόσεις που έχουν μετρηθεί σε διαφορετικές κλίμακες χρησιμοποιούμε τις Z-τιμές που ορίζονται ως η απόσταση μιας παρατήρησης από το μέση τιμή του δείγματος εκφρασμένη σε μονάδες τυπικής απόκλισης, δηλαδή:

$$Z_i = \frac{x_i - \overline{x}}{S}$$

4.4 Μέτρα διασποράς ή μεταβλητότητας (Measures of variation)

Τα μέτρα διασποράς δίνουν πληροφορίες για τη μεταβλητότητα δηλαδή το "άπλωμα" των παρατηρήσεων σε ένα σύνολο δεδομένων. Με αυτά δίνεται η δυνατότητα να εξαχθούν άμεσα συμπεράσματα σχετικά με την συμπεριφορά των υποκείμενων τιμών, χωρίς ιδιαίτερες πολυπλοκότητες.
Τα βασικότερα μέτρα διασποράς ή μεταβλητότητας, όπως αλλιώς ονομάζονται, είναι:

4.4.1 Εύρος (Range)

Το εύρος R ενός δείγματος ορίζεται ως η διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής αυτού. Δηλαδή:

$$R = x_{\max} - x_{\min}$$

4.4.2 Ενδοτεταρτημοριακό εύρος (Interquartile range)

Υπολογίζεται ως διαφορά μεταξύ του $3^{ου}$ από το $1^{ου}$ τεταρτημόριου, η οποία δεν επηρεάζεται από ακραίες τιμές. Δηλαδή:

$$Q = Q_3 - Q_1$$

4.4.3 Διακύμανση (Variance)

Η διακύμανση S^2 είναι ένα μέτρο διασποράς το οποίο στηρίζεται στην έννοια της απόστασης μιας παρατήρησης από τη μέση τιμή των παρατηρήσεων. Στην περίπτωση δείγματος n παρατηρήσεων με δειγματικό μέσο $\bar{x}$ η δειγματική διακύμανση ορίζεται ως η μέση τιμή των τετραγώνων των αποκλίσεων των n τιμών του δείγματος από το δειγματικό μέσο. Δηλαδή:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Όταν έχουμε πίνακα συχνοτήτων ή ομαδοποιημένα δεδομένα, η διακύμανση ορίζεται από τη σχέση:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 v_i$$

όπου $x_1, x_2, \ldots, x_n$ οι τιμές της μεταβλητής (ή τα κέντρα των κλάσεων) με αντίστοιχες συχνότητες $v_1, v_2, \ldots, v_n$.

Εάν το n αντικατασταθεί με το (n-1) τότε το S^2 είναι η αμερόληπτη εκτιμήτρια της διακύμανσης.

4.4.4 Τυπική απόκλιση (Standard deviation)

Η διακύμανση είναι μια πολύ σημαντική παράμετρος διασποράς αλλά δεν εκφράζεται με τις μονάδες με τις οποίες εκφράζονται οι παρατηρήσεις. Για παράδειγμα, αν οι παρατηρήσεις εκφράζονται σε cm, η διακύμανση εκφράζεται σε cm^2. Αν όμως πάρουμε τη θετική τετραγωνική ρίζα της διακύμανσης, θα έχουμε ένα μέτρο διασποράς που θα εκφράζεται με την ίδια μονάδα μέτρησης του χαρακτηριστικού. Η ποσότητα αυτή λέγεται τυπική απόκλιση S και υπολογίζεται από τη σχέση:

$$S = \sqrt{S^2}$$

4.5 Μέτρα σχετικής μεταβλητότητας

4.5.1 Συντελεστής μεταβλητότητας (coefficient of variation)

Ο συντελεστής μεταβλητότητας CV είναι ένα μέγεθος που είναι «αδιάστατο» και υπολογίζεται ως ποσοστό επί τοις εκατό. Δηλαδή δεν λαμβάνει υπόψη το ύψος του μέσου όρου και για αυτό το λόγο δύναται να συγκρίνει κατανομές που είναι μεταξύ τους διαφορετικές, δηλαδή ομάδων τιμών, που είτε εκφράζονται σε διαφορετικές μονάδες μέτρησης είτε εκφράζονται στην ίδια μονάδα μέτρησης, αλλά έχουν σημαντικά διαφορετικές μέσες τιμές. Υπολογίζεται από τη σχέση:

$$CV = \frac{S}{\bar{x}} \cdot 100\%$$

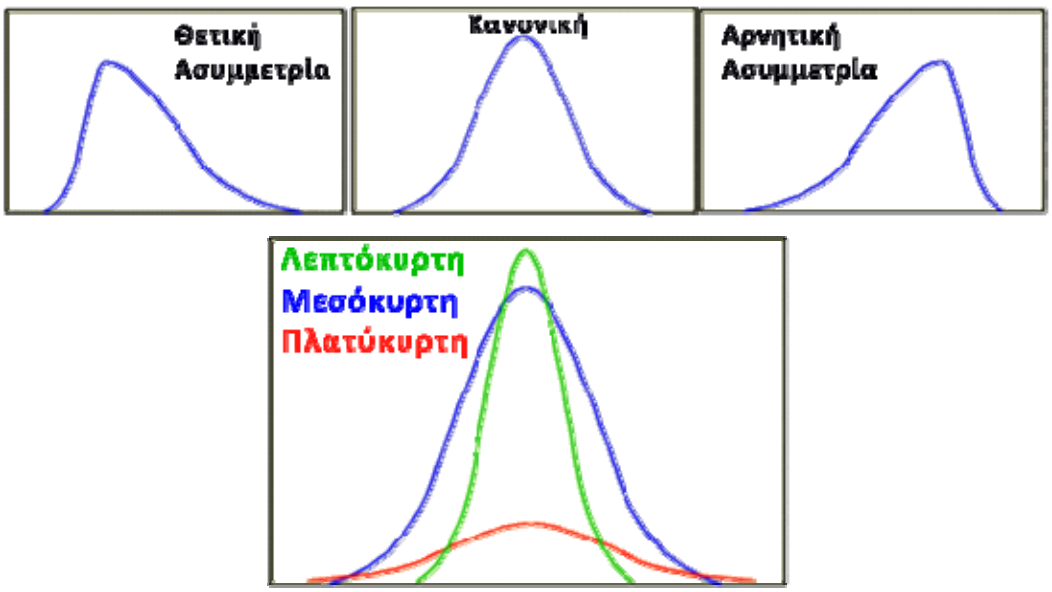
4.6 Μέτρα ασυμμετρίας και κύρτωσης

Εκτός από τα μέτρα κεντρικής τάσης και διασποράς μιας κατανομής, τα οποία δίνουν μία πρώτη εικόνα της μορφής της, πολύ σημαντικά για τη βελτίωση αυτής της εικόνας είναι και τα μέτρα ασυμμετρίας και κύρτωσης. Δηλαδή κατά πόσο και προς ποια κατεύθυνση

αποκλίνει η κατανομή των δεδομένων σε σχέση με την κανονική κατανομή καθώς και το πόσο πεπλατυσμένη (οξύτητα της κορυφής της) αυτή είναι.

Η ασυμμετρία μπορεί να είναι θετική ή αρνητική. Θετικά ασυμμετρική είναι μια κατανομή όταν παρουσιάζει εξόγκωση προς τα αριστερά και επιμήκυνση του δεξιού της άκρου. Η μεγάλη συγκέντρωση των παρατηρήσεων βρίσκεται στις μικρές τιμές της μεταβλητής, δηλαδή έχει λίγες μεγάλες τιμές, επομένως η μέση τιμή βρίσκεται δεξιά της επικρατούσας τιμής που είναι η κορυφή της κατανομής. Αντίθετα, αρνητικά ασυμμετρική είναι μία κατανομή όταν παρουσιάζει εξόγκωση προς τα δεξιά και επιμήκυνση του αριστερού της άκρου. Η μεγάλη συγκέντρωση των παρατηρήσεων βρίσκεται στις μεγάλες τιμές της μεταβλητής, δηλαδή έχει λίγες μικρές τιμές, επομένως η μέση τιμή βρίσκεται αριστερά της επικρατούσας τιμής που είναι η κορυφή της κατανομής. Η διάμεσος είναι πάντα μεταξύ της μέσης τιμής και της επικρατούσας. Εάν υπάρχει ουσιώδης διαφορά μεταξύ αυτών των τριών μέτρων τότε έχουμε ένδειξη ασυμμετρίας. Εάν η επικρατούσα, η διάμεσος και η μέση τιμή είναι ίσες τότε έχουμε πλήρη συμμετρία.

Δύο μονοκόρυφες κατανομές μπορεί να έχουν τον ίδιο αριθμητικό μέσο, την ίδια τυπική απόκλιση, να είναι συμμετρικές αλλά παρόλα αυτά να διαφέρουν ως προς την οξύτητα της κορυφής τους. Με βάση το χαρακτηριστικό αυτό μια κατανομή μπορεί να θεωρηθεί: Λεπτόκυρτη, Πλατύκυρτη, Μεσόκυρτη.



4.6.1 Συντελεστής ασυμμετρίας του Pearson (Pearson's skewness coefficient)

Ο K. Pearson πρότεινε ως συντελεστή ασυμμετρίας τη σχέση:

$$S_p = \frac{\overline{x} - T_0}{S}$$

το οποίο είναι ανεξάρτητο από τη μονάδα μέτρησης λόγω της διαίρεσης της διαφοράς της επικρατούσας τιμής από τη μέση με τη την τυπική απόκλιση.

Όταν δεν είναι γνωστή η επικρατούσα τιμή τότε μπορούμε να υπολογίσουμε τη κύρτωση από τη σχέση:

$$S_M = \frac{3(\overline{x} - M)}{S}$$

και στις δύο συναρτήσεις ισχύει ότι εάν:

$$S_M \text{ ή } S_P = \begin{cases} > 0, \text{ θετική ασυμμετρία} \\ = 0, \text{ συμμετρία} \\ < 0, \text{ αρνητική ασυμμετρία} \end{cases}$$

4.6.2 Τυποποιημένος συντελεστής ασυμμετρίας (standardized coefficient of skewness)

Ένα ακόμη μέτρο ελέγχου ασυμμετρίας της κατανομής n παρατηρήσεων είναι ο τυποποιημένος συντελεστής ασυμμετρίας β_3 του Pearson και υπολογίζεται από τη σχέση:

$$\beta_3 = \frac{m_3}{S^3}$$

όπου:

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

και ισχύει ότι εάν:

$$\beta_3 = \begin{cases} > 0, \text{ θετική ασυμμετρία} \\ = 0, \text{ συμμετρία} \\ < 0, \text{ αρνητική ασυμμετρία} \end{cases}$$

4.6.3 Συντελεστής κύρτωσης του Pearson (Pearson's kurtosis coefficient)

Για τον έλεγχο της οξύτητας της κορυφής της κατανομής των δεδομένων η παρατηρήσεων χρησιμοποιείται ο συντελεστής κύρτωσης β_4 του Pearson και υπολογίζεται από τη σχέση:

$$\beta_4 = \frac{m_4}{S^4} - 3$$

όπου:

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$$

και ισχύει ότι εάν:

$$\beta_4 = \begin{cases} > 0, \text{ λεπτόκυρτη} \\ = 0, \text{ μεσόκυρτη} \\ < 0, \text{ πλατύκυρτη} \end{cases}$$

4.7 Γραφικές Μέθοδοι (Γραφικές Παραστάσεις ή Διαγράμματα)

Γράφημα ονομάζεται η γραφική αναπαράσταση μίας ή περισσοτέρων μεταβλητών με απώτερο σκοπό την οπτική κατανόηση της σχέσης ανάμεσα σε δύο ή περισσότερες μεταβλητές καθώς και το σχήμα της κατανομής των εμπλεκόμενων συνεχών μεταβλητών.
Παρόλο που οι γραφικές παραστάσεις σε σχέση με τους πίνακες παρέχουν πιο σαφή και άμεση εικόνα του χαρακτηριστικού, είναι πολύ πιο ενδιαφέρουσες και ελκυστικές δεν προσφέρουν περισσότερη πληροφορία από εκείνη που περιέχεται στους αντίστοιχους πίνακες συχνοτήτων.
Υπάρχουν διάφοροι τρόποι γραφικής παρουσίασης, ανάλογα με το είδος των δεδομένων που έχουμε, αλλά κύριος στόχος των γραφικών μεθόδων παραμένει η άντληση όσο το δυνατό περισσότερων πληροφοριών από τα δεδομένα και όχι η ερμηνεία αυτών.
Για τη σωστή γραφική αναπαράσταση των δεδομένων πρέπει τα διαγράμματα, όπως και οι στατιστικοί πίνακες, να συνοδεύονται από:

- Τον τίτλο.
- Την κλίμακα με τις τιμές των μεγεθών που απεικονίζονται.
- Το υπόμνημα που επεξηγεί συνήθως τις τιμές της μεταβλητής.
- Την πηγή των δεδομένων.

Πίνακας 3

Ειδικός Οδηγός Επιλογής Γραφικής Μεθόδου

Τύπος γραφήματος	Τύπος μεταβλητών	Διάγραμμα
Μιας διάστασης (μεταβλητής)	Ποιοτική	Ραβδόγραμμα (Bar-chart) Κυκλικό ή Πίτας (Pie-chart)
	Ποσοτική	Ιστόγραμμα (Histogram) Διάγραμμα μίσχου-φύλλου (Steam and Leaf plot) Διάγραμμα πλαισίου και απολήξεων (Box-plot)
Δύο διαστάσεων (μεταβλητών)	Δύο ποσοτικές	Διάγραμμα σημείων (Scatter-plot)
	Δύο ποιοτικές	Ραβδόγραμμα (Bar -plot)
	Μια ποσοτική και μια ποιοτική	Διάγραμμα πλαισίου και απολήξεων (Box-plot) Διάγραμμα σφαλμάτων (Error bars)
Πολλών διαστάσεων (Πολυμεταβλητά)	Ποσοτικές ή Διατάξιμες	Πίνακας διαγραμμάτων σημείων (Scatter-plot matrix)

4.7.1 Θηκόγραμμα ή Διάγραμμα πλαισίου και απολήξεων (Box plot ή Box and whisker plot)

Το 1977 προτάθηκε από τον αμερικανό στατιστικό John Tukey μία γραφική μέθοδο παρουσίασης των δεδομένων που προσδιορίζει πέντε μέτρα: της μέγιστης και ελάχιστης τιμής (άρα και το εύρους), του πρώτου και τρίτου τεταρτημορίου (άρα και του ενδοτεταρτημοριακού εύρους) καθώς και της διαμέσου γι αυτό και ονομάζεται σύνοψη των πέντε αριθμών.

Ο τρόπος κατασκευής του έχει ως παρακάτω:

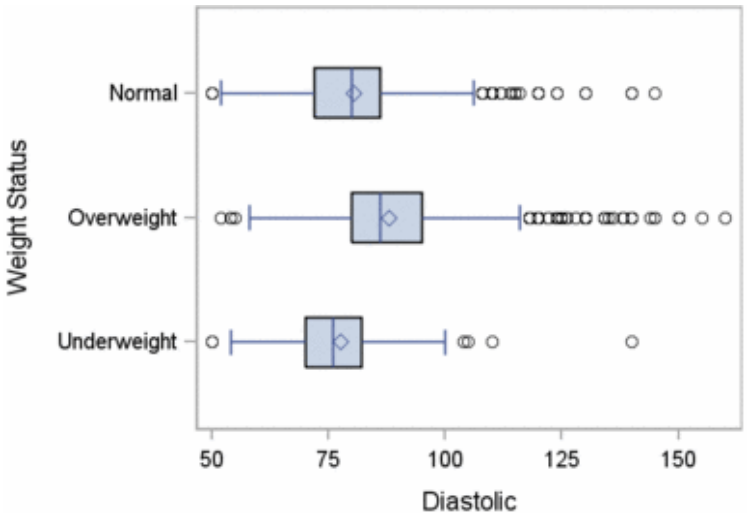
- Υπολογισμός του πρώτου (Q_1) και τρίτου (Q_3) τεταρτημορίου καθώς και της διαμέσου ($T_0=Q_2$).
- Υπολογισμός του ενδοτεταρτημοριακού εύρους ($Q=Q_3-Q_1$) με την αφαίρεση του και τρίτου (Q_3) από το πρώτο (Q_1) τεταρτημόριο.
- Δημιουργία ενός πλαισίου με άκρα το πρώτο (Q_1) και τρίτο (Q_3) τεταρτημόριο.
- Κατασκευή της διαμέσου εντός του πλαισίου με μία διαχωριστική γραμμή.

- Οποιαδήποτε παρατήρηση βρίσκεται περισσότερο από 1,5 Q χαμηλότερα από το Q_1 ή 1,5 Q υψηλότερα από το Q_3 θεωρείται ακραία τιμή (outlier). Εντοπίζουμε τη χαμηλότερη τιμή που δεν είναι ακραία και την ενώνουμε με το πλαίσιο με μία γραμμή η οποία ονομάζεται απόληξη. Μαρκάρουμε τη θέση αυτής της τιμής με μία μικρή κάθετη γραμμή, κάθετη στην απόληξη. Ομοίως, ενώνουμε με το πλαίσιο την υψηλότερη τιμή που δεν είναι ακραία με μία απόληξη (και μαρκάρουμε τη θέση αυτής της τιμής με μία μικρή κάθετη γραμμή, κάθετη στην απόληξη).

- Το μαρκάρισμα των ακραίων τιμών πραγματοποιείται με μικρούς κύκλους και τελείες. "Εξαιρετικά" ακραίες τιμές δηλαδή, αυτές που βρίσκονται περισσότερο από 3 Q προς τα αριστερά ή δεξιά από το Q_1 ή το Q_3 αντίστοιχα, σημειώνονται με μία τελεία. "Ήπιες" ακραίες τιμές δηλαδή, αυτές που βρίσκονται περισσότερο από 1,5 Q προς τα αριστερά ή δεξιά από το Q_1 ή το Q_3, αλλά δεν είναι "εξαιρετικά" ακραίες, σημειώνονται με μικρό κύκλο (πολλές φορές οι "εξαιρετικά" και οι "Ήπιες" ακραίες τιμές δεν διαχωρίζονται και σημειώνονται όλες με μικρό κύκλο).

- Το διάγραμμα μπορεί να παρουσιαστεί και κάθετα κατασκευάζοντας το με το με τον ίδιο τρόπο εάν το περιστρέψουμε κατά 90^o.

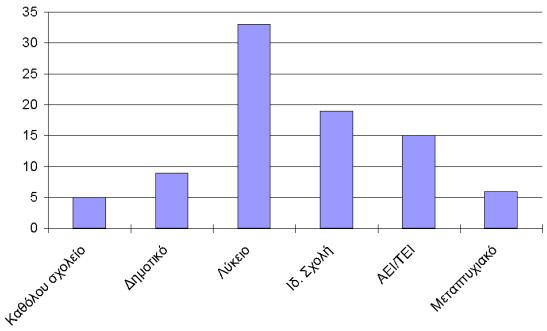
Στο παρακάτω διάγραμμα παρουσιάζεται το θηκόγραμμα με τη “διαστολική πίεση 87 ασθενών ανάλογα με την κατάσταση του σωματικού τους βάρους”.



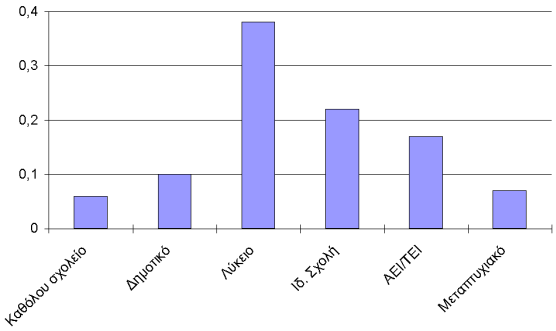
Διάγραμμα 1: Διαστολική πίεση 87 ασθενών ανάλογα με την κατάσταση του σωματικού τους βάρους.

4.7.2 Ραβδόγραμμα (Bar-chart)

Το ραβδόγραμμα (bar-chart) χρησιμοποιείται για τη γραφική απεικόνιση των τιμών μιας ποιοτικής μεταβλητής. Το ραβδόγραμμα αποτελείται από ορθογώνιες στήλες όπου οι βάσεις τους βρίσκονται πάνω στον οριζόντιο ή τον κατακόρυφο άξονα και κάθε μία αντιστοιχεί σε μία τιμή της μεταβλητής Χ. Το ύψος κάθε ορθογώνιας στήλης είναι ίσο με την αντίστοιχη συχνότητα (ραβδόγραμμα συχνοτήτων) ή σχετική συχνότητα (ραβδόγραμμα σχετικών συχνοτήτων) ενώ η απόσταση μεταξύ των στηλών και το μήκος των βάσεών τους καθορίζονται αυθαίρετα. Στον πίνακα 2 έχουμε την κατανομή συχνοτήτων της μεταβλητής Χ: "μορφωτικό επίπεδο ασθενών" και στα διαγράμματα 2(α), (β), που ακολουθούν, τα αντίστοιχα ραβδογράμματα συχνοτήτων και σχετικών συχνοτήτων.



