ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι ΘΕΣΣΑΛΟΝΙΚΗΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΩΝ ΕΦΑΡΜΟΓΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

Το Στατιστικό Πρόγραμμα Ελεύθερου Λογισμικού PSPP: Σχεδιασμός Εγχειριδίου Λειτουργίας και Παραγωγή Εκπαιδευτικού υλικού

Φοιτητής: Αθανάσιος Τριανταφύλλου

Αρ.Μητρώου: 05/2971

Επιβλέπων Καθηγητής: Βασίλης Κώστογλου

Θεσσαλονίκη 2013

Περιεχόμενα

1. Εισαγωγή

Η στατιστική αποτελεί αναμφισβήτητα τον ακρογωνιαίο λίθο για την επιστημονική έρευνα, βασιζόμενη στην παρατήρηση φαινομένων προκειμένου να συλλέξει και να αναλύσει πληροφορίες και μετρήσεις που προέρχονται από αυτές, και στη συνέχεια να προχωρήσει στην ερμηνεία των αποτελεσμάτων με χρήσιμα και γενικεύσιμα συμπεράσματα. Οι ερευνητές διεξάγουν μελέτες ή έρευνες, με σκοπό να δώσουν απαντήσεις σε συγκεκριμένα ερωτήματα που σχετίζονται με μια συγκεκριμένη ομάδα ή ομάδες ατόμων ή αντικειμένων, οι λεγόμενες οντότητες.

Από τα παραπάνω, προκύπτει η ανάγκη για τη χρήση των βέλτιστων εργαλείων, που παρέχουν τις καλύτερες δυνατότητες επεξεργασίας και ανάλυσης των δεδομένων, των λεγόμενων στατιστικών πακέτων. Ένα εξ αυτών είναι και το PSPP, διαδεδομένο λόγω της ελεύθερης διανομής του, την δωρεάν δηλαδή απόκτησή του αλλά την απαγόρευση της εμπορικής εκμετάλλευσης ή της ενσωμάτωσής του κώδικά του σε αυτόν άλλων προγραμμάτων. Πίσω από την κατασκευή του βρίσκεται η ομάδα του GNU project, με μαζική παραγωγή προγραμμάτων ελευθέρου κώδικα, ενώ η πρώτη του εμφάνιση έγινε το 1998. Από τότε ακολούθησαν περισσότερες από 17 εκδόσεις του, προσθέτοντας πάντα περισσότερα χαρακτηριστικά και διορθώνοντας προηγούμενα. Η υλοποίησή του έγινε σε γλώσσα C, πράγμα που διευκόλυνε στην κατασκευή του. Παρέχει τη δυνατότητα χρήσης του γραφικού του περιβάλλοντος (psppire) αλλά και της παραδοσιακής γραμμής εντολών (PSPP).

Με μια γρήγορη επισκόπηση, θα μπορούσαμε να αναφέρουμε τις βασικές ικανότητες του PSPP: μπορεί να εκτελέσει πολλούς μετασχηματισμούς δεδομένων συμπεριλαμβανομένης της μέτρησης(count), του recode, της στάθμισης(weighting), το χειρισμό των ελλειπουσών τιμών, τον υπολογισμό της περιγραφικής στατιστικής, τον υπολογισμό των crosstabs, των T-tests (ανεξάρτητα δείγματα T-test, ζεύγη δειγμάτων T-test και μονό T-test), της ANOVA, της γραμμικής παλινδρόμησης (linear regression), της διμεταβλητής συσχέτισης (bivariate correlation), της ανάλυσης παραγόντων(factor analysis), της εύρεσης των περιθωρίων αξιοπιστίας, και ορισμένων μη παραμετρικών τεστ (Chi-square και διωνυμικό) και της ROC καμπύλης, πολλά από τα οποία θα αναλυθούν παρακάτω. Ο αριθμός των στατιστικών υπολογισμών που μπορεί να εκτελέσει διαρκώς

αλλάζει και προσαρμόζεται. Το PSPP σχεδιάστηκε για να είναι εξαιρετικά ελαφρύ και γρήγορο, τόσο στην εγκατάσταση όσο και στη διαχείρισή του.

Στους ανταγωνιστές ανήκουν πολλές άλλες εφαρμογές οι οποίες έχουν η κάθε μία τα δικά της προτερήματα. Ενδεικτικά αναφέρουμε το R project, το οποίο είναι και αυτό ελεύθερο λογισμικό, στηρίζεται όμως στη γλώσσα R, και τείνει να επιβληθεί σε μεγάλο εύρος χρηστών. Άλλοι πολύ μεγάλοι ανταγωνιστές είναι το SPSS, το οποίο και θεωρείται από πολλούς η πρωτότυπη εφαρμογή και ανήκει στην IBM, και επιπλέον, το Statgraphics, το Statistical, το Splus κ.ά.

Στις επόμενες παραγράφους δε θα ασχοληθούμε με την στατιστική μεθοδολογία. Συγκεκριμένα, στο δεύτερο κεφάλαιο κάνουμε λόγο για τις βασικές έννοιες της στατιστικής, στο τρίτο ακολουθεί μια περιγραφή του πακέτου PSPP με αναφορά στα πλεονεκτήματα και τα χαρακτηριστικά του. Το γραφικό περιβάλλον και οι εντολές του PSPP περιγράφονται αναλυτικά στο τέταρτο κεφάλαιο, ενώ στο επόμενο γίνεται αναλυτική επίλυση μιας μελέτης περίπτωσης. Στο έκτο και τελευταίο κεφάλαιο υπάρχει ένας αριθμός ασκήσεων προς επίλυση για εξάσκηση με το στατιστικό πακέτο.

2. Βασικές Έννοιες της Στατιστικής

2.1 Τι είναι στατιστική;

Πρώτα ας κάνουμε μια διευκρίνιση: Ως *δεδομένα* (data) μπορούμε να ορίσουμε τα γεγονότα ή τις παρατηρήσεις που μπορούν να καταγραφούν. Δηλαδή, είναι τιμές κάποιων χαρακτηριστικών που ανήκουν σε συγκεκριμένες οντότητες. Τα δεδομένα για να είναι χρήσιμα πρέπει να έχουν ακρίβεια, πληρότητα, σχετικότητα και διαθεσιμότητα. Η *πληροφορία* (information) είναι τα δεδομένα που έχουν επεξεργαστεί και έχουν μορφή ερμηνεύσιμη και χρήσιμη στους τελικούς χρήστες έτσι ώστε να αποκτούν νόημα και αξία.

Γενικά τον όρο στατιστική μπορούμε να ορίσουμε ως:

- *την επιστήμη της συλλογής, της οργάνωσης, της ερμηνείας και παρουσίασης των δεδομένων του πραγματικού κόσμου.*

Είναι πολλές οι φορές που η στατιστική κλήθηκε να λύσει προβλήματα βασιζόμενα σε κοινωνικά δεδομένα, ιατρικές έρευνες ή στη βιομηχανική έρευνα. Πλέον όμως η χρήση της είναι πολύ πιο ευρεία.

Σαν υπόβαθρο, χρησιμοποιεί την επιστήμη της θεωρίας πιθανοτήτων, ένας κλάδος ακόμα των μαθηματικών, προκειμένου να εκφράσει τη στατιστική συμπερασματολογία.

Ένα από τα βασικά προβλήματα σε ένα πλήθος επιστημονικών πεδίων που έρχεται να επέμβει η στατιστική είναι ότι συνήθως το πλήθος των δεδομένων τους είναι τόσο μεγάλο που το ανθρώπινο μυαλό αδυνατούσε να επεξεργαστεί. Έτσι και προέκυψε η ανάγκη για αντικατάσταση του αρχικού πλήθους των δεδομένων από άλλα λιγότερα χωρίς όμως να χάνεται η αρχική πληροφορία ή καλύτερα απαιτούνταν η εύρεση μετρικών και στατιστικών που θα αντικατόπτριζαν τη σύσταση, την ποιότητα και τις ιδιότητες του αρχικού συνόλου δεδομένων.

Περαιτέρω έννοιες:

- Πληθυσμός (population) είναι το σύνολο στο οποίο αναφέρεται η έρευνα όπως ορίζεται από τους σκοπούς της ίδιας της έρευνας. Είναι εμφανές ότι ένας πληθυσμός μπορεί να είναι από πολύ μικρός (π.χ. οι έφηβοι σε μία τάξη ενός σχολείου), έως πολύ μεγάλος (π.χ. το σύνολο όλων των εφήβων του πλανήτη). Τα υποκείμενα που εξετάζονται εδώ είναι οι έφηβοι.

- Το *δείγμα (sample)* είναι ένα υποσύνολο του πληθυσμού που πρέπει να αντιπροσωπεύει ολόκληρο τον πληθυσμό.
- Η *δειγματική μονάδα(sample unit)* είναι η βασική μονάδα της έρευνας.
- *Απογραφή (Census)* είναι η καταγραφή ολόκληρου του πληθυσμού.

Ένα δείγμα είναι *αντιπροσωπευτικό* του πληθυσμού όταν έχει χρησιμοποιηθεί η διαδικασία της *τυχαίας δειγματοληψίας* (random sampling) για την απόκτηση του. Η τυχαία δειγματοληψία απαιτεί κάθε υποκείμενο του πληθυσμού να έχει την ίδια πιθανότητα να επιλεγεί. Και εδώ, ένα δείγμα μπορεί να είναι από πολύ μικρό έως και πολύ μεγάλο: όσο μεγαλύτερο είναι το δείγμα, τόσο πιο αντιπροσωπευτικό θα είναι, καθώς αυξάνεται ο αριθμός των υποκειμένων που επιλέγονται από τον πληθυσμό. Ο κλάδος της στατιστικής που με αφετηρία το δείγμα, προσπαθεί να βγάλει συμπεράσματα για τον πληθυσμό, αποτελεί τη *συμπερασματική στατιστική*.

Ανεξάρτητα από την ευκολία που προσφέρει, και όσο αντιπροσωπευτικό και αν θεωρείται ένα οποιοδήποτε δείγμα, δε μπορεί να πετύχει την πιστή αναπαράσταση του πληθυσμού. Έτσι, παίρνουμε σαν δεδομένο ότι θα υπάρχει μία παρέκκλιση, ή με άλλα λόγια ένα ποσοστό λάθους, που αποδεικνύεται αν συγκρίνουμε το στατιστικό δείκτη που προέρχεται από το δείγμα, με την αντίστοιχη τιμή της παραμέτρου του πληθυσμού. Αυτό το ποσοστό λάθους ονομάζεται *σφάλμα δειγματοληψίας* (sampling error), και αποτελεί ένα από τα κύρια προβλήματα για την εξαγωγή γενικών συμπερασμάτων για τον πληθυσμό, όταν έχουμε διαθέσιμα μόνο κάποια δείγματα. Το σφάλμα δειγματοληψίας δεν αναιρεί τη χρησιμότητα της δειγματοληψίας, απλά αποτελεί απόδειξη ότι ένας στατιστικός δείκτης αποτελεί μόνο μία εκτίμηση της αντίστοιχης παραμέτρου του πληθυσμού, κάτι το οποίο δε πρέπει να ξεχνάμε κατά την ανάλυση.

Συνήθως για να λάβει κάποιος δείγμα του πληθυσμού, οι βασικοί λόγοι είναι οι ακόλουθοι:

- Οικονομικοί λόγοι, αφού το κόστος που προκύπτει από την εξέταση όλου του πληθυσμού είναι πολύ μεγάλο
- Χρονοβόρες διαδικασίες από μεριάς του ερευνητή.
- Προβληματική ανάλυση πολύ μεγάλου όγκου δεδομένων, αν και αυτό τείνει να εξαλειφθεί, καθώς ολοένα και αναπτύσσονται λογισμικά για τεράστιες βάσεις δεδομένων(VLDB).
- Η αποτελεσματικότητα ενός σωστά επιλεγμένου δείγματος μπορεί να είναι ισάξια με αυτή της εξέτασης του πληθυσμού.

Συνήθως μάλιστα, χρησιμοποιείται ο όρος παράμετρος (parameter) για να περιγράψει δεδομένα που αναφέρονται στον πληθυσμό, και ο όρος στατιστικός δείκτης (statistic) για τα δεδομένα που έχουν να κάνουν με ένα δείγμα. Τέλος με το δείγμα ασχολείται η επαγωγική στατιστική (inferential statistics) όπως και με τον έλεγχο στατιστικών υποθέσεων.

Το PSPP είναι ένα εργαλείο για τη στατιστική ανάλυση των δεδομένων του δείγματος. Μπορεί να χρησιμοποιηθεί για να εξηγήσει τις διαφορές σε ένα υποσύνολο των δεδομένων από την άποψη ενός άλλου υποσυνόλου και να μάθει κανείς εάν δικαιολογούνται ορισμένες απόψεις σχετικά με τα δεδομένα.

2.2 Μεταβλητές

Μεταβλητή ονομάζεται οποιοδήποτε χαρακτηριστικό ή κατάσταση παρουσιάζει αλλαγή, ή έχει διαφορετική τιμή για διαφορετικά υποκείμενα τα οποία ανήκουν σε ένα πληθυσμό ή ένα δείγμα.

Οι μεταβλητές πάντα απεικονίζονται ως οι στήλες ενώ στις γραμμές έχουμε τις λεγόμενες περιπτώσεις *(cases),* τις παρατηρήσεις δηλαδή που επισημάνθηκαν για την κάθε μεταβλητή. Κάθε γραμμή αποτελεί μια εγγραφή για ένα συγκεκριμένο υποκείμενο, ενώ κάθε κελί αποτελεί μια παρατήρηση.

Υπάρχουν δύο είδη μεταβλητών: Η μεταβλητή που ελέγχει ο ερευνητής και ονομάζεται ανεξάρτητη *(independent variable),* ενώ αυτή που αποτελεί το αντικείμενο παρατήρησης ονομάζεται εξαρτημένη *(dependent variable).* Το ζητούμενο για τον ερευνητή είναι να εξετάσει την επίδραση των διαφορετικών τιμών της ανεξάρτητης μεταβλητής στις τιμές της εξαρτημένης.

Τα δεδομένα του δείγματος μπορούν να χωριστούν σε δύο κατηγορίες: τα ποσοτικά και τα ποιοτικά. Ακολουθεί ο ορισμός:

Ονομάζουμε *ποιοτικά ή κατηγορικά ή ονομαστικά δεδομένα (qualitative, kategorical)* τα χαρακτηριστικά που δεν εκφράζουν κάτι μετρήσιμο. Μπορεί να είναι οργανωμένα απλά σαν ονομαστικές κατηγορίες (nominal data), όπως για παράδειγμα το φύλο, ή σαν ταξινομημένες κατηγορίες (ordinal data).Τέτοια παραδείγματα είναι το χρώμα ματιών, το επάγγελμα, το θρήσκευμα που είναι όλα ποιοτικά δεδομένα.

Όταν η ποιοτική μεταβλητή παίρνει μόνο δύο τιμές, όπως για παράδειγμα είναι το φύλο ενός ατόμου, τότε η απάντηση είναι ΝΑΙ-ΟΧΙ σε ένα ερωτηματολόγιο και λέγεται διχοτομική ή δίτιμη μεταβλητή.

Από την άλλη πλευρά, όταν τα δεδομένα παίρνουν μόνο αριθμητικές τιμές τότε ονομάζονται *ποσοτικά (quantitative)* και ο δειγματοχώρος τους είναι ένα υποσύνολο των πραγματικών αριθμών.

Αν η τυχαία μεταβλητή παίρνει πεπερασμένο ή αριθμήσιμο πλήθος τιμών τότε λέγεται *διακριτή ή απαριθμητή (discrete)*, ενώ αν παίρνει τιμές σε ένα διάστημα (α,β) με $-\infty \leq a < \beta \leq \infty$ θα λέγεται *συνεχής*.

Σχετικά παραδείγματα αναφέρουμε, το πλήθος τηλεφωνικών κλήσεων, το πλήθος των βακτηριδίων σε δειγματοληπτικό έλεγχο, είναι διακριτές μεταβλητές, ενώ ο χρόνος ζωής ενός ανταλλακτικού, η θερμοκρασία κ.ά. είναι *συνεχείς*.

Στο PSPP, οι μεταβλητές ορίζονται είτε ως αριθμητικές (numeric) είτε ως αλφαριθμητικά (string) κατά αντιστοιχία με τις ποσοτικές/ποιοτικές που αναφέρονται στη θεωρεία.

2.3 Τα Σύνολα Δεδομένων

Το PSPP λειτουργεί με τα δεδομένα να οργανώνονται σε σύνολα δεδομένων. Ένα σύνολο δεδομένων αποτελείται από ένα σύνολο *μεταβλητών*, οι οποίες λαμβάνονται μαζί και σχηματίζουν ένα *λεξικό(dictionary)*, και μία ή περισσότερες *περιπτώσεις*(cases) τα οποία αποτελούνται από την κάθε γραμμή, και η καθεμία από τις οποίες έχει μία τιμή για την κάθε μεταβλητή.

Σε κάθε δεδομένη στιγμή το PSPP δουλεύει πάνω σε ένα συγκεκριμένο σύνολο δεδομένων, που ονομάζεται ενεργό σύνολο δεδομένων. Οι περισσότερες PSPP εντολές λειτουργούν μόνο με το ενεργό σύνολο δεδομένων. Εκτός από το ενεργό σύνολο δεδομένων, το PSPP υποστηρίζει επίσης οποιοδήποτε αριθμό πρόσθετων ανοικτών συνόλων δεδομένων.

2.3.1 Οι Ιδιότητες των Μεταβλητών

Παρακάτω θα παραθέσουμε τις απαραίτητες ιδιότητες των μεταβλητών που μπορούν να χρησιμοποιηθούν από το PSPP:

- *Όνομα:* Είναι ένα αναγνωριστικό, έως 64 bytes. Κάθε μεταβλητή πρέπει να έχει ένα διαφορετικό όνομα. Μερικά ονόματα μεταβλητών του συστήματος μπορεί να

αρχίζουν από «$», αλλά όσες μεταβλητές καθορίζονται από το χρήστη, τα ονόματά τους δεν μπορούν να αρχίζουν με το «$». Ο τελικός χαρακτήρας στο όνομα μιας μεταβλητής δεν πρέπει να είναι η τελεία «.», διότι με ένα τέτοιο αναγνωριστικό μπορεί να παρερμηνευθεί, αφού με την τελεία ο συμβολομεταφραστής καταλαβαίνει το τέλος μιας εντολής και όχι έναν οποιοδήποτε χαρακτήρα. Επίσης, ο τελικός χαρακτήρας σε ένα όνομα μεταβλητής δεν θα πρέπει να είναι ο «_», δηλαδή η κάτω παύλα, επειδή κάποια τέτοια αναγνωριστικά χρησιμοποιούνται για ειδικούς σκοπούς από τις διαδικασίες του PSPP. Όπως με όλα τα αναγνωριστικά του PSPP, τα ονόματα των μεταβλητών δεν είναι case-sensitive.

- *Τύπος:* Μπορεί να είναι αριθμητικός ή αλφαριθμητικός(string).
- *Πλάτος:* Ορίζεται μόνο για τις String μεταβλητές. Οι μεταβλητές String μπορούν να έχουν πλάτος μέχρι 8 χαρακτήρες οπότε και ονομάζονται *σύντομες μεταβλητές string(short string variables).* Αυτές οι σύντομες μεταβλητές συμβολοσειράς μπορεί να χρησιμοποιηθούν σε λίγες περιπτώσεις όπου οι μεγάλες μεταβλητές συμβολοσειράς, δηλαδή εκείνες με πλάτος μεγαλύτερο από 8, δεν επιτρέπονται.
- *Θέση:* Οι μεταβλητές σε ένα λεξικό διατάσσονται σε μια συγκεκριμένη σειρά.
- *Αρχικοποίηση:* Είτε η επαναρχικοποίηση που γίνεται με μηδενικά, είτε τα κενά διαστήματα, και τρίτον υπάρχει η επιλογή να μείνουν με τις υπάρχουσες τιμές.
- *Οι απούσες τιμές:* Όπου αυτό χρειαστεί, δηλαδή μέχρι και τρεις τιμές, ή ένα συγκεκριμένο εύρος τιμών, ή μια συγκεκριμένη τιμή συν το εύρος, μπορεί να οριστεί από το χρήστη ως user-missing values, που έχουν το νόημα των ανεπιθύμητων τιμών. Υπάρχουν επίσης οι μεταβλητές με την ετικέτα system-missing που έχουν εκχωρηθεί σε μια παρατήρηση, όταν δεν υπάρχει άλλη προφανής τιμή για την παρατήρηση αυτή. Οι παρατηρήσεις με τις τιμές που λείπουν αποκλείονται αυτομάτως από οποιαδήποτε ανάλυση. Οι user-missing values αντιστοιχούν σε τιμές για τα δεδομένα, ενώ αντιθέτως, οι system missing ετικέτες δεν είναι έχουν καμιά αξία.
- *Οι ετικέτες των μεταβλητών:* Είναι μια συμβολοσειρά που περιγράφει τη μεταβλητή.
- *Οι ετικέτες των τιμών:* Προαιρετικά χρησιμοποιούνται για να συνδέουν κάθε δυνατή τιμή της μεταβλητής με ένα αλφαριθμητικό.
- *Διάταξη εκτύπωσης:* Περιλαμβάνει την εμφάνιση του πλάτους, της μορφοποίησης, και (για τις αριθμητικές μεταβλητές) τον αριθμό των δεκαδικών ψηφίων. Αυτό το

χαρακτηριστικό δεν επηρεάζει πώς τα δεδομένα αποθηκεύονται, μόνο τον τρόπο που εμφανίζονται. Παράδειγμα: πλάτος 8, με 2 δεκαδικά ψηφία.