Διάγραμμα 2(α): Ραβδόγραμμα συχνοτήτων του μορφωτικού επιπέδου 87 ασθενών.



Διάγραμμα 2(β): Ραβδόγραμμα σχετικών συχνοτήτων του μορφωτικού επιπέδου 87 ασθενών.

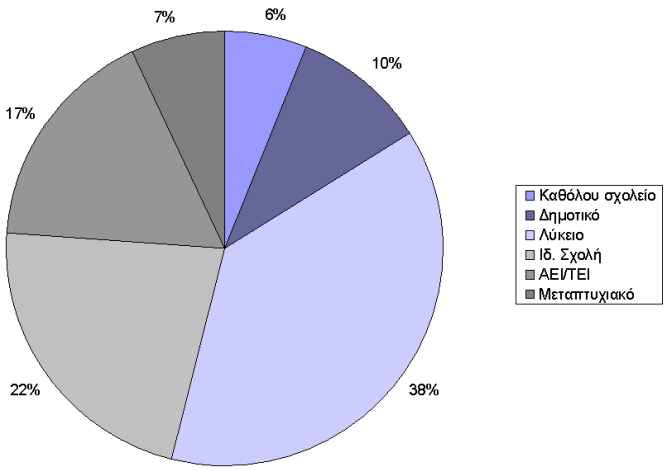
4.7.3 Κυκλικό ή Πίτας (Pie-chart)

Το κυκλικό διάγραμμα (pie-chart) χρησιμοποιείται για τη γραφική παράσταση τόσο των ποιοτικών όσο και των ποσοτικών δεδομένων, όταν οι διαφορετικές τιμές της μεταβλητής είναι σχετικά λίγες. Το κυκλικό διάγραμμα είναι ένας κύκλος υποδιαιρεμένος σε κυκλικούς τομείς και κάθε ένας αντιπροσωπεύει μία κατηγορία της εξεταζόμενης μεταβλητής. Τα εμβαδά (ή ισοδύναμα τα τόξα) των υποδιαιρέσεων είναι ανάλογα προς τις αντίστοιχες

συχνότητες n_i ή τις σχετικές συχνότητες f_i των τιμών x_i της μεταβλητής. Το μήκος α_i του τόξου στο κυκλικό διάγραμμα συχνοτήτων είναι:

$$\alpha_i = \frac{n_i}{n} * 360^o = f_i * 360^o \text{, όπου } i = 1,2,...,\kappa .$$

Στο διάγραμμα 3 παριστάνεται το κυκλικό διάγραμμα σχετικών συχνοτήτων του "μορφωτικού επιπέδου ασθενών" για τα δεδομένα του πίνακα 2.



Διάγραμμα 3: Κυκλικό διάγραμμα σχετικών συχνοτήτων του μορφωτικού επιπέδου 87 ασθενών.

4.7.4 Ιστόγραμμα (Histogram)

Το ιστόγραμμα (histogram) χρησιμοποιείται για τη γραφική παράσταση ποσοτικών δεδομένων όταν οι διαφορετικές τιμές της μεταβλητής έχουν ομαδοποιηθεί σε k κλάσεις (συνήθως ίδιου πλάτους). Το ιστόγραμμα αποτελείται από ορθογώνιες στήλες όπου οι βάσεις τους βρίσκονται πάνω στον οριζόντιο ή τον κατακόρυφο άξονα και κάθε μία αντιστοιχεί σε μία κλάση της μεταβλητής Χ. Το ύψος κάθε ορθογώνιας στήλης είναι ίσο με την αντίστοιχη συχνότητα (ιστόγραμμα συχνοτήτων) ή σχετική συχνότητα (ιστόγραμμα σχετικών συχνοτήτων). Η απόσταση μεταξύ των στηλών είναι μηδενική καθώς οι κλάσεις είναι η μία συνέχεια της προηγούμενης και το μήκος των βάσεών τους (δηλαδή των κλάσεων) καθορίζεται αυθαίρετα ή με τον τρόπο της παραγράφου 2.7.

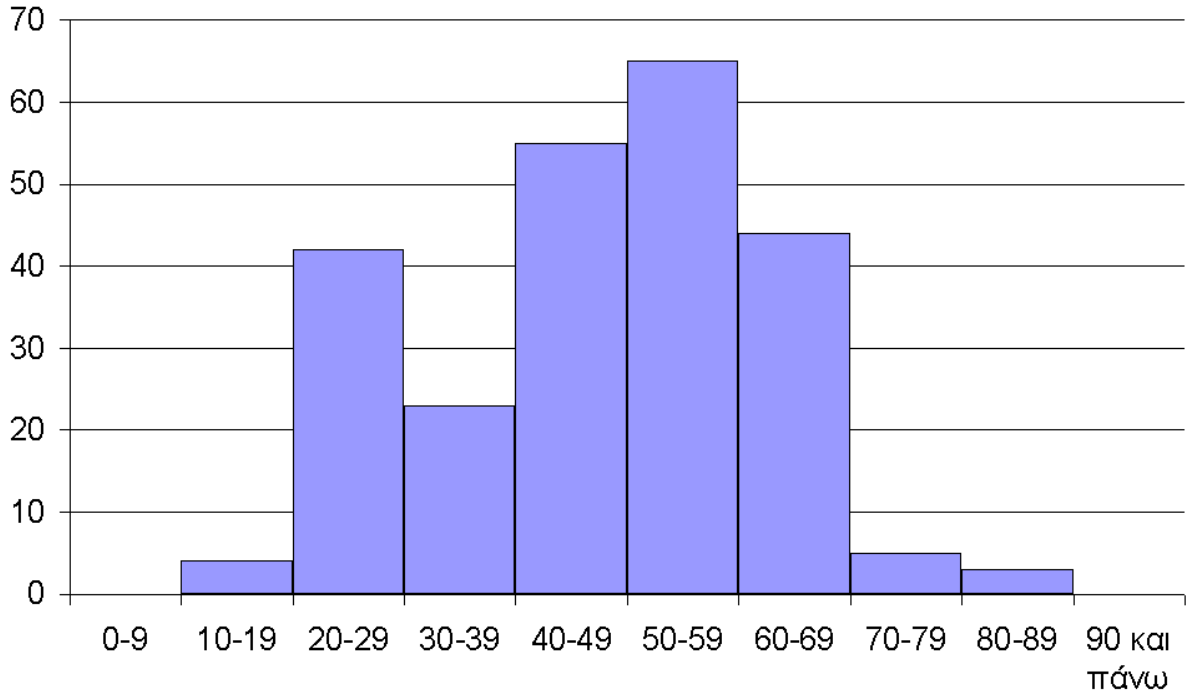
Στην πραγματικότητα η συχνότητα που έχει σχέση με κάθε διάστημα σε ένα ιστόγραμμα αντιπροσωπεύεται όχι από το ύψος της στήλης αλλά από την επιφάνειά της. Έτσι, όλο το εμβαδό-επιφάνεια του ιστογράμματος έχει άθροισμα 100% ή 1 δηλαδή το ποσοστό της συχνότητας ή της σχετικής συχνότητας που αντιστοιχεί σε αυτό το διάστημα (κλάση) ισούται με το αντίστοιχο ποσοστό επί της συνολικής επιφάνειας του ιστογράμματος. Επομένως, κατά την κατασκευή του ιστογράμματος θέλει ιδιαίτερη προσοχή όταν τα διαστήματα δεν είναι ίσα διότι η επιφάνεια κάθε στήλης είναι αυτή που αντιπροσωπεύει τη σχετική αναλογία των παρατηρήσεων σε ένα διάστημα οπότε και το ύψος θα πρέπει να αλλάζει ανάλογα με το πλάτος έτσι ώστε η επιφάνεια κάθε στήλης να παραμένει στη σωστή αναλογία.

Στον πίνακα 4 έχουμε την κατανομή συχνοτήτων της μεταβλητής Χ: "ηλικία ασθενών" και στο διάγραμμα 4, που ακολουθεί, το αντίστοιχο ιστόγραμμα συχνοτήτων.

Πίνακας 4

Κατανομή συχνοτήτων για την ηλικία 241 ασθενών

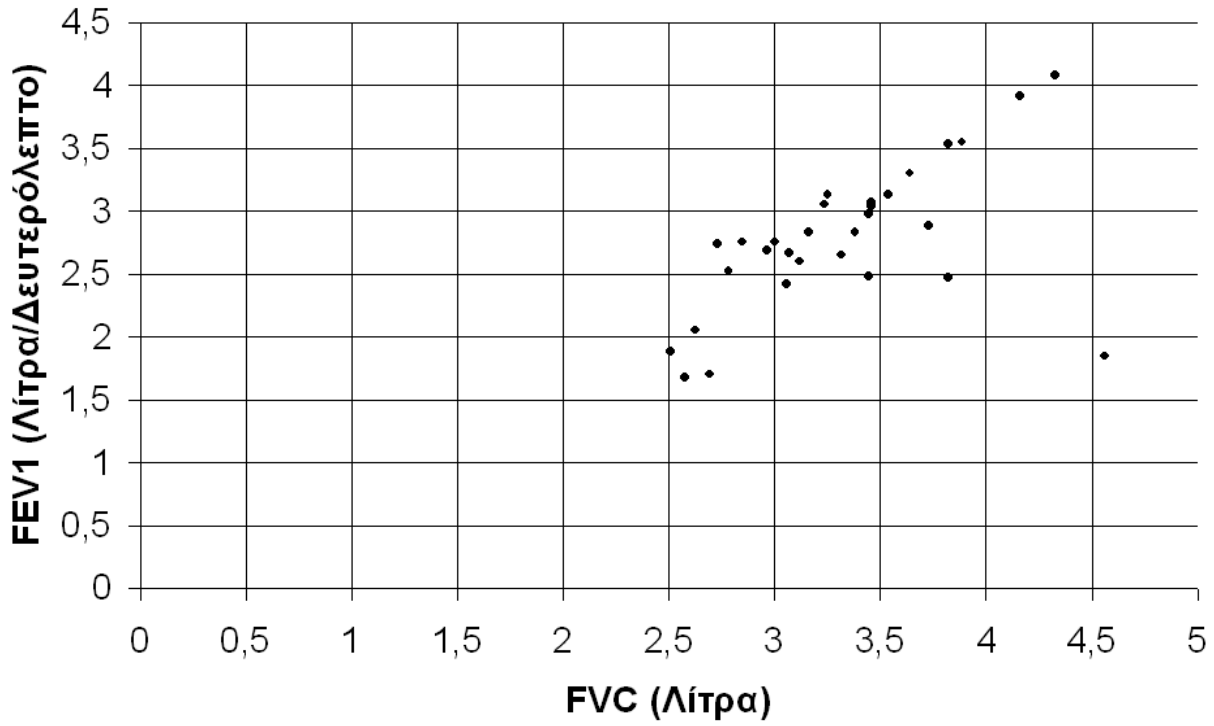
i	Κλάσεις	Συχνότητα n_i	Σχετική συχνότητα f_i
1	0-9	0	0,000
2	10-19	4	0,017
3	20-29	42	0,174
4	30-39	23	0,095
5	40-49	55	0,228
6	50-59	65	0,270
7	60-69	44	0,183
8	70-79	5	0,021
9	80 -89	3	0,012
10	90 και πάνω	0	0
Σύνολο:		241	1



Διάγραμμα 4: Ιστόγραμμα συχνοτήτων 241 ασθενών κατά ηλικιακή ομάδα.

4.7.5 Διάγραμμα σημείων ή διασποράς (Scatter-plot)

Το διάγραμμα σημείων (Scatter-plot) χρησιμοποιείται για τη γραφική παράσταση δύο ποσοτικών μεταβλητών. Αποτελείται από σημεία (κουκίδες) σε ένα σύστημα δύο αξόνων (δισδιάστατο) όπου κάθε ένα σημείο αντιπροσωπεύει ένα ζεύγος τιμών. Η εξαρτημένη (απαντητική ή μετρήσιμη) μεταβλητή ορίζεται στον κάθετο άξονα και η ανεξάρτητη (ελέγχου ή επεξηγηματική) στον οριζόντιο. Το διάγραμμα σημείων δείχνει τη σχέση που υπάρχει μεταξύ των δύο μεταβλητών, πιο συγκεκριμένα: α. την ένταση, β. το σχήμα, γ. την κατεύθυνση και δ. την παρουσία ακραίων τιμών. Στο διάγραμμα 5, που ακολουθεί, παρουσιάζεται το διάγραμμα σημείων των "δεικτών πνευμονικής λειτουργίας FVC και FEV1" 30 παιδιών μίας ακαδημίας ποδοσφαίρου [FVC (forced vital capacity) = βίαιη ζωτική χωρητικότητα - ο όγκος του αέρα που εκπνέεται μετά από πλήρη εισπνοή, FEV1 (Forced Expiratory Volume in 1 Second) = – ο όγκος αέρα που εκπνέεται μετά από μέγιστη προσπάθεια μέσα στο 1ο δευτερόλεπτο].



Διάγραμμα 5: Διάγραμμα σημείων των δεικτών πνευμονικής λειτουργίας FVC και FEV1 30 παιδιών μίας ακαδημίας ποδοσφαίρου.

4.7.6 Διάγραμμα μίσχου-φύλλου (Steam and Leaf plot)

Το διάγραμμα μίσχου-φύλλου (Steam and Leaf plot) εξελίχθηκε από τον Arthur Bowley στις αρχές του προηγούμενου αιώνα και χρησιμοποιείται για τη γραφική παράσταση ποσοτικών δεδομένων όταν οι διαφορετικές τιμές της μεταβλητής (παρατηρήσεις) δεν είναι πολλές αλλά είναι το πολύ 150 περίπου. Το διάγραμμα αυτό δίνει την δυνατότητα ανασύστασης και ανάκλησης των μετρήσεων των αρχικών δεδομένων του δείγματος με ακρίβεια, χωρίς να οδηγεί σε απώλεια πληροφοριών, πράγμα το οποίο δεν επιτυγχάνεται με το ιστόγραμμα ή τους πίνακες συχνοτήτων. Ο τρόπος κατασκευής του έχει ως παρακάτω:

- Ταξινόμηση των παρατηρήσεων κατά αύξουσα σειρά.
- Φτιάχνουμε δύο στήλες χωρισμένες μεταξύ τους με μία κάθετη γραμμή, την αριστερή που είναι ο μίσχος και η δεξιά που είναι το φύλλο.

- Καταγραφή όλων των αριθμών όπου συνήθως ως φύλλο κάθε στοιχείου λαμβάνεται το τελευταίο ή τα δύο τελευταία ψηφία της τιμής της παρατήρησης και ως κορμός το πρώτο ή τα εναπομείναντα πρώτα ψηφία. π.χ 134 – 1|34 ή 13|4. Εάν υπάρχουν δεκαδικά στρογγυλοποιούνται.

Από το διάγραμμα μίσχου-φύλλου μπορούμε εύκολα να δούμε τη συγκέντρωση των παρατηρήσεων (συχνότητες), τη μορφή της κατανομής, την εμφάνιση τυχόν ακραίων παρατηρήσεων, την επισήμανση της απουσίας συγκεκριμένων τιμών. Στο διάγραμμα 6, που ακολουθεί, παρουσιάζεται το διάγραμμα μίσχου-φύλλου της "Συστολικής πίεσης" 30 παιδιών μίας ακαδημίας ποδοσφαίρου.

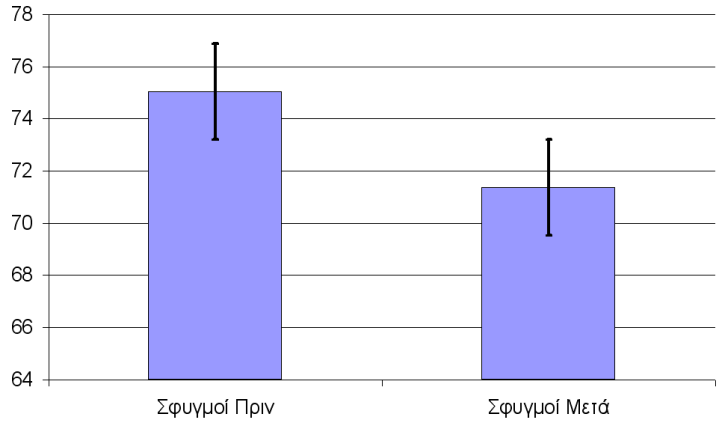
Μίσχος	**Φύλλο**
9	02
10	00048
11	000002466668
12	00000
13	008
14	00
15	
16	
17	0

Διάγραμμα 6: Διάγραμμα μίσχου-φύλλου της "Συστολικής πίεσης" 30 παιδιών μίας ακαδημίας ποδοσφαίρου.

4.7.7 Διάγραμμα σφαλμάτων (Error bars)

Το διάγραμμα σφαλμάτων (Error bars) χρησιμοποιείται για τη γραφική παράσταση μίας ποσοτικής μεταβλητής στα διάφορα επίπεδα μίας άλλης ποιοτικής. Αναπαριστά το 100(1-α)% διάστημα εμπιστοσύνης γύρο από το μέσο (συνήθως α=0,05), αλλά εναλλακτικά μπορεί να χρησιμοποιηθεί και για την τυπική απόκλιση του μέσου (δηλαδή του τυπικού σφάλματος). Το σχήμα τους είναι μια ράβδος κατανεμημένη ισομερώς γύρο από το τη μέση τιμή των παρατηρήσεων και η τυπική τους χρήση είναι για την οπτική απεικόνιση του σφάλματος ή κατάλοιπου (error=η διαφορά της θεωρητικής ορθής τιμής από την μετρηθείσα) ή της αβεβαιότητας των μετρήσεων με λίγα λόγια της διασποράς των μετρήσεων γύρο από τη μέση τιμή. Στο διάγραμμα 7, που ακολουθεί, παρουσιάζεται το

διάγραμμα σφαλμάτων της μέσης τιμής των “σφυγμών καρδιάς πριν και μετά κάποιων τεχνικών για βελτίωση της αναπνοής” 30 παιδιών μίας ακαδημίας ποδοσφαίρου.



Διάγραμμα 7: Διάγραμμα σφαλμάτων της μέσης τιμής των “σφυγμών καρδιάς πριν και μετά κάποιων τεχνικών για βελτίωση της αναπνοής” 30 παιδιών μίας ακαδημίας ποδοσφαίρου.

4.8 Μέθοδος ομαδοποίησης ποσοτικών (συνεχών) δεδομένων

Όταν η μεταβλητή είναι συνεχής και το αντίστοιχο πλήθος των τιμών της αρκετά μεγάλο η κατασκευή του πίνακα συχνοτήτων καθώς και των αντίστοιχων διαγραμμάτων είναι δύσκολη. Οπότε, σωστό είναι να κατασκευαστούν ομάδες-κλάσεις των τιμών της μεταβλητής. Έτσι, θα πρέπει να απαντηθούν δύο ερωτήματα: ποιος θα είναι ο αριθμός των κλάσεων k και ποιο το μέγεθος w κάθε κλάσης. Η διαδικασία έχει ως παρακάτω:

- Ο αριθμός των κλάσεων k μπορεί να υπολογιστεί από τον κανόνα του Strunges, δηλαδή:

$$k = 1 + 3{,}322 * \log_{10} n$$

όπου, πάντα στρογγυλοποιείται το k στον αμέσως επόμενο ακέραιο.

- Για την κατασκευή κλάσεων ίσου πλάτους w διαιρείται το εύρος R με τον αριθμό των κλάσεων k, δηλαδή:

$$w = \frac{R}{k}$$

όπου, πάντα στρογγυλοποιείται το w στον αμέσως επόμενο ακέραιο.

Στη συνέχεια ξεκινώντας από τη χαμηλότερη τιμή της μεταβλητής και προσθέτοντας το πλάτος w δημιουργούνται k κλάσεις. Υπάρχουν κάποιες αρχές που πρέπει να έχουμε υπόψη:

- Κάθε τιμή ανήκει μόνο σε μία κλάση.
- Μια κλάση περιέχει τις τιμές που είναι μεγαλύτερες ή ίσες της μικρότερης τιμής και μικρότερες της μεγαλύτερης, δηλαδή οι κλάσεις είναι της μορφής [,).
- Οι παρατηρήσεις κάθε κλάσης θεωρούνται όμοιες, οπότε μπορούν να "αντιπροσωπευθούν" από τις κεντρικές τιμές, τα κέντρα δηλαδή κάθε κλάσης.