- *Διάταξη εγγραφής:* Παρόμοια με το προηγούμενο.
- *Προσαρμοσμένα χαρακτηριστικά:* Ορίζονται από το χρήστη οι συσχετίσεις μεταξύ ονομάτων και τιμών.

Επίσης υπάρχουν και μεταβλητές που ορίζονται από το PSPP, οι λεγόμενες μεταβλητές συστήματος. Το PSPP έχει επτά τέτοιες μεταβλητές οι οποίες δεν είναι σαν τις κοινές μεταβλητές, αφού για παράδειγμα πάντα δε μπορούν να αποθηκευτούν αλλά μπορούν να χρησιμοποιούνται σε εκφράσεις.

2.4 Περιγραφική Στατιστική

Κατά τη διάρκεια της διαδικασίας της περιγραφικής στατιστικής, σκοπός μας είναι η περιγραφή του συνόλου των δεδομένων που έχουμε αντλήσει από τα στοιχεία του δείγματος.

Παρακάτω παραθέτουμε ένα ορισμό της περιγραφικής στατιστικής:

- *Περιγραφική στατιστική είναι ο κλάδος της στατιστικής που ασχολείται με την οργάνωση, τη συγκέντρωση και την περιγραφή ενός συνόλου δεδομένων.*Το σύνολο αυτό των δεδομένων είναι ο πληθυσμός.

Οι πιο διαδεδομένες τεχνικές είναι ο υπολογισμός της μέσης τιμής (mean) και της τυπικής απόκλισης (standard deviation).

Για τη στατιστική περιγραφή ενός δείγματος μπορούμε να χρησιμοποιήσουμε τις παρακάτω τρεις μεθόδους:

1. Στατιστικούς πίνακες - κατανομές συχνοτήτων
2. Γραφικές παραστάσεις
3. Στατιστικά μέτρα

2.4.1 Στατιστικοί Πίνακες

Οι πλέον διαδεδομένοι πίνακες συνοπτικής παρουσίασης μεταβλητών, που χρησιμοποιούνται και από το PSPP, είναι οι πίνακες κατανομής συχνοτήτων *(frequency tables)*. Η μορφή ενός τέτοιου πίνακα εξαρτάται από το είδος της μεταβλητής. Συγκεκριμένα:

1. *Ποιοτικές μεταβλητές*. Σε διαδοχικές στήλες καταγράφονται με τη σειρά:
    i. όλες οι τιμές - κατηγορίες της ποιοτικής μεταβλητής *(values)*,
    ii. οι συχνότητες εμφάνισης των τιμών *(frequencies)*,
    iii. τα ποσοστά των τιμών στο σύνολο όλων των παρατηρήσεων *(percents)*,
    iv. τα έγκυρα ποσοστά, δηλαδή τα ποσοστά στο σύνολο των έγκυρων παρατηρήσεων, που μένουν μετά από απόρριψη των χαμένων τιμών *(valid percents)* και τέλος
    v. τα αθροιστικά ποσοστά *(cummulative percents)*.
2. *Ποσοτικές ασυνεχείς μεταβλητές*. Χρησιμοποιείται ο ίδιος ακριβώς πίνακας με τις ποιοτικές μεταβλητές με τη διαφορά ότι οι τιμές της μεταβλητής είναι διακριτές μετρήσεις (αριθμοί) και όχι κατηγορίες.
3. *Ποσοτικές συνεχείς μεταβλητές*. Και σε αυτή την περίπτωση χρησιμοποιείται ο ίδιος πίνακας κατανομής συχνοτήτων με την προϋπόθεση ότι οι τιμές της μεταβλητής έχουν ομαδοποιηθεί σε κλάσεις. Ο αριθμός *k*των κλάσεων είναι αυθαίρετος και εξαρτάται από τον αριθμό *n*των δεδομένων μας (ένας εμπειρικός τύπος για τον αριθμό των κλάσεων είναι αυτός του Sturges: $k=1+3.2\log n$).

2.4.2 Γραφικές Παραστάσεις

Η συνοπτική παρουσίαση των δεδομένων με πίνακες δεν είναι αρκετή για να δώσει μια αρχική εικόνα κάποιας μεταβλητής, ειδικά όταν οι πίνακες είναι μεγάλοι και οι άνθρωποι στους οποίους απευθύνονται οι πληροφορίες δεν είναι ειδικοί. Για μία πιο παραστατική και ευανάγνωστη παρουσίαση δημιουργούμε τις γραφικές παραστάσεις. Το είδος της γραφικής παράστασης που θα επιλέξουμε για την απεικόνιση των δεδομένων εξαρτάται από το είδος της μεταβλητής. Δυστυχώς το PSPP δεν μας δίνει πολλές επιλογές γραφικών παραστάσεων. Συγκεκριμένα, για τις ποιοτικές και τις ποσοτικές ασυνεχείς μεταβλητές

επιλέγουμε τα κυκλικά διαγράμματα *(piecharts)* ενώ για τις ποσοτικές συνεχείς μεταβλητές τα ιστογράμματα *(histograms)*.

2.4.3 Στατιστικά μέτρα

Τα στατιστικά μέτρα (statistics - measures) είναι αριθμοί που υπολογίζονται από τα δεδομένα και η τιμή τους αντιπροσωπεύει κάποια τάση ή συμπεριφορά του δείγματος. Θα μπορούσαμε να πούμε ότι είναι αριθμοί - μέτρα που περιγράφουν το σύνολο του δείγματος με τον ίδιο τρόπο που ένα στερεό σώμα όσο πολύπλοκο και αν είναι στην δομή του μπορεί να περιγραφεί συνοπτικά μόνο από το βάρος του και από τον όγκο του. Τα στατιστικά μέτρα διακρίνονται στα *μέτρα κεντρικής τάσης (measures of central tendency)*, στα *μέτρα μεταβλητότητας (measures of dispersion)* και στα *μέτρα σχήματος (measures of shape)*. Τις τρεις αυτές κατηγορίες περιγράφουμε αναλυτικά στη συνέχεια.

1. *Μέτρα κεντρικής τάσης:* Χρησιμοποιούνται για να περιγράψουν τη θέση του συνόλου των δεδομένων. Δηλαδή, η τιμή τους είναι η πιο αντιπροσωπευτική για να σχηματίσει κάποιος μία πρώτη εικόνα για το ύψος των τιμών των παρατηρήσεων μίας μεταβλητής. Τα πλέον γνωστά μέτρα κεντρικής τάσης είναι τρία:

    i. Η *μέση τιμή (mean).* Γνωστή και σαν αριθμητικός μέσος όρος, είναι το άθροισμα των τιμών όλων των παρατηρήσεων διαιρεμένο με το πλήθος των παρατηρήσεων, δηλαδή:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

    όπου N ο αριθμός των περιπτώσεων και Xi η τιμή της μεταβλητής για την i-οστή περίπτωση. Η μέση τιμή υπολογίζεται και ερμηνεύεται στατιστικά στις ποσοτικές μεταβλητές. Απαιτείται ιδιαίτερη προσοχή στην περίπτωση ποιοτικών μεταβλητών που οι τιμές τους έχουν κωδικοποιηθεί με αριθμούς, όπου η μέση τιμή δεν έχει νόημα. Εξαίρεση αποτελεί η περίπτωση όπου η ποιοτική μεταβλητή παίρνει δύο μόνο τιμές (π.χ. "ΝΑΙ" και "ΟΧΙ") οι οποίες έχουν κωδικοποιηθεί με τους αριθμούς 0 και 1. Η μέση τιμή σε αυτή την περίπτωση παριστά το ποσοστό των περιπτώσεων που έχουν δηλωθεί σαν 1 στα δεδομένα.

ii. *Η διάμεσος (median).* Είναι η τιμή της μεταβλητής για την οποία το 50% των τιμών είναι μεγαλύτερο από αυτή και το υπόλοιπο 50% μικρότερο. Αν το πλήθος των παρατηρήσεων είναι περιττό, η διάμεσος είναι η μεσαία παρατήρηση, όταν αυτές διαταχθούν σε αύξουσα σειρά. Στην περίπτωση που το πλήθος είναι άρτιο, η διάμεσος υπολογίζεται από τη μέση τιμή των δύο μεσαίων παρατηρήσεων. Χρησιμοποιείται αποκλειστικά για ποσοτικές μεταβλητές.

iii. *Η επικρατούσα τιμή (mode)*. Είναι η πιο συχνά εμφανιζόμενη τιμή (ή και τιμές) στο δείγμα. Είναι δυνατόν να υπάρχουν περισσότερες από μία επικρατούσες τιμές, το PSPP όμως εμφανίζει μόνο τη μία από αυτές. Χρησιμοποιείται και για τις ποσοτικές αλλά και για τις ποιοτικές μεταβλητές.

2. *Μέτρα μεταβλητότητας:* Χρησιμοποιούνται για να περιγράψουν τις διαφορές που υπάρχουν ανάμεσα στις τιμές μίας μεταβλητής. Δηλαδή, προσδιορίζουν αν οι παρατηρήσεις είναι συγκεντρωμένες γύρω από μία αντιπροσωπευτική τιμή ή παρουσιάζουν μεγάλη διασπορά. Τα πιο γνωστά μέτρα μεταβλητότητας είναι:

   i. Το *εύρος (range)*. Είναι η διαφορά ανάμεσα στη μέγιστη (maximum) και στην ελάχιστη (minimum) από τις παρατηρούμενες τιμές.

   ii. Η *διακύμανση ή διασπορά (variance)*. Είναι το πλέον συνηθισμένο μέτρο μεταβλητότητας. Υπολογίζεται αθροίζοντας τα τετράγωνα των διαφορών από τη μέση τιμή όλων των παρατηρήσεων και στη συνέχεια διαιρώντας το άθροισμα με το πλήθος των παρατηρήσεων ελαττωμένο κατά ένα. Ο τύπος υπολογισμού δηλαδή είναι:

$$s^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}$$

iii. Η *τυπική απόκλιση (standard deviation).* Είναι η τετραγωνική ρίζα της διακύμανσης. Έχει πιο φανερή σημασία από τη διακύμανση γιατί εκφράζεται σε μονάδες μέτρησης ίδιες με αυτές των παρατηρήσεων, ενώ η διακύμανση στα τετράγωνα των μονάδων αυτών.

3. *Μέτρα σχήματος:* Είναι μέτρα που περιγράφουν το σχήμα της μεταβλητής όταν αυτή παρασταθεί με ιστόγραμμα. Υπολογίζονται τα δύο μέτρα σχήματος:

   i. Η *λοξότητα (skewness).* Για πολλές μεταβλητές το μεγαλύτερο πλήθος των παρατηρήσεων συγκεντρώνεται σε μία κεντρική τιμή. Όσο αυξάνει η απόσταση από την μεσαία αυτή τιμή, μειώνεται η συχνότητα των παρατηρήσεων. Αν αυτή η συμπεριφορά είναι ίδια για τις μεγάλες και για τις μικρές τιμές, η μεταβλητή παρουσιάζει στο σχήμα της μία συμμετρία ως προς την κεντρική τιμή. Τέτοιο παράδειγμα είναι η κανονική κατανομή. Υπάρχουν όμως μεταβλητές που δεν είναι συμμετρικές και εμφανίζουν μία "ουρά" προς τη μία μόνο κατεύθυνση. Αυτές ονομάζονται λοξές (skewed). Ανάλογα προς τα πού εμφανίζεται η "ουρά" (μεγάλες ή μικρές τιμές), η μεταβλητή ονομάζεται θετικά λοξή (λοξή προς τα δεξιά) ή αρνητικά λοξή (λοξή προς τα αριστερά). Αν λοιπόν η λοξότητα που υπολογίζει το PSPP είναι 0, αυτό σημαίνει ότι έχουμε συμμετρική κατανομή συχνοτήτων. Τιμή θετική ή αρνητική δείχνει αντίστοιχα θετικά ή αρνητικά λοξή κατανομή συχνοτήτων.

   ii. Η *κύρτωση (kurtosis).* Είναι ο βαθμός στον οποίο οι παρατηρήσεις συνωστίζονται γύρω από μία κεντρική τιμή. Η κύρτωση της κανονικής κατανομής είναι ίση με 0. Μία τιμή θετική δείχνει κατανομή συχνοτήτων λεπτόκυρτη που σημαίνει ότι η γραφική παράσταση της μεταβλητής παρουσιάζει "αιχμηρή" κορυφή και πολλές ακραίες τιμές. Αντίθετα, μία αρνητική τιμή δείχνει κατανομή πλατύκυρτη δηλαδή "αμβλεία" κορυφή και σχετικά κοντές ουρές.

2.5 Συμπερασματική Στατιστική

Το τμήμα που αναλαμβάνει τις εκτιμήσεις των παραμέτρων του πληθυσμού ονομάζεται Συμπερασματική Στατιστική (Inferential Statistics).

Ο στατιστικός έλεγχος υποθέσεων (hypothesis testing) είναι μια συμπερασματική διαδικασία/μέθοδος που προσφέρει η Στατιστική Συμπερασματολογία και βρίσκει εφαρμογή σε στοχαστικά προβλήματα απόφασης μεταξύ δύο εναλλακτικών υποθέσεων. Η μία υπόθεση έχει επικρατήσει να συμβολίζεται με Η0 και ονομάζεται μηδενική υπόθεση (null hypothesis), και η άλλη με Η1 και ονομάζεται εναλλακτική υπόθεση (alternative hypothesis). Αναγκαία προϋπόθεση για τη σωστή εφαρμογή των στατιστικών ελέγχων και κυρίως για τη σωστή ερμηνεία των αποτελεσμάτων τους, είναι η κατανόηση της λογικής και του νοήματός τους, για τους ελέγχους αυτούς κάνουμε διάφορα στατιστικά τεστ.

2.6 Μετασχηματισμός Δεδομένων και Επιλογή Περιπτώσεων

Είναι πολύ συνηθισμένο σε ένα σύνολο δεδομένων που έχουν συλλεχθεί για στατιστική ανάλυση, να υπάρχουν μεταβλητές που χρειάζονται μετασχηματισμούς (transformations). Ο όρος αυτός χρησιμοποιείται για να δηλώσει είτε τη μετατροπή των τιμών των ίδιων των μεταβλητών, είτε τη δημιουργία νέων μεταβλητών από τις ήδη υπάρχουσες. Σε κάθε περίπτωση, κάποιος κανόνας (συνάρτηση) απαιτείται για τον υπολογισμό των νέων τιμών με βάση τις παλαιές.

Σε αυτή την κατηγορία ανήκουν δυο βασικές εντολές του PSPP, η *recode* και η *compute*. Η εντολή recode χρησιμοποιείται για την επανακωδικοποίηση των τιμών μίας μεταβλητής, που ήδη υπάρχει. Χρησιμοποιείται κυρίως για την κατηγοριοποίηση ποσοτικών συνεχών μεταβλητών, αλλά και για τη σύμπτυξη τιμών από ποιοτικές ή ποσοτικές ασυνεχείς μεταβλητές. Η εντολή compute χρησιμοποιείται για τη δημιουργία νέων μεταβλητών σαν αποτέλεσμα υπολογισμών πάνω σε μεταβλητές που ήδη υπάρχουν.

Επιπλέον, πολλές φορές στα δεδομένα μας υπάρχουν κατηγορίες και υποσύνολα που καθορίζονται από τις τιμές κάποιων μεταβλητών. Στις περιπτώσεις αυτές μας ενδιαφέρει να απομονώσουμε κάποια συγκεκριμένη κατηγορία και να εργαστούμε μόνο με τα δεδομένα που ανήκουν στο υποσύνολο αυτό. Σαν παράδειγμα μπορούμε να αναφέρουμε την επιθυμία μας να εργαστούμε μόνο με τα δεδομένα που αναφέρονται στους υπαλλήλους

που εργάζονται στην έρευνα αγοράς. Υπάρχουν λοιπόν εντολές οι οποίες επιλέγουν (απομονώνουν ή φιλτράρουν) ένα συγκεκριμένο σύνολο περιπτώσεων, σύμφωνα με κάποιο καθορισμένο κριτήριο, και επιτρέπουν την ανάλυση μόνο αυτών των περιπτώσεων.

Στην κατηγορία αυτή η επιλογή των περιπτώσεων γίνεται με δύο τρόπους: (α) επιλογή περιπτώσεων που ικανοποιούν κάποια λογική συνθήκη και (β) τυχαία επιλογή περιπτώσεων. Χρησιμοποιείται η εντολή Select Cases που ενσωματώνει και τους δύο παραπάνω τρόπους.

Στο 4^{ο} κεφάλαιο θα γίνει αναλυτική περιγραφή των παραπάνω εντολών μέσω του γραφικού περιβάλλοντος του PSPP.

2.7 Ο Χειρισμός των ΄Υποπτων Τιμών

Αρχικά, για τον εντοπισμό των λανθασμένων τιμών μπορούμε να χρησιμοποιήσουμε την περιγραφική ανάλυση που δίνει μια σύντομη περίληψη της συνολικής εικόνας των τιμών των μεταβλητών. Πιο συγκεκριμένα, μπορεί να χρησιμοποιηθεί έλεγχος των ελαχίστων τιμών, που μερικές φορές μπορεί να είναι ικανός να εντοπίσει λάθη όπως αρνητικές τιμές σε φυσικά μεγέθη: για παράδειγμα σε μεταβλητές όπως είναι το ύψος ή το βάρος κλπ. Από την άλλη πλευρά, η χρήση των μεγίστων τιμών ως κριτήριο ελέγχου μπορεί να γίνει μέσω της τυπικής απόκλισης και σε σύγκριση με τη μέση τιμή: για παράδειγμα, μια τιμή με απόσταση από τη μέση τιμή 5 τυπικές αποκλίσεις είναι αρκετά ύποπτη για λάθη.

Όταν είναι δυνατόν, τα ύποπτα δεδομένα θα πρέπει να ελέγχονται και να μετράται εκ νέου. Ωστόσο, αυτό μπορεί να μην είναι πάντα εφικτό, οπότε ο ερευνητής μπορεί να αποφασίσει αν θα αγνοήσει αυτές τις μεταβλητές. Το PSPP έχει τη δυνατότητα να δίνει στα δεδομένα το χαρακτηρισμό «sysmis». Η ιδιαίτερη αυτή τιμή-χαρακτηρισμός, θα πρέπει να μην λαμβάνεται υπόψη στην επικείμενη ανάλυση.

2.8 Έλεγχος Αξιοπιστίας

Ένας από τους σημαντικούς ελέγχους που πρέπει να εκτελούνται στις περιπτώσεις που τα δεδομένα αντλούνται από ερωτηματολόγιο είναι ο υπολογισμός της *αξιοπιστίας(checking data consistency)*. Αυτό μπορεί να εξασφαλίσει στον ερευνητή ότι τα ερωτηματολόγια έχουν

συμπληρωθεί συνειδητά και κατόπιν σκέψης. Για παράδειγμα, συχνά συμβαίνει στα ερωτηματολόγια οι ετικέτες των μεταβλητών-ερωτημάτων να ζητούν πολύ παρόμοιες απαντήσεις. Θα περίμενε κανείς, συνεπώς, τις τιμές αυτών των μεταβλητών (μετά την εκ νέου κωδικοποίηση) να συνάδουν στενά η μία με την άλλη, και μπορούμε να το ελέγξουμε με την εντολή αξιοπιστίας.

Ο *συντελεστής άλφα του Cronbach* υποδηλώνει υψηλό βαθμό αξιοπιστίας ανάμεσα στις μεταβλητές που εξετάζουμε. Είναι κοινώς αποδεκτό ότι ο συντελεστής άλφα του Cronbach, θα πρέπει να παίρνει τιμές από 0,7 ή υψηλότερη για να αποδεικνύεται η αξιοπιστία των στοιχείων. Έτσι, ενδιάμεσες τιμές όπως είναι 0,80 αποδεικνύουν ότι οποιαδήποτε «διορθωτική ενέργεια» και αν εκτελέσαμε, όπως η εκ νέου κωδικοποίηση, δικαιώνονται.

2.9 Έλεγχος Υποθέσεων

Μία από τις πιο θεμελιώδεις πρακτικές της στατιστικής ανάλυσης είναι ο έλεγχος υποθέσεων. Οι ερευνητές συνήθως πρέπει να ελέγξουν τις υποθέσεις σχετικά με ένα σύνολο δεδομένων. Για παράδειγμα, είναι πιθανόν να θέλουμε να ελέγξουμε αν ένα σύνολο δεδομένων προέρχεται από την ίδια κατανομή με ένα άλλο, ή αν η μέση τιμή ενός συνόλου δεδομένων διαφέρει σημαντικά από μια συγκεκριμένη τιμή.

Οι έλεγχοι αυτοί ξεκινάν κάνοντας μια μηδενική υπόθεση. Τις περισσότερες φορές, είναι μια υπόθεση που και είναι εύκολο κανείς να καταλάβει ότι είναι ψευδής. Για παράδειγμα, εάν υπάρχει η υποψία ότι το A είναι μεγαλύτερο από το B θέτουμε ως μηδενική υπόθεση την A = B.

Το *p-value* είναι μια επαναλαμβανόμενη έννοια στους ελέγχους υποθέσεων. Η τιμή της p-value αντιπροσωπεύει έναν δείκτη της αξιοπιστίας ενός αποτελέσματος. Όσο υψηλότερη η p-value, λιγότεροι μπορούμε να πιστέψουμε ότι η παρατηρηθείσα σχέση μεταξύ των μεταβλητών στο δείγμα είναι ένας αξιόπιστος δείκτης της σχέσης μεταξύ των αντίστοιχων μεταβλητών στον πληθυσμό.

Συγκεκριμένα, η p-value αντιπροσωπεύει την πιθανότητα του λάθους που περιλαμβάνεται στην αποδοχή του παρατηρηθέντος αποτελέσματός μας τόσο έγκυρου, δηλαδή όσο "η αντιπροσώπευση του πληθυσμού". Παραδείγματος χάριν, μια p-value του 0.05 (δηλ., 1/20) δείχνει ότι υπάρχει μια πιθανότητα 5% ότι η σχέση μεταξύ των μεταβλητών που βρίσκονται στο δείγμα μας να είναι "ψευδής". Με άλλα λόγια, υποθέτοντας ότι στον

πληθυσμό δεν υπήρξε καμία σχέση μεταξύ εκείνων των μεταβλητών, και επαναλαμβάναμε τα πειράματα μας, θα μπορούσαμε να αναμείνουμε ότι περίπου σε κάθε 20 επαναλήψεις του πειράματος θα υπήρχε ένα στον οποίο η σχέση μεταξύ των εν λόγω μεταβλητών θα ήταν ίση ή ισχυρότερη από ότι στους δικούς μας υπολογισμούς.

Σε πολλούς τομείς της έρευνας, η p-value του 0.05 είναι συνήθως η διαχωριστική γραμμή ως αποδεκτό "επίπεδο λάθους".

2.10 Έλεγχος Μέσων Τιμών

Στις έρευνες που περιλαμβάνουν δεδομένα με ποσοτικές μεταβλητές, αντιμετωπίζουμε συχνά το πρόβλημα της σύγκρισης μέσων τιμών. Έτσι, άλλες φορές υπάρχει η ανάγκη να συγκριθούν οι μέσες τιμές της ίδιας ποσοτικής μεταβλητής που όμως προέρχεται από δύο ανεξάρτητους πληθυσμούς, άλλοτε πρέπει να συγκριθούν δύο ποσοτικές μεταβλητές που προέρχονται από τον ίδιο πληθυσμό και οι τιμές τους λαμβάνονται σε ζεύγη, και ακόμη, υπάρχει η περίπτωση της σύγκρισης της μέσης τιμής ποσοτικής μεταβλητής με ένα σταθερό αριθμό. Τις τρεις αυτές περιπτώσεις θα αναλύσουμε στη συνέχεια.

2.10.1 Σύγκριση Μέσων Τιμών Ανεξαρτήτων Πληθυσμών

Αρκετά συχνά μας ενδιαφέρει να ελέγξουμε αν δύο ομοειδείς ποσοτικές μεταβλητές, που προέρχονται από ανεξάρτητους μεταξύ τους πληθυσμούς, διαφέρουν κατά μέση τιμή. Αν δηλαδή, οι μέσες τιμές τους είναι ίσες ή διαφέρουν σημαντικά. Για παράδειγμα, θα ενδιέφερε σε μια έρευνα να ελέγξουμε αν οι αξιολογήσεις των εργαζομένων στον τομέα της έρευνας αγοράς διαφέρουν κατά μέσο όρο από εκείνες των εργαζομένων στον τομέα της διαφήμισης. Όπως είναι φυσικό για κάθε τέτοιου είδους έλεγχο, όπου είναι πρακτικά αδύνατο να υπολογίσουμε τις μέσες τιμές των πληθυσμών, χρησιμοποιούμε δείγματα που παίρνουμε από τους δύο πληθυσμούς. Τα δείγματα αυτά δεν είναι αναγκαστικά ίδιου μεγέθους. Από αυτά τα δείγματα υπολογίζουμε τις δειγματικές μέσες τιμές και τις δειγματικές διασπορές και στη συνέχεια, εκτελούμε ένα στατιστικό έλεγχο, το γνωστό και πολύ δημοφιλές *t-test*.

Εκφράζοντας τις παραπάνω αρχές με στατιστική ορολογία, υποθέτουμε ότι έχουμε δύο ανεξάρτητους πληθυσμούς με μέσες τιμές μ_1 και μ_2 και διασπορές σ_1^2 και σ_2^2 . Η υπόθεση που έχουμε να ελέγξουμε είναι:

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

με εναλλακτική την

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

Για το σκοπό αυτό, χρησιμοποιούμε δύο δείγματα από τους δύο πληθυσμούς με μεγέθη n_1 και n_2. Από τα δείγματα αυτά υπολογίζουμε τις δειγματικές μέσες τιμές x_1 και x_2 και τις 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'st-test*. Η διαδικασία που ακολουθείται αποτελείται από τα παρακάτω βήματα:

Βήμα 1. Διατύπωση της μηδενικής και εναλλακτικής υπόθεσης:

H_0: $\mu_1 - \mu_2 = 0$

H_α: $\mu_1 - \mu_2 \neq 0$

Βήμα 2. Σύγκριση διασπορών των δύο δειγμάτων. Εδώ ελέγχουμε την υπόθεση:

H_0: $\sigma_1^2 = \sigma_2^2$

H_α: $\sigma_1^2 \neq \sigma_2^2$

Ο έλεγχος γίνεται με το στατιστικό μέτρο F, που είναι ο λόγος της μεγαλύτερης δειγματικής διασποράς προς τη μικρότερη. Αν η στάθμη σημαντικότητας για το F*(significance ή two tail probability)* είναι μικρή (<0.05 συνήθως), τότε η υπόθεση H_0 απορρίπτεται, δηλαδή θεωρούμε ότι οι δύο διασπορές παρουσιάζουν σημαντική διαφορά.

Βήμα 3. Εδώ διακρίνουμε δύο περιπτώσεις:

Περίπτωση 1η: Οι δύο διασπορές των πληθυσμών βρέθηκαν ίσες στο Βήμα 2.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 t στατιστικό από τον τύπο :

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

και η στάθμη σημαντικότητας (*significance ή two tail probability*) 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 στάθμη σημαντικότητας 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 2η:Οι δύο διασπορές των πληθυσμών βρέθηκαν άνισες (απορρίφθηκε δηλαδή η μηδενική υπόθεση στο Βήμα 2). Στην περίπτωση αυτή, το στατιστικό μέτρο t υπολογίζεται από τον τύπο:

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

Στη συνέχεια υπολογίζεται η στάθμη σημαντικότητας 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 εκτιμώνται εδώ με πιο πολύπλοκο τρόπο.

2.10.2 Σύγκριση Μέσων Τιμών σε Ζευγάρια Παρατηρήσεων

Σε πολλές περιπτώσεις, ιδίως σε έρευνες όπου τα δεδομένα προέρχονται από «κλειστούς» και ελεγχόμενους πειραματισμούς, έχουμε αντί για ανεξάρτητα δείγματα, ζευγάρια παρατηρήσεων. Για παράδειγμα σε ένα ιατρικό πείραμα, όπου ερευνούμε την επίδραση ενός νέου φαρμάκου που καταπολεμά την υπέρταση, είναι πολύ φυσικό να επιλέξουμε ένα δείγμα ασθενών και να μετρήσουμε την πίεσή τους πριν και μετά από τη λήψη του φαρμάκου. Στην περίπτωση αυτή έχουμε ζευγάρια (pairs) από παρατηρήσεις πάνω στα ίδια άτομα του δείγματος. Σε άλλες περιπτώσεις είναι δυνατό να θεωρήσουμε σα

ζευγάρια, παρατηρήσεις που λαμβάνονται σε ζευγάρια υποκειμένων, όπως για παράδειγμα σε δίδυμα, σε αγρούς γειτονικούς κλπ. Έχουμε λοιπόν δύο δείγματα Χ και Υ ζευγαρωτά, ή συσχετισμένα όπως συνήθως λέγονται, μεγέθους n. Η στατιστική υπόθεση την οποία πρέπει να ελέγξουμε είναι

$$H_0: \mu_X - \mu_Y = 0$$
$$H_a: \mu_X - \mu_Y \neq 0$$

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

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

Η στάθμη σημαντικότητας του t θα είναι αυτή που θα κρίνει την απόρριψη της μηδενικής υπόθεσης. Έτσι, σημαντικότητα μικρή (<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.

2.10.3 Σύγκριση Μέσης Τιμής Πληθυσμού με Δεδομένη Τιμή

Σε πολλές περιπτώσεις επιθυμούμε να συγκρίνουμε την (άγνωστη) μέση τιμή μ ενός πληθυσμού με μία δεδομένη τιμή μ_0 που την γνωρίζουμε από εμπειρία, από προηγούμενες μελέτες κλπ. Μας ενδιαφέρει δηλαδή να ελέγξουμε την υπόθεση:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

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 :

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

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

2.11 Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση είναι μια τεχνική που χρησιμοποιείται για να ερευνήσει εάν και πώς μια μεταβλητή σχετίζεται γραμμικά με άλλες. Αν μια μεταβλητή έχει βρεθεί να είναι γραμμικά συναφής, τότε αυτό το χαρακτηριστικό της μπορεί να χρησιμοποιηθεί για να προβλέψει τις μελλοντικές τιμές της μεταβλητής. Αυτό μπορεί να συμβαίνει χρησιμοποιώντας περισσότερες από μία μεταβλητές. Αυτός είναι και ο στόχος/χρήση της γραμμικής παλινδρόμησης. Χρησιμοποιείται λοιπόν προκειμένου να:

- Προβλέψουμε τις τιμές μιας μεταβλητής με βάση τις τιμές μίας ή περισσότερων άλλων μεταβλητών.
- Αποφασίσουμε αν κάποια μεταβλητή είναι ‘καλή’ για την πρόβλεψη κάποιας άλλης μεταβλητής.
- Να βρούμε ποιό είναι το ποσοστό της διακύμανσης των τιμών μιας μεταβλητής που μπορεί να εξηγηθεί από τις τιμές μίας μεμονωμένης ή ενός συνόλου μεταβλητών.
- Να φτιάξουμε καινούρια μοντέλα και να ελέγξουμε υπάρχουσες θεωρίες (theory building, theory testing)

Ένα τέτοιο παράδειγμα μπορεί να είναι μια εταιρεία που θέλει να προβλέψει το χρόνο για την επισκευή εξοπλισμού, έτσι ώστε να βελτιωθεί η ακρίβεια των παραγγελιών τους. Έστω ότι υπάρχει η υποψία ότι ο χρόνος που απαιτείται για την επισκευή θα μπορούσε να σχετίζεται με το χρόνο μεταξύ των αποτυχημένων προσπαθειών και τον κύκλο εργασιών για την επισκευή του εξοπλισμού. Το p-value σε αυτές τις περιπτώσεις είναι της τάξης του 0,1. Η εντολή παλινδρόμησης στο PSPP εκτός από το τεστάρισμα της συσχέτισης των μεταβλητών, μπορεί να προσδιορίσει και την γραμμική ή όχι σχέση μεταξύ τους.

2.12 Διαχείριση των Χαμένων Τιμών (user missing values)

Το PSPP περιλαμβάνει ειδική διαδικασία για τις άγνωστες αριθμητικές τιμές που βρίσκονται μέσα στα δεδομένα: στις παρατηρήσεις που λείπουν αποδίδεται μια ειδική τιμή, που ονομάζεται «system-missing value». Αυτή η ετικέτα υποδεικνύει στην πραγματικότητα την απουσία τιμής, ότι δηλαδή η πραγματική τιμή είναι άγνωστη και πολύ συχνά αυτές προέρχονται από ερωτηματολόγια, όταν ο ερωτώμενος συμπληρώνει το ερωτηματολόγιο μόνος του ή το ίδιο το ερωτηματολόγιο είναι δεν είναι καθόλου καλά σχεδιασμένο.

Έτσι, το PSPP περιλαμβάνει διαδικασίες που αυτόματα αποκλείουν από οποιαδήποτε ανάλυση αυτές τις παρατηρήσεις ή τις περιπτώσεις που έχουν τιμές που λείπουν. Οι επιμέρους λεπτομέρειες για τον χειρισμό των απουσών τιμών εξαρτώνται από την περίπτωση και συχνά μπορούν να ελέγχονται από το χρήστη.

Η παραπάνω διαδικασία ορίζεται μόνο για τις αριθμητικές μεταβλητές. Οι μεταβλητές τύπου String έχουν πάντα μια καθορισμένη τιμή, ακόμη και αν είναι μια σειρά από κενούς χαρακτήρες. Οι μεταβλητές, είτε αριθμητικές είτε String, μπορούν να οριστούν ως «user-missing values». Σε κάθε τέτοια τιμή αντιστοιχεί μια πραγματική τιμή για τη μεταβλητή. Ωστόσο, τις περισσότερες φορές οι user-missing τιμές αντιμετωπίζονται κατά τον ίδιο τρόπο όπως και οι «system-missing value».

2.13 Πίνακες Συνάφειας, Ανεξαρτησία και Ομοιογένεια

Συνήθως στα δεδομένα που έχουμε από μία έρευνα, υπάρχουν περισσότερες από μία ποιοτικές μεταβλητές. Στην περίπτωση αυτή, μας ενδιαφέρει να παρουσιάσουμε συνοπτικά σε πίνακα την κοινή κατανομή δύο τέτοιων μεταβλητών, και στη συνέχεια, να ελέγξουμε αν υπάρχει κάποια σχέση μεταξύ τους. Σαν παραδείγματα θα μπορούσαμε να αναφέρουμε τη συσχέτιση που πιθανόν να υπάρχει ανάμεσα στις μεταβλητές "επάγγελμα" και "εκλογική προτίμηση", "κάπνισμα" και "αναπνευστικές ασθένειες", "περιοχή σχολείου" και "επιτυχία στις εξετάσεις" κλπ.

2.13.1 Κοινή Κατανομή Δύο Ποιοτικών Μεταβλητών. Πίνακες Συνάφειας

Υποθέτουμε ότι στα δεδομένα υπάρχουν δύο ποιοτικές μεταβλητές Α και Β των οποίων οι τιμές είναι οι κατηγορίες A_1, A_2, ..., A_R και B_1, B_2, ..., B_C αντίστοιχα. Αυτό που μας ενδιαφέρει αρχικά, είναι να κατασκευάσουμε ένα κοινό πίνακα κατανομής συχνοτήτων για τις δύο μεταβλητές, δηλαδή να καθορίσουμε πόσες περιπτώσεις *(cases)* ανήκουν σε κάθε διασταύρωση των κατηγοριών A_i και B_j. Ένας τέτοιος πίνακας ονομάζεται *πίνακας συνάφειας* (*contingency table*) των μεταβλητών Α και Β και έχει την παρακάτω γενική μορφή.

	Μεταβλητή Β				
Μεταβλητή Α	B_1	B_2	...	B_C	**Σύνολα γραμμών**
A_1	n_{11}	n_{12}	...	n_{1C}	r_1
A_2	n_{21}	n_{22}	...	n_{2C}	r_2
...	...	...	...	...	...
A_R	n_{R1}	n_{R2}	...	n_{RC}	R_R
Σύνολα στηλών	c_1	c_2	...	c_C	n

Παρατηρούμε ότι οι κατηγορίες του χαρακτηριστικού Α γράφονται στις γραμμές (rows) του πίνακα, οι κατηγορίες του Β στις στήλες (columns)και κάθε συνδυασμός των τιμών A_i και B_j σχηματίζει ένα κελί (cell). Στα κελιά αυτά καταγράφονται οι κοινές συχνότητες των τιμών A_i και B_j, δηλαδή πόσες φορές εμφανίζονται στα δεδομένα οι τιμές A_i και B_j μαζί. Η τελευταία στήλη του πίνακα με τίτλο "Σύνολα γραμμών" περιέχει τα αθροίσματα των συχνοτήτων κάθε γραμμής και παριστά την κατανομή συχνοτήτων μόνο της μεταβλητής Α. Αντίστοιχα, η τελευταία γραμμή με τίτλο "Σύνολα στηλών" παριστά την κατανομή συχνοτήτων της μεταβλητής Β. Πρέπει να σημειωθεί εδώ ότι η επιλογή της μεταβλητής που θα παρασταθεί στις γραμμές είναι αυθαίρετη.

2.13.2 Έλεγχος Ανεξαρτησίας και Ομοιογένειας

Μετά από την κατασκευή του πίνακα συνάφειας, που μας δίνει μία συνοπτική περιγραφή της κοινής κατανομής των δύο ποιοτικών μεταβλητών, αυτό που πρέπει να εξεταστεί είναι αν οι δύο μεταβλητές είναι ανεξάρτητες(αν δηλαδή η κάθε μία από τις μεταβλητές δεν επηρεάζει την κατανομή της άλλης). Ο έλεγχος της ανεξαρτησίας γίνεται με το στατιστικό x^2. Το ίδιο στατιστικό χρησιμοποιείται για τον έλεγχο μίας άλλης έννοιας, της ομοιογένειας(αν δηλαδή οι τιμές - κατηγορίες της μίας μεταβλητής θεωρηθούν υποπληθυσμοί που κατανέμονται με τον ίδιο τρόπο για όλες τις τιμές της άλλης). Πρέπει να σημειωθεί ότι η ομοιογένεια και η ανεξαρτησία είναι τις περισσότερες φορές ισοδύναμες έννοιες.

Εκφράζοντας τα παραπάνω με στατιστική ορολογία, υποθέτουμε ότι οι δύο μεταβλητές του δείγματος προέρχονται από δύο ποιοτικά χαρακτηριστικά Α και Β του αρχικού πληθυσμού και μας ενδιαφέρει να ελέγξουμε την υπόθεση της ανεξαρτησίας:

H_0: Τα χαρακτηριστικά Α και Β είναι ανεξάρτητα μεταξύ τους
με εναλλακτική την
H_α: Τα χαρακτηριστικά Α και Β είναι εξαρτημένα.

ή ισοδύναμα την υπόθεση της ομοιογένειας:

H_0: Οι υποπληθυσμοί A_1, A_2, ..., A_R του χαρακτηριστικού Α παρουσιάζουν κατανομές ομοιογενείς ως προς το Β.
με εναλλακτική την
H_α: Οι υποπληθυσμοί A_1 , A_2 , ..., A_R δεν παρουσιάζουν ομοιογένεια ως προς το Β.

Για τον έλεγχο και των δύο αυτών υποθέσεων υπολογίζεται το στατιστικό του Pearson, που αναφέραμε προηγουμένως. Για τον υπολογισμό του εκτιμώνται πρώτα οι λεγόμενες *αναμενόμενες συχνότητες (expected frequencies)* των κελιών (οι συχνότητες δηλαδή που θα είχαμε αν τα χαρακτηριστικά ήταν πραγματικά ανεξάρτητα). Το στατιστικό x^2 είναι ουσιαστικά ένα μέτρο της απόστασης των αναμενόμενων συχνοτήτων από τις πραγματικές συχνότητες. Συγκεκριμένα, οι αναμενόμενες συχνότητες E_{ij} για κάθε κελί υπολογίζονται από τον τύπο:

E_{ij} = (άθροισμα i γραμμής) $\times$ (άθροισμα j στήλης) / (γενικό άθροισμα n)

και στη συνέχεια το x^2 από τον τύπο:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Μία άλλη παράμετρος που πρέπει να υπολογιστεί είναι οι *βαθμοί ελευθερίας (degrees of freedom)*. Αν έχουμε στον πίνακα συνάφειας R γραμμές και C στήλες, οι βαθμοί ελευθερίας είναι (R-1)(C-1). Η παράμετρος αυτή προσδιορίζει τη θεωρητική κατανομή x^2 που ακολουθεί το στατιστικό x^2.

Με αυτά τα δεδομένα, υπολογίζεται κατόπιν (με προσεγγιστικές μεθόδους) η *σημαντικότητα (significance)* του ελέγχου, που ουσιαστικά είναι η πιθανότητα λάθους όταν απορρίπτουμε τη μηδενική υπόθεση. Η σημαντικότητα αυτή πρέπει να είναι αρκετά μικρή, ώστε η απόρριψη της μηδενικής απόφασης να είναι ασφαλής. Ένα γενικά αποδεκτό όριο σφάλματος για την απόρριψη της H_0 είναι το 0.05. Αποφασίζουμε λοιπόν να απορρίπτουμε την H_0 όταν η σημαντικότητα είναι μικρότερη από 0.05.

Ένα σημείο που πρέπει να προσέξουμε στον έλεγχο, είναι το ποσοστό των κελιών με αναμενόμενη συχνότητα μικρότερη του 5. Αν το ποσοστό αυτό ξεπερνά το 20%, τότε ο έλεγχος που έχουμε εφαρμόσει δεν είναι αξιόπιστος και πρέπει να προχωρήσουμε σε σύμπτυξη κατηγοριών της μίας τουλάχιστο μεταβλητής και να επαναλάβουμε τον έλεγχο.

3. **Το Πρόγραμμα PSPP**

3.1 Τα Πλεονεκτήματα του PSPP

Από τα βασικά πλεονεκτήματα του PSPP είναι όπως προείπαμε η ευκολία στην εγκατάσταση και το πόσο ελαφρύ είναι. Μάλιστα αναφέρεται ότι υποστηρίζει τη χρήση πάνω από 1 δισεκατομμύριο περιπτώσεων και 1 δισεκατομμύριο αντίστοιχα μεταβλητές, πράγμα που σημαίνει την εκτέλεση στατιστικών διεργασιών σε πολύ μεγάλα στατιστικά σύνολα δεδομένων.

Πολλοί το θεωρούν ως την αντιγραφή του προγράμματος SPSS και όχι άδικα αφού φαίνεται να έχει ληφθεί υπόψη κατά το σχεδιασμό του, από μια άλλη οπτική βέβαια: υπάρχει πλήρης συμβατότητα με τα αρχεία που παράγονται από το δεύτερο. Έτσι, και ως προς τα αρχεία της σύνταξης, με κατάληξη .sps, και αυτά των δεδομένων, τα .sav, μπορεί να υπάρξει συνέχεια και επεξεργασία των ίδιων διεργασιών.

Ως προς τη σύγκρισή του με το SPSS, η βασική διαφορά είναι ότι δεν έχει άδειες χρήσης (αφού είναι δωρεάν).