Σημαντική σημείωση: Σε πολλές περιπτώσεις, τα αριθμητικά αποτελέσματα που εξάγονται είναι διαφορετικά από πρόγραμμα σε πρόγραμμα. Αυτό συμβαίνει διότι δεν έχουν όλα τα προγράμματα τον ίδιο βαθμό ακρίβειας, αλλά και ο τρόπος με τον οποίο υπολογίζουν ορισμένα μεγέθη δεν είναι πάντα ο ίδιος. Πχ η τυπική απόκλιση, θα μπορούσε να μετρηθεί με αρκετούς διαφορετικούς τρόπους, ανάλογα με τις υποθέσεις που κάναμε. Το ίδιο συμβαίνει και με την ασυμμετρία. Ενώ η έννοια της «ασυμμετρίας» ή η έννοια της «σκέδασης» είναι κατανοητή, ο τρόπος που υπολογίζονται διαφέρει. Για μια πλήρη κατανόηση των διαφορετικών μεθόδων υπολογισμού, απαιτείται γνώση μαθηματικών και στατιστικής. Για τις δικές σας ανάγκες, απλά είναι απαραίτητο να γνωρίζετε την μέθοδο που χρησιμοποιείται κάθε φορά από το υπολογιστικό μας πρόγραμμα και να ανατρέχετε στην βοήθεια. Εκεί συνήθως δίνονται οι πλήρεις μαθηματικοί τύποι που χρησιμοποιούνται. Σε περίπτωση που κάτι δεν σας είναι κατανοητό, καλύτερα θα ήταν να συμβουλευθείτε κάποιον ειδικό που έχει πιο προχωρημένες γνώσεις στατιστικής. Μια άλλη ασφαλής λύση είναι να αναφέρεται το πρόγραμμα από το οποίο εξάγετε τα αποτελέσματά σας, ώστε αυτά να μπορούν να επιβεβαιωθούν αν παρουσιασθεί ανάγκη.

Παράδειγμα Περιγραφικής Στατιστικής χωρίς SPSS

Στη συνέχεια ακολουθεί ένα παράδειγμα και αφορά βιολογικά ζητήματα που εμφανίζονται στην πράξη. Τα δεδομένα και τα αποτελέσματα των παραδειγμάτων που ακολουθούν, ελήφθησαν αυτούσια, απλοποιήθηκαν, τροποποιήθηκαν ή επεκτάθηκαν για τις ανάγκες του παρόντος από το βιβλίο: «Εφαρμοσμένη Ιατρική Στατιστική», Λαζαρίδης-Λαζαρίδου,1999, Αθήνα.

Παράδειγμα

Εξετάσαμε την απτογλοβίνη 10 ατόμων που φέρονται ως υγιείς και πήραμε τις επόμενες τιμές: 1,19-1,2-1,14-1,16-1,13-1,1-1,15-1,08-1,13-1,18.

Ζητείται να βρεθεί ο μέσος όρος, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας. Βρείτε επίσης, το εύρος του δείγματος και τη διάμεσο.

Απάντηση

Από τα δεδομένα του παραδείγματος και ακολουθώντας την βήμα προς βήμα διαδικασία ή με την βοήθεια ηλεκτρονικού υπολογιστή βρίσκουμε ότι η μέση τιμή είναι **μ=1,14**.

Στη συνέχεια υπολογίζεται η τυπική απόκλιση **S=0,0384**.

Η διακύμανση ισούται με το τετράγωνο της τυπικής απόκλισης (τετράγωνο σημαίνει ο πολλαπλασιασμός ενός μεγέθους με τον εαυτό του) οπότε προκύπτει: **S2 =0,00147**.

Ο συντελεστής μεταβλητότητας προκύπτει από την διαίρεση της τυπικής απόκλισης με το μέσο όρο, οπότε: **V=0,0384/1,146=3,4%**. Επομένως, σύμφωνα με το ανωτέρω αποτέλεσμα, κατά μέσο όρο, οι τιμές του δείγματος διασπείρονται στο 3,4% σε σχέση με την μέση τιμή.

Το εύρος του δείγματος προκύπτει από την αφαίρεση της μεγαλύτερης και της μικρότερης τιμής. Οπότε ισούται με **1,2-1,08=0,12**

Η διάμεσος, κατά τα γνωστά ισούται με **Median=1,145**. Προσέξτε ότι στο συγκεκριμένο παράδειγμα ο αριθμός των παρατηρήσεων είναι άρτιος (10 άτομα) οπότε χρειάζεται να βρείτε το μέσο όρο τον δύο κεντρικών τιμών ή να συμβουλευτείτε ηλεκτρονικό υπολογιστή.

Παράδειγμα Περιγραφικής Στατιστικής με SPSS

Στο παράδειγμα που ακολουθεί θα χρησιμοποιηθεί το αρχείο «patient_los.sav» του SPSS. Το συγκεκριμένο αρχείο περιέχει δεδομένα ασθενών που είχαν εισαχθεί στο νοσοκομείο με πιθανό έμφραγμα του μυοκαρδίου. Κάθε εγγραφή (περίπτωση) αντιστοιχεί σε ένα ξεχωριστό ασθενή και περιέχει πολλές μεταβλητές που σχετίζονται με τη διαμονή τους στο νοσοκομείο. Με τα διαθέσιμα δεδομένα θα φτιαχτούν πίνακες συχνοτήτων, θα υπολογιστούν διάφορα περιγραφικά στατιστικά μέτρα καθώς και γραφήματα. Σημαντική βοήθεια θα προσφέρουν οι Πίνακες 1 και 3 της θεωρίας.

Ανάλυση συνεχών ποσοτικών μεταβλητών.

Οι συνεχείς ποσοτικές μεταβλητές είναι 2: «age» και «cost». Η διαδικασία έχει όπως παρακάτω:

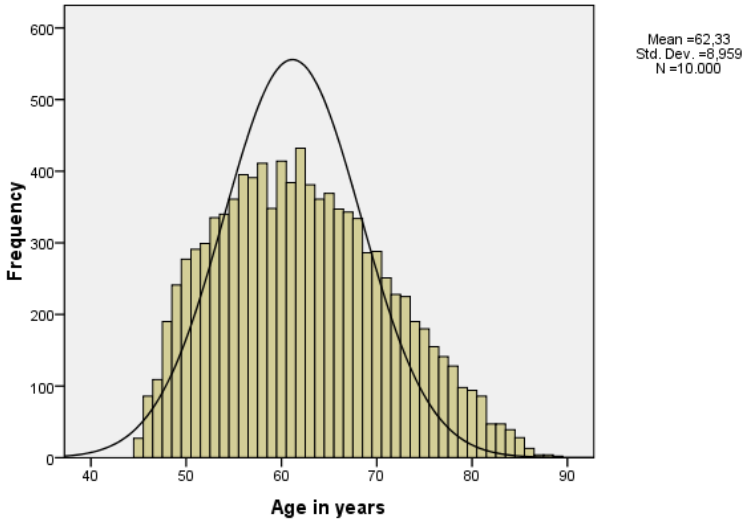
- Από τη γραμμή μενού του SPSS επιλογή: [Analyze=>Descriptive Statistics=>Frequencies…].
- Από το αναδυθέν παράθυρο, από τη λίστα με τις μεταβλητές, επιλογή πρώτα με κλικ της μεταβλητής «Age» και πάτημα του κουμπιού  και στη συνέχεια της «Cost» για να μεταφερθούν στις επιλεγείσες μεταβλητές.
- Πάτημα του κουμπιού [Statistics…] και επιλογή όλων.
- Πάτημα του κουμπιού [Charts…] και επιλογή του «Histograms» και του «with normal curve».
- Πάτημα του κουμπιού [OK].

Τα αποτελέσματα είναι τα παρακάτω.

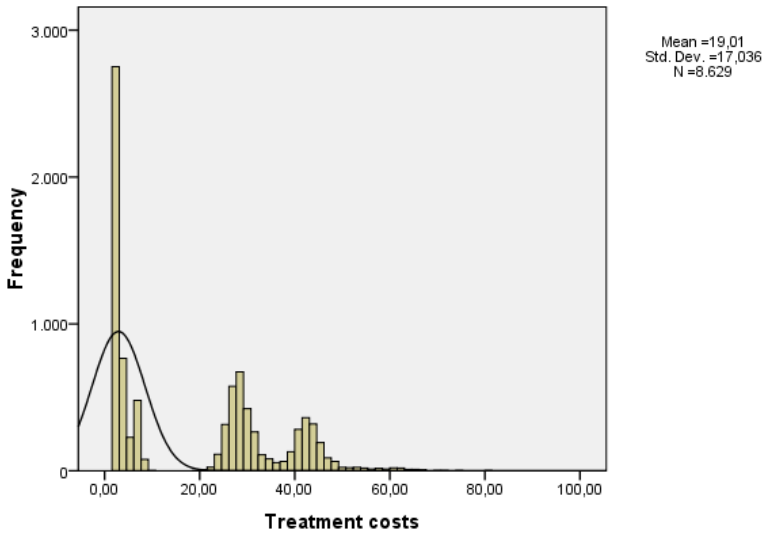
Statistics

		Age in years	Treatment costs
N	Valid	10000	8629
	Missing	0	1371
Mean		62,33	19,0057
Std. Error of Mean		,090	,18340
Median		62,00	22,0000
Mode		62	1,85
Std. Deviation		8,959	17,03644
Variance		80,256	290,240
Skewness		,308	,407
Std. Error of Skewness		,024	,026
Kurtosis		-,600	-1,199
Std. Error of Kurtosis		,049	,053
Range		44	78,67
Minimum		45	1,63
Maximum		89	80,30
Sum		623316	164000,16
Percentiles	25	55,00	1,8900
	50	62,00	22,0000
	75	69,00	30,9600

Age in years



Treatment costs



Ανάλυση ποιοτικών μεταβλητών.

Οι ποιοτικές μεταβλητές είναι αρκετές. Θα αναλυθούν οι: «agecat» , «gender», «bp» και «smoker». Η διαδικασία έχει όπως παρακάτω:

- Από τη γραμμή μενού του SPSS επιλογή: [Analyze=>Descriptive Statistics=>Frequencies…].
- Από το αναδυθέν παράθυρο, από τη λίστα με τις μεταβλητές, επιλογή πρώτα με κλικ της μεταβλητής «agecat» και πάτημα του κουμπιού  και στη συνέχεια των υπολοίπων για να μεταφερθούν στις επιλεγείσες μεταβλητές.
- Πάτημα του κουμπιού [Charts…] και επιλογή του «Bar Chart» .
- Πάτημα του κουμπιού [OK].

Τα αποτελέσματα είναι τα παρακάτω.

Statistics

		Age category	Gender	Blood pressure	Smoker
N	Valid	10000	10000	10000	10000
	Missing	0	0	0	0

Frequency Table

Age category

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	45-54	2195	22,0	22,0	22,0
	55-64	3878	38,8	38,8	60,7
	65-74	2861	28,6	28,6	89,3
	75+	1066	10,7	10,7	100,0
	Total	10000	100,0	100,0	

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	5029	50,3	50,3	50,3
	Female	4971	49,7	49,7	100,0
	Total	10000	100,0	100,0	

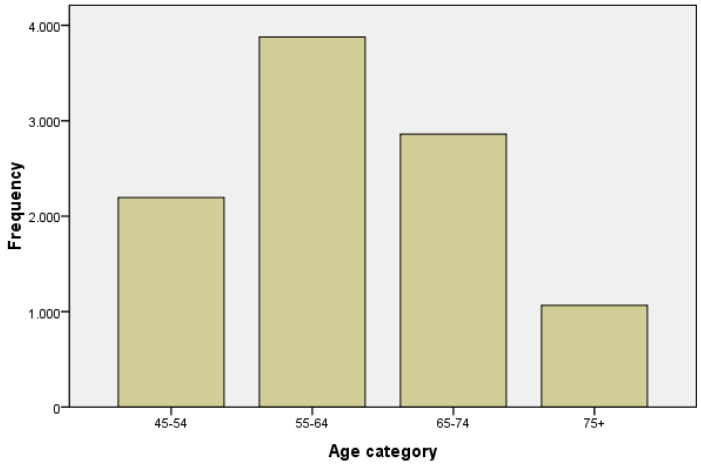
Blood pressure

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hypotension	1207	12,1	12,1	12,1
	Normal	6134	61,3	61,3	73,4
	Hypertension	2659	26,6	26,6	100,0
	Total	10000	100,0	100,0	

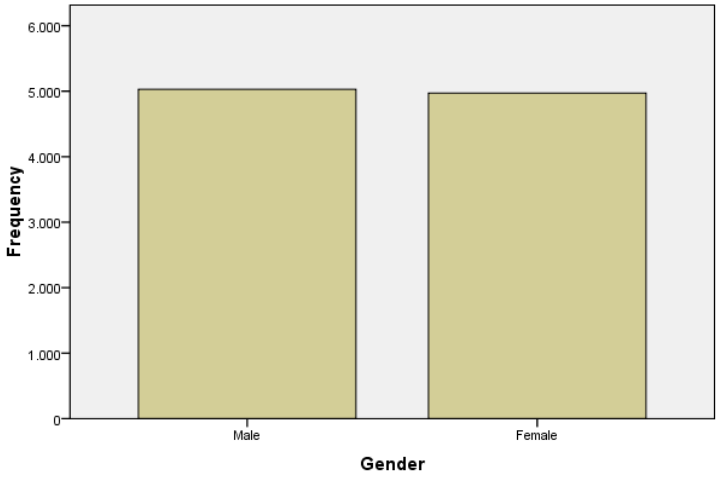
Smoker

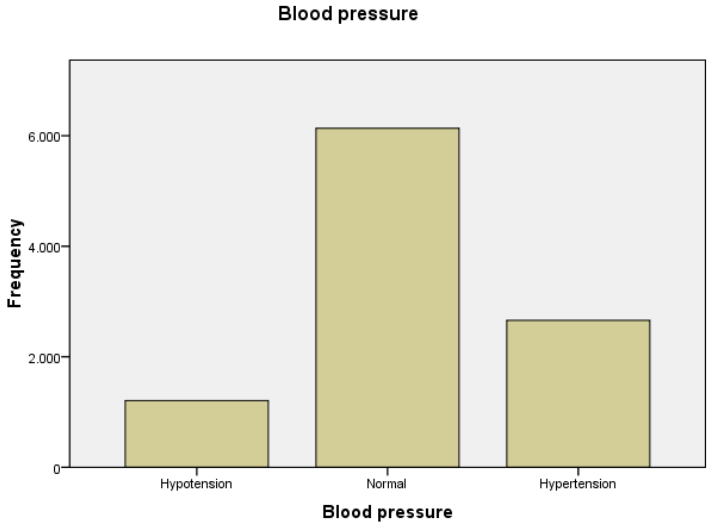
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	7725	77,3	77,3	77,3
	Yes	2275	22,8	22,8	100,0
	Total	10000	100,0	100,0	

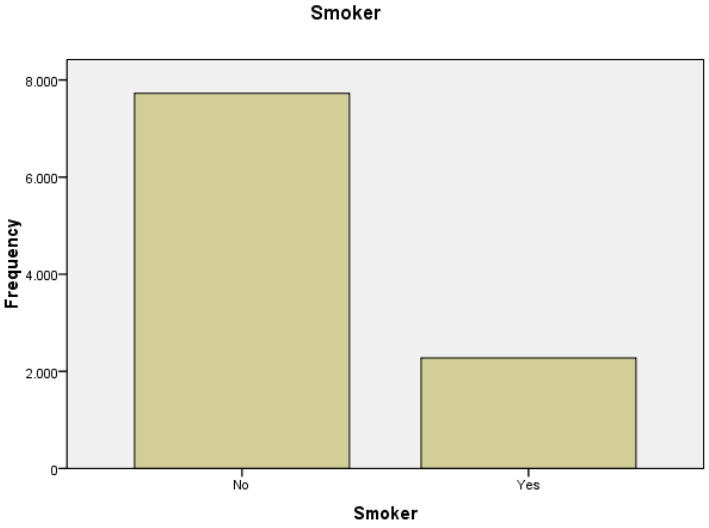
Age category



Gender







<u>Δημιουργία πίνακα διπλής εισόδου ποιοτικών μεταβλητών.</u>

Θα αναλυθούν οι μεταβλητές: «agecat» και «gender». Η διαδικασία έχει όπως παρακάτω:

- Από τη γραμμή μενού του SPSS επιλογή: [Analyze=>Descriptive Statistics=>Crosstabs…].
- Από το αναδυθέν παράθυρο, από τη λίστα με τις μεταβλητές, επιλογή πρώτα με κλικ της μεταβλητής «agecat» και τοποθέτησή της στη λίστα [Row(s)] και στη συνέχεια της μεταβλητής «gender» και τοποθέτησή της στη λίστα [Column(s)].
- Επιλογή του «Display clustered bar charts» και από-επιλογή του « Suppress tables»..
- Πάτημα του κουμπιού [OK].

Τα αποτελέσματα είναι τα παρακάτω.

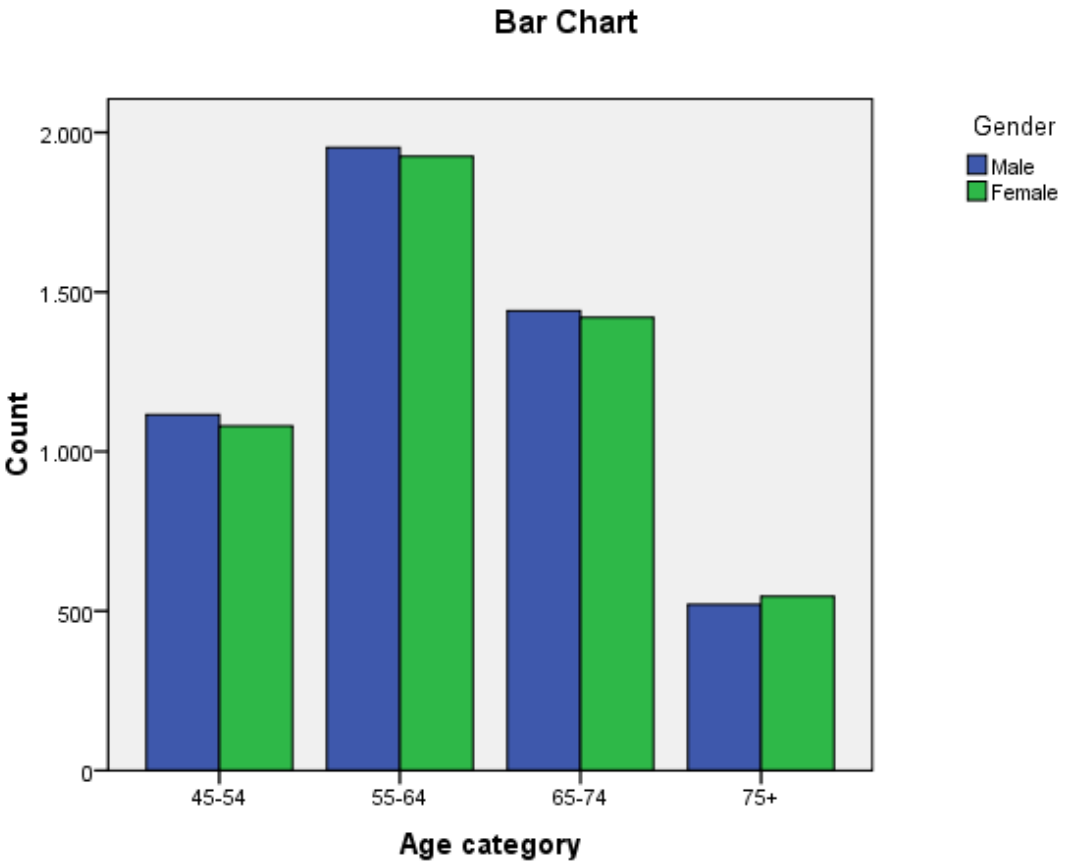
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age category * Gender	10000	100,0%	0	,0%	10000	100,0%

Age category * Gender Crosstabulation

Count

		Gender		
		Male	Female	Total
Age category	45-54	1115	1080	2195
	55-64	1953	1925	3878
	65-74	1441	1420	2861
	75+	520	546	1066
Total		5029	4971	10000



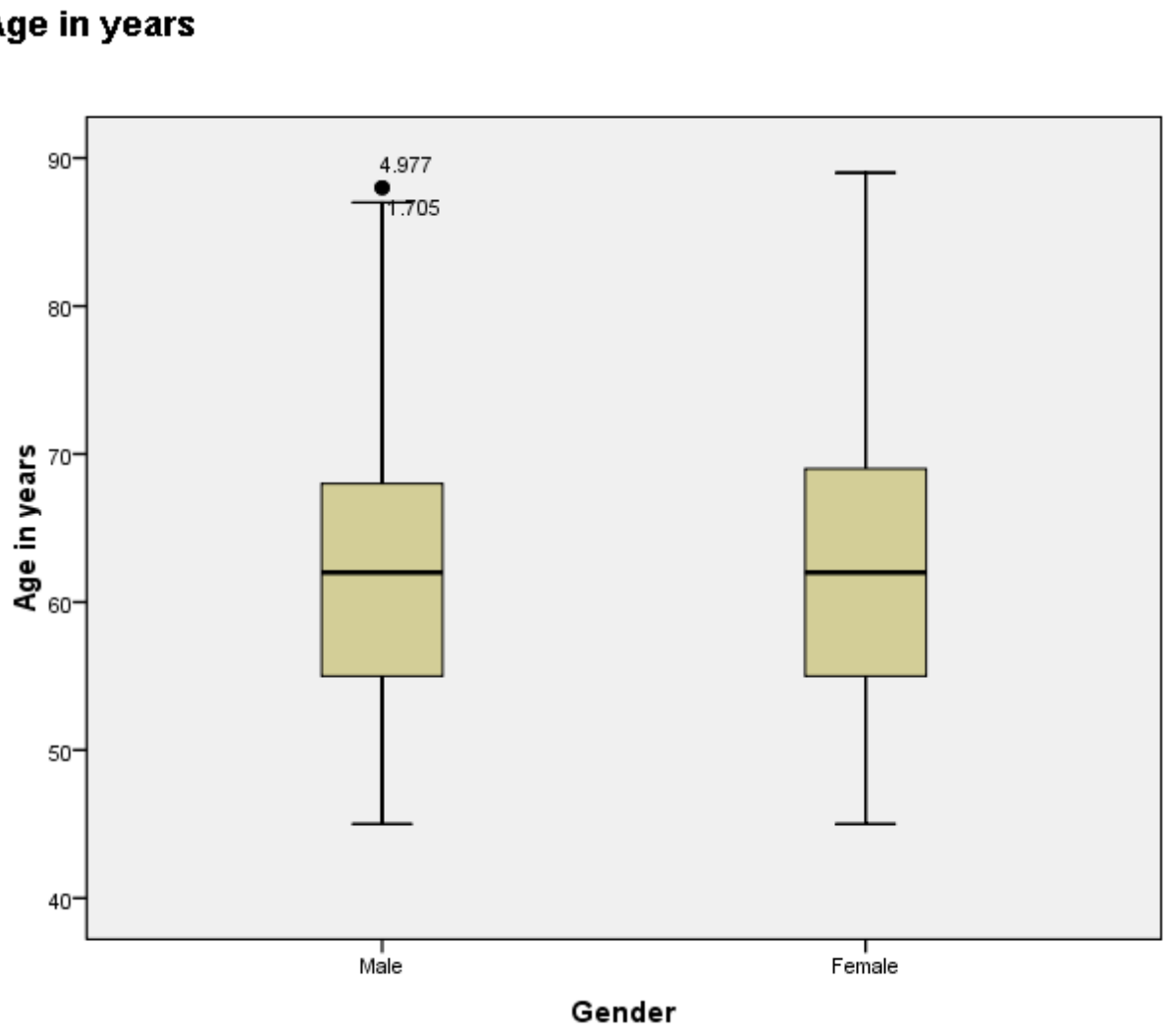
<u>Δημιουργία διαγράμματος πλαισίου-απολήξεων.</u>

Θα αναλυθούν οι μεταβλητές: «age» και «gender». Η διαδικασία έχει όπως παρακάτω:

- Από τη γραμμή μενού του SPSS επιλογή: [Graphs=>Legasy Dialogs=>Boxplot…].

- Από το αναδυθέν παράθυρο επιλογή «Simple» και «Summaries for group of cases» και πάτημα του κουμπιού [Define].
- Από το αναδυθέν παράθυρο, από τη λίστα με τις μεταβλητές, επιλογή πρώτα με κλικ της μεταβλητής «age» και τοποθέτησή της στο πεδίο [Variable] και στη συνέχεια της μεταβλητής «gender» και τοποθέτησή της στο πεδίο [Category Axis].
- Πάτημα του κουμπιού [OK].

Το αποτέλεσμα είναι το παρακάτω γράφημα.



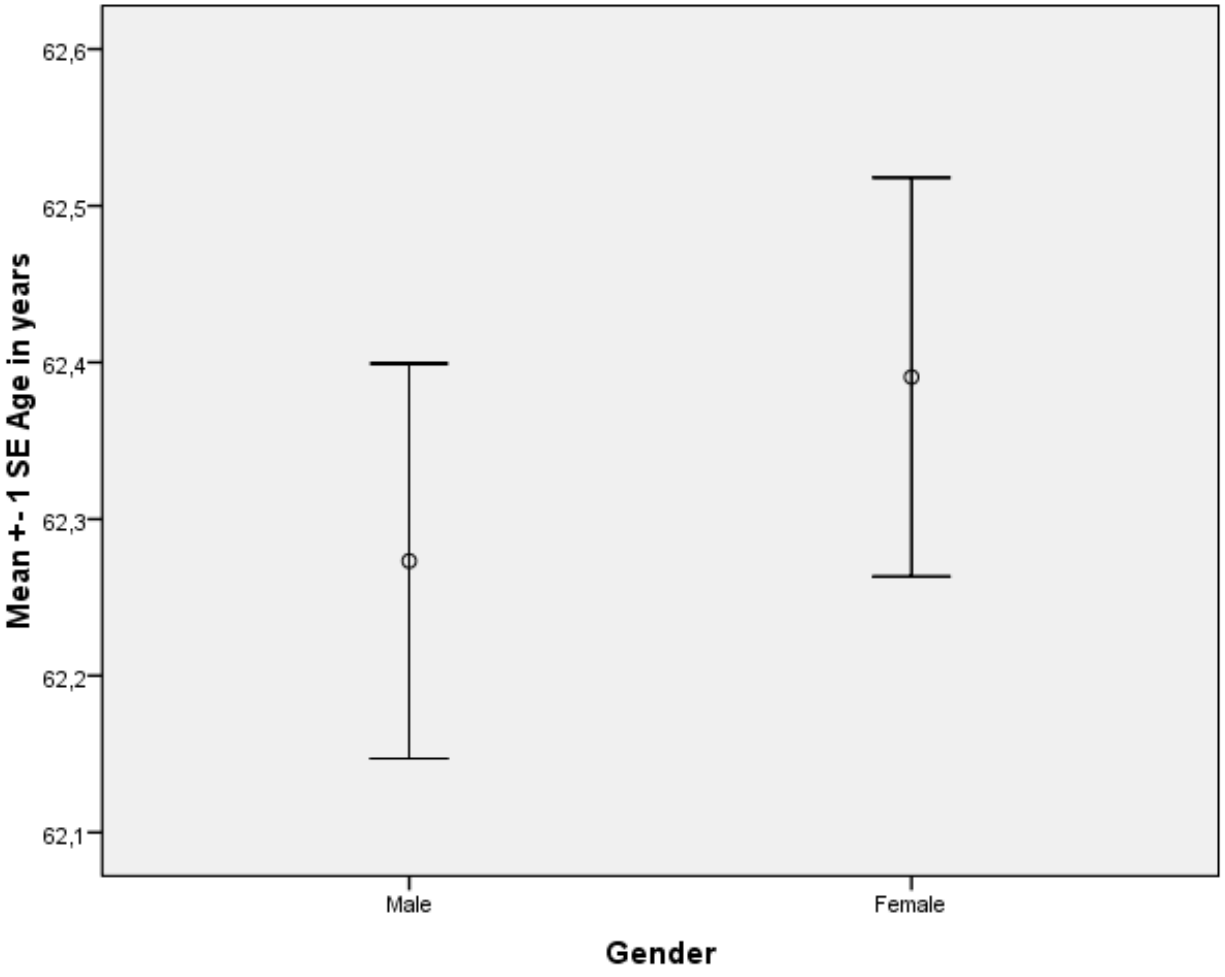
<u>Δημιουργία διαγράμματος σφαλμάτων.</u>

Θα αναλυθούν οι μεταβλητές: «age» και «gender». Η διαδικασία έχει όπως παρακάτω:

- Από τη γραμμή μενού του SPSS επιλογή: [Graphs=>Legasy Dialogs=>Error Bar…].

- Από το αναδυθέν παράθυρο επιλογή «Simple» και «Summaries for group of cases» και πάτημα του κουμπιού [Define].
- Από το αναδυθέν παράθυρο, από τη λίστα με τις μεταβλητές, επιλογή πρώτα με κλικ της μεταβλητής «age» και τοποθέτησή της στο πεδίο [Variable] και στη συνέχεια της μεταβλητής «gender» και τοποθέτησή της στο πεδίο [Category Axis].
- Στο πεδίο «Bars Represent» επιλέγω «Standard error of mean» και στο πεδίο «Multiplier» συμπληρώνω «1».
- Πάτημα του κουμπιού [OK].

Το αποτέλεσμα είναι το παρακάτω γράφημα.



5ο ΚΕΦΑΛΑΙΟ
Κατανομές Πιθανοτήτων (Probability Distributions)

Εισαγωγή

Κάθε στοιχείο του δειγματικού χώρου (δηλαδή τα απλά ενδεχόμενα) αντιστοιχίζεται σε ένα πραγματικό ή φυσικό αριθμό από τους πολλούς πιθανούς (δηλαδή ποσοτικοποιείται) μέσω μίας συνάρτησης που δεν είναι γνωστές οι τιμές της εκ των προτέρων (αβεβαιότητα) και ονομάζεται τυχαία μεταβλητή. Η ποσοτικοποίηση της αβεβαιότητας γίνεται με τη χρήση πιθανοτήτων δηλαδή μιας συνάρτησης (συνάρτηση πυκνότητας πιθανότητας) που δείχνει πόσο «συχνά» αναμένεται να παρουσιαστούν οι τιμές της τυχαίας μεταβλητής. Το σύνολο των πιθανοτήτων που αντιστοιχούν σε όλες τις τιμές μιας τυχαίας μεταβλητής ονομάζεται κατανομή πιθανότητας της τυχαίας μεταβλητής η οποία χαρακτηρίζεται από παραμέτρους που πρέπει να εκτιμηθούν.

Όπως υπάρχουν τυχαίες μεταβλητές που είναι συνεχείς ή διακριτές αντίστοιχα υπάρχουν και θεωρητικές κατανομές που είναι συνεχείς ή διακριτές και προσεγγίζουν τη κατανομή της υπό μελέτη μεταβλητής. Η συνάρτηση πυκνότητας πιθανότητας, τυχαίας μεταβλητής X, των συνεχών κατανομών συμβολίζεται με $f(x)$ και των διακριτών με $P(x)$.

Οι συνηθέστερες συνεχείς κατανομές είναι: κανονική, ομοιόμορφη, εκθετική, γάμα, βήτα, Cauchy, weibull.

Οι συνηθέστερες διακριτές κατανομές είναι: Bernoulli, Poisson, διωνυμική, γεωμετρική, υπεργεωμετρική.

Λόγω ακριβώς της μεγάλης τους χρησιμότητας τα χαρακτηριστικά και οι ιδιότητες των κατανομών αυτών έχουν μελετηθεί διεξοδικά και τα αποτελέσματα διαφόρων υπολογισμών που χρησιμοποιούνται συχνά έχουν συγκεντρωθεί σε εύχρηστους πίνακες.

5.1 Κανονική κατανομή ή κατανομή Gauss (Normal distribution)

Η σημαντικότερη κατανομή πιθανότητας θεωρείται η κανονική για τρεις κυρίως λόγους:

- Πολλά πειράματα εκφράζονται μέσω τυχαίων μεταβλητών που ακολουθούν κανονική κατανομή.
- Πολλές ποσότητες που συναντάμε στη φύση ακολουθούν την κανονική κατανομή.
- Προσεγγίζει ικανοποιητικά άλλες κατανομές, λόγω του Κεντρικού Οριακού Θεωρήματος.
- Πολλές τεχνικές που χρησιμοποιούνται στην επαγωγική στατιστική στηρίζονται σ' αυτήν.

Ορίζουμε ότι μία τυχαία συνεχής μεταβλητή X, που μπορεί να πάρει τιμές σε όλη την ευθεία των πραγματικών αριθμών, ακολουθεί την κανονική κατανομή με παραμέτρους μ και σ2 [συμβολίζεται $X \square N(\mu,\sigma^2)$] εάν η συνάρτηση πυκνότητας πιθανότητας είναι ίση με:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Η κανονική κατανομή:

- Είναι συμμετρική γύρω από το μ
- Είναι μονοκόρυφη
- Έχει κωνοειδές σχήμα
- Είναι ασύμπτωτη ως προς τον άξονα χ
- Η μέση τιμή, η διάμεσος και η επικρατούσα συμπίπτουν.
- Έχει συντελεστή ασυμμετρίας και κυρτότητας ίσο με 0.

Εάν ορίσουμε ως Z την τυποποιημένη μεταβλητή που αντιστοιχεί στη X δηλαδή:

$$Z=\frac{X-\mu}{\sigma}$$

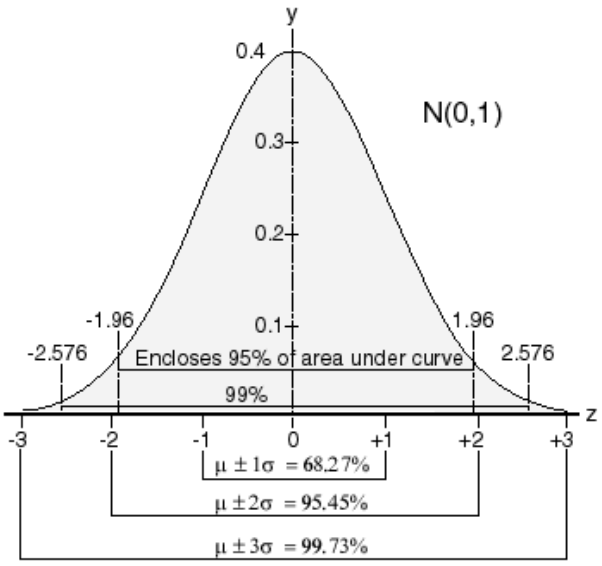
τότε η μέση τιμή της Z είναι ίση με 0 και η διασπορά της με 1. Η συνάρτηση πυκνότητας πιθανότητας της Z είναι:

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

και ονομάζεται τυποποιημένη ή τυπική κανονική κατανομή και συμβολίζεται $Z \square N(0,1)$.

Ισχύει για την κανονική κατανομή, και κατ' επέκταση για την τυποποιημένη κανονική κατανομή, ότι περίπου:

- Το 68% των παρατηρήσεων βρίσκεται εντός του διαστήματος $\bar{x}\pm1\sigma$.
- Το 95% των παρατηρήσεων βρίσκεται εντός του διαστήματος $\bar{x}\pm2\sigma$.
- Το 99,7% των παρατηρήσεων βρίσκεται εντός του διαστήματος $\bar{x}\pm3\sigma$.



Η διαδικασία κατασκευής ενός 95% διαστήματος εμπιστοσύνης (95% Confidence Interval) για τη μέση τιμή του πληθυσμού, δηλαδή η πιθανότητα η μέση τιμή του πληθυσμού να ανήκει σ' αυτό το διάστημα, έχει όπως παρακάτω:

- Υπολογισμός της τετραγωνικής ρίζας του πλήθους των τιμών n.
- Διαίρεση της τυπικής απόκλισης με το προηγούμενο αποτέλεσμα.
- Πολλαπλασιασμός του προηγούμενου αποτελέσματος με το 1,96.
- Αφαίρεση του προηγούμενου αποτελέσματος από τον μέσο όρο για τον υπολογισμό του «κάτω άκρου του διαστήματος εμπιστοσύνης».
- Πρόσθεση του προ-προηγούμενου αποτελέσματος στο μέσο όρο για τον υπολογισμό του «άνω άκρου του διαστήματος εμπιστοσύνης».

Για την κατασκευή ενός 99% διαστήματος εμπιστοσύνης αντικαθίσταται το 1,96 με τον αριθμό 2,57.

6ο ΚΕΦΑΛΑΙΟ
Επαγωγική Στατιστική (Συμπερασματολογία)

Εισαγωγή

Επαγωγική στατιστική είναι ο κλάδος της στατιστικής που ασχολείται με τις μεθόδους μεταφοράς των πληροφοριών από ένα δείγμα (μερικό) στον πληθυσμό (γενικό) και την εξαγωγή των αντίστοιχων συμπερασμάτων. Πιο συγκεκριμένα περιλαμβάνει:

- Την εκτιμητική, η οποία περιλαμβάνει τις μεθόδους εκτίμησης των παραμέτρων του πληθυσμού (σημειακή εκτίμηση) καθώς και της εκτίμησης διαστημάτων εμπιστοσύνης γι' αυτές, με βάση τα αντίστοιχα στατιστικά στοιχεία του δείγματος.
- Τον έλεγχο υποθέσεων, δηλαδή την επιβεβαίωση (ή απόρριψη) των ισχυρισμών για τις τιμές των παραμέτρων του πληθυσμού. Τα στάδια του ελέγχου είναι:
    - Ορισμός της μηδενικής υπόθεσης H_0.
    - Ορισμός της εναλλακτικής υπόθεσης H_1.
    - Ορισμός του επιπέδου σημαντικότητας α (δηλαδή την πιθανότητα απόρριψης της H_0 ενώ είναι σωστή, συνήθως α=0,05).
    - Ορισμός του στατιστικού τεστ (συνάρτηση ελέγχου) που θα χρησιμοποιηθεί.
    - Υπολογισμός του p-value. (χώρος απόρριψης της H_0).
    - Έλεγχος εάν το p-value είναι μικρότερο από το επίπεδο σημαντικότητας α (οπότε και απορρίπτουμε την H_0) ή μεγαλύτερο (οπότε και δεχόμαστε την H_0).
    - Συμπεράσματα.
- Τα στατιστικά μοντέλα, με τα οποία πραγματοποιείται η εκτίμηση της τιμής ή του διαστήματος εμπιστοσύνης μιας τυχαίας εξαρτημένης μεταβλητής όταν γνωρίζουμε τις τιμές ορισμένων άλλων τυχαίων ανεξάρτητων μεταβλητών.

Παρουσίαση των στατιστικών τεστ (συναρτήσεων ελέγχου) παραμετρικών και μη παραμετρικών που μπορούν να χρησιμοποιηθούν ανάλογα με τον αριθμό, μία ή δύο, και τον τύπο της μεταβλητής παρουσιάζονται στους δύο επόμενους πίνακες αντίστοιχα. Για να

πραγματοποιηθεί το παραμετρικό τεστ απαιτείται η μεταβλητή να ακολουθεί την κανονική κατανομή ή να έχει πάνω από 30 παρατηρήσεις.

Τύπος μεταβλητής	Τύπος μεταβλητής - Μία μεταβλητή	
	Συνεχής	One Sample t-test
	Διάταξης	Sign test
	Ονομαστική	X2 καλής προσαρμογής
	Δυαδική	Διωνυμικό τεστ

Τύπος μεταβλητής	Τύπος μεταβλητής - Δύο μεταβλητές				
		Συνεχής	Διάταξης	Ονομαστική	Δυαδική
	Συνεχής	Γραμμική παλινδρόμηση, Pearson's Γραμμική συσχέτιση	Kendall's tau, συντελεστής Gamma	One-way ANOVA	Independent Samples t-test
				One-way ANOVA (επαναλαμβανόμενες μετρήσεις)	Paired Samples t-test
	Διάταξης		Kendall's tau, συντελεστής Gamma	Kruskal-Wallis test	Mann-Whitney U test, Kolmogorov-Smirnov test
				Friedman's test	Sign test κατά ζεύγη
	Ονομαστική			X2 τεστ ανεξαρτησίας	X2 τεστ ανεξαρτησίας
					Maxwell's test
	Δυαδική				Fisher's exact test
					Liddell's exact test
	Ανεξάρτητες μεταβλητές				
	Εξαρτημένες μεταβλητές				

Τα αντίστοιχα μη παραμετρικά τεστ για κάποια βασικά παραμετρικά είναι όπως στον παρακάτω πίνακα:

Παραμετρικός έλεγχος	**Μη παραμετρικός**
Independent Samples t-test	Man-Whitney U-test
T –paired test	Wilcoxon test
Ανάλυση διακύμανσης (ANOVA)	Kruskal-Wallis test

Όταν οι εμπλεκόμενες μεταβλητές είναι 3 και πάνω για να εκτιμηθεί η τιμή της εξαρτημένης βάσει των τιμών των ανεξάρτητων χρησιμοποιούνται στατιστικά εμπειρικά μοντέλα όπως στον παρακάτω πίνακα.