Ένα ακόμα πλεονέκτημα είναι η χρήση του σε διαφορετικά περιβάλλοντα (cross platform) πάνω σε διαφορετικά λειτουργικά συστήματα όπως είναι τα Windows και το Linux αλλά και άλλο ελεύθερο λογισμικό με άδεια GPLv3 και μετά, ενισχύοντας τη διαλειτουργικότητα του προϊόντος.

Τέλος, μπορεί να υποστηρίξει αρκετούς και διαφορετικούς τύπους αρχείων τόσο για την είσοδο όσο και για την έξοδο των αποτελεσμάτων, εξυπηρετώντας τις ανάγκες του χρήστη και αυξάνοντας τις παραμέτρους της ικανοποίησης και αποδοτικότητας.

Ένα βασικό μειονέκτημα του πακέτου αυτού είναι ότι δεν υποστηρίζει αναπαραστάσεις σε πολλά διαγράμματα και γραφικές παραστάσεις.

3.2 Τα Χαρακτηριστικά του PSPP

Μπορεί να εξάγει τους παραγόμενους πίνακες και τα αποτελέσματα σε κείμενο, είτε με τη μορφή PostScript αρχείου, είτε σαν PDF, ακόμα και HTML ή μορφή συμβατή με το OpenOffice.

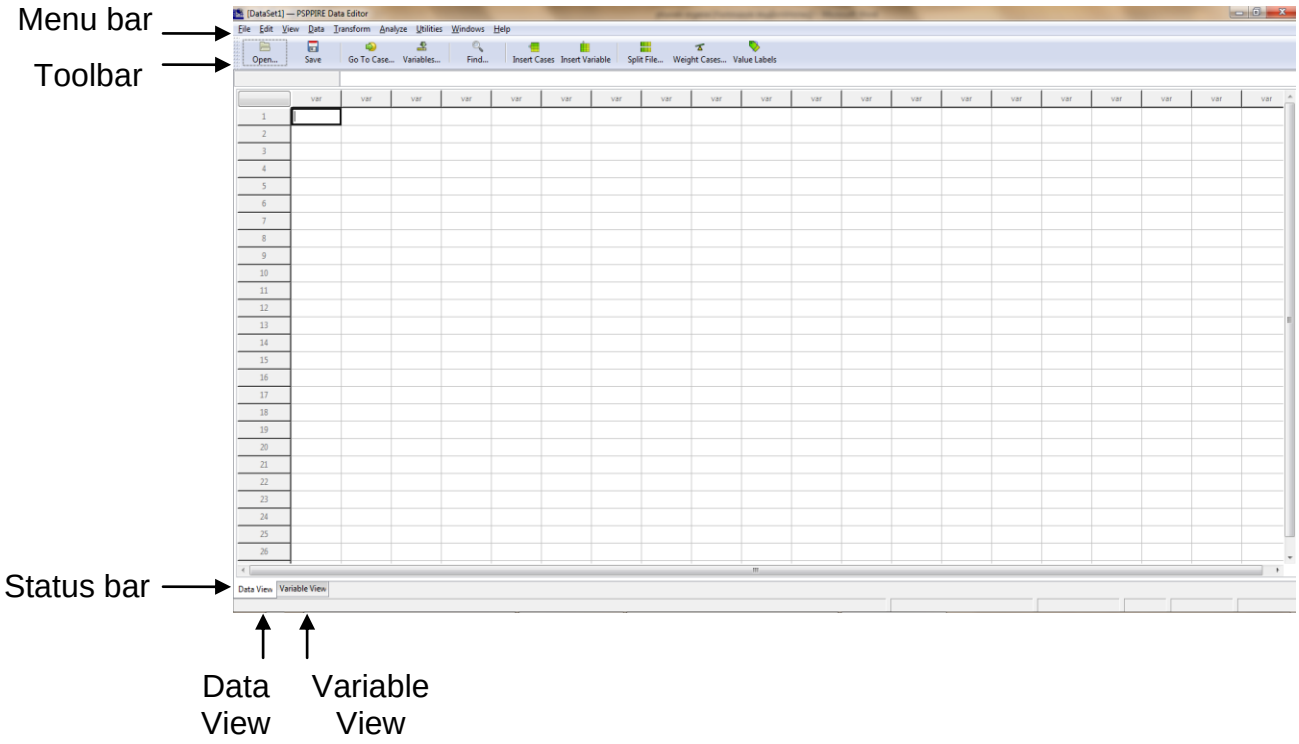
Ως προς την είσοδό του, μπορεί να υποστηρίξει τα αρχεία του Gnumeric (κάτι ανάλογο με το Excel της Microsoft), δημιουργημένο από το ProjectGnome ακόμα ένα project ελεύθερου και δωρεάν λογισμικού, αρχεία από το ίδιο το Excel, αρχεία .dat που είναι και τα πιο απλά, αρχεία δεδομένων διαχωρισμένων με κόμματα (τα CSV), Postgres αρχεία και OpenDocument, τα λεγόμενα και ODF αρχεία, ένα format αρχείου δημιουργημένο από τη SunMicrosystems, που χρησιμοποιεί XML και υποστηρίζεται απόλυτα από τη σουίτα των OpenOffice και φυσικά τα .sav αρχεία του SPSS.

4. Γραφικό Περιβάλλον και Εντολές του PSPP

4.1 Παράθυρα του PSPP

4.1.1 Data Editor Window

Στο παράθυρο αυτό εμφανίζονται τα περιεχόμενα ενός αρχείου. Εδώ μπορούμε να δημιουργήσουμε αρχεία δεδομένων, ή να τροποποιήσουμε ένα ήδη υπάρχον (Σχήμα 4.1).



Αρχικά το παράθυρο δεδομένων είναι άδειο. Κάθε γραμμή των δεδομένων θα περιέχει στοιχεία-δεδομένα για ένα συγκεκριμένο άτομο/ αντικείμενο. Κάθε στήλη αντιστοιχεί σε μια μεταβλητή.

4.1.1.1 Menu bar

Το *menu bar* περιέχει κυλιόμενα μενού, μέσω των οποίων μπορούμε να εκτελέσουμε εντολές ως εξής:

1. Κάνουμε κλικ στο όνομα του στοιχείου, όπως File.
2. Κάνουμε κλικ στην εντολή που θέλουμε από το συγκεκριμένο μενού, όπως New.
3. Αν η εντολή απαιτεί να δηλώσουμε κάποιες επιπλέον πληροφορίες, τότε το PSPP ανοίγει ένα πλαίσιο διαλόγου, διαφορετικά εκτελεί την εντολή άμεσα.

4.1.1.2 Status bar και Toolbar

Το *status bar* δηλώνει την τρέχουσα κατάσταση του επεξεργαστή του PSPP. Αν ο επεξεργαστής εκτελεί κάποια εντολή, εμφανίζει το όνομα της εντολής και έναν μετρητή ο οποίος δείχνει τον αριθμό του στοιχείου που επεξεργάζεται από το σύνολο των δεδομένων. Επίσης παρέχει πληροφορίες όπως ποια εντολή χρησιμοποιείται, αν υπάρχει κάποιο φίλτρο, αν είναι ενεργή κάποια στάθμιση των δεδομένων.

Το *toolbar* βρίσκεται ακριβώς κάτω από το menu bar και εμφανίζει εικονίδια χρήσιμα για λειτουργίες που χρησιμοποιούνται συχνά. Για μια σύντομη περιγραφή του κάθε εικονιδίου αρκεί να μετακινήσουμε το δείκτη του ποντικιού πάνω σε κάθε εικονίδιο.

4.1.1.3 Data και Variable View

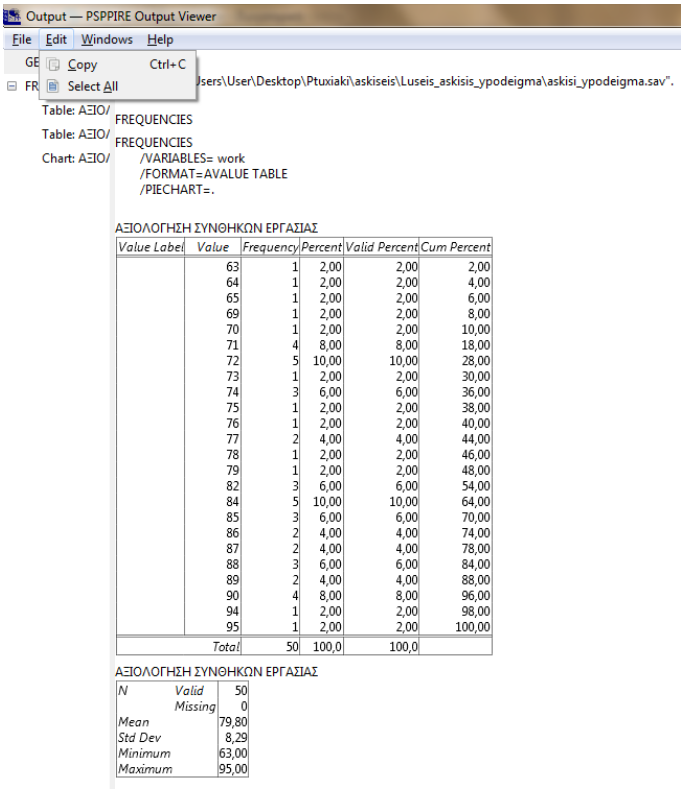
Στο παράθυρο δεδομένων του PSPPεμφανίζονται δυο βασικά παράθυρα:

1. Το *Data View*, όπου εισάγουμε, βλέπουμε και επεξεργαζόμαστε τα δεδομένα.
2. Το *Variable View*, όπου ορίζουμε πως θα εμφανίζονται τα δεδομένα (τα δεκαδικά ψηφία, τον τρόπο εμφάνισης της ημερομηνίας, το εύρος των στηλών). Επίσης στο παράθυρο αυτό ορίζουμε τα ονόματα των μεταβλητών, κάποιες επεξηγηματικές ετικέτες είτε για τις μεταβλητές είτε για τις τιμές δεδομένων που θέλουμε να εμφανίζονται στα αποτελέσματα και τον τρόπο που θα εμφανίζονται οι ελλιπείς τιμές.

4.1.2 Output Viewer Window

Στο παράθυρο αυτό παρουσιάζονται τα στατιστικά αποτελέσματα, οι πίνακες και τα γραφήματα κάθε ανάλυσης (Σχήμα 4.2). Ένα τέτοιο παράθυρο ανοίγει αυτόματα κάθε φορά

που εκτελείται μία διαδικασία για να δημιουργηθεί ένα αποτέλεσμα. Στο παράθυρο αυτό είναι δυνατή η επεξεργασία, η αντιγραφή και η εξαγωγή των αποτελεσμάτων σε διάφορες μορφές αρχείων (.pdf, .txt, .html κ.α.).



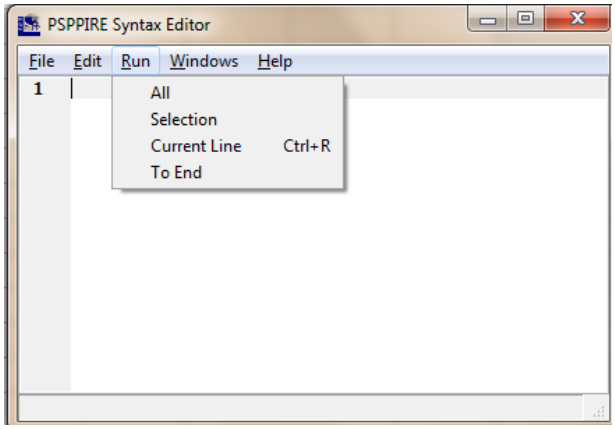
Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	63	1	2,00	2,00	2,00
	64	1	2,00	2,00	4,00
	65	1	2,00	2,00	6,00
	69	1	2,00	2,00	8,00
	70	1	2,00	2,00	10,00
	71	4	8,00	8,00	18,00
	72	5	10,00	10,00	28,00
	73	1	2,00	2,00	30,00
	74	3	6,00	6,00	36,00
	75	1	2,00	2,00	38,00
	76	1	2,00	2,00	40,00
	77	2	4,00	4,00	44,00
	78	1	2,00	2,00	46,00
	79	1	2,00	2,00	48,00
	82	3	6,00	6,00	54,00
	84	5	10,00	10,00	64,00
	85	3	6,00	6,00	70,00
	86	2	4,00	4,00	74,00
	87	2	4,00	4,00	78,00
	88	3	6,00	6,00	84,00
	89	2	4,00	4,00	88,00
	90	4	8,00	8,00	96,00
	94	1	2,00	2,00	98,00
	95	1	2,00	2,00	100,00
	Total	50	100,0	100,0	

N	Valid	50
	Missing	0
Mean		79,80
Std Dev		8,29
Minimum		63,00
Maximum		95,00

Σχήμα 4.2 Παράθυρο Output Viewer

4.1.3 Syntax Editor Window

Οι εντολές καθώς και οι επιλογές που γίνονται σε ένα παράθυρο διαλόγου μιας διεργασίας μπορούν να επικολληθούν σε ένα Syntax Editor παράθυρο και να εμφανιστούν με τη μορφή εγγεγραμμένων εντολών. Οι εντολές αυτές μπορούν να τροποποιηθούν για να χρησιμοποιηθούν κάποιες επιπλέον δυνατότητες οι οποίες δεν εμφανίζονται στα παράθυρα επιλογών. Το παράθυρο αυτό μπορεί να αποθηκευτεί προκειμένου να χρησιμοποιηθεί σε μετέπειτα αναλύσεις (Σχήμα 4.3).



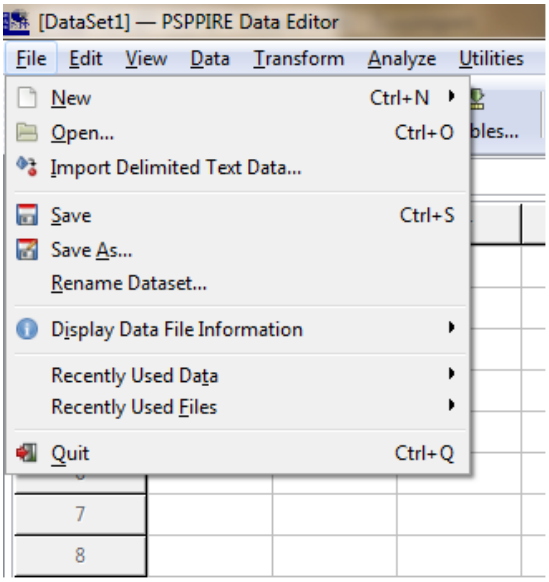
Σχήμα 4.3 Παράθυρο Syntax Editor

4.2 Οι Εντολές (διεργασίες) του PSPP μέσω των Καρτελών-Επιλογών

Στο menu bar του PSPP υπάρχουν κάποιες καρτέλες-επιλογές, οι οποίες περιέχουν το σύνολο των εντολών. Παρακάτω θα δούμε τις καρτέλες αυτές και για κάθε μια κάποιες βασικές της λειτουργίες.

4.2.1 Η Καρτέλα File

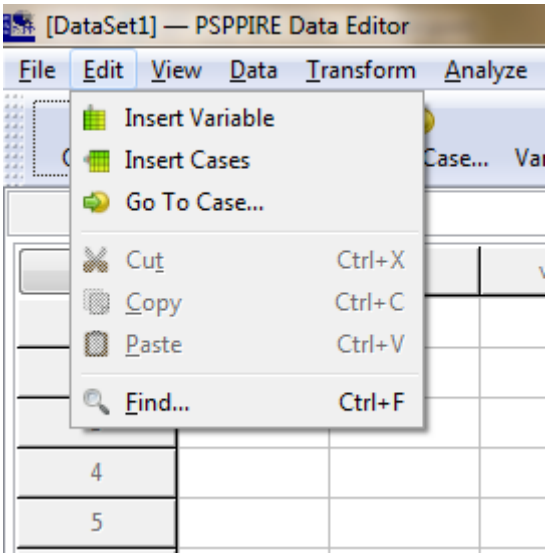
Έχει λειτουργίες για το άνοιγμα ενός αρχείου, το κλείσιμο, τη μετονομασία ολόκληρου του project και την αποθήκευση του. Επίσης μπορεί να γίνει ανάγνωση διαφόρων τύπων αρχείων (όπως .dat, .sav, .odf) τα οποία δημιουργούνται από άλλα προγράμματα λογισμικού. Μπορεί να χρησιμοποιηθεί για την ανάγνωση ενός ASCII αρχείου δεδομένων, καθώς και για την δημιουργία ενός νέου αρχείου εντολών (command file) ή την ανάκτηση ενός ήδη υπάρχον (Σχήμα 4.4).



Σχήμα 4.4 Καρτέλα File

4.2.2 Η Καρτέλα Edit

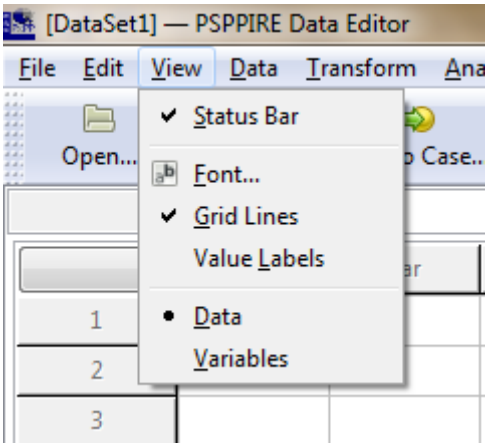
Έχει εντολές εισαγωγής νέας μεταβλητής, μετάβασης σε μεταβλητή, αντιγραφής, επικόλλησης και αποκοπής καθώς και αναζήτησης δεδομένων (Σχήμα 4.5).



Σχήμα 4.5 Καρτέλα Edit

4.2.3 Η Καρτέλα View

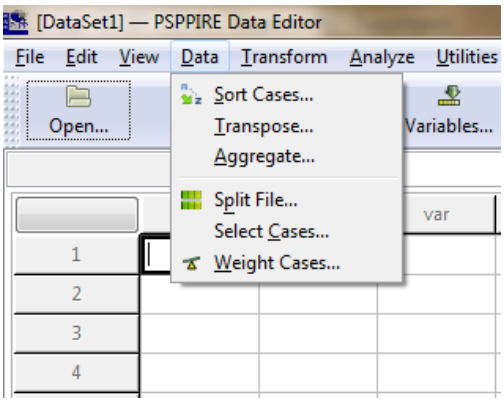
Έχει επιλογές για ενεργοποίηση ή απενεργοποίηση της status bar, μορφοποίησης των κελιών με εμφάνιση πλέγματος ή χωρίς και επιλογής γραμματοσειράς. Επίσης διαθέτει εναλλαγή προβολής των δεδομένων μεταξύ data view και variable view (Σχήμα 4.6).



Σχήμα 4.6 Καρτέλα View

4.2.4 Η Καρτέλα Data

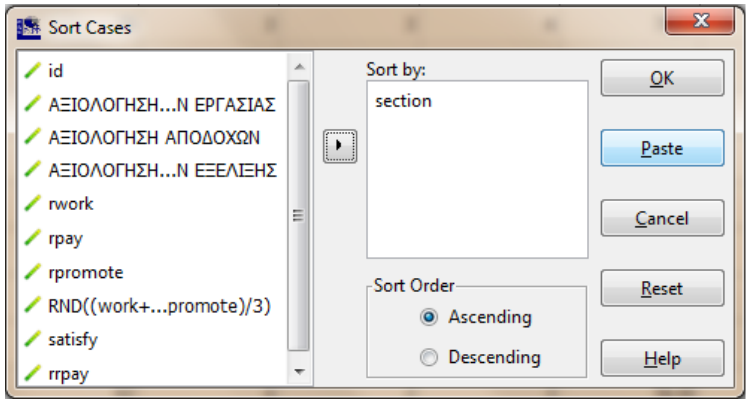
Χρησιμοποιούμε το μενού Data για να κάνουμε καθολικές αλλαγές σε PSPP αρχεία δεδομένων, όπως μεταφορά μεταβλητών και περιπτώσεων ή δημιουργώντας υποσύνολα των περιπτώσεων για ανάλυση. Αυτές οι αλλαγές είναι προσωρινές και δεν επηρεάζουν το μόνιμο αρχείο εκτός και αν αποθηκεύσουμε το αρχείο μας με τις αλλαγές. Παρακάτω θα δούμε μία μία τις εντολές που περιλαμβάνει η συγκεκριμένη καρτέλα (Σχήμα 4.7).



Σχήμα 4.7 Καρτέλα Data

- **Sort Cases**

Η εντολή ταξινομεί το ενεργό σύνολο δεδομένων κατά μία ή περισσότερες μεταβλητές (Σχήμα 4.8). Είναι απαραίτητος ο προσδιορισμός του *Sort by*, και της λίστας των μεταβλητών που θέλουμε να ταξινομήσουμε. Εξ ορισμού, οι μεταβλητές είναι ταξινομημένες κατά αύξουσα σειρά, αν θέλουμε όμως αυτή τη ρύθμιση μπορούμε να την αλλάξουμε. Σε περιπτώσεις ισοβαθμίας, όταν δηλαδή δύο ή παραπάνω παρατηρήσεις έχουν ίσες τιμές ως προς τη μεταβλητή που ταξινομούνται, τότε κρατάνε την ίδια θέση πριν και μετά την ταξινόμηση. Η ταξινόμηση δηλαδή είναι μια διαδικασία που αν αναιρεθεί δεν χαλάει τη σειρά των παρατηρήσεων. Γι' αυτό και λέμε ότι η Sort Cases είναι μόνο για αλλαγές στην ανάγνωση.



Σχήμα 4.8 Εντολή Sort Cases

- **Transpose**

Δημιουργεί ένα νέο αρχείο δεδομένων στο οποίο οι γραμμές και οι στήλες του αρχικού αρχείου μεταφέρονται έτσι ώστε οι περιπτώσεις (σειρές) να γίνονται μεταβλητές και οι μεταβλητές (στήλες) να γίνονται υποθέσεις. Τα νέα ονόματα μεταβλητών δημιουργούνται αυτόματα και εμφανίζονται σε μια λίστα.

- Μια νέα μεταβλητή τύπου String περιέχει το αρχικό όνομα της μεταβλητής, case_lbl, το οποίο δημιουργείται αυτόματα.
- Εάν το ενεργό σύνολο δεδομένων περιέχει ένα αναγνωριστικό(ID) ή όνομα μεταβλητής με μοναδικές τιμές, μπορεί να χρησιμοποιηθεί ως όνομα της μεταβλητής, και οι τιμές της, μπορούν να χρησιμοποιηθούν ως ονόματα μεταβλητών. Αν πρόκειται για μια αριθμητική μεταβλητή, το όνομα της μεταβλητής αρχίζει με το γράμμα V, ακολουθούμενο από την αριθμητική τιμή.

- **Aggregate**

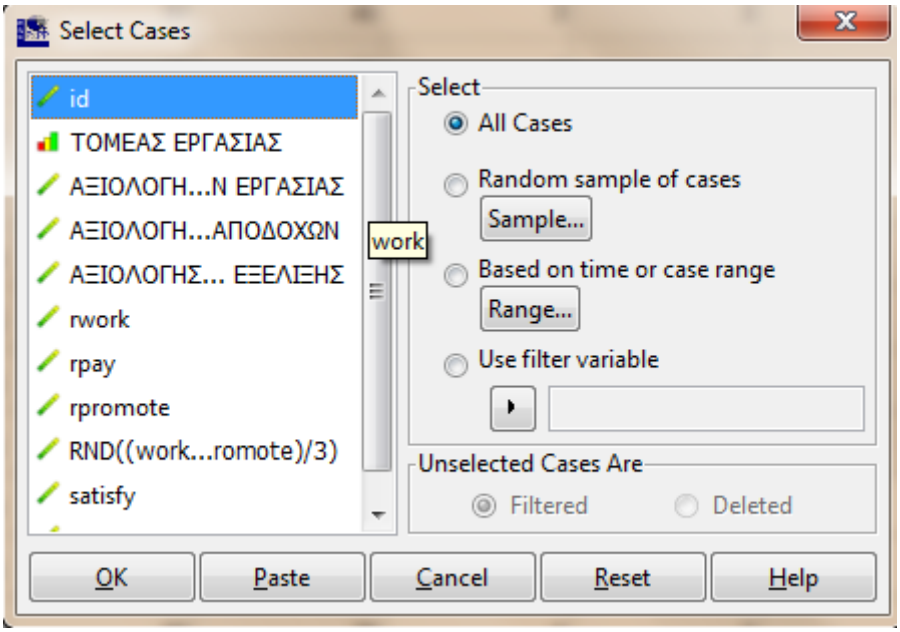
  Η διαδικασία του aggregate συνοψίζει τις ομάδες των περιπτώσεων σε μεμονωμένες περιπτώσεις. Οι περιπτώσεις/παρατηρήσεις χωρίζονται σε ομάδες που έχουν τις ίδιες τιμές για μία ή περισσότερες μεταβλητές που ονομάζονται μεταβλητές «break values». Διάφορες λειτουργίες είναι διαθέσιμες για τη συνοπτική παρουσίαση των περιεχομένων των περιπτώσεων. Τα αποτελέσματα μπορούν να αποθηκευτούν προστιθέμενα στο ίδιο αρχείο δεδομένων, αντικαθιστώντας τα προηγούμενα ή και σε διαφορετικό αρχείο.

- **Split File**

  Η εντολή επιτρέπει σε πολλαπλές σειρές δεδομένων που συνυπάρχουν σε ένα αρχείο δεδομένων να μπορούν να αναλυθούν ξεχωριστά με στατιστικές εντολές. Καταρχήν πρέπει να οριστεί η λίστα με τα ονόματα των μεταβλητών που θα αναλυθούν ξεχωριστά. Οι ομάδες των κοντινών παρατηρήσεων που έχουν τις ίδιες τιμές για αυτές τις μεταβλητές θα αναλυθούν με στατιστικές εντολές σαν μία ομάδα. Ανεξάρτητη ανάλυση εκτελείται για κάθε ομάδα παρατηρήσεων, και οι τιμές των μεταβλητών για την ομάδα προβάλλονται στο αποτέλεσμα μαζί με την ανάλυση. Υπάρχουν όμως και κάποιοι περιορισμοί. Για παράδειγμα, οι ομάδες που μπορούν να σχηματιστούν πρέπει να αποτελούνται μόνο από γειτονικές-διπλανές παρατηρήσεις. Σε περίπτωση όμως που χρειάζεται αυτή η διάσπαση, μπορείτε να χρησιμοποιήσετε την μεταβλητή, όπου, όπως είπαμε οι τιμές της δεν είναι διπλανές στο αρχείο δεδομένων, αφού πρώτα διατάξουμε τα δεδομένα της εν λόγω μεταβλητής ώστε να προκαλέσουμε το επιθυμητό αποτέλεσμα. Η συγκεκριμένη εντολή πρέπει να καθοριστεί «OFF» προκειμένου να συνεχίσετε την ανάλυση του ενεργού συνόλου δεδομένων ως ένα ενιαίο σύνολο δεδομένων.

- **Select Cases**

  Το PSPP μας δίνει την δυνατότητα να επιλέξουμε ένα μέρος των δεδομένων για περαιτέρω ανάλυση, εξαιρώντας τις υπόλοιπες περιπτώσεις (Σχήμα 4.9). Επιλέγοντας την εντολή επιλογής περιπτώσεων από την καρτέλα Dataεμφανίζεται ένα παράθυρο διαλόγου στο οποίο πρέπει να δηλώσουμε τον τρόπο με τον οποίο επιθυμούμε να γίνει η επιλογή των περιπτώσεων. Συγκεκριμένα:



Σχήμα 4.9 Εντολή Select Cases

All cases: Δηλώνουμε ότι θα εργαστούμε με όλες τις περιπτώσεις των δεδομένων.

Random sample of cases: Δηλώνουμε ότι θα εργαστούμε με ένα τυχαίο δείγμα περιπτώσεων από τα δεδομένα μας. Για να καθορίσουμε με ποιο τρόπο θα επιλεγεί το τυχαίο δείγμα πρέπει να πατήσουμε το πλήκτρο *Sample*. Εδώ έχουμε τις εξής δυνατότητες:
Approximately _% of all cases: Πρέπει να εισάγουμε έναν ακέραιο αριθμό από 1 ως 99 σαν ποσοστό έτσι ώστε να προκύψει προσεγγιστικά το ζητούμενο ποσοστό των περιπτώσεων.
Exactly _ cases from the first _ cases: Πρέπει να εισάγουμε έναν ακέραιο που να δηλώνει το ακριβές μέγεθος του δείγματος και έναν δεύτερο που θα δηλώνει από πόσες πρώτες περιπτώσεις θα επιλεγεί το δείγμα. Ο δεύτερος αυτός ακέραιος δεν πρέπει να είναι μεγαλύτερος από τον συνολικό αριθμό περιπτώσεων στο αρχείο δεδομένων. Αν εισάγουμε τον ίδιο ακέραιο n και στις δύο θέσεις, θα επιλεγούν οι n πρώτες περιπτώσεις του αρχείου δεδομένων.

Based on time or case range: Δηλώνουμε ότι θα εργαστούμε μόνο με τις περιπτώσεις αυτές που ανήκουν σε ένα διάστημα καθορισμένων τιμών. Πατώντας το

πλήκτρο *Range* μπορούμε να ορίσουμε την πρώτη και τελευταία περίπτωση του διαστήματος με το οποίο θέλουμε να εργαστούμε.

Use filter variable: Δηλώνουμε ότι επιθυμούμε να χρησιμοποιήσουμε τις τιμές μιας υπάρχουσας αριθμητικής μεταβλητής για να ορίσουμε το φιλτράρισμα των περιπτώσεων. Αρκεί να επιλέξουμε μία μεταβλητή από τη λίστα και τότε οι περιπτώσεις που αντιστοιχούν στην τιμή 0 της μεταβλητής αυτής θα αποκλειστούν από την ανάλυση.

Η τελευταία επιλογή που μένει να κάνουμε είναι για τον τρόπο με τον οποίο θα διαχειριστεί το PSPP τις περιπτώσεις που δεν έχουν επιλεγεί για επεξεργασία (Unselected Cases Are).
Filtered: Οι περιπτώσεις που δεν έχουν επιλεγεί απλά αγνοούνται κατά τη διάρκεια οποιασδήποτε επεξεργασίας αλλά το αρχείο δεδομένων διατηρείται όπως ήταν. Στην περίπτωση αυτή η ακύρωση της επιλογής είναι απλούστατη και γίνεται από το ίδιο πλαίσιο διαλόγου αλλάζοντας τις κατάλληλες παραμέτρους.
Deleted: Οι μη επιλεγμένες περιπτώσεις διαγράφονται από το αρχείο δεδομένων. Για να τις επανακτήσουμε πρέπει να ανοίξουμε πάλι το αρχικό αρχείο. Απαιτείται μεγάλη προσοχή διότι αν μετά από αυτή την επιλογή γίνει αποθήκευση του αρχείου δεδομένων τα αρχικά δεδομένα θα αλλάξουν αφού θα έχουν διαγραφεί περιπτώσεις.

Αν εκτελέσουμε την εντολή Select Cases και αν επιπλέον ζητήσουμε οι μη επιλεγμένες περιπτώσεις να φιλτραριστούν (Filtered), τότε θα δημιουργηθεί μία νέα μεταβλητή με όνομα filter_$ η οποία θα έχει τιμές 1 (επιλεγμένες περιπτώσεις) και 0 (μη επιλεγμένες περιπτώσεις). Απαιτείται λοιπόν να προσέχουμε την μεταβλητή αυτή να μην αλλοιωθεί από λάθος.

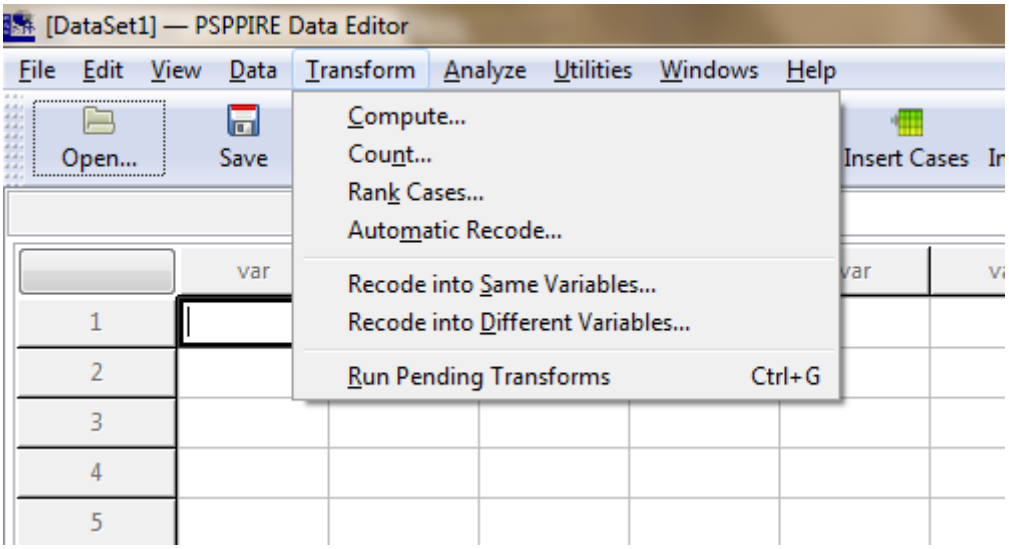
- **Weight Cases**

Η εντολή εκχωρεί στις παρατηρήσεις διαφορετικά βάρη, αλλάζοντας την κατανομή συχνότητας του ενεργού συνόλου δεδομένων. Όταν γίνει απενεργοποίηση της εντολής, οι μετέπειτα στατιστικές διαδικασίες θα σταθμίζουν όλες τις παρατηρήσεις εξίσου. Για παράδειγμα, ένας θετικός ακέραιος w ως συντελεστής στάθμισης για κάθε παρατήρηση θα δώσει την ίδια στατιστική, αποτέλεσμα που θα

είχαμε αν επαναλαμβάναμε την υπόθεση w φορές. Λειτουργεί δηλαδή σαν συχνότητα εμφάνισης. Ένας παράγοντας στάθμισης ίσος με 0 αντιμετωπίζεται για στατιστικούς λόγους, όπως αν η παρατήρηση δεν υπήρχε καν στην είσοδο. Οι τιμές της στάθμισης δεν χρειάζεται να είναι ακέραιοι, αλλά μπορεί να είναι και αρνητικές ή ακόμα και system missing values, όπου σε αυτή την περίπτωση, για τη μεταβλητή στάθμισης θα ερμηνευθεί όπως και ο συντελεστής στάθμισης 0. Οι user missing values δεν έχουν καμιά ειδική αντιμετώπιση.

4.2.5 Η Καρτέλα Transform

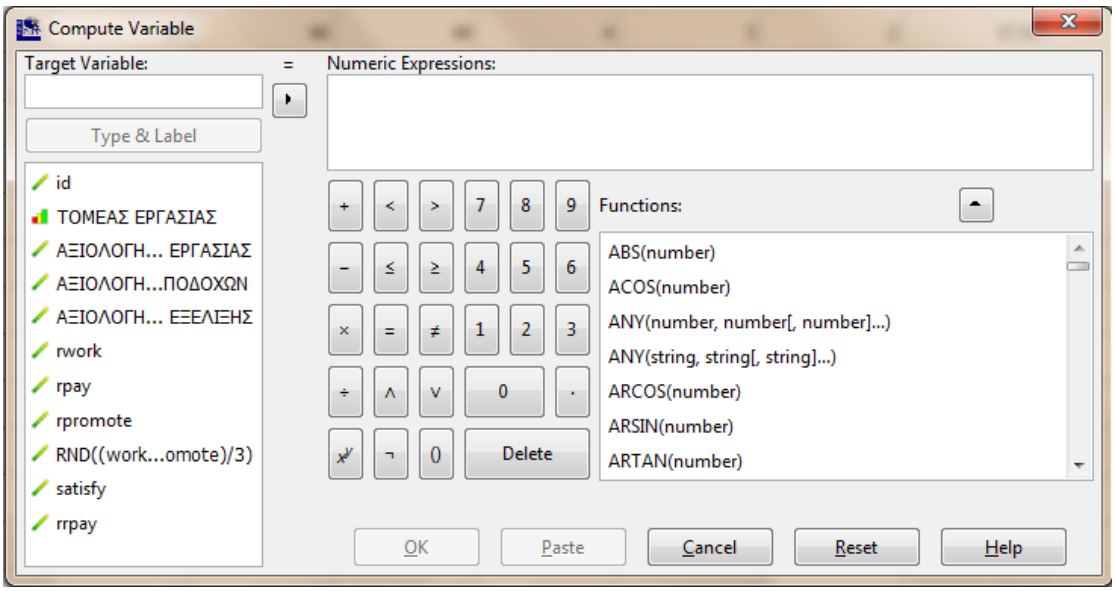
Το μενού *Transform,* το χρησιμοποιούμε για να κάνουμε αλλαγές σε επιλεγμένες μεταβλητές στο αρχείο δεδομένων και να υπολογίσουμε νέες μεταβλητές με βάση τις τιμές των υφιστάμενων. Αυτές οι αλλαγές είναι προσωρινές και δεν επηρεάζουν το μόνιμο αρχείο, εκτός αν στη συνέχεια αποθηκεύσουμε το αρχείο με τις αλλαγές που πραγματοποιήσαμε. Οι πιο σημαντικές εντολές που περιέχονται στη συγκεκριμένη καρτέλα είναι η *compute*, η *count* και η *recode* των οποίων τις λειτουργίες θα περιγράψουμε παρακάτω (Σχήμα 4.10).



Σχήμα 4.10 ΚαρτέλαTransform

- **Compute**

Η εντολή χρησιμοποιείται για τη δημιουργία τιμών νέων μεταβλητών, με την εκτέλεση υπολογισμών πάνω σε τιμές μεταβλητών που ήδη υπάρχουν. Στο παράθυρο διαλόγου που εμφανίζεται (Σχήμα 4.11) υπάρχουν κάποιες επιλογές που θα αναλύσουμε:



Σχήμα 4.11 Παράθυρο διαλόγου εντολής Compute

Target Variable: Είναι το όνομα της μεταβλητής που θα δεχθεί τις τιμές που προκύπτουν από έναν υπολογισμό. Είναι δυνατό να είναι μία ήδη υπάρχουσα μεταβλητή ή μία νέα μεταβλητή. Εξ ορισμού οι νέες μεταβλητές είναι αριθμητικές.

Numeric Expresssion: Είναι η έκφραση που χρησιμοποιείται για τον υπολογισμό των τιμών της μεταβλητή-στόχου (target variable). Η έκφραση αυτή είναι δυνατό να περιέχει ονόματα μεταβλητών, σταθερές, αριθμητικούς τελεστές και συναρτήσεις. Στο πλαίσιο κειμένου μπορούμε να πληκτρολογήσουμε και να επεξεργασθούμε την έκφραση. Επίσης μπορούμε να χρησιμοποιήσουμε το πινακίδιο υπολογισμών (calculator pad), τη λίστα των μεταβλητών (variable list) και τη λίστα των συναρτήσεων (functions) για να επικολλήσουμε στοιχεία στην έκφραση που εδώ δομούμε. Το πινακίδιο υπολογισμών περιέχει αριθμούς, αριθμητικούς τελεστές και σχεσιακούς τελεστές. Η λίστα των συναρτήσεων περιέχει περισσότερες από εβδομήντα ενσωματωμένες συναρτήσεις, που διακρίνονται σε αριθμητικές,

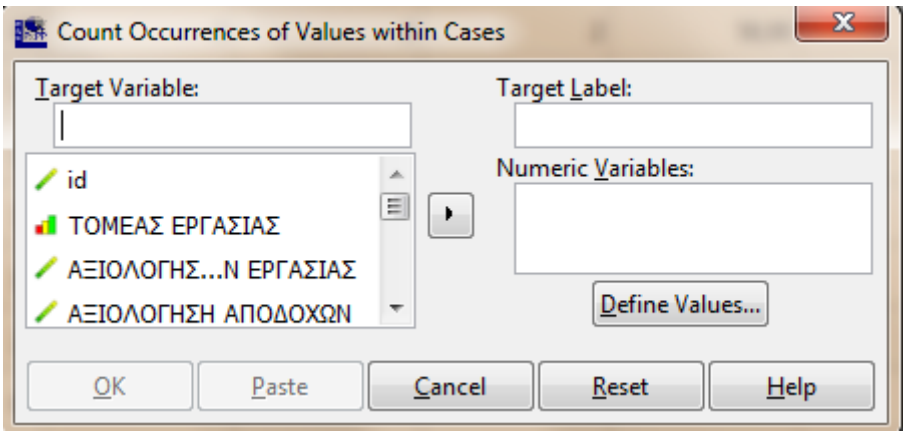
στατιστικές, λογικές ημερομηνίας και ώρας, συναρτήσεις χαμένων τιμών και αλφαριθμητικές συναρτήσεις.

Type and Label: Ενεργοποιείται αφού δοθεί όνομα στην *Target Variable*. Στην επιλογή αυτή συναντάμε τις παρακάτω επιλογές:

- *Label*: Για να αναθέσουμε μία ετικέτα σε μεταβλητή έχουμε δύο δυνατότητες.
  Label: Εισάγουμε μία ετικέτα με μέγιστο δυνατό μήκος τους 120 χαρακτήρες.
  Use expression as label: Οι πρώτοι 110 χαρακτήρες της έκφρασης υπολογισμού χρησιμοποιούνται ως ετικέτα.
- *Type:* Για να αναθέσουμε τον τύπο δεδομένων της μεταβλητής έχουμε δύο δυνατότητες.
  Numeric: Είναι η εξ ορισμού ρύθμιση
  String: Αλφαριθμητική μεταβλητή.
  Width: Εισάγουμε το μέγιστο μήκος, μόνο στην προηγούμενη περίπτωση της αλφαριθμητικής μεταβλητής.

➢ **Count:**

Δημιουργεί ή αντικαθιστά μια αριθμητική μεταβλητή-στόχο που μετράει την εμφάνιση μιας τιμής που ικανοποιεί κάποια κριτήρια ή σύνολο τιμών με τη χρήση μίας ή περισσότερων μεταβλητών δοκιμής για κάθε περίπτωση. Οι τιμές μεταβλητές-στόχοι για τη διαδικασία COUNT είναι πάντα μη αρνητικοί ακέραιοι (Σχήμα 4.12).



Σχήμα 4.12 Παράθυρο διαλόγου εντολής Count

Από το παράθυρο διαλόγου ο χρήστης μπορεί να εισάγει ένα όνομα μεταβλητής στόχου (*Target Variable*), έπειτα να επιλέξει δύο ή περισσότερες μεταβλητές του

ίδιου τύπου (αριθμητικά ή String) και τέλος μέσα από την επιλογή για καθορισμό τιμών *(Define Values*) να καθορίσει ποια τιμή ή τιμές θα πρέπει να υπολογίζονται.

- **Rank Cases**

Με την εντολή *Rank Cases* αναθέτουμε τάξεις μεγέθους στα δεδομένα των στηλών που επιλέγουμε. Στο αρχικό παράθυρο διαλόγου, τοποθετούμε τις μεταβλητές των οποίων οι τιμές θα χρησιμοποιηθούν. Μετά από κάθε μεταβλητή πρέπει να ορίσουμε τον τρόπο που θα κατατάσσεται, σε αύξουσα (προεπιλογή) ή φθίνουσα σειρά. Στο πλαίσιο *By,* εισάγουμε μία λίστα μεταβλητών που θα χρησιμεύουν ως μεταβλητές ομάδας. Στην περίπτωση αυτή, οι μεταβλητές συγκεντρώνονται σε ομάδες, και κατατάσσονται στην κάθε ομάδα. Από το *Rank Types,* μπορούμε να επιλέξουμε αν θέλουμε απλά τάξεις μεγέθους για τα δεδομένα ή ποσοστά. Τέλος, μέσω της επιλογής *Ties* επιλέγουμε την τάξη μεγέθους σε περίπτωση ισοβαθμούντων τάξεων μεγέθους. Η προεπιλογή είναι ο μέσος όρος των τάξεων.