Ερώτημα	**Τύπος Τ.Μ.**	**Στατιστικός έλεγχος**
Πώς επηρεάζεται η τιμή της εξαρτημένης μεταβλητής από τις τιμές των ανεξάρτητων. Μπορούμε να προβλέψουμε την τιμή της εξαρτημένης μεταβλητής;	Συνεχείς	1) Γραμμική απλή ή πολλαπλή παλινδρόμηση 2) Μη γραμμική παλινδρόμηση
	Η εξαρτημένη μεταβλητή είναι δίτιμη οι ανεξάρτητες οποιουδήποτε τύπου	Λογιστική παλινδρόμηση (logistic regression)
	Η εξαρτημένη μεταβλητή είναι συνεχής, οι ανεξάρτητες κατηγορικές.	Ανάλυση διακύμανσης ANOVA) κατά παράγοντες (Factors)
	Η εξαρτημένη μεταβλητή είναι συνεχής, οι ανεξάρτητες κατηγορικές και συνεχείς.	ANOVA MANOVA
	Η εξαρτημένη μεταβλητή είναι κατηγορική, οι ανεξάρτητες κατηγορικές και συνεχείς.	Κατηγορική Λογιστική παλινδρόμηση (Nominal logistic regression)

6.1 Ιατρική και Στατιστική

Το επιστημονικό μοντέλο σκέψης της Ιατρικής βασίζεται στην παρατήρηση, στην υπόθεση, στο πείραμα, στην τεκμηρίωση και τα συμπεράσματα.

Βασικό βήμα της όλης διαδικασίας, είναι η διατύπωση επιστημονικών υποθέσεων που αφορούν τις ιδιότητες ενός πληθυσμού. Οι υποθέσεις αυτές τροποποιούνται ύστερα από την επιστημονική παρατήρηση.

Στενά συνδεόμενη με την έννοια της υπόθεσης είναι η έννοια της αβεβαιότητας ή της πιθανότητας. Δηλαδή ποια είναι η πιθανότητα αυτή η υπόθεση που διατυπώθηκε να έχει ισχύ στον πραγματικό πληθυσμό. Τα συνηθισμένα, ιστορικά, επίπεδα πιθανότητας αβεβαιότητας είναι το 0,01 ή και 0,05 ιδιαίτερα η τελευταία ως οριακή. Στην πραγματικότητα πάντως επαφίεται στην εμπειρία και τη διαίσθηση του ερευνητή να ορίσει το ακριβές επίπεδο σημαντικότητας (p value) που αυτός θεωρεί αποδεκτό προκειμένου να απορρίψει ή να αποδεχθεί μια υπόθεση. Για παράδειγμα, αν η υπόθεση μας είναι ότι δύο θεραπευτικές αγωγές που χειρίζονται τις β λιποπρωτείνες ορού έχουν το ίδιο αποτέλεσμα και το επίπεδο σημαντικότητας είναι p=0,006, αυτό σημαίνει ότι υπάρχει πιθανότητα 6/1000 η υπόθεση που κάναμε να είναι αληθινή. Είναι όμως αυτό αξιόπιστο; Επειδή το 0,006 είναι αρκετά μικρό συμπεραίνουμε ότι η αρχική μας υπόθεση δεν έχει ισχύ και την απορρίπτουμε. Αν αντίθετα το επίπεδο σημαντικότητας ήταν p=0,06, πράγμα που σημαίνει ότι υπάρχει πιθανότητα 6% να είναι αληθινή, τότε θα μπορούσαμε να την αποδεχθούμε. Γίνεται προφανές, ότι η απόρριψη ή αποδοχή μιας υπόθεσης εξαρτάται από το πόσο «αυστηροί» είμαστε σχετικά με την πιθανότητα. Δηλαδή, πόσο μεγάλη θέλουμε να είναι η πιθανότητα αποδοχής; Στην συντριπτική πλειοψηφία, μια πιθανότητα άνω του 0,05 κάνει μια υπόθεση αληθινή κυρίως για λόγους κοινής μεθοδολογίας, αλλά αυτό δεν είναι απαραίτητα δεσμευτικό.

Αυτό που είναι απαραίτητα σημαντικό να γνωρίζετε όταν διαβάζεται στατιστικά αποτελέσματα, είναι:

- Ποια είναι η υπόθεση που εξετάζεται.
- Ποια είναι η πιθανότητα αποδοχής αυτής της πιθανότητας (p-value).
- Τα όριο εντός των οποίων κινείται η αληθινή τιμή στον πληθυσμό.

Όλα τα υπόλοιπα ή ενδιάμεσα αποτελέσματα έχουν σημασία για ένα Στατιστικό, αλλά όχι για κάποιον Ιατρό. Σε πολλές περιπτώσεις η υπόθεση μπορεί να είναι λίγο ασαφής ή να μην γίνεται άμεσα κατανοητή, με αποτέλεσμα να μην μπορείτε να ερμηνεύσετε το αποτέλεσμα του p-value. Για αυτό το λόγο πρέπει κανείς να είναι ιδιαίτερα προσεκτικός. Πάντως τα περισσότερα προγράμματα στον υπολογιστή δίνουν το επίπεδο σημαντικότητας και την υπόθεση που εξετάζεται, καθώς και τα όρια των πραγματικών τιμών για δοσμένη πιθανότητα.

7ο ΚΕΦΑΛΑΙΟ
Σύγκριση Μέσων Τιμών - Διαδικασία t-test

Εισαγωγή

Πολλές φορές επιβάλλεται σ' ένα αρχείο δεδομένων να πραγματοποιηθεί σύγκριση των μέσων τιμών ποσοτικών μεταβλητών. Έτσι, άλλες φορές υπάρχει η ανάγκη σύγκρισης των μέσων τιμών της ίδιας ποσοτικής μεταβλητής δύο ανεξάρτητων πληθυσμών (Independent-Samples t-test), άλλοτε η ανάγκη σύγκρισης των μέσων τιμών δύο ποσοτικών μεταβλητών που προέρχονται από τον ίδιο πληθυσμό (Paired-Samples t-test), και ακόμη, υπάρχει η ανάγκη σύγκρισης της μέσης τιμής μίας ποσοτικής μεταβλητής με ένα σταθερό αριθμό (One-Sample t-test).

7.1 Σύγκριση μέσων τιμών ανεξάρτητων πληθυσμών

Πολύ συχνά απαιτείται να ελεγχθεί εάν δύο ομοειδείς ποσοτικές μεταβλητές, που προέρχονται από δύο ανεξάρτητους μεταξύ τους πληθυσμούς, διαφέρουν κατά μέση τιμή ως προς μία ανεξάρτητη μεταβλητή (δύο κατηγοριών), εάν δηλαδή, οι μέσες τιμές τους είναι ίσες ή διαφέρουν σημαντικά. Όπως είναι φυσικό για κάθε τέτοιου είδους έλεγχο είναι πρακτικά αδύνατο να υπολογιστούν τις μέσες τιμές των πληθυσμών οπότε χρησιμοποιούνται δείγματα που παίρνονται από τους δύο πληθυσμούς. Από τα δείγματα αυτά, που δεν είναι υποχρεωτικά του ίδιου μεγέθους, υπολογίζονται οι δειγματικές μέσες τιμές και οι δειγματικές διασπορές και στη συνέχεια εκτελείται στατιστικός έλεγχος. Για παράδειγμα, για να ελεγχθεί εάν η μέση διαστολική πίεση (εξαρτημένη μεταβλητή) διαφέρει μεταξύ της ομάδας των αντρών και της ομάδας των γυναικών (ανεξάρτητη μεταβλητή) χρησιμοποιείται το t-test για ανεξάρτητα δείγματα (Independent-Samples t-test).

Εκφράζοντας τις παραπάνω αρχές με στατιστική ορολογία, έστω δύο ανεξάρτητοι πληθυσμοί με μέσες τιμές μ_1 και μ_2 και διασπορές σ_1^2 και σ_2^2. Η υπόθεση που πρέπει να ελεγχθεί είναι:

H_0: Οι μέσες τιμές των δύο πληθυσμών δε διαφέρουν σημαντικά ($\mu_1 - \mu_2 = 0$).

H_1: Οι μέσες τιμές των δύο πληθυσμών διαφέρουν σημαντικά ($\mu_1 - \mu_2 \neq 0$).

Για το σκοπό αυτό, χρησιμοποιούνται δύο δείγματα από τους δύο πληθυσμούς με μεγέθη $n_1>30$ και $n_2>30$. Από τα δείγματα αυτά υπολογίζονται οι δειγματικές μέσες τιμές $\overline{x}_1$ και $\overline{x}_2$ και οι δειγματικές διασπορές s_1^2 και s_2^2. Σε αυτά τα στατιστικά μέτρα βασίζεται ο στατιστικός έλεγχος που είναι γνωστός με το όνομα Student's t-test. Η διαδικασία αποτελείται από τα παρακάτω βήματα:

- [Βήμα 1]: Διατύπωση της μηδενικής και της εναλλακτικής υπόθεσης.

  $$\mathbf{H_0}: \mu_1 - \mu_2 = 0$$

  $$\mathbf{H_1}: \mu_1 - \mu_2 \neq 0$$

- [Βήμα 2]: Ορισμός του επιπέδου σημαντικότητας α (significance level).

  Ορίζεται σφάλμα τύπου I η απόρριψη της μηδενικής υπόθεσης δεδομένου ότι είναι σωστή. Η πιθανότητα του σφάλματος συμβολίζεται με α και

  $$\alpha = P[\text{απόρριψη της } H_0 | H_0 \text{ είναι σωστή}].$$

  Ορίζεται σφάλμα τύπου II η αποδοχή της μηδενικής υπόθεσης δεδομένου ότι είναι λάθος. Η πιθανότητα του σφάλματος συμβολίζεται με β και

  $$\beta = P[\text{αποδοχή της } H_0 | H_0 \text{ είναι λάθος}].$$

  Συνήθως το επίπεδο σημαντικότητας είναι α=0,05 και το διάστημα εμπιστοσύνης (Confidence Interval) ισούται με CI=100*(1-α)%.

- [Βήμα 3]: Σύγκριση διασπορών των δύο δειγμάτων. Εδώ ελέγχεται η υπόθεση:

  $$\mathbf{H_0}: \sigma_1^2 = \sigma_2^2$$

  $$\mathbf{H_1}: \sigma_1^2 \neq \sigma_2^2$$

  Ο έλεγχος γίνεται με το στατιστικό μέτρο F (Levene's Test for Equality of Variances). Αν το επίπεδο σημαντικότητας (p-value ή significance) για το F είναι μικρό (<0.05 συνήθως), τότε η υπόθεση H_0 απορρίπτεται και μπορούμε να υποθέσουμε ότι οι δύο διασπορές παρουσιάζουν σημαντική διαφορά.

- [Βήμα 4]: Εδώ διακρίνονται δύο περιπτώσεις:
  - [Περίπτωση 1η]: Οι δύο διασπορές των πληθυσμών βρέθηκαν ίσες στο [Βήμα 3]. Στην περίπτωση αυτή υπολογίζεται η κοινή διασπορά (pooled variance) των δύο δειγμάτων ως εκτιμητής της κοινής διασποράς των δύο πληθυσμών. Ο τύπος υπολογισμού είναι :

    $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Στη συνέχεια υπολογίζεται το στατιστικό μέτρο t από τον τύπο :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

και το επίπεδο σημαντικότητας α σύμφωνα με τους βαθμούς ελευθερίας της κατανομής που ακολουθεί το t (οι βαθμοί ελευθερίας υπολογίζονται από τη σχέση n_1+n_2-2). Αν η σημαντικότητα είναι μικρή (συνήθως <0.05) τότε η μηδενική υπόθεση της ισότητας των δύο μέσων τιμών απορρίπτεται (στατιστικά σημαντική διαφορά). Στην αντίθετη περίπτωση μπορούμε να υποθέσουμε ότι οι δύο πληθυσμοί δεν διαφέρουν σημαντικά ως προς τη μέση τιμή τους.

- [Περίπτωση $2^η$]: Οι δύο διασπορές των πληθυσμών βρέθηκαν άνισες (απορρίφθηκε δηλαδή η μηδενική υπόθεση στο [Βήμα 3]). Στην περίπτωση αυτή, το στατιστικό μέτρο t υπολογίζεται από τον τύπο

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Στη συνέχεια υπολογίζεται η σημαντικότητα του t. Αν αυτή είναι μικρή (συνήθως <0.05) τότε η μηδενική υπόθεση της ισότητας των δύο μέσων τιμών απορρίπτεται (στατιστικά σημαντική διαφορά). Στην αντίθετη περίπτωση μπορούμε να υποθέσουμε ότι οι δύο πληθυσμοί δε διαφέρουν σημαντικά ως προς τη μέση τιμή τους. Οι βαθμοί ελευθερίας της κατανομής του t είναι:

$$DF = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{(s_1^2 / n_1)^2 / (n_1 - 1) + (s_2^2 / n_2)^2 / (n_2 - 1)}$$

7.2 Σύγκριση μέσων τιμών ζευγών παρατηρήσεων (εξαρτημένων πληθυσμών)

Σε πολλές περιπτώσεις, ιδίως σε έρευνες όπου τα δεδομένα προέρχονται από «κλειστούς» και ελεγχόμενους πειραματισμούς, έχουμε αντί για ανεξάρτητα δείγματα, ζεύγη παρατηρήσεων (εξαρτημένα δείγματα). Για παράδειγμα σε ένα ιατρικό πείραμα, όπου ερευνάται η επίδραση ενός νέου φαρμάκου που καταπολεμά την υπέρταση, είναι πολύ φυσικό να επιλεχτεί ένα δείγμα ασθενών και να μετρηθεί η πίεσή τους πριν και μετά από τη λήψη του φαρμάκου. Στην περίπτωση αυτή υπάρχουν ζεύγη (pairs) μετρήσεων, στους ίδιους συμμετέχοντες του δείγματος,

σε δύο διαφορετικές χρονικές στιγμές. Σε άλλες περιπτώσεις είναι δυνατό να θεωρηθούν σα ζεύγη μετρήσεις που λαμβάνονται, από τους ίδιους συμμετέχοντες του δείγματος, για δύο μεταβλητές χρησιμοποιώντας την ίδια κλίμακα μέτρησης.

Υπάρχουν δύο δείγματα (μετρήσεις) X και Y κατά ζεύγη, ή συσχετισμένα όπως συνήθως λέγονται, μεγέθους n. Η στατιστική υπόθεση την οποία πρέπει να ελεγχθεί είναι:

$$\mathbf{H_0}: \mu_X - \mu_Y = 0$$

$$\mathbf{H_1}: \mu_X - \mu_Y \neq 0$$

Η μέθοδος που ακολουθείται είναι αρχικά η δημιουργία ενός νέου δείγματος (νέας μεταβλητής) D = X - Y το οποίο έχει τιμές τις διαφορές των κατά ζεύγη τιμών των δύο δειγμάτων και στη συνέχεια, ο υπολογισμός του στατιστικού t από τον τύπο:

$$t = \frac{\overline{D}}{s_d / \sqrt{n}}$$

Η σημαντικότητα του t θα είναι αυτή που θα κρίνει την απόρριψη της μηδενικής υπόθεσης. Έτσι, σημαντικότητα μικρή (p-value<0.05 συνήθως) μας οδηγεί στην απόρριψη της H_0, δηλαδή στο συμπέρασμα της στατιστικά σημαντικής διαφοράς στις μέσες τιμές των δειγμάτων.

Ένας συντελεστής που έχει ουσιαστική σημασία στον παραπάνω έλεγχο, είναι ο συντελεστής συσχέτισης (correlation) ανάμεσα στα δύο δείγματα καθώς και η σημαντικότητά του. Συντελεστής συσχέτισης θετικός δείχνει ότι η επιλογή της μεθόδου των ζευγαρωτών παρατηρήσεων που κάναμε, ήταν αποδοτική.

7.3 Σύγκριση μέσης τιμής πληθυσμού με δεδομένη τιμή

Σε πολλές περιπτώσεις χρειάζεται να συγκριθεί η (άγνωστη) μέση τιμή μ ενός πληθυσμού με μία δεδομένη τιμή μ_0 γνωστή από εμπειρία, από προηγούμενες μελέτες κλπ. Η υπόθεση που θα ελεγχτεί είναι η:

$$\mathbf{H_0}: \mu = \mu_0$$

$$\mathbf{H_1}: \mu \neq \mu_0$$

Ο έλεγχος γίνεται αφού ληφθεί δείγμα μεγέθους n με στατιστικά μέτρα $\overline{x}$ και s^2 και στη συνέχεια με τη βοήθεια του στατιστικού t που υπολογίζεται από τον τύπο:

$$t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

Η υπόθεση H_0 απορρίπτεται (δηλαδή υπάρχει στατιστικά σημαντική διαφορά) αν η σημαντικότητα του t βρεθεί πολύ μικρή (συνήθως <0.05), διαφορετικά τη δεχόμαστε..

7.4 Προϋποθέσεις για την εφαρμογή του t-test

Υπάρχουν κάποιες σημαντικές προϋποθέσεις για την εφαρμογή του t-test:

- [Ανεξάρτητα δείγματα]:
    - Κανονικότητα (Normality): δηλαδή οι κατανομές των εξαρτημένων μεταβλητών ακολουθούν την κανονική κατανομή. Το t-test είναι αρκετά ανθεκτικό στις αποκλίσεις από την κανονικότητα ειδικά εάν το δείγμα δεν είναι μικρό. Εάν έχουμε μεγάλο δείγμα βάσει του Κεντρικού Οριακού Θεωρήματος (ΚΟΘ), το οποίο λέει ότι καθώς το μέγεθος του δείγματος τείνει στο άπειρο, η κατανομή των δεδομένων τείνει στην κανονική, μπορούμε να παραβλέψουμε αυτήν την προϋπόθεση. Η θεωρία αναφέρει ότι αν για κάθε ομάδα έχουμε μέγεθος δείγματος μεγαλύτερο του 30 (δηλαδή τουλάχιστον 30 ασθενείς στην ομάδα Θεραπείας και τουλάχιστον 30 ασθενείς στην ομάδα Placebo, όπως στο παράδειγμά μας) είναι ικανοποιητικό για να μας εξασφαλίσει την ισχύ του θεωρήματος και μπορεί να εφαρμοστεί το t-test παραβλέποντας την κανονικότητα. Εάν το δείγμα είναι μικρό (μικρότερο από 30 περιπτώσεις) ή/και δεν ισχύει η κανονικότητα θα πρέπει να εφαρμοστεί το μη παραμετρικό ισοδύναμό του το Mann-Whitney U test. Επίσης, ο έλεγχος της κανονικότητας μπορεί να πραγματοποιηθεί με το μη-παραμετρικό Kolmogorov-Smirnov test καθώς και γραφικά (ιστόγραμμα-διάγραμμα διασποράς).
    - Ανεξαρτησία παρατηρήσεων (Independent Observations): δηλαδή επιλογή ανεξάρτητου τυχαίου δείγματος από τον πληθυσμό.
    - Ίσες διακυμάνσεις (Equal Variances): δηλαδή η διακυμάνσεις των κατανομών των ανεξάρτητων δειγμάτων να είναι ίσες. Εάν δεν είναι ίσες (Levene's test) τότε επιλέγεται το αποτέλεσμα του διορθωμένου t-test ($2^{η}$ γραμμή του πίνακα αποτελεσμάτων).

- [Εξαρτημένα δείγματα]:
    - Οι παρατηρήσεις για κάθε ζεύγος θα πρέπει να πραγματοποιήθηκαν υπό τους ίδιους όρους και συνθήκες.
    - Κανονικότητα (Normality): η μεταβλητή με τη μέση διαφορά των ζευγών θα πρέπει να κατανέμεται κανονικά. Εάν το δείγμα είναι μικρό (μικρότερο από 30 περιπτώσεις) ή/και δεν ισχύει η κανονικότητα θα πρέπει να εφαρμοστεί το μη παραμετρικό ισοδύναμό του το Wilcoxon test.
    - Η διακύμανση των εξαρτημένων μεταβλητών μπορεί να είναι ίση ή άνιση.
- [Ένα δείγμα με δεδομένη τιμή]:
    - Κανονικότητα (Normality): δηλαδή η κατανομή της εξαρτημένης μεταβλητής πρέπει να ακολουθεί την κανονική κατανομή (μη παραμετρικό ισοδύναμο τεστ στο SPSS δεν υπάρχει).
    - Ανεξαρτησία παρατηρήσεων (Independent Observations): δηλαδή επιλογή ανεξάρτητου τυχαίου δείγματος από τον πληθυσμό.
    - Ακραίες τιμές (Outliers).

7.5 Μονόπλευρος έλεγχος t-test (1-tailed) (ή μονής κατεύθυνσης)

Το SPSS σε πολλές περιπτώσεις δεν επιτρέπει τον καθορισμό για το αν η υπόθεση που θα ελεγχθεί θα είναι μονόπλευρη ή αμφίπλευρη (μονής ή διπλής κατεύθυνσης), αλλά προχωρεί στον έλεγχο υπόθεσης διπλής κατεύθυνσης (2-tailed). Στις περιπτώσεις εκείνες όπου η διατύπωση υπόθεσης είναι μονής κατεύθυνσης, θα πρέπει να προσαρμοστεί κατάλληλα το επίπεδο στατιστικής σημαντικότητας.