- **Automatic Recode**

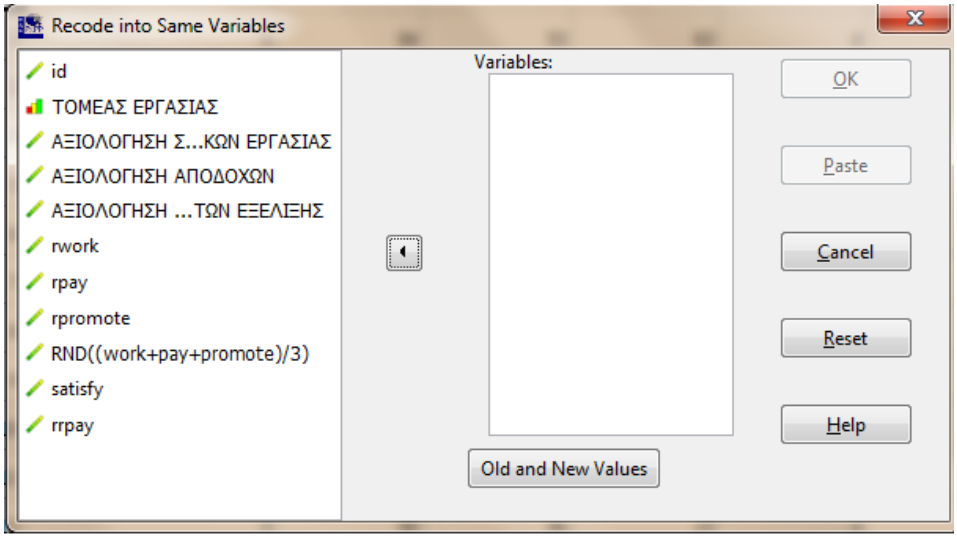
Χρησιμοποιείται για να αλλάξετε τις τιμές της μιας μεταβλητής σε άλλες τιμές. Μπορεί να πάρει n τιμές μιας μεταβλητής και να τις κάνει να αντιστοιχούν σε νέες τιμές από το 1 ως n σε μια νέα αριθμητική μεταβλητή με νέο όνομα που το ορίζουμε εμείς. Για παράδειγμα , μπορεί να θέλετε να μετατρέψετε τις τιμές μιας μεταβλητής τύπου String (π.χ. Μαθηματικά , Αγγλικά) σε μια αριθμητική μεταβλητή (π.χ. μαθηματικά γίνεται 1 , αγγλικά γίνεται 2 , κλπ.). Ο ευκολότερος τρόπος είναι να αφήσετε το PSPP να το κάνει αυτόματα (Automatic Recode) για εσάς. Θα ταξινομήσει τις τιμές μιας μεταβλητής-String από το υψηλότερο στο χαμηλότερο (ή το αντίστροφο) και στη συνέχεια θα εκχωρήσει συνεχόμενους αριθμούς στις τιμές . Προσθέτει επίσης την τιμή συμβολοσειράς ως ετικέτα στην αριθμητική τιμή.

- **Recode**

Με αυτή την εντολή είναι δυνατή η επανακωδικοποίηση μιας μεταβλητής που ήδη υπάρχει, με δύο διαφορετικές δυνατότητες: Είτε τη μετατροπή της ίδιας μεταβλητής με διαγραφή των παλαιών τιμών και αντικατάστασή τους από τις νέες είτε τη

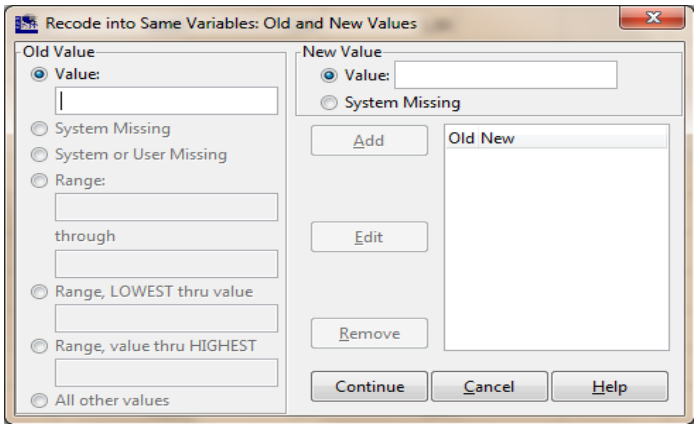
δημιουργία νέας μεταβλητής με τιμές τις κωδικοποιημένες τιμές. Τις δύο αυτές δυνατότητες θα παρουσιάσουμε στη συνέχεια.

- *Επανακωδικοποίηση μεταβλητών με αντικατάσταση των τιμών τους (Recode into Same Variables).* Στο παράθυρο διαλόγου επιλέγουμε από τη λίστα των μεταβλητών αυτή που πρόκειται να μετασχηματίσουμε και τη μεταφέρουμε μέσα στο πλαίσιο Variables. Αν επιλέξουμε περισσότερες από μία μεταβλητές, πρέπει να είναι του ίδιου τύπου (Σχήμα 4.13).



Σχήμα 4.13 Αρχικό παράθυρο διαλόγου εντολής Recode into Same Variables

Στη συνέχεια, για να προσδιορίσουμε τις νέες τιμές της μεταβλητής, ενεργοποιούμε την επιλογή *Old and New Values*. Εδώ για κάθε τιμή (ή εύρος τιμών) που θέλουμε να επανακωδικοποιήσουμε, προσδιορίζουμε την παλαιά τιμή (Old value) και τη νέα τιμή (New value) και στη συνέχεια ενεργοποιούμε το πλήκτρο Add. Είναι δυνατή η επανακωδικοποίηση πολλών παλαιών τιμών σε μία μοναδική τιμή, ενώ δεν είναι δυνατή η επανακωδικοποίηση μίας μοναδικής παλαιάς τιμής σε πολλές νέες τιμές (Σχήμα 4.14).



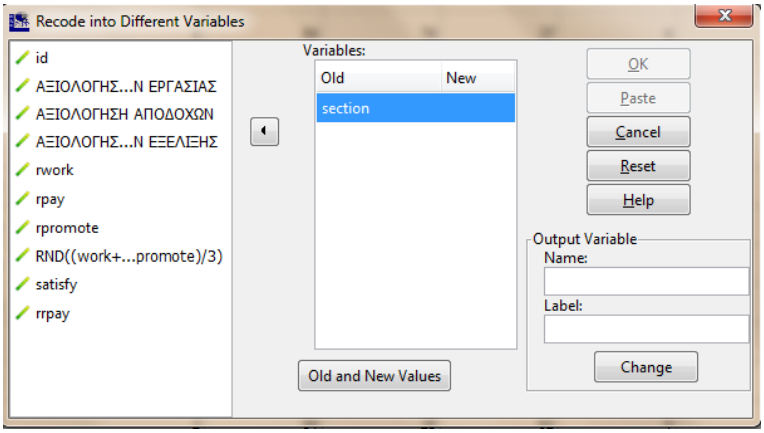
Σχήμα 4.14 Παράθυρο διαλόγου για προσδιορισμό παλαιών και νέων τιμών της μεταβλητής

- *Επανακωδικοποίηση μεταβλητών με δημιουργία νέων μεταβλητών (Recode into Different Variables).* Οι επιλογές που παρουσιάζονται στο παράθυρο διαλόγου είναι οι εξής (Σχήμα 4.15) :

  Input Variable ->Output Variable: Εδώ εμφανίζονται οι μεταβλητές που επιλέγονται από τη λίστα. Ένα ερωτηματικό στη στήλη output variable δηλώνει ότι ένα όνομα είναι απαραίτητο για την κάθε μεταβλητή.

  Output variable: Εδώ ορίζουμε και δίνουμε (προαιρετικά) ετικέτα στη νέα μεταβλητή που δέχεται τις μετασχηματισμένες τιμές.

  - *Name*: Επιλέγουμε τη μεταβλητή εισόδου από τον κατάλογο των επιλεγμένων μεταβλητών και στη συνέχεια, πληκτρολογούμε ένα όνομα για τη μεταβλητή εξόδου που της αντιστοιχεί. Τα ονόματα των μεταβλητών ακολουθούν τους γνωστούς κανόνες.
  - *Label*: Οι ετικέτες που εισάγουμε εδώ μπορούν να έχουν μέχρι και 120 χαρακτήρες.



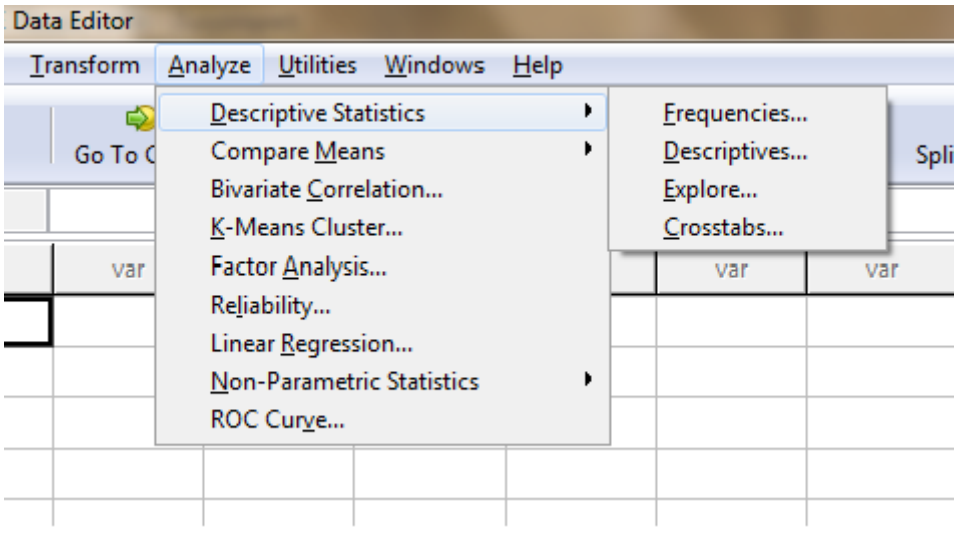
Σχήμα 4.15 Αρχικό παράθυρο διαλόγου εντολής Recode into Different Variables

Μετά την εισαγωγή του ονόματος της μεταβλητής και της ετικέτας της, ενεργοποιούμε το πλήκτρο Changeπροκειμένου να εισάγουμε το όνομα στον κατάλογο “*output variable*” δίπλα στο όνομα της “*input variable*” στην οποία αντιστοιχεί.

Old and New Values: Η ενεργοποίηση του πλήκτρου αυτού έχει ως αποτέλεσμα την εμφάνιση ενός παραθύρου διαλόγου παρόμοιο της επιλογής *Recode into Same Variables* μέσω του οποίου γίνεται η επανακωδικοποίηση των τιμών των μεταβλητών προσδιορίζοντας την παλαιά και την νέα τιμή.

4.2.6 Η Καρτέλα Analyze

Το μενού των εντολών που περιλαμβάνει η συγκεκριμένη καρτέλα ασχολείται με την βασική στατιστική ανάλυση των στοιχείων. Η έμφαση δίνεται στη χρήση του στατιστικού πακέτου για την απόκτηση των αποτελεσμάτων και όχι στην ερμηνεία των εννοιών (Σχήμα 4.16).



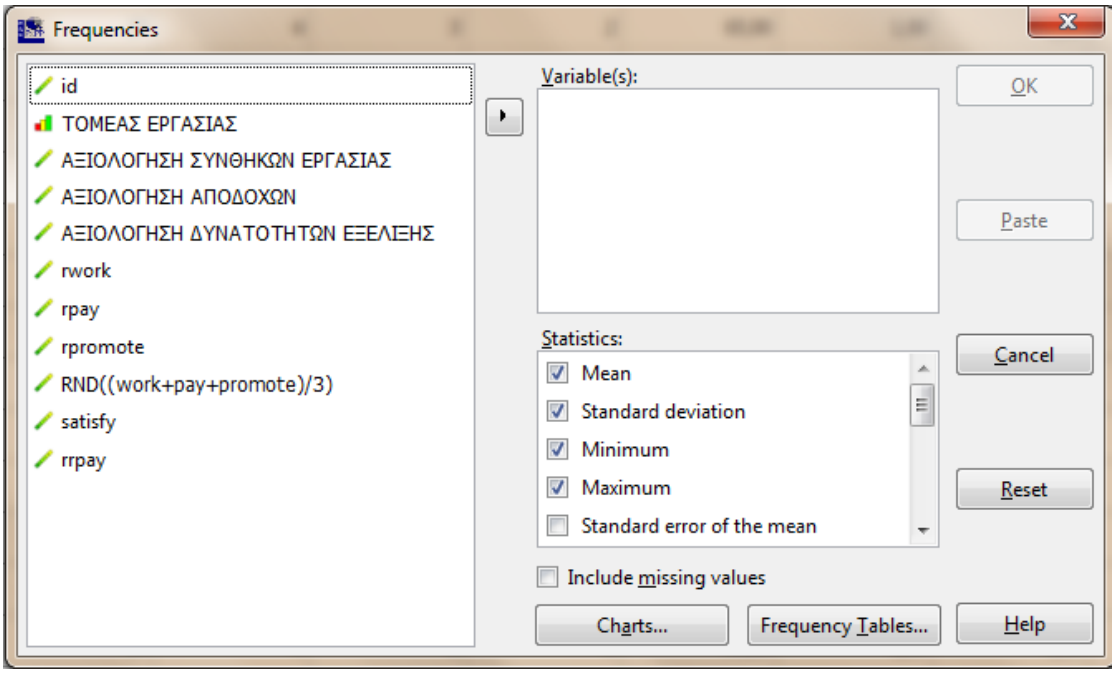
Σχήμα 4.16 Καρτέλα Analyze

4.2.6.1 Descriptive Statistics (Περιγραφική Στατιστική)

- **Frequencies**

Η διαδικασία εμφανίζει πίνακες κατανομής συχνοτήτων και γραφικές παραστάσεις (ραβδογράμματα και ιστογράμματα) και υπολογίζει στατιστικά μέτρα. Η διαδικασία εκτελείται εφόσον δηλωθεί μία τουλάχιστον μεταβλητή. Τα ελάχιστα αποτελέσματα (default output) που παράγει είναι ένας πίνακας κατανομής συχνοτήτων για κάθε μεταβλητή. Χρησιμοποιείται κυρίως για περιγραφική στατιστική κατηγοριοποιημένων μεταβλητών δηλαδή είτε ποιοτικών είτε ποσοτικών διακριτών.

Στο παράθυρο διαλόγου που εμφανίζεται πρέπει να δηλώσουμε τις μεταβλητές στις οποίες θέλουμε να γίνει περιγραφική στατιστική. Αυτό γίνεται με την επιλογή της κάθε μεταβλητής και τη μεταφορά της στη λίστα *Variable(s)*με το βοηθητικό βέλος. Αρκεί να επιλέξουμε μία τουλάχιστο μεταβλητή έτσι ώστε να ενεργοποιηθούν οι επιλογές OK και Paste στο παράθυρο (Σχήμα 4.17).



Σχήμα 4.17 Αρχικό παράθυρο διαλόγου της εντολής Frequencies

Επιλέγοντας το τετραγωνίδιο *Frequency Tables* στα αποτελέσματα θα εμφανιστεί πίνακας κατανομής , για κάθε μεταβλητή που ορίσαμε.Ο πίνακας αυτός έχει σαν τίτλο το όνομα της μεταβλητής συνοδευόμενο από την ετικέτα με την οποία δηλώθηκε από την εντολή *Labels* και αποτελείται από τις εξής στήλες:

- *Value Label:* Οι ετικέτες της κάθε τιμής όπως αυτές έχουν δηλωθεί από την εντολή labels.
- *Value:* Οι τιμές της μεταβλητής όπως αυτές είναι καταχωρημένες στα δεδομένα. Οι αριθμητικές τιμές είναι διαταγμένες σε αύξουσα σειρά ενώ οι αλφαριθμητικές διατάσσονται αλφαβητικά. Οι χαμένες τιμές εμφανίζονται και αυτές.
- *Frequency*: Οι συχνότητες εμφάνισης της κάθε τιμής.
- *Percent:* Το % ποσοστό εμφάνισης της κάθε τιμής.
- *Valid Percent:* Το έγκυρο % ποσοστό εμφάνισης της κάθε τιμής (στο σύνολο των μη χαμένων τιμών).
- *Cum Percent:* Το αθροιστικό ποσοστό της κάθε τιμής (στο σύνολο των μη χαμένων τιμών).

Στο ίδιο παράθυρο μπορούμε να επιλέξουμε την σειρά των τιμών στον πίνακα καθώς και την εμφάνιση ενός δείκτη.

Statistics: Επιλέγουμε τα στατιστικά μέτρα που θέλουμε να υπολογίσουμε για κάθε μεταβλητή. Τα στατιστικά μέτρα που είναι διαθέσιμα εδώ είναι τα μέτρα κεντρικής τάσης, τα μέτρα μεταβλητότητας, τα μέτρα σχήματος και ποσοστιαία σημεία της κατανομής της μεταβλητής.

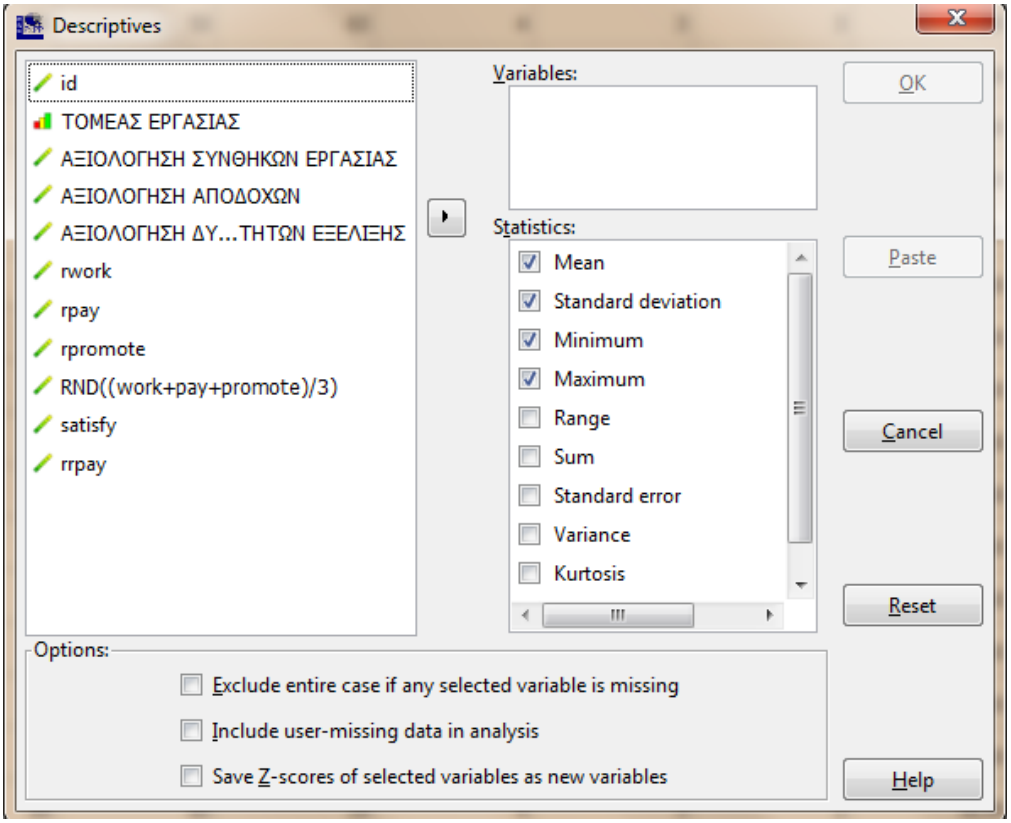
Charts: Το επιλέγουμε για να ζητήσουμε τη δημιουργία γραφικών παραστάσεων (διαγράμματα πίτας και ιστογράμματα). Αν επιλέξουμε τη δημιουργία ιστογράμματος και την επιλογή *Super impose normal curve* δίνει τη δυνατότητα να σχεδιαστεί μία καμπύλη της κανονικής κατανομής επάνω από το ιστόγραμμα έτσι ώστε να μπορούμε να κρίνουμε αν τα δεδομένα μας προέρχονται από κανονική κατανομή.

- **Descriptives**

Υπολογίζει και εμφανίζει στατιστικά μέτρα για ποσοτικές, συνεχείς κυρίως, μεταβλητές για τις οποίες δεν χρειάζεται να κατασκευαστεί πίνακας κατανομής συχνοτήτων.

Στο παράθυρο διαλόγου (Σχήμα 4.18) πρέπει να ορίσουμε τις μεταβλητές στις οποίες θέλουμε να υπολογιστούν τα περιγραφικά στατιστικά μέτρα. Αρκεί και εδώ να

μεταφέρουμε μία τουλάχιστο μεταβλητή στη λίστα μεταβλητών *Variable(s)* έτσι ώστε να ενεργοποιηθούν οι επιλογές OK και Paste στο παράθυρο.



Σχήμα 4.18 Αρχικό παράθυρο διαλόγου της εντολής Descriptives

Statistics: Επιλέγουμε για να δηλώσουμε ποια στατιστικά μέτρα επιθυμούμε να υπολογιστούν. Πρέπει να σημειώσουμε εδώ, ότι από τα μέτρα κεντρικής τάσης είναι διαθέσιμα μόνο η μέση τιμή (*Mean*) και το άθροισμα (*Sum*).

- **Explore**

Με την συγκεκριμένη εντολή μπορούμε να επιτύχουμε την πιο πλούσια και πλήρη περιγραφική στατιστική των παρατηρήσεων μιας ποσοτικής μεταβλητής στις διάφορες κατηγορίες κάποιας ποιοτικής.

Από το αρχικό παράθυρο διαλόγου, μετακινούμε την ποσοτική μεταβλητή που θέλουμε να περιγράψουμε στο πλαίσιο *Dependent List* και την ποιοτική μεταβλητή στο πλαίσιο *Factor List.*

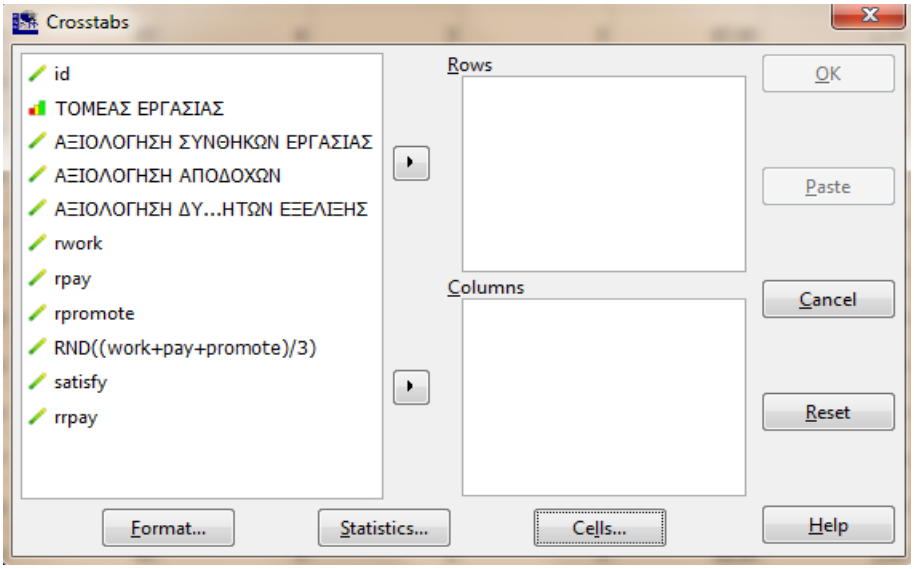
Εξ ορισμού η διαδικασία Explore παράγει ένα πλήθος στατιστικών αποτελεσμάτων όπως τα στατιστικά μέτρα, το φυλλογράφημα (Stem-Leaf) και το θηκογράφημα (Box Plot).

Σημαντικός είναι και ο υπολογισμός διαστημάτων εμπιστοσύνης για το μέσο ενός πληθυσμού (95% confidence interval for mean (lower bound, upper bound)) που παρουσιάζεται μαζί με τα στατιστικά μέτρα.

- **Crosstabs**

Εμφανίζει πίνακες συνάφειας για ζεύγη ποιοτικών μεταβλητών και υπολογίζει διάφορα στατιστικά μέτρα και ελέγχους ανεξαρτησίας τους.

Στο αρχικό παράθυρο διαλόγου (Σχήμα 4.19) πρέπει να δηλώσουμε τις μεταβλητές για τις οποίες θέλουμε να υπολογιστούν οι πίνακες συνάφειας. Αυτό γίνεται με την επιλογή της κάθε μεταβλητής και τη μεταφορά της στη λίστα *Row(s)* ή *Column(s)* με το αντίστοιχο βοηθητικό βέλος. Με τον τρόπο αυτό καθορίζουμε ποιες από τις μεταβλητές θέλουμε να εμφανιστούν στις γραμμές και ποιες στις στήλες. Αρκεί να επιλέξουμε μία τουλάχιστο μεταβλητή για κάθε λίστα, έτσι ώστε να ενεργοποιηθούν οι επιλογές OK και Paste στο παράθυρο. Στην περίπτωση που έχουμε δηλώσει περισσότερες από μία μεταβλητές θα δημιουργηθούν πολλοί πίνακες, ένας για κάθε μία από τις μεταβλητές που έχουν δηλωθεί στη λίστα Rows σε συνδυασμό με κάθε μία από αυτές που έχουν δηλωθεί στη λίστα Columns.



Σχήμα 4.19 Αρχικό παράθυρο διαλόγου της εντολής Crosstabs

Statistics: Επιλέγουμε τον υπολογισμό διαφόρων στατιστικών μέτρων, που αφορούν τη σχέση των δύο μεταβλητών του κάθε πίνακα συνάφειας που κατασκευάζεται. Το στατιστικό μέτρο που μας ενδιαφέρει εδώ είναι το *Chi-square*. Επιλέγοντάς το εκτελείται και ο έλεγχος ανεξαρτησίας x^2.

Cells: Καθορίζουμε τα περιεχόμενα των κελιών του πίνακα συνάφειας. Απαιτείται τουλάχιστον ένα για την κατασκευή του πίνακα. Τα στοιχεία που μπορούν να εμφανιστούν στο κάθε κελί είναι: *Counts* και *Expected* (Παρατηρούμενες και αναμενόμενες συχνότητες), *Row*, *Column* και *Total* (Ποσοστά γραμμής, στήλης ή συνολικά) και *Residuals*, A*dj. Standardized* (Υπόλοιπα - δηλαδή διαφορές παρατηρούμενων και αναμενόμενων τιμών - απλές διαφορές, τυποποιημένες και προσαρμοσμένες τυποποιημένες).

Format: Καθορίζουμε την εμφάνιση του πίνακα συνάφειας. Έχουμε την δυνατότητα να εμφανίσουμε τις γραμμές σε αύξουσα (*Ascending)* ή φθίνουσα σειρά των τιμών της αντίστοιχης μεταβλητής, εκτύπωσης των πινάκων (print tables) και εμφάνισης της τιμής *pivot*.

4.2.6.2 Compare Means

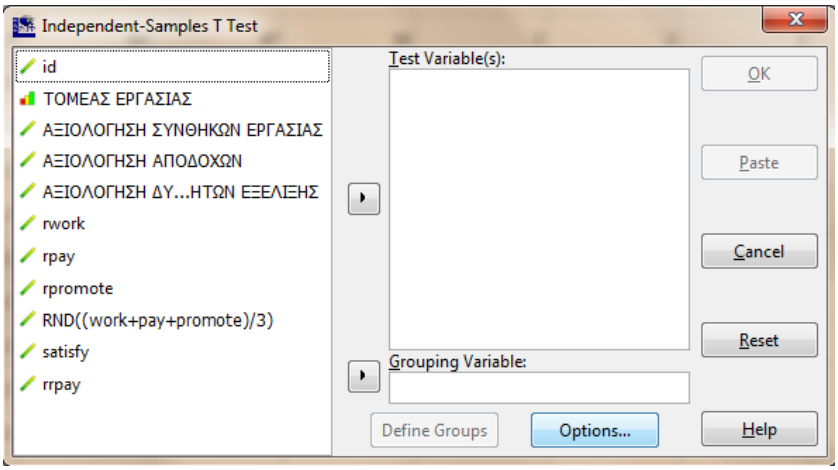
Τον έλεγχο μέσων τιμών (*Compare Means*) τον περιγράψαμε αναλυτικά σε προηγούμενο κεφάλαιο (παράγραφος 2.10), εδώ θα δούμε τις τρεις κατηγορίες ελέγχου και τις διάφορες επιλογές που μας προσφέρει το γραφικό περιβάλλον του PSPP.

- **Independent Samples T-Test (Σύγκριση Μέσων Τιμών Ανεξαρτήτων Πληθυσμών)**

  Αυτή η λειτουργία χρησιμοποιείται για να ελεγχθεί αν δύο ομάδες τιμών έχουν την ίδια μέση τιμή με αυτή του πληθυσμού.

  Στο αρχικό παράθυρο διαλόγου (Σχήμα 4.20) εμφανίζονται σε μία λίστα όλες οι μεταβλητές του αρχείου δεδομένων. Επιλέγουμε μία ή περισσότερες ποσοτικές μεταβλητές και τις μεταφέρουμε στη λίστα *Test Variable(s)* (κάθε μία παράγει έναν έλεγχο). Επιλέγουμε στη συνέχεια μία ποιοτική μεταβλητή για τη θέση *Grouping Variable* και προχωρούμε στον προσδιορισμό των δύο ομάδων (ανεξάρτητων

πληθυσμών), στις οποίες διαχωρίζει το αρχείο δεδομένων, με την ενεργοποίηση του κουμπιού *Define groups*.



Σχήμα 4.20 Αρχικό παράθυρο διαλόγου της εντολής Independent Samples T-Test

Εδώ οι επιλογές μας είναι δύο:

- *Use specified values:* Είναι η εξ ορισμού επιλογή, σύμφωνα με την οποία εισάγουμε μία τιμή για το *Group 1*και μία για το *Group 2*, που αντιστοιχούν στις δύο κατηγορίες της μεταβλητής που διαχωρίζει σε δύο ανεξάρτητους πληθυσμούς τα δεδομένα. Οι περιπτώσεις με άλλες τιμές αποκλείονται από την ανάλυση.
- *Cut point:* Όλες οι περιπτώσεις με τιμές μεγαλύτερες ή ίσες της τιμής που ορίζουμε ως cut point ανήκουν στον ένα πληθυσμό, ενώ οι υπόλοιπες περιπτώσεις ανήκουν στον δεύτερο πληθυσμό.

Options: Έχουμε την δυνατότητα να αλλάξουμε τη στάθμη σημαντικότητας(σφάλμα) σύμφωνα με την οποία θα εφαρμόζεται ο στατιστικός έλεγχος, ή ακόμη να διαχειρισθούμε με ειδικό τρόπο τις χαμένες τιμές.

Confidence Interval: Ένα 95% εμφανίζεται ως το εξ ορισμού επίπεδο εμπιστοσύνης για τον έλεγχο αυτό. Μπορούμε να ζητήσουμε ένα διαφορετικό επίπεδο εμπιστοσύνης, εισάγοντας εδώ μία τιμή μεταξύ 1 και 99. Έτσι, αν για παράδειγμα θέλουμε σφάλμα 1%, πρέπει να εισάγουμε τον αριθμό 99.

Missing values: Είναι δυνατές οι παρακάτω επιλογές:

- *Exclude cases analysis-by analysis*. Είναι η εξ ορισμού ρύθμιση, που αποκλείει τις χαμένες τιμές μίας μεταβλητής (test ή grouping variable) από την ανάλυση της συγκεκριμένης αυτής μεταβλητής.
- *Exclude cases listwise*. Αποκλείει τις χαμένες τιμές μίας μεταβλητής (test ή grouping variable) από όλες τις αναλύσεις.

➢ **Paired Samples T-Test (Σύγκριση Μέσων Τιμών σε Ζευγάρια Παρατηρήσεων)**

Μπορεί κάποιος να χρησιμοποιήσει αυτή την λειτουργία όταν έχουμε αντί για ανεξάρτητα δείγματα, ζευγάρια παρατηρήσεων.

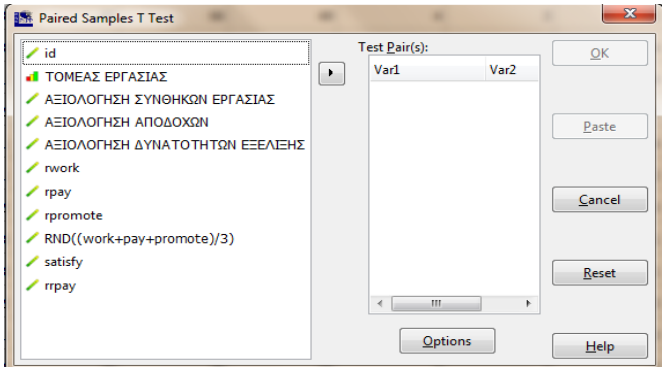
Στο αρχικό παράθυρο (Σχήμα 4.21) εμφανίζονται σε μία λίστα όλες οι μεταβλητές του αρχείου δεδομένων, από τις οποίες επιλέγουμε το ζευγάρι που θα χρησιμοποιήσουμε στην ανάλυση. Η επιλογή αυτή γίνεται ως εξής:

1. Επιλογή της πρώτης μεταβλητής (*Var1*), στο πλαίσιο *Test Pair(s).*
2. Επιλογή της δεύτερης μεταβλητής (*Var2*), που επίσης εμφανίζεται στο πλαίσιο*Test Pair(s).*

Αν θέλουμε περισσότερα ζευγάρια μεταβλητών, συνεχίζουμε με τον ίδιο τρόπο τη διαδικασία της επιλογής των μεταβλητών.

Η ανάλυση εκτελείται με εξ ορισμού παραμέτρους, όπως για παράδειγμα το σφάλμα 5%,αν ενεργοποιήσουμε την ένδειξη OK. Αν όμως θέλουμε να τροποποιήσουμε τις αρχικές αυτές παραμέτρους, τότε ενεργοποιούμε τα *Options.*

Options: Με τις επιλογές μας στο παράθυρο (που έχει ακριβώς την ίδια μορφή με το παράθυρο της διαδικασίας *Independent Samples T-Test* που περιγράψαμε αναλυτικά εκεί), έχουμε τη δυνατότητα να τροποποιήσουμε τη στάθμη σημαντικότητας του ελέγχου.

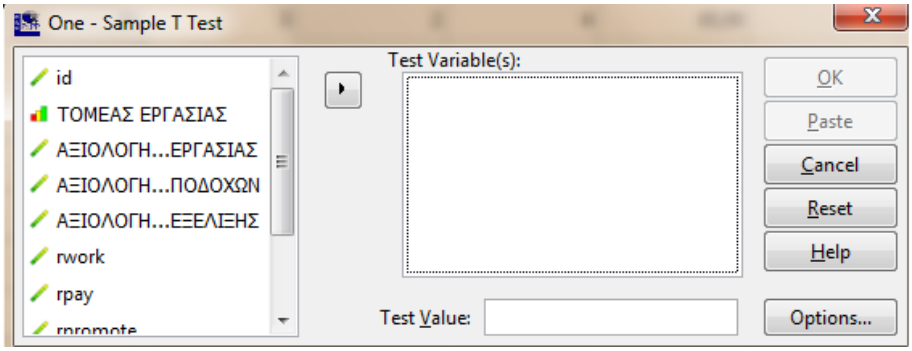


Σχήμα 4.21 Αρχικό παράθυρο διαλόγου της εντολής Paired Samples T-Test

➢ **One Sample T-Test (Σύγκριση Μέσης Τιμής Πληθυσμού με Δεδομένη Τιμή)**

Η λειτουργία αυτή χρησιμοποιείται για τη σύγκριση της μέσης τιμής ενός πληθυσμού ως προς μια υποθετική τιμή.

Στο αρχικό παράθυρο διαλόγου (Σχήμα 4.22) εμφανίζονται σε μία λίστα όλες οι μεταβλητές του αρχείου δεδομένων, από τις οποίες επιλέγουμε τη μεταβλητή, της οποίας τη μέση τιμή θέλουμε να συγκρίνουμε με μία δεδομένη τιμή και την μεταφέρουμε στο πλαίσιο *Test Variable(s).*Την τιμή για την οποία θα γίνει η σύγκριση εισάγουμε στη συνέχεια στη θέση *Test Value.* Η ανάλυση εκτελείται με εξ ορισμού παραμέτρους αν ενεργοποιήσουμε την ένδειξη OK. Αν όμως θέλουμε να τροποποιήσουμε τις αρχικές αυτές παραμέτρους, τότε ενεργοποιούμε τα *Options* και οδηγούμαστε στο παράθυρο διαλόγου που ήδη έχουμε περιγράψει στις άλλες κατηγορίες ελέγχου.



Σχήμα 4.22 Αρχικό παράθυρο διαλόγου της εντολής One Sample T-Test

➢ **One Way ANOVA**

Η διαδικασία εκτελεί μια μονόδρομη ανάλυση διακύμανσης των μεταβλητών χρησιμοποιώντας σαν κριτήριο μια απλή ανεξάρτητη μεταβλητή. Χρησιμοποιείται για να συγκρίνει τις μέσες τιμές ενός πληθυσμού που χωρίζεται σε περισσότερες από δύο ομάδες.

Στο παράθυρο διαλόγου της εντολής, οι εξαρτημένες μεταβλητές που θα αναλυθούν, πρέπει να δοθούν στο πλαίσιο *Dependent Variable(s)*. Στο πλαίσιο *Factor* τοποθετείται η ανεξάρτητη μεταβλητή με την οποία θα γίνει η σύγκριση.

Στο πλαίσιο *Statistics* υπάρχουν δύο επιλογές:

- Η *Descriptives* που εμφανίζει τα περιγραφικά στατιστικά για τις ομάδες που υπολογίζονται με βάση την ανεξάρτητη μεταβλητή.

- Η *Homogeneity* που εκτελεί το Levene test για την ομοιογένεια της διασποράς τόσο για τις μεταβλητές όσο και τις ομάδες τους.

Η επιλογή *Contrast*μπορεί να χρησιμοποιηθεί όταν διαβλέπετε ορισμένες διαφορές μεταξύ των ομάδων. Πρέπει να ακολουθείται από μια λίστα αριθμών που θα χρησιμοποιηθούν ως συντελεστές για την ομάδα που πρόκειται να ελεγχθεί. Ο αριθμός των συντελεστών αυτών πρέπει βέβαια να είναι σε αντιστοιχία με τον αριθμό των διακριτών ομάδων (ή τιμές της ανεξάρτητης μεταβλητής). Εάν το συνολικό άθροισμα των συντελεστών δεν είναι μηδέν, τότε το PSPP θα εμφανίσει μια προειδοποίηση, αλλά παρόλα αυτά θα προχωρήσει με την ανάλυση. Η Contrast μπορεί να δοθεί έως και 10 φορές για να καθορίσετε διαφορετικά τεστ σύγκρισης.

4.2.6.3 Bivariate Correlation

Η διαδικασία παράγει πίνακες του συντελεστή συσχέτισης του Pearson για ένα σύνολο μεταβλητών. Ο συντελεστής γραμμικής συσχέτισης του Pearson (r) παίρνει τις τιμές: $-1 \leq r \leq +1$.

Όσο το r βρίσκεται πιο κοντά στο +1(-1), τόσο πιο ισχυρή θετική (αρνητική) συσχέτιση υπάρχει. Όσο το r βρίσκεται πιο κοντά στο 0, τόσο πιο ασθενής συσχέτιση υπάρχει. Συνήθως θεωρούμε ότι η συσχέτιση είναι:

- Ισχυρή έως πολύ ισχυρή, όταν $|r| > 0,7$
- Μέτρια έως ικανοποιητική, όταν $0,5 < |r| < 0,7$
- Ασθενής έως μέτρια, όταν $|r| < 0,5$

Στο αρχικό παράθυρο διαλόγου, επιλέγουμε τις μεταβλητές για τις οποίες θέλουμε να υπολογίσουμε τον συντελεστή συσχέτισης του Pearson. Στο πλαίσιο *Test of Significance*, επιλέγουμε πώς οι αναφερόμενες τιμές σημαντικότητας θα εκτυπωθούν. Έχουμε δυο επιλογές, την *One-tailed* και *Two-tailed*. Σαν προεπιλογή χρησιμοποιούμε την Two-tailed. Έπειτα μπορούμε μέσω της επιλογής *Flag significant correlations* να επιλέξουμε αν θέλουμε να εμφανίζονται οι συντελεστές συσχέτισης με έναν ή δυο αστερίσκους (*) για τις σημαντικές τιμές 0,05 και 0,01 αντίστοιχα.

4.2.6.4 K – Means Cluster Analysis

Η διαδικασία αυτή επιχειρεί να εντοπίσει σχετικά ομοιογενείς ομάδες περιπτώσεων με βάση επιλεγμένα χαρακτηριστικά, χρησιμοποιώντας έναν αλγόριθμο που μπορεί να χειριστεί μεγάλο αριθμό υποθέσεων. Ωστόσο, ο αλγόριθμος απαιτεί να καθορίσουμε τον αριθμό των *clusters*. Τα clusters είναι μια κατηγορία τεχνικών που χρησιμοποιούνται για να ταξινομούν τις περιπτώσεις σε ομάδες που είναι σχετικά ομοιογενή ή ετερογενή μεταξύ τους, με βάση ένα καθορισμένο σύνολο μεταβλητών.

Στο παράθυρο διαλόγου της διαδικασίας, αρχικά επιλέγουμε τις μεταβλητές που θα συμπεριλάβουμε στην ανάλυση και έπειτα καθορίζουμε τον αριθμό των clusters. Ο αριθμός των clusters πρέπει να είναι τουλάχιστον 2 και όχι μεγαλύτερος από το πλήθος των περιπτώσεων του αρχείου δεδομένων.

4.2.6.5 Factor Analysis

Η παραγοντική ανάλυση *(Factor Analysis)* είναι μια στατιστική μέθοδος που έχει σκοπό να βρει την ύπαρξη παραγόντων κοινών ανάμεσα σε μια ομάδα μεταβλητών. Με αυτή την μεθοδολογία καταφέρνουμε:

- Να μειώσουμε τις διαστάσεις του προβλήματος
- Να δημιουργήσουμε καινούργιες μεταβλητές, τους παράγοντες, τις οποίες μπορούμε να τις θεωρήσουμε ως κάποιες μη μετρήσιμες μεταβλητές, όπως ελκυστικότητα ενός προϊόντος στο Marketing κ.α.
- Να εξηγήσουμε τις συσχετίσεις που υπάρχουν στα δεδομένα, για τις οποίες έχουμε υποθέσει ότι οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων που δημιούργησαν τα δεδομένα.