Έτσι, εάν πρέπει να ελεγχθεί η υπόθεση:

$$\mathbf{H_0}: \mu = \mu_0$$

$$\mathbf{H_1}: \mu > \mu_0$$

Πρέπει να υπολογιστεί το:

$$p-value_{>} = \begin{cases} p-value_{\neq} / 2, & T(x) \geq 0 \\ 1 - p-value_{\neq} / 2, & T(x) < 0 \end{cases}$$

όπου $p\text{-value}_{\neq}$ είναι το p-value του αμφίπλευρου ελέγχου.

Ανάλογα πραγματοποιείται ο υπολογισμός για να ελεγχθεί η υπόθεση:

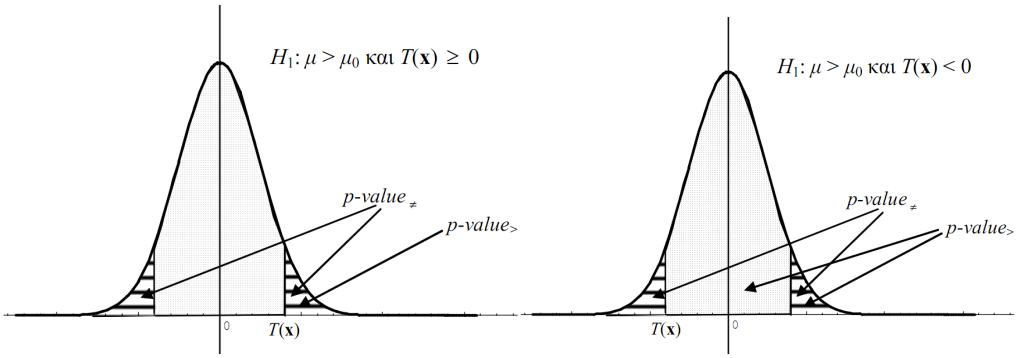
$$\mathbf{H_0}: \mu = \mu_0$$

$$\mathbf{H_1}: \mu < \mu_0$$

Οπότε πρέπει να υπολογιστεί το:

$$p-value_{<} = \begin{cases} p-value_{\neq} / 2, \; T(x) < 0 \\ 1 - p-value_{\neq} / 2, \; T(x) \geq 0 \end{cases}$$

Τα παραπάνω είναι εύκολο να επαληθευτούν και γραφικά, πχ. για $\mathbf{H_1}: \mu > \mu_0$:



Παραδείγματα Σύγκρισης Μέσων Τιμών (Διαδικασίας t-test) με SPSS

Παράδειγμα 1

Σε μία κλινική μελέτη για την αξιολόγηση ενός φαρμάκου για την αρτηριακή πίεση επιλέχθηκε τυχαία δείγμα 75 ασθενών με υπέρταση όπου και χωρίστηκαν τυχαία σε μια ομάδα placebo (40 άτομα) και σε μια ομάδα θεραπείας (35 άτομα). Η ομάδα placebo λάμβανε ένα χάπι χωρίς δραστική ουσία ημερησίως για δύο μήνες ενώ της θεραπείας ένα χάπι ημερησίως με τη νέα δραστική ουσία κατά της υψηλής πίεσης για το ίδιο χρονικό διάστημα. Μετρήθηκαν τόσο η συστολική όσο και η διαστολική πίεση όλων των ασθενών στην αρχή και στο τέλος της παρούσας φάσης της κλινικής μελέτης (όπως φαίνεται στον παρακάτω πίνακα).

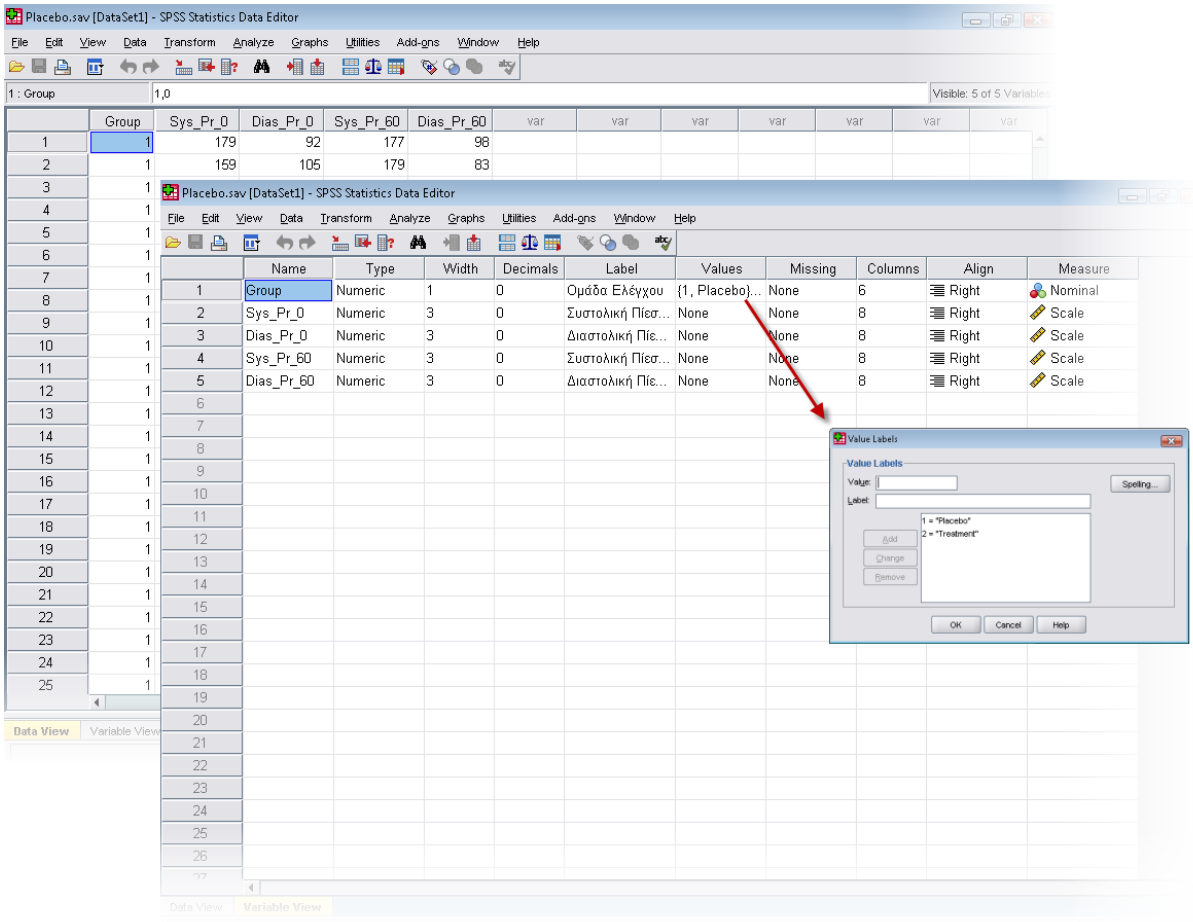
	Group	Sys_Pr_0	Dias_Pr_0	Sys_Pr_60	Dias_Pr_60
1	1	179	92	177	98
2	1	159	105	179	83
3	1	167	101	159	82
4	1	153	88	166	96
5	1	151	102	147	110
6	1	143	104	158	92
7	1	159	102	168	105
8	1	162	106	175	88
9	1	172	101	194	86
10	1	181	93	181	76
11	1	186	84	167	89
12	1	163	95	168	92
13	1	153	95	163	84
14	1	181	104	160	97
15	1	184	110	150	79
16	1	148	106	188	80
17	1	180	95	187	90
18	1	168	96	177	84
19	1	165	103	167	97
20	1	175	102	198	101
21	1	191	105	175	83
22	1	149	93	209	70
23	1	191	106	182	73
24	1	188	104	165	85
25	1	201	103	151	91
26	1	178	106	162	90
27	1	162	98	155	73
28	1	185	96	162	86
29	1	205	103	178	93
30	1	142	91	150	67
31	1	179	94	177	92
32	1	152	100	169	97
33	1	158	104	160	109
34	1	169	102	176	83
35	1	172	100	151	109
36	1	166	93	148	93
37	1	150	100	175	99
38	1	161	107	171	84
39	1	189	97	172	103
40	1	171	99	148	95
41	2	203	97	130	90
42	2	163	112	142	75
43	2	153	103	147	94
44	2	158	107	148	88
45	2	185	101	153	85
46	2	193	103	166	92
47	2	171	80	159	85
48	2	159	92	156	73
49	2	180	98	171	82
50	2	163	103	165	98
51	2	150	105	165	94
52	2	176	103	163	98
53	2	155	105	168	83
54	2	179	106	140	68
55	2	172	113	163	93
56	2	188	99	153	94
57	2	169	98	164	93
58	2	185	101	146	83
59	2	177	100	155	73
60	2	160	105	149	93
61	2	181	102	163	87
62	2	182	102	167	79
63	2	169	99	176	76
64	2	171	109	148	85
65	2	165	99	154	90
66	2	192	113	160	93
67	2	165	103	148	80
68	2	177	96	170	84
69	2	163	104	151	66
70	2	166	91	162	99
71	2	190	93	152	81
72	2	165	99	163	94
73	2	194	114	159	91
74	2	170	90	163	78
75	2	161	105	160	89

Σύγκριση μέσων τιμών ανεξάρτητων πληθυσμών στο SPSS

Είναι στατιστικά σημαντική η διαφορά της μέσης συστολικής και διαστολικής πίεσης στις δύο ομάδες ασθενών τόσο στην αρχή όσο και στο τέλος της παρούσας φάσης της κλινικής μελέτης (σε ε.σ. 5%);

Πραγματοποιείται η εισαγωγή των δεδομένων στον Data View έχοντας υπόψη ότι οι γραμμές του πίνακα αντιστοιχούν σε περιπτώσεις (στο παράδειγμα μας οι ασθενείς) και οι στήλες σε μεταβλητές (στο παράδειγμα μας η ομάδα, η συστολική και η διαστολική πίεση στην αρχή και στο τέλος της ΚΜ). Ο Data View και ο Variable View θα έχουν τη μορφή της παρακάτω εικόνας (βλ. Εικόνα 7.1).

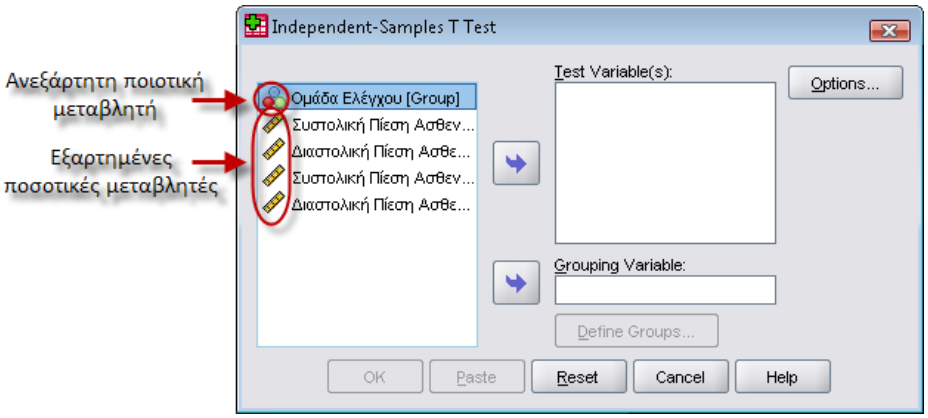
Στη συνέχεια, πραγματοποιείται η κωδικοποίηση της «Ομάδας» όπου ορίζονται οι «Placebo» με τον κωδικό '1' και οι «Θεραπείας» με τον κωδικό '2'.



Εικόνα 7.1: Οι Data και Variable View μετά την εισαγωγή των δεδομένων.

Η διαδικασία σύγκρισης των μέσων τιμών ανεξάρτητων πληθυσμών έχει ως εξής:

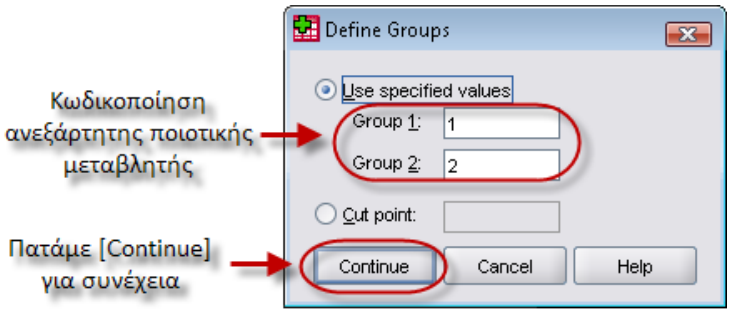
- Επιλογή από τη γραμμή μενού του [Analyze → Compare Means → Independent-Samples T Test…], οπότε και αναδύεται το παρακάτω παράθυρο διαλόγου (βλ. Εικόνα 7.2) όπου εμφανίζονται σε μία λίστα όλες οι μεταβλητές του αρχείου δεδομένων.



Εικόνα 7.2: Παράθυρο διαλόγου του Independent-Samples T Test.

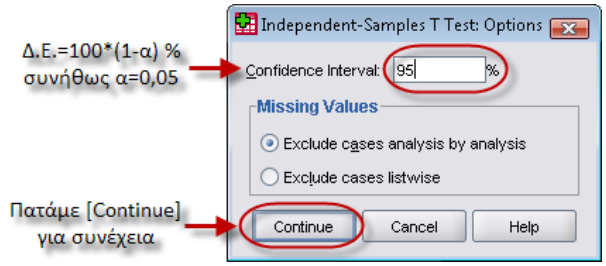
- Επιλογή μίας ή περισσότερων ποσοτικών μεταβλητών και μεταφορά τους στη λίστα [Test Variable(s)] (κάθε μία παράγει έναν έλεγχο). Στο παράδειγμα μας επιλέγονται και οι 4 ποσοτικές μεταβλητές.
- Επιλογή στη συνέχεια μίας κωδικοποιημένης ποιοτικής μεταβλητής και μεταφορά της στο [Grouping Variable] όπου εμφανίζονται δύο λατινικά ερωτηματικά δίπλα από το όνομα της μεταβλητής ομαδοποίησης. Στα ερωτηματικά αυτά πρέπει να αντιστοιχηθούν οι κωδικοί των ομάδων (αν δεν τους θυμόμαστε ανατρέχουμε στην προβολή [Variable View]).
- Πάτημα του κουμπιού [Define groups…] (βλ. Εικόνα 7.3) για τον προσδιορισμό των δύο ομάδων (ανεξάρτητων πληθυσμών). Εδώ οι επιλογές είναι δύο:
    - [Use specified values]: Είναι η εξ ορισμού επιλογή, σύμφωνα με την οποία εισάγονται οι κωδικοί των δύο ομάδων, ένας για το [Group 1] και ένας για το [Group 2], που αντιστοιχούν στις δύο κατηγορίες της μεταβλητής που διαχωρίζει σε δύο ανεξάρτητους πληθυσμούς τα δεδομένα. Στο παράδειγμα μας βάζουμε '1' που είναι η κωδικοποίηση της ομάδας Placebo και '2' για την ομάδα της Θεραπείας. Οι περιπτώσεις με άλλες τιμές αποκλείονται από την ανάλυση.
    - [Cut point]: Όλες οι περιπτώσεις με τιμές μεγαλύτερες ή ίσες της τιμής που ορίζεται ως cut point ανήκουν στον ένα πληθυσμό, ενώ οι υπόλοιπες περιπτώσεις ανήκουν στο δεύτερο πληθυσμό.

Στη συνέχεια πάτημα του κουμπιού [Continue].



Εικόνα 7.3: Παράθυρο διαλόγου του Define Groups.

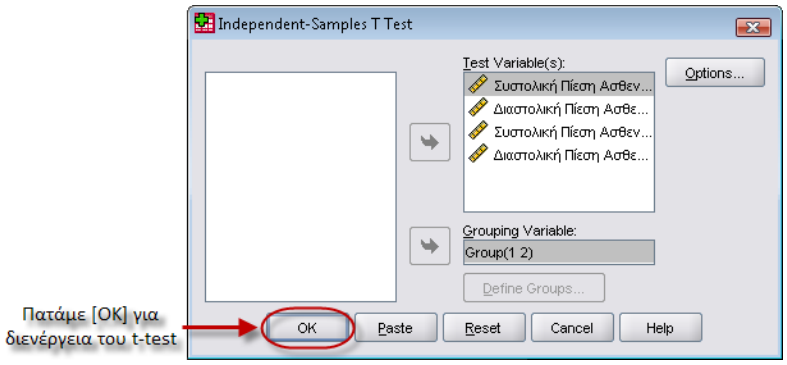
- Με την επιστροφή στο αρχικό παράθυρο της διαδικασίας, πάτημα του κουμπιού [Options…] προκειμένου να αλλαχθεί το επίπεδο σημαντικότητας (σφάλμα τύπου I) σύμφωνα με το οποίο θα εφαρμοστεί ο στατιστικός έλεγχος και να διαχειρισθούν οι ελλείπουσες τιμές (Missing Values) (βλ. Εικόνα 7.4).



Εικόνα 7.4: Παράθυρο διαλόγου του Options.

  - [Confidence Interval]: Το εξ ορισμού Διάστημα Εμπιστοσύνης είναι το 95% (δηλαδή υπάρχει 95% πιθανότητα το διάστημα εμπιστοσύνης να περιλαμβάνει την πραγματική και άγνωστη παράμετρο μ) για όλους τους ελέγχους άρα το επίπεδο σημαντικότητας θα είναι 95%=100*(1-α)% δηλαδή α=0,05. Υπάρχει η δυνατότητα να ορισθεί ένα διαφορετικό Διάστημα Εμπιστοσύνης άρα και επίπεδο σημαντικότητας εισάγοντας εδώ μία τιμή μεταξύ 1 και 99. Έτσι, αν για παράδειγμα το επιθυμητό σφάλμα τύπου I είναι 1%, πρέπει να εισαχθεί ο αριθμός 99.
  - [Missing values]:Είναι δυνατές οι παρακάτω επιλογές :
    - [Exclude cases analysis-by analysis]: Είναι η εξ ορισμού ρύθμιση, αποκλείει από κάθε ανάλυση που πραγματοποιείται τις περιπτώσεις (cases) όπου έχουμε ελλείπουσες τιμές σε μία συγκεκριμένη μεταβλητή.
    - [Exclude cases listwise]: Αποκλείει από όλες τις αναλύσεις τις περιπτώσεις (cases) όπου έχουμε ελλείπουσες τιμές σε μία συγκεκριμένη μεταβλητή.

- Πάτημα του κουμπιού [OK] στο παράθυρο διαλόγου της διαδικασίας Independent-Samples T Test για διενέργειά του (βλ. Εικόνα 7.5).



Εικόνα 7.5: Παράθυρο διαλόγου του Independent-Samples T Test στο τέλος της διαδικασίας.

Στους παρακάτω δύο πίνακες παρουσιάζονται τα αποτελέσματα του Independent-Samples T Test στο SPSS.

Group Statistics

	Ομάδα Ελέγχου	N	Mean	Std. Deviation	Std. Error Mean
Συστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	Placebo	40	169,70	15,875	2,510
	Treatment	35	172,86	13,007	2,199
Διαστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	Placebo	40	99,63	5,763	,911
	Treatment	35	101,43	7,018	1,186
Συστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	Placebo	40	169,13	14,403	2,277
	Treatment	35	157,11	9,872	1,669
Διαστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	Placebo	40	89,60	10,524	1,664
	Treatment	35	85,89	8,598	1,453

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Συστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	Equal variances assumed	1,696	,197	-,934	73	,354	-3,157	3,381	-9,896	3,582
	Equal variances not assumed			-,946	72,708	,347	-3,157	3,337	-9,808	3,494
Διαστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	Equal variances assumed	,176	,676	-1,222	73	,226	-1,804	1,476	-4,746	1,139
	Equal variances not assumed			-1,206	65,946	,232	-1,804	1,496	-4,790	1,183
Συστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	Equal variances assumed	3,594	,062	4,152	73	,000	12,011	2,893	6,245	17,776
	Equal variances not assumed			4,254	69,230	,000	12,011	2,823	6,379	17,643
Διαστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	Equal variances assumed	,984	,324	1,659	73	,101	3,714	2,239	-,749	8,177
	Equal variances not assumed			1,681	72,681	,097	3,714	2,209	-,689	8,118

Εάν το Sig. του Levene's Test είναι Sig.>=0,05 τότε κοιτάμε την πρώτη γραμμή

Εάν το Sig. του Levene's Test είναι Sig.<0,05 τότε κοιτάμε τη δεύτερη γραμμή

Στατιστικά σημαντική διαφορά

Ο πρώτος πίνακας δίνει περιγραφικά στοιχεία για τις δύο ομάδες που συγκρίναμε, δηλαδή δίνει ξεχωριστά το μέσο όρο, την τυπική απόκλιση και το τυπικό σφάλμα της μέσης τιμής για τις δύο ομάδες για κάθε εξαρτημένη μεταβλητή.

Ο δεύτερος πίνακας παρουσιάζει τα αποτελέσματα της ανάλυσης. Οι δύο πρώτες στήλες παρουσιάζουν τα αποτελέσματα του στατιστικού κριτηρίου Levene για την ισότητα των διακυμάνσεων των δύο ομάδων, που είναι και μία από τις προϋποθέσεις για την εφαρμογή των παραμετρικών στατιστικών κριτηρίων (και του κριτηρίου t). Για να εφαρμοστεί το t-test θα

πρέπει το αποτέλεσμα από το κριτήριο Levene να είναι στατιστικά μη σημαντικό (υποδηλώνοντας με αυτό τον τρόπο ότι οι δύο διακυμάνσεις είναι ίσες - δηλαδή το Sig. θα πρέπει να είναι μεγαλύτερο ή ίσο του 0,05, όπως συμβαίνει στο παράδειγμά μας), οπότε και διαβάζεται η πρώτη γραμμή. Αν το αποτέλεσμα του κριτηρίου Levene είναι στατιστικώς σημαντικό, τότε θα πρέπει να διαβαστεί η δεύτερη γραμμή του πίνακα (equal variances not assumed). Το SPSS κάνει μια διόρθωση των τιμών του t- test και σε αυτή την περίπτωση θα αναφερθεί: t(72,708)= -0,946, p=0,347. Επομένως, η διαφορά των μέσων όρων της Συστολικής Πίεσης των ασθενών στην αρχή της Κλινικής Μελέτης δεν είναι στατιστικώς σημαντική (ρ=0, 347). Στο παράδειγμα μας όλα τα του στατιστικά κριτήρια Levene είναι μη στατιστικώς σημαντικά οπότε και διαβάζονται οι πρώτες γραμμές. Στατιστικά σημαντική διαφορά των μέσων τιμών υπάρχει μόνο στην Συστολική Πίεση των ασθενών στο τέλος της Κλινικής Μελέτης όπου t(73)= 4,152, p<0,05, (η μηδενική υπόθεση απορρίπτεται), δηλαδή οι μέσοι των δύο πληθυσμών από τα οποία προήλθαν τα δύο δείγματα διαφέρουν στατιστικά σημαντικά σε επίπεδο στατιστικής σημαντικότητας α=5%. Ειδικότερα, οι ασθενείς της ομάδας Θεραπείας (M = 157,11, SD = 9,872) παρουσιάζουν μικρότερη Συστολική Πίεση στο τέλος της Κλινικής Μελέτης από ότι οι ασθενείς της ομάδας Placebo (M = 169,13, SD = 14,403), οπότε μπορούμε να υποθέσουμε με 95% βεβαιότητα ότι η μέση Συστολική Πίεση στο τέλος της Κλινικής Μελέτης της ομάδας Θεραπείας δεν είναι ίση με αυτήν της Placebo.

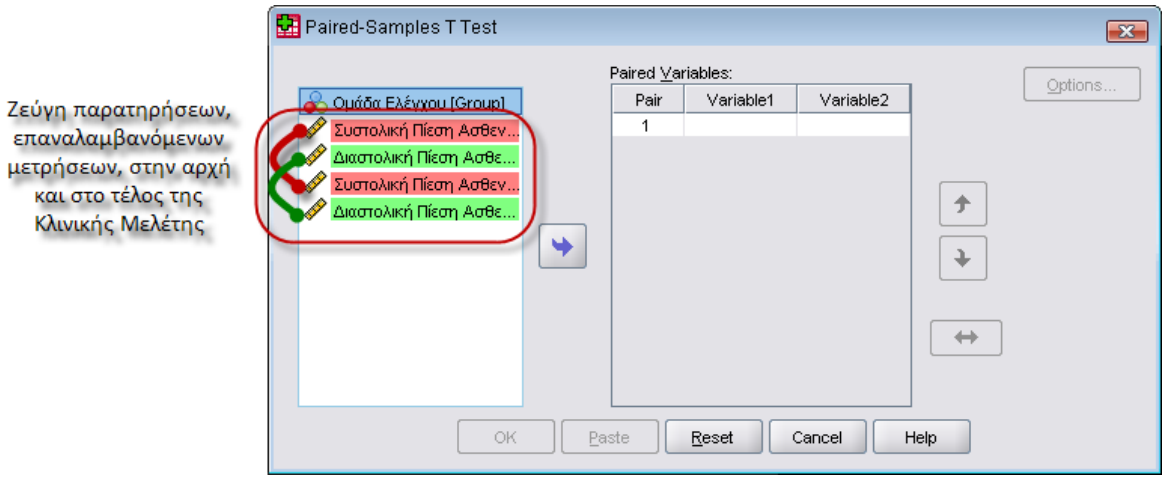
Σύγκριση μέσων τιμών ζευγών παρατηρήσεων (εξαρτημένων πληθυσμών) στο SPSS

Δημιουργείται ένα νέο αρχείο δεδομένων SPSS, από τα δεδομένα του παραδείγματος της προηγούμενης ενότητας, με την ομάδα Θεραπείας μόνο (δηλαδή ασθενείς που έλαβαν θεραπεία). Υπάρχει θετική μέση επίδραση του συγκεκριμένου σκευάσματος (σε ε.σ. 5%) στην αρτηριακή πίεση (συστολική και διαστολική) των ασθενών;

Η διαδικασία σύγκρισης των μέσων τιμών εξαρτημένων πληθυσμών έχει ως εξής:

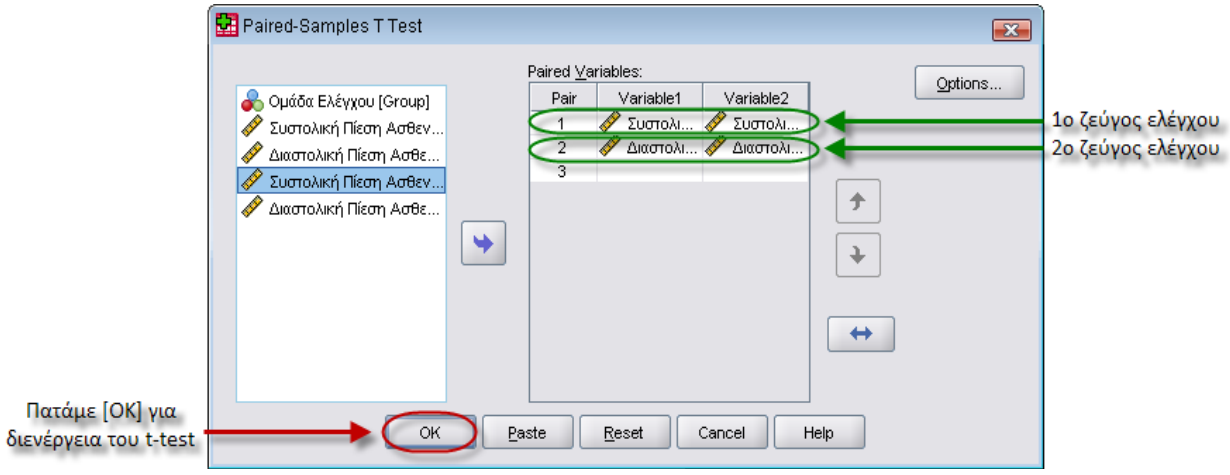
- Επιλογή από τη γραμμή μενού του [Analyze → Compare Means → Paired-Samples T Test …], οπότε και αναδύεται το παρακάτω παράθυρο διαλόγου (βλ. Εικόνα 7.6) όπου

εμφανίζονται σε μία λίστα όλες οι μεταβλητές του αρχείου δεδομένων, από τις οποίες επιλέγεται το ζεύγος ή τα ζεύγη που θα χρησιμοποιηθούν στην ανάλυση.



Εικόνα 7.6: Παράθυρο διαλόγου του Paired-Samples T Test.

- Επιλογή μίας ή δύο (ζεύγος) ποσοτικής/ές μεταβλητής/ές και μεταφορά τους στη λίστα [Paired Variables] στο πεδίο [Variable1] ή/και [Variable2] (κάθε ζεύγος παράγει έναν έλεγχο). Στο παράδειγμα μας δημιουργούνται δύο ζεύγη το Συστολική Πίεση στην αρχή και στο τέλος της ΚΜ και το Διαστολική Πίεση στην αρχή και στο τέλος της ΚΜ.
- Η επιλογή [Options...] είναι ακριβώς η ίδια με το Independent-Samples T Test.
- Πάτημα του κουμπιού [OK] στο παράθυρο διαλόγου της διαδικασίας Paired-Samples T Test για διενέργειά του (βλ. Εικόνα 7.7).



Εικόνα 7.7: Παράθυρο διαλόγου του Paired-Samples T Test στο τέλος της διαδικασίας.

Στους παρακάτω τρεις πίνακες παρουσιάζονται τα αποτελέσματα του Paired-Samples T Test στο SPSS.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Συστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	172,86	35	13,007	2,199
	Συστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	157,11	35	9,872	1,669
Pair 2	Διαστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	101,43	35	7,018	1,186
	Διαστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	85,89	35	8,598	1,453

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Συστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ & Συστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	35	-,136	,435
Pair 2	Διαστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ & Διαστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	35	,097	,578

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Συστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ - Συστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	15,743	17,369	2,936	9,777	21,709	5,362	34	,000
Pair 2	Διαστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ - Διαστολική Πίεση Ασθενή (mmHg) στο τέλος της ΚΜ	15,543	10,556	1,784	11,917	19,169	8,711	34	,000

Στατιστικά σημαντική διαφορά

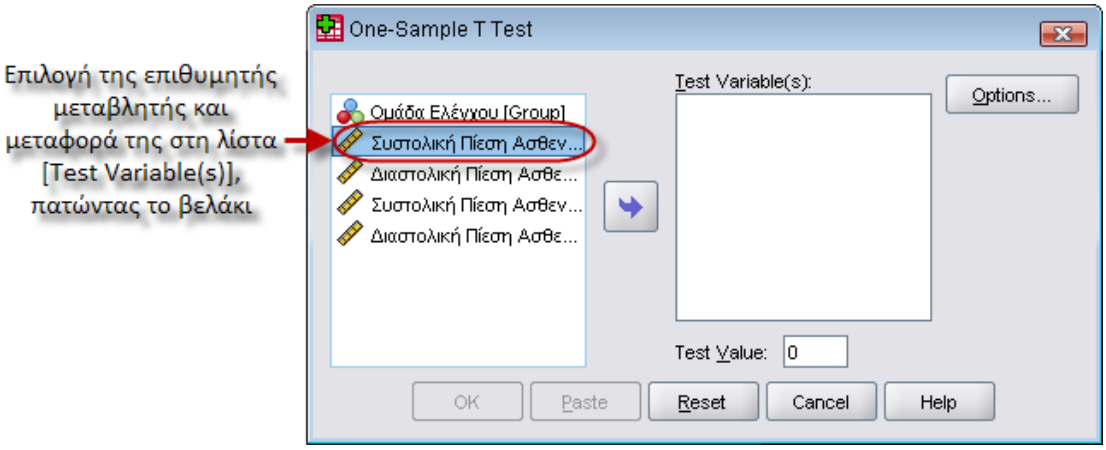
Ο πρώτος πίνακας δίνει περιγραφικούς δείκτες (οι μέσοι όροι και οι τυπικές αποκλίσεις είναι οι σημαντικότεροι) και ο δεύτερος τη συνάφεια για τα ζεύγη των μεταβλητών. Ο τρίτος πίνακας δίνει την τιμή του t (στήλη t), τους βαθμούς ελευθερίας (στήλη df) και το p-value [στήλη Sig.(2-tailed)]. Σύμφωνα με τους κανόνες μορφοποίησης της APA, το αποτέλεσμα αυτό θα παρουσιαστεί ως εξής: t(34)=5,362, p-value<0,.001 και t(34)=8,711, p-value<0,.001 αντίστοιχα για τα δύο ζεύγη. Η τιμή του p-value δείχνει ότι στο παράδειγμά μας υπάρχει στατιστικώς σημαντική διαφορά μεταξύ των μέσων όρων των δύο μεταβλητών και στα δύο ζεύγη άρα μπορούμε να υποθέσουμε με 95% βεβαιότητα ότι το φαρμακευτικό σκεύασμα έχει θετική μέση επίδραση στην αρτηριακή πίεση (συστολική και διαστολική).

Σύγκριση μέσης τιμής πληθυσμού με δεδομένη τιμή στο SPSS

Στα δεδομένα του παραδείγματος της πρώτης ενότητας χρειάζεται να ελεγχθεί εάν η μέση συστολική πίεση των ασθενών πριν την έναρξη της κλινικής μελέτης είναι ίση με 165 (σε ε.σ. 5%).

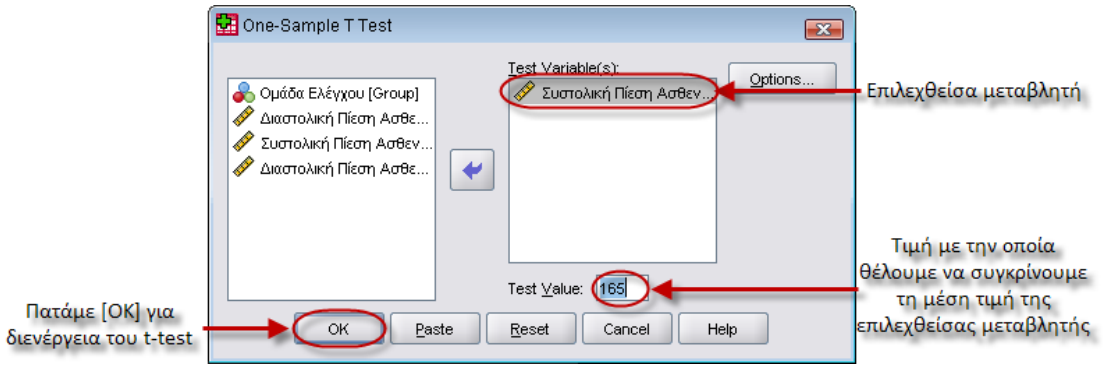
Η διαδικασία έχει ως εξής:

- Επιλογή από τη γραμμή μενού του [Analyze → Compare Means → One-Sample T Test …], οπότε και αναδύεται το παρακάτω παράθυρο διαλόγου (βλ. Εικόνα 7.8) όπου εμφανίζονται σε μία λίστα όλες οι μεταβλητές του αρχείου δεδομένων.



Εικόνα 7.8: Παράθυρο διαλόγου του One-Sample T Test.

- Επιλογή της επιθυμητής μεταβλητής, στο παράδειγμα μας τη συστολική πίεση ασθενών στην αρχή της μελέτης, και μεταφορά της στη λίστα [Test Variable(s)].
- Στο πεδίο [Test Value] γράφεται η τιμή με την οποία θα συγκριθεί η μέση τιμή της μεταβλητής που επιλέχτηκε.
- Η επιλογή [Options…] είναι ακριβώς η ίδια με το Independent-Samples T Test.
- Πάτημα του κουμπιού [OK] στο παράθυρο διαλόγου της διαδικασίας One-Sample T Test για διενέργειά του (βλ. Εικόνα 7.9).



Εικόνα 7.9: Παράθυρο διαλόγου του One-Sample T Test στο τέλος της διαδικασίας.

Στους παρακάτω δύο πίνακες παρουσιάζονται τα αποτελέσματα του Paired-Samples T Test στο SPSS.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Συστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	75	171,17	14,597	1,685

One-Sample Test

	Test Value = 165					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Συστολική Πίεση Ασθενή (mmHg) στην αρχή της ΚΜ	3,663	74	,000	6,173	2,81	9,53

Στατιστικά σημαντική διαφορά

Ο πρώτος πίνακας δίνει περιγραφικούς δείκτες (οι μέσοι όροι και οι τυπικές αποκλίσεις είναι οι σημαντικότεροι). Ο δεύτερος πίνακας δίνει την τιμή του t (στήλη t), τους βαθμούς ελευθερίας (στήλη df) και το p-value [στήλη Sig.(2-tailed)]. Σύμφωνα με τους κανόνες μορφοποίησης της APA, το αποτέλεσμα αυτό θα παρουσιαστεί ως εξής: t(74)=3,663, p-value<0,001. Η τιμή του p-value δείχνει ότι στο παράδειγμά μας υπάρχει στατιστικώς σημαντική διαφορά μεταξύ του μέσου όρου της εξαρτημένης μεταβλητής και της τιμής με την οποία συγκρίθηκε (0,001<0,05). Επομένως, μπορούμε να υποθέσουμε με 95% βεβαιότητα ότι η μέση συστολική πίεση στην αρχή της κλινικής μελέτης του πληθυσμού δεν μπορεί να είναι ίση με 165.

Το 95% διάστημα εμπιστοσύνης (2,81 , 9,53) αφορά στη μεταβλητή που προκύπτει αν αφαιρεθεί από κάθε τιμή του δείγματος η τιμή 165 (που ελέγχεται ως πιθανός πληθυσμιακός μέσος). Το διάστημα εμπιστοσύνης δεν περιλαμβάνει την τιμή 0 και, κατά συνέπεια, η διαφορά της συστολικής πίεσης στην αρχή της κλινικής μελέτης από το 165 δεν μπορεί να θεωρηθεί ότι μπορεί στον πληθυσμό να πάρει την τιμή 0: η μέση συστολική πίεση στην αρχή της κλινικής μελέτης του πληθυσμού συνεπώς δεν μπορεί να είναι 165. Επίσης, οι θετικές τιμές που περιέχει το διάστημα υποδηλώνουν ότι ο μέσος του πληθυσμού θα είναι μεγαλύτερος του 165.

Μονόπλευρος έλεγχος t-test (1-tailed) (ή μονής κατεύθυνσης)

Στα δεδομένα του παραδείγματος της πρώτης ενότητας χρειάζεται να ελεγχθεί εάν η μέση συστολική πίεση των ασθενών πριν την έναρξη της κλινικής μελέτης είναι ίση ή μεγαλύτερη με 165 (σε ε.σ. 5%).

Η υπόθεση είναι:

$$\mathbf{H_0}: \mu = 165$$

$$\mathbf{H_1}: \mu > 165$$

Δηλαδή: $p-\text{value}_{>} = p-\text{value}_{\neq} / 2 = 0,0004661 / 2 = 0,00023305$, διότι $T(x)=3,663 \geq 0$, οπότε και θα απορρίπτονταν η H_0.

Εάν η υπόθεση ήταν:

$$\mathbf{H_0}: \mu = 165$$

$$\mathbf{H_1}: \mu < 165$$

Τότε: $p-\text{value}_{<} = 1 - p-\text{value}_{\neq} / 2 = 1 - 0,0004661 / 2 = 0,9997$, διότι $T(x)=3,663 \geq 0$, οπότε και δεν θα απορρίπτονταν η H_0.

Στη συνέχεια ακολουθούν ορισμένα παραδείγματα που αφορούν βιολογικά ζητήματα που εμφανίζονται στην πράξη. Τα δεδομένα και τα αποτελέσματα των παραδειγμάτων που ακολουθούν, ελήφθησαν αυτούσια, απλοποιήθηκαν, τροποποιήθηκαν ή επεκτάθηκαν για τις ανάγκες του παρόντος από το βιβλίο: «Εφαρμοσμένη Ιατρική Στατιστική», Λαζαρίδης-Λαζαρίδου,1999, Αθήνα.

Παράδειγμα 2

Έχουμε ένα συγκεκριμένο δείγμα υγιών ατόμων στα οποία εξετάσαμε την απτογλοβίνη. Ο μέσος του δείγματος ισούται με μ=1,146, ενώ η διακύμανση, $\mathbf{S^2 = 0,00147.}$ Το 95% διάστημα εμπιστοσύνης του πραγματικού μέσου είναι τα **1,118-1,173.** Ζητείται να βρεθεί το κατάλληλο μέγεθος δείγματος ώστε το διάστημα εμπιστοσύνης να έχει πλάτος 0,01.

Απάντηση

Το παραπάνω ερώτημα είναι ιδιαίτερα σημαντικό και συναντάται συχνά σε περιπτώσεις προσδιορισμού του μεγέθους του δείγματος. Για να απαντήσουμε με απόλυτη ακρίβεια χρειάζεται να γίνει χρήση της κατανομής Student, μιας κατανομής διαφορετικής από την

κανονική. Για λόγους απλότητας, ας υποθέσουμε ότι το συγκεκριμένο δείγμα αποτελείται από περισσότερα των 30 ατόμων, οπότε μπορούμε ικανοποιητικά να αντικαταστήσουμε την συγκεκριμένη κατανομή με την «κανονική». Αρχικά παρατηρούμε ότι το εύρος του διαστήματος εμπιστοσύνης ισούται με (1,118-1,173) 0,055. Επομένως, αφού θέλουμε καλύτερη «στόχευση», χρειάζεται να μεγαλώσουμε το μέγεθος του δείγματος.
Εργαζόμαστε ως εξής:

1. **Υπολόγισε το τετράγωνο του επιθυμητού διαστήματος εμπιστοσύνης**

$0,01^2$=0,0001

2. **Διαίρεσε τη διακύμανση με το (1)**

0,00147/0,0001=**14,7**

3. **Πολλαπλασίασε το (2) με τον αριθμό 3,8416**

14,7*3,8416**= 56,47**

4. **Στρογγυλοποίησε στον αμέσως μεγαλύτερο ακέραιο**

57

Επομένως χρειαζόμαστε δείγμα μεγαλύτερο των 57 ατόμων για να έχουμε διάστημα εμπιστοσύνης της τάξης του 0,01. Ας σημειωθεί πάλι, ότι στην περίπτωση που το δείγμα ήταν μικρό (λιγότερο των 30 παρατηρήσεων), θα χρειαζόσασταν την Κατανομή Student, ο υπολογισμός όμως της οποίας είναι περισσότερο περίπλοκος για αυτό το εισαγωγικό στάδιο. Σε μια τέτοια περίπτωση, ο αριθμός του απαιτούμενου δείγματος θα ήταν ελαφρά μεγαλύτερος.

Παράδειγμα 3

Εξετάσαμε την ποσότητα Ιωδίου συνδεδεμένη με τις πρωτείνες (PBI) 16 ατόμων που θεωρήθηκαν υγιείς και πήραμε τις επόμενες τιμές.

487,5	473,3	480,7	505,1	499,7	519,8	480,3	485,2
500,1	470,3	503,1	504,9	499,4	500,2	494,7	488,3

Μπορούμε να δεχθούμε με πιθανότητα 95% ότι η μέση τιμή πληθυσμού είναι ίση με 500;

Απάντηση

Το παραπάνω ερώτημα αποτελεί καίριο ερώτημα στη Βιοστατιστική και εμπίπτει στο ευρύτερο τμήμα που ονομάζεται «έλεγχος υποθέσεων». Αυτός ο όρος υπονοεί ότι βάσει στατιστικών μεθόδων επιδιώκουμε να δεχθούμε ή να απορρίψουμε κάποια υπόθεση που διατυπώθηκε. Όπως έχει αναφερθεί, είναι σημαντικό να ορισθεί σωστά η υπόθεση μας. Η υπόθεση αυτή συμβολίζεται με Ho (hypothesis 0). Αν η υπόθεση αυτή αποδειχθεί αναληθής, τότε κάποια άλλη υπόθεση θα αντιπροσωπεύει την πραγματικότητα. Στην περίπτωση μας η εναλλακτική υπόθεση (H1) είναι: H1= Η μέση τιμή του πληθυσμού σε Ιώδιο δεν ισούνται με 500.

Επομένως:

Ho= Η μέση τιμή του πληθυσμού σε Ιώδιο ισούνται με 500

H1= Η μέση τιμή του πληθυσμού σε Ιώδιο δεν ισούνται με 500.

Σύμφωνα με το πρόγραμμα SPSS, παίρνουμε τα παρακάτω αποτελέσματα:

Test Value=500					
				95% Confidence Interval of the	
t	df	**sig (2-tailed)**	Mean Difference	Lower	Upper
-2,029	15	**0,061**	-6,7125	**-13,7644**	**0,3394**

Επίσης, με την βοήθεια υπολογιστή ή με την βήμα προς βήμα διαδικασία, υπολογίζουμε ότι: Mean=493,2875 και S=13,234.

Όπως έχουμε αναφέρει, βασική προϋπόθεση για την αντιμετώπιση του προβλήματος, είναι η σωστή κατανόηση της Ho. Στο συγκεκριμένο παράδειγμα, για λόγους απλότητας, η υπόθεση αυτή είναι ξεκάθαρη και έχει διατυπωθεί σαφώς για έναν μη στατιστικό επιστήμονα.

Για λόγους οπτικής βοήθειας, τα σημαντικά που πρέπει να γνωρίζουμε, έχουν τονισθεί με **bold** στον πίνακα αποτελεσμάτων.

Το δεύτερο που πρέπει να γνωρίζουμε είναι η πιθανότητα η υπόθεση μας να είναι αληθής. Η πιθανότητα αυτή ισούται με 0,061 (sig-2 tailed). Επειδή το 0,061 η 6,1% είναι μεγαλύτερο από το 0,05 ή 5%, μπορούμε να αποδεχθούμε την υπόθεση μας. Έτσι συμπεραίνουμε, ότι η μέση τιμή του Ιωδίου ανέρχεται στα 500.

Το τρίτο που θέλουμε να ξέρουμε είναι το διάστημα εμπιστοσύνης. Σύμφωνα με τα αποτελέσματα, το διάστημα αυτό δεν είναι άμεσα εμφανές διότι στα αποτελέσματα παίρνουμε:

Lower	Upper
-13,7644	**0,3394**

Αυτό απλά σημαίνει ότι το κάτω άκρο του διαστήματος εμπιστοσύνης είναι 13,7644 μονάδες κάτω από την υπόθεση που κάναμε (Ho=500) και το άνω άκρο, 0,3394 μονάδες άνω της Ho. Ήτοι, το 95% διάστημα εμπιστοσύνης είναι το (486,23 έως 500,34).

Ένας πρακτικός, οπτικός κανόνας που διευκολύνει κάποιον και δίνει μεγαλύτερη κατανόηση των αποτελεσμάτων, είναι ότι όταν το lower και το Upper, έχουν το ίδιο πρόσημο, τότε η αρχική μας υπόθεση δεν ισχύει. Και αυτό διότι αν και τα δύο είναι αρνητικά, σημαίνει ότι με πιθανότητα 95%, το διάστημα εμπιστοσύνης είναι μικρότερο της τιμής που θέσαμε ως υπόθεση και δεν την περιλαμβάνει εντός του. Το αντίστοιχο ισχύει αν και τα δύο είναι θετικά. Τότε σημαίνει ότι διάστημα εμπιστοσύνης περιέχει τιμές μεγαλύτερες της τιμής που θέσαμε ως υπόθεση και δεν την περιλαμβάνει, επίσης.

Παράδειγμα 4

Σε έξι άτομα εφαρμόσθηκε μια θεραπευτική αγωγή για την ρύθμιση του μαγνησίου του ορού. Μετρήθηκε το μαγνήσιο πριν και μετά την αγωγή και ελήφθησαν τα αποτελέσματα.

Πριν	2,12	1,75	2,42	1,53	1,1	1,7
Μετά	1,83	1,62	2,33	1,4	0,75	1,71

Ζητείται να διαπιστωθεί αν η εφαρμογή της θεραπευτικής αγωγής συντέλεσε στην μείωση της ποσότητας Μαγνησίου.

Απάντηση

Η υπόθεση που εξετάζουμε καθώς και η εναλλακτική σε αυτή την περίπτωση είναι:

Ho= Η μέση τιμή της ποσότητας ασβεστίου είναι ίση πριν και μετά την αγωγή
H1=Η μέση τιμή της ποσότητας ασβεστίου είναι μεγαλύτερη πριν την αγωγή

Είναι προφανές ότι η Η1 είναι η λογική εναλλακτική σε αυτή την περίπτωση, καθώς δεν θα αναμέναμε η μέση τιμή της ποσότητας ασβεστίου να είναι μεγαλύτερη μετά την θεραπεία. Προφανώς ή θα ήταν ίση (οπότε η θεραπεία θα ήταν αναποτελεσματική) ή θα ήταν μικρότερη (οπότε η θεραπεία θα ήταν αποτελεσματική).
Από το πρόγραμμα SPSS παίρνουμε τα παρακάτω αποτελέσματα:

Paired sample Test					
				95% Confidence Interval of the	
t	df	**sig (2-tailed)**	Mean Difference	Lower	Upper
3,007	5	**0,03**	0,1633	**0,02369**	**0,303**

Σύμφωνα με τα αποτελέσματα η υπόθεση μας έχει πιθανότητα να είναι αληθής 0,03 ή 3%. Βάσει αυτού μπορούμε να την απορρίψουμε, γιατί δεν φτάνει το 5%. Οπότε οι δύο θεραπείες δεν έχουν τους ίδιους μέσους. Άρα η δεύτερη θεραπεία είναι πράγματι αποτελεσματική και μείωσε την ποσότητα μαγνησίου.
Το διάστημα εμπιστοσύνης της διαφοράς των μέσων ανέρχεται μεταξύ 0,02369 και 0,303 με πιθανότητα 95%. Αυτό σημαίνει ότι πριν τη θεραπεία η ποσότητα μαγνησίου ήταν αυξημένη σε σχέση με ύστερα, από 0,02369 έως 0,303 με πιθανότητα 95%.

Όπως και στο προηγούμενο παράδειγμα το άνω και κάτω όριο του αληθινού διαστήματος εμπιστοσύνης είναι ομόσημα (και τα δύο θετικά) οπότε η υπόθεση δεν μπορεί να είναι

αληθής. Παρατηρήστε όμως ότι σε ένα άλλο επίπεδο πιθανότητας (1%) η υπόθεση θα γίνονταν αποδεκτή. Δηλαδή το 3% είναι μεγαλύτερο του 1%, οπότε θα μπορούσαμε να αποδεχθούμε την υπόθεση. Σε αυτή την περίπτωση τα όρια του διαστήματος εμπιστοσύνης θα ήταν μεγαλύτερα και οι ακραίες τιμές του διαστήματος θα ήταν ετερόσημες (η μια θετική και η άλλη αρνητική) για να υπακούουν στον εμπειρικό κανόνα στον οποίο αναφερθήκαμε.

Σημαντική σημείωση: Σε πολλές περιπτώσεις, τα αριθμητικά αποτελέσματα που εξάγονται είναι διαφορετικά από πρόγραμμα σε πρόγραμμα. Αυτό συμβαίνει διότι δεν έχουν όλα τα προγράμματα τον ίδιο βαθμό ακρίβειας, αλλά και ο τρόπος με τον οποίο υπολογίζουν ορισμένα μεγέθη δεν είναι πάντα ο ίδιος. Πχ η τυπική απόκλιση, θα μπορούσε να μετρηθεί με αρκετούς διαφορετικούς τρόπους, ανάλογα με τις υποθέσεις που πραγματοποιούνται. Το ίδιο συμβαίνει και με την ασυμμετρία. Ενώ η έννοια της «ασυμμετρίας» ή η έννοια της «σκέδασης» είναι κατανοητή, ο τρόπος που υπολογίζονται διαφέρει. Για μια πλήρη κατανόηση των διαφορετικών μεθόδων υπολογισμού, απαιτείται γνώση μαθηματικών και στατιστικής. Για τις δικές σας ανάγκες, απλά είναι απαραίτητο να γνωρίζετε την μέθοδο που χρησιμοποιείται κάθε φορά από το υπολογιστικό μας πρόγραμμα και να ανατρέχετε στην βοήθεια. Εκεί συνήθως δίνονται οι πλήρεις μαθηματικοί τύποι που χρησιμοποιούνται. Σε περίπτωση που κάτι δεν σας είναι κατανοητό, καλύτερα θα ήταν να συμβουλευθείτε κάποιον ειδικό που έχει πιο προχωρημένες γνώσεις στατιστικής. Μια άλλη ασφαλής λύση είναι να αναφέρεται το πρόγραμμα από το οποίο εξάγετε τα αποτελέσματά σας, ώστε αυτά να μπορούν να επιβεβαιωθούν αν παρουσιασθεί ανάγκη.

Βιβλιογραφία

1. Πεντόγαλος ΓΗ. Εισαγωγή στην Ιστορία της Ιατρικής. Παρατηρήσεις. Θεσσαλονίκη 1983.

2. Segar I. Knaurs Buch der Modernen Soziologie. Ed. Droemershe Verlagsanstalt Thl. Knaur Nachf. Munchen. Zurich (μετάφρ. στα Ελληνικά Μαστοράκη Τ.) σσ. 39-52. Εκδ. Μπουκουμάνη. Αθήνα 1977.

3. Kline M. Mathematics in Western culture. Oxford University Press. Oxford. (μετ. στα Ελληνικά Μαρκέτος Σ.) σσ. 152-224. Εκδ. Κώδικας. Θεσσαλονίκη 1981.

4. Κατσουγιαννόπουλος ΒΧ. Βασική Ιατρική Στατιστική. Εκδ. Αδελφοί Κυριακίδη, Θεσσαλονίκη 1993.

5. Hald A. A history of probability and statistics and their applications before 1750. Ed. John Wiley and Sons. New York 1990.

6. Κατσουγιαννόπουλος ΒΧ. Υγιεινή και Κοινωνική Ιατρική. Κοινωνική Ιατρική. Τόμ. 2. Εκδ. Αδελφοί Κυριακίδη, Θεσσαλονίκη 1994.

7. Armitage P and Berry G. Statistical methods in medical research. Second Edition. Ed. Blackwell scientific publications. Oxford 1987.

8. McPherson G. Statistics in Scientific Investigation. Its basis, Application and Interpretation. Ed. Springer-Verlag. New York 1990.

9. Kramer MS. Clinical Epidemiology and Biostatistics. A primer Clinical Investigators and Decision-Makers. Ed. Springer-Verlag. Berlin 1988.

10. Altman DG. Practical Statistics for Medical Research. Ed. Chapman end Hill. London 1992.

11. Σπάρρος Λ. Αξιολόγηση κλινικών και εργαστηριακών ευρημάτων. Εκδ. Εργαστήριο Υγιεινής και Επιδημιολογίας Ιατρικής Σχολής Πανεπιστημίου Αθηνών. Αθήνα 1987.

12. Σπάρρος Λ. Σφάλματα μέτρησης. Διαγνωστικές πιθανότητες. Εισήγηση στο σεμινάριο: Μεθοδολογία της Έρευνας. Τόμος Πρακτικών. Ιωάννινα 1992.

13. Τριχόπουλος Δ. Ιατρική Στατιστική. Αρχαί και βασικαί μέθοδοι Βιο-Ιατρικής Στατιστικής. Εκδ. Παρισιάνος, Αθήνα 1975.

14. Lwanga SK and Lemeshow S. Sample size determination in health studies. A practical manual. Ed. World Health Organization. Geneva 1991.

15. Daly LE, Bourke GJ and McGilvray J. Interpretation and uses of medical statistics. Fourth Edition. Ed. Blackwell Scientific Publications. Oxford 1991.

16. Everitt BS. Statistical methods for medical investigations. Ed. Oxford University Press. New York. Edward Arnold London.

17. Sokal RR and Rohlf FJ. Biometry. Second edition. W.H. Freeman and Co. San Francisco 1981.

18. Moore DS. Η Στατιστική: Η Επιστήμη των Δεδομένων. Κεφ. Στο Consortium for Mathematics and its applications. Τα σύγχρονα Μαθηματικά στη ζωή μας. σσ107-210. Εκδ. W.H. Freeman and Co., Γιαλλέλης, Μανωλάκης. Αθήνα 1990.

19. Παπαευαγγέλου Γ. Πόσο απαραίτητη είναι η βιοστατιστική στην ιατρική. Ιατρική 1978, 34:563-564.

20. Dougherty CM and Bur RL. Comparison of the heart rate variability in survivors and nonsurvivors od sudden cardiac arrest. Am J Cardiol 1992, 70:436-47.

21. Αθυρος Β και Κοντόπουλος Α. Η μεταβλητότητα της καρδιακής συχνότητας. Ελλην. Ιατρ. 1993, 59(6): 463-469.

22. Κατσουγιαννόπουλος ΒΧ. Η επιδημιολογική θεώρησις του αίτιου και αποτελέσματος. Επιστημονική Επετηρίδα Ιατρικής Σχολής ΑΠΘ. 1977, 197-205, Θεσσαλονίκη.

23. Spilker B. Bias and Confounding Factors. Chap. in Guide to Clinical Trials. Pp. 21-26. Ed. Raven Press. New York 1991.

24. Κούβελας Η. Πρακτικά και θεωρητικά προβλήματα στην έρευνα για το νευρικό σύστημα. Επ. Κοιν. Ερ. 1982, 44-47: 74-82

25. Αναστασόπουλος Π, Χουλιάρα Σ και Δημολιάτης Γ. Οι αιτιακές σχέσεις υπό το φως των θεωριών του Χάους. Εισήγηση στο σεμινάριο: Μεθοδολογία της Έρευνας. Τόμος Πρακτικών. Ιωάννινα 1992.

26. Spilker B. Clinical Significance versus Statistical Significance of Abnormal Data. Chap. in Guide to Clinical Trials. pp. 536-542. Ed. Raven Press. New York 1991.

27. Ahlborm A and Norell S. Introduction to Modern Epidemiology. Ed. Epidemiology Resources Inc. 1990.

28. Τριχόπουλος Δ. Επιδημιολογία. Αρχές, Μέθοδοι, Εφαρμογές. Εκδ. Παρισιάνος, Αθήνα 1982.

29. Feinstein AR. Clinical Biostatistics. On the sensitivity, specificity and discrimination of diagnostic tests. Clinical Pharmacology and Therapeutics. 1975, 17 (1): 104-116.

30. Μπένος ΑΣ. Εισαγωγή στην Κοινωνική Ιατρική. Εκδ. Εργαστήριο Υγιεινής Ιατρικού Τμήματος Α.Π.Θ. Θεσσαλονίκη 1994.

31. Chambers JM, Cleveland WS, Kleiner B. and Turkey PA. Graphical methods for data analysis, In ser. Statistics – Probability series. Bell Telephone Laboratories Inc. Ed. Wadsworth and Brooks – Cole Publishing Company. Advanced Books and Software. Pacific Grove, California 1983.

32. Moses LE. Graphical methods in statistical analysis, Annual Review of Public Health 1987, 8: 309-353.

33. Hennekens CH and Buring JE. Selection of an Appropriate Test of Statistical Significance. Appendix in Epidemiology in Medicine. Ed. Little, Brown and Company. Boston. Toronto 1987.

34. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986, I: 307-310.

35. Ilstrup DM. Statistical Methods in Microbiology. Clin Microbiol Rev 1990, 3(3): 219-226

36. Burton AH, Dean JA and Dean AG. Software for data management and analysis in epidemiology. World Health Forum 1990, 11 (1): 75-77.

37. Maloney JV jr. The trouble with patient monitoring. Ann. Surg. 1968, 168: 605-610.

38. Rose J. Computers in medicine. Proceedings fo the Second Symposium held at Blackburn College of Technology and Design. Ed. John Wright and sons Ltd. Bristol 1972.

39. Smith PK. Computer applications in cardiothoracic surgery. Chap. In : Sabiston DC jr and Spencer FC (eds.), Surgery of the chest. Pp. 317-334. Ed. W.B. Saunders 1990.

40. Piele DT. Introductory Statistics with Spreadsheets. Ed. Addison –Wisley. 1990

41. Gardner MJ and Altman DG. Statistics with Confidence. Confidence Intervals and statistical guidelines. Ed. British Medical Journal. London 1989.

42. Gardner MJ. Gardner SB and Winter PD, Confidence Interval Analysis (CIA). Microcomputer Program Manual. Version 1.0. Ed. British Medical Journal. London 1989.

43. Hintze JL. Number Cruncher Statistical System Manual. Version 5.5 (2/90). Kaysville, Utah 1990.

44. Elliot AC. Kwikstat. Condensed Version of Manual. Version 1.0. Ed. Mission Technologies. Cedar Hill. Texas 1986.

45. Dean AG, Dean JA, Burton AH and Dicker RC. Epi Info. Ver. 5.A Word Processing, Database and Statistics System for Epidemiology on Microcomputers. Ed. Center of Disease Control , Anlanta, Georgia. World Health Organization, Geneva 1990.

46. Norusis MJ. SPSS for Windows, User's Guide. Ed. SPSS Inc. Chicago 2007.

47. Κατσουγιάννη Κ. Οι ηλεκτρονικοί υπολογιστές στην Ιατρική στατιστική και επιδημιολογία: Στατιστικά πακέτα-χρήση και δυνατότητες. Εισ. Στο 1ο Πανελλήνιο Συμπόσιο Εφαρμογών Ηλεκτρονικών Υπολογιστών. Εκδ. Ελληνική Ακτινολογική Εταιρεία. Αθήνα 1984.

48. Κωνσταντινίδης Θ.Κ. Η συμβολή της στατιστικής μεθοδολογίας στην Ιατρική. Εισ. στο Συμπόσιο: Ιατρική Έρευνα Ελληνική Ιατρική 1994, 60 (Παράρτ. 1) : 27-37

Εργαστήριο Υγιεινής και Προστασίας Περιβάλλοντος
Τμήματος Ιατρικής Δημοκρίτειου Πανεπιστημίου Θράκης