Το αξιοσημείωτο σε αυτού του είδους την ανάλυση είναι, ότι προσπαθεί να εξηγήσει περισσότερο τη δομή παρά την μεταβλητότητα (ποσοστό διακύμανσης).

Στο αρχικό παράθυρο διαλόγου, τοποθετούμε τις μεταβλητές που πρόκειται να πάρουν μέρος στην ανάλυση στο πλαίσιο *Variables.* Μέσω της επιλογής *Extraction* καθορίζουμε τον τρόπο με τον οποίο οι παράγοντες, δηλαδή οι συνιστώσες, εξάγονται από τα δεδομένα. Η προεπιλεγμένη μέθοδος είναι η ανάλυση κύριων συνιστωσών (*Principal Components Analysis*) και συνήθως αυτή χρησιμοποιείται. Η επιλογή *Rotation* χρησιμοποιείται για να

καθορίσει τη μέθοδο με την οποία θα εξάγεται η λύση. Υπάρχουν τρεις μέθοδοι: η *Varimax*(που είναι η προεπιλογή), η *Equimax* και η *Quartimax*. Αν δεν επιθυμείτε αυτή η διαδικασία να εκτελεστεί, τότε η *None* θα αποτρέψει την εντολή από την εκτέλεση οποιασδήποτε περιστροφής των δεδομένων.

4.2.6.6 Reliability

Η διεργασία εκτελεί ανάλυση αξιοπιστίας στα δεδομένα. Για την μέτρηση της αξιοπιστίας χρησιμοποιούμε κυρίως το συντελεστή *Alpha του Cronbach* αντί του συντελεστή διχοτομικής αξιοπιστίας (*Split-half coefficient*).

- Το μοντέλο *split* χωρίζει την κλίμακα σε δύο μέρη και εξετάζει τη συσχέτιση μεταξύ των μερών.
- Ο συντελεστής *Cronbach's a* ή δείκτης εσωτερικής συνάφειας υπολογίζεται με βάση τις συσχετίσεις μεταξύ των items(δεδομένων) της κλίμακας:

$$\text{Cronbach's a} = \frac{a}{a-1}\left(1 - \frac{a}{a+2b}\right)$$

a = αριθμός items

b = άθροισμα των συσχετίσεων μεταξύ των items

Θεωρητικά μπορεί να κυμαίνεται από το – άπειρο έως το 1 (μόνο οι θετικές τιμές έχουν νόημα).

Ενδεικτικές τιμές αξιοπιστίας:

< 0.6	η κλίμακα είναι αναξιόπιστη
0.6	το ελάχιστο αποδεκτό όριο (μή αποδεκτό για κλίμακες με πολλά items)
0.7	επαρκές, αλλά όχι καλό
0.8	καλύτερο
0.95	πολύ υψηλή αξιοπιστία (μάλλον σπάνιο)

Στο αρχικό παράθυρο διαλόγου, επιλέγουμε το σύνολο των μεταβλητών που θα συμπεριλάβουμε στην ανάλυση και τις τοποθετούμε στο πλαίσιο *Items*. Στην επιλογή *Model* καθορίζουμε το είδος της ανάλυσης που θα ακολουθήσουμε. Εάν η Alpha έχει καθοριστεί, τότε η Alpha του Cronbach υπολογίζεται για τις τιμές. Εάν έχει καθοριστεί η Split, τότε οι μεταβλητές χωρίζονται σε 2 υποσύνολα. Μέσω της επόμενης επιλογής μπορούμε να καθορίσουμε πόσες μεταβλητές θα βρίσκονται στην πρώτη υποομάδα.

4.2.6.7 Linear Regression

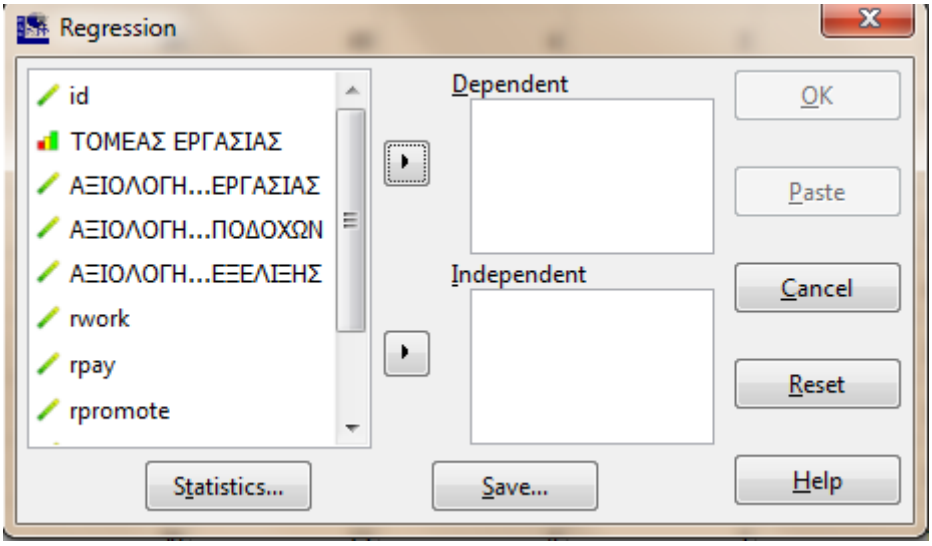
Η Γραμμική Παλινδρόμηση *(Linear Regression)* αποτελεί μία στατιστική μέθοδο η οποία αποσκοπεί στον προσδιορισμό ενός μαθηματικού μοντέλου για την περιγραφή / ερμηνεία / πρόβλεψη των τιμών ενός χαρακτηριστικού (μεταβλητής) σε σχέση με τις τιμές ενός πλήθους άλλων χαρακτηριστικών (μεταβλητών).

Για την πραγματοποίηση της απλής γραμμικής παλινδρόμησης, στο αρχικό παράθυρο διαλόγου της διαδικασίας (Σχήμα 4.23), στο πλαίσιο *Dependent* μετακινούμε την εξαρτημένη μεταβλητή (μεταβλητή κριτήριο) και στο πλαίσιο *Independent,* την ανεξάρτητη μεταβλητή (ή προβλεπτική μεταβλητή). Στην περίπτωση που θέλουμε να πραγματοποιήσουμε πολλαπλή παλινδρόμηση, στο πλαίσιο αυτό τοποθετούμε το σύνολο των ανεξάρτητων μεταβλητών που επιθυμούμενα συμπεριλάβουμε στην ανάλυση.

Η επιλογή *Statistics* καθορίζει τα στατιστικά στοιχεία που θα πρέπει να αναγράφονται:

- Το R υπολογίζει τον λόγο των αθροισμάτων των τετραγώνων που προκύπτουν από το μοντέλο προς το σύνολο των αθροισμάτων των τετραγώνων της εξαρτημένης μεταβλητής
- Η Coeff υπολογίζει έναν πίνακα που περιέχει τους εκτιμημένους από το μοντέλο συντελεστές και τα τυπικά λάθη (standard errors)
- Anova δίνει την ανάλυση του πίνακα διακύμανσης για το μοντέλο
- Bcov υπολογίζει τον πίνακα συνδιακύμανσης για τους εκτιμώμενους συντελεστές στο μοντέλο.

Τέλος, η υποεντολή *Save* σώζει τα κατάλοιπα(residuals) ή τις προβλεπόμενες τιμές από το προσαρμοσμένο μοντέλο του ενεργού συνόλου δεδομένων.



Σχήμα 4.23 Αρχικό παράθυρο διαλόγου της εντολής Linear Regression

4.2.6.8 Non-Parametric Statistics

Τα μη-παραμετρικά tests κάνουν ελάχιστες υποθέσεις σχετικά με την υποκείμενη κατανομή των δεδομένων. Ακολουθεί μια περιληπτική ανάλυση των tests που είναι διαθέσιμα μέσω του πακέτου PSPP.

- **Chi-Square**

  Το test Chi-Square ή πιο απλά στα ελληνικά x^2, παράγει ένα στατιστικό για τις διαφορές μεταξύ των αναμενόμενων και των παρατηρούμενων συχνοτήτων των κατηγοριών μιας μεταβλητής. Προαιρετικά, ένα εύρος τιμών *(Expected Range)* μπορεί να προσδιοριστεί προερχόμενο από μια λίστα μεταβλητών. Εάν δοθεί αυτό το εύρος, τότε οι μη ακέραιες τιμές χάνουν το δεκαδικό τους μέρος, και οι τιμές εκτός του καθορισμένου εύρους εξαιρούνται από την ανάλυση. Στο πλαίσιο *Expected Values* καθορίζονται οι αναμενόμενες τιμές για κάθε κατηγορία. Θα πρέπει να υπάρχει ακριβώς μία μη μηδενική αναμενόμενη τιμή για κάθε παρατηρούμενη κατηγορία, ή αλλιώς η επιλογή *All categories equal* θα πρέπει να έχει καθοριστεί.

- **Binomial**

  Το διωνυμικό test συγκρίνει την παρατηρούμενη κατανομή μιας δίτιμης μεταβλητής με εκείνη της διωνυμικής κατανομής. Η μεταβλητή p (*Test Proportion*) καθορίζει το ποσοστό του τεστ της διωνυμικής κατανομής. Η προεπιλεγμένη τιμή είναι 0,5. Εάν μια μόνο τιμή εμφανίζεται, που προέκυψε από μια λίστα μεταβλητών (επιλογή Get from data), τότε η τιμή αυτή χρησιμοποιείται ως όριο για να διαχωρίσει

σε ομάδες τις παρατηρούμενες τιμές. Οι τιμές που είναι ίσες ή μικρότερες από την τιμή κατωφλίου *(Cut point)* διαμορφώνουν την πρώτη κατηγορία. Οι τιμές που είναι μεγαλύτερες από το όριο-κατώφλι θα αποτελέσουν τη δεύτερη κατηγορία. Αν δύο τιμές εμφανίζονται , τότε θα χρησιμοποιούνται ως τιμές τις οποίες οι μεταβλητές θα πρέπει να λάβουν για να είναι στην αντίστοιχη κατηγορία. Σε όλες τις υπόλοιπες περιπτώσεις όπου για τις οποίες μια μεταβλητή δεν έχει τιμή ίση με καμία από τις καθορισμένες τιμές, δεν λαμβάνουν μέρος στο τεστ για τη συγκεκριμένη μεταβλητή. Εάν δεν εμφανίζονται καθόλου τιμές, τότε η μεταβλητή πρέπει να παίρνει δίτιμες τιμές. Εάν υπάρχουν περισσότερες από δύο διακριτές, non-missing τιμές για τη μεταβλητή που εξετάζεται στο τεστ τότε αυτή η κατάσταση είναι προβληματική και προκαλεί το τύπωμα μηνύματος λάθους.

- **Runs**

Το test εξετάζει αν μια ακολουθία δεδομένων έχει τυχαία διάταξη. Λειτουργεί εξετάζοντας πόσες φορές η τιμή μιας μεταβλητής υπερβαίνει ένα συγκεκριμένο όριο. Μπορεί να καθοριστεί στο πλαίσιο *Cut Point* του παράθυρου διαλόγου, είτε ως ένας αριθμός (*custom*) είτε ως μέση τιμή (*mean*), διάμεσος (*median*) ή μέσω μιας τιμής της *mode*. Όταν εφαρμοστεί και ο περιορισμός της τιμής που έχουμε θέσει σαν όριο, τότε προκύπτει και ο κατάλογος των μεταβλητών, οι τιμές των οποίων θα ελεγχθούν. Σαν αποτέλεσμα προβάλλεται ο αριθμός των υπερβάσεων της δοθείσας τιμής, η ασυμπτωτική σημαντικότητα που υπολογίζεται με βάση το μήκος των στοιχείων.

- **1-Sample K-S**

Το τεστ Kolmogorov-Smirnov για ένα μόνο δείγμα*(sample),* χρησιμοποιείται για να ελέγξουμε αν ισχύει ή όχι ότι ένα σύνολο δεδομένων προέρχονται από μια συγκεκριμένη κατανομή. Τέσσερις κατανομές μόνο εξετάζονται και αυτές είναι: η κανονική *(Normal),* η ομοιόμορφη *(Uniform),* η *Poisson* και η εκθετική *(Exponential).*

Ιδανικά θα έπρεπε να δίνει ο χρήστης τις παραμέτρους της κατανομής βάσει των οποίων θέλει να τεστάρει τα δεδομένα. Για παράδειγμα, για την κανονική κατανομή θέλουμε τη μέση τιμή και την τυπική απόκλιση , για την ομοιόμορφη κατανομή θα πρέπει να δοθεί το ελάχιστο και μέγιστο. Ωστόσο, εάν παραλείπονται οι παράμετροι αυτοί θα υπολογιστούν από τα δεδομένα. Ο υπολογισμός των παραμέτρων μειώνει την ισχύ του τεστ και θα πρέπει να αποφεύγεται αν αυτό είναι δυνατόν.

- **2 Related Samples**

Υπάρχουν 3 διαφορετικά είδητests από τα οποία μπορούμε να επιλέξουμε, ώστε να πραγματοποιηθεί έλεγχος μεταξύ ενός ζεύγους μεταβλητών:

- *Wilcoxon:* ελέγχει τις διαφορές μεταξύ των μέσων όρων που αναφέρονται οι μεταβλητές. Δεν κάνει υποθέσεις σχετικά με τις διακυμάνσεις των δειγμάτων, ωστόσο, από την αρχή υποθέτουμε ότι η κατανομή είναι συμμετρική.
- *Sign:* εξετάζει τις διαφορές μεταξύ των μέσων όρων των μεταβλητών που δίνονται. Για το τεστ δεν γίνονται υποθέσεις σχετικά με την κατανομή των δεδομένων.
- *McNemar:* χρησιμοποιείται για να αναλύσουμε τη σημαντικότητα της διαφοράς μεταξύ ζευγών των υπό-δειγμάτων που οι διαφορετικοί συνδυασμοί του αποτελούν το αρχικό δείγμα προς εξέταση. Τα δεδομένα σε κάθε μεταβλητή θα πρέπει να είναι δίτιμα, με την έννοια ότι πρέπει να χωρίζονται σε δύο κατηγορίες μη επικαλυπτόμενες, ώστε το άθροισμα των υπό-δειγμάτων να δίνει πάντα το αρχικό δείγμα. Εάν υπάρχουν περισσότερες από δύο διακριτές μεταβλητές θα παρουσιαστεί σφάλμα και το τεστ δεν θα τρέξει.

- **K-Related Samples**

Και εδώ, υπάρχουν 3 διαφορετικά είδη tests από τα οποία μπορούμε να επιλέξουμε, ώστε να πραγματοποιηθεί έλεγχος μεταξύ μεταβλητών:

- *Friedman:* χρησιμοποιείται για τη δοκιμή των διαφορών μεταξύ των επαναλαμβανόμενων μέτρων όταν δεν υπάρχουν ένδειξεις ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Η διαδικασία εκτυπώνει το άθροισμα των βαθμών για κάθε μεταβλητή, το αποτέλεσμα και τη σημαμαντικότητά του.
- *Kendall's W:* ερευνά αν ένας αυθαίρετος αριθμός σχετικών δειγμάτων προέρχονται από τον ίδιο πληθυσμό. Είναι ακριβώς το ίδιο με τη δοκιμή *Friedman* εκτός από το ότι υπολογίζει το πρόσθετο στατιστικό W. Κατά την εκτέλεση του test, ο *Συντελεστής Συμφωνίας του Kendall*τυπώνεται *(Kendall's Coefficient of Concordance).* Έχει εύρος τιμών [0,1], όπου η τιμή μηδέν

δείχνει ότι δεν υπάρχει συμφωνία μεταξύ των δειγμάτων, ενώ η μονάδα δείχνει πλήρη συμφωνία.

- *Cochran's Q:* χρησιμοποιείται για τη δοκιμή για τις διαφορές μεταξύ των τριών ή περισσοτέρων ομάδων. Η τιμή του Q θα τυπώνεται καθώς και η σημαντικότητα που θα υπολογίζεται ασυμπτωτικά βασιζόμενη στην *chi-square* κατανομή.

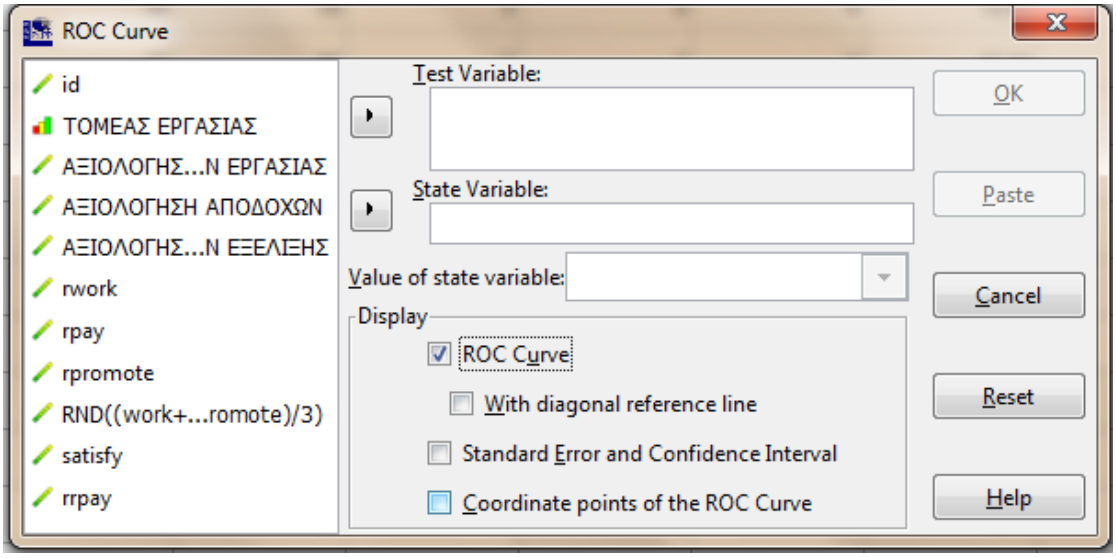
4.2.6.9 ROC Curve

Η εντολή χρησιμοποιείται για να δείξει διαγραμματικά τα χαρακτηριστικά και τη λειτουργία της καμπύλης ενός συνόλου δεδομένων, και να γίνουν εκτιμήσεις για την περιοχή κάτω από την καμπύλη. Αυτό είναι χρήσιμο για την ανάλυση της αποτελεσματικότητας μιας μεταβλητής ως προγνωστικός παράγοντας για την ουσιαστική κατάστασή της.

Στο παράθυρο διαλόγου της εντολής (Σχήμα 4.24), ορίζεται αρχικά η λίστα των μεταβλητών πρόβλεψης (πλαίσιο *Test Variable*). Η μεταβλητή *state*είναι η μεταβλητή της οποίας οι τιμές αντιπροσωπεύουν την πραγματική κατάσταση και η αξία *(value of state variable)* αυτής αντιπροσωπεύει την θετική κατάσταση.

Στο πλαίσιο *Display* περιέχονται οι εξής επιλογές σχεδίασης:

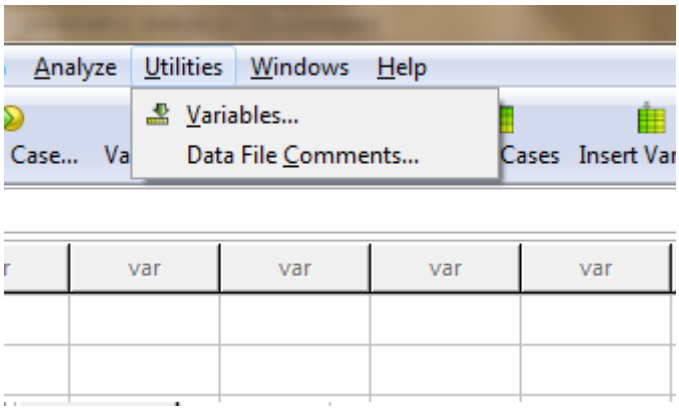
- *ROC Curve:* επιλέγουμε αν θέλουμε η καμπύλη ROC να σχεδιαστεί ή όχι. Αν επιθυμούμε, μπορούμε να επιλέξουμε αν θέλουμε να σχεδιαστεί η διαγώνια γραμμή αναφοράς*(Diagonal reference line),* η οποία βοηθάει στην καλύτερη οπτικοποίηση των αποτελεσμάτων.
- *Standard Error and Confidence Interval:* επιλέγουμε αν θα εκτυπωθεί ή όχι το τυπικό σφάλμα της περιοχής κάτω από την καμπύλη, καθώς και η ίδια η περιοχή.
- *Coordinate points of the ROC Curve:* επιλέγουμε αν επιθυμούμε την εκτύπωση ενός πίνακα συντεταγμένων της καμπύλης ROC.



Σχήμα 4.24 Παράθυρο διαλόγου της εντολής ROC

4.2.7 Η Καρτέλα Utilities

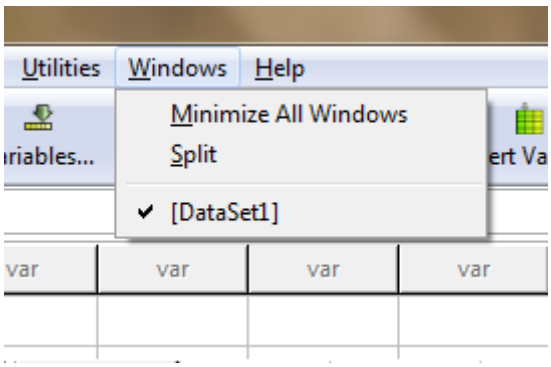
Περιέχει πληροφορίες για κάθε μεταβλητή, και προσάρτηση σχολίων σε κάθε μεταβλητή (Σχήμα 4.25).



Σχήμα 4.25 Η Καρτέλα Utilities

4.2.8 Η Καρτέλα Windows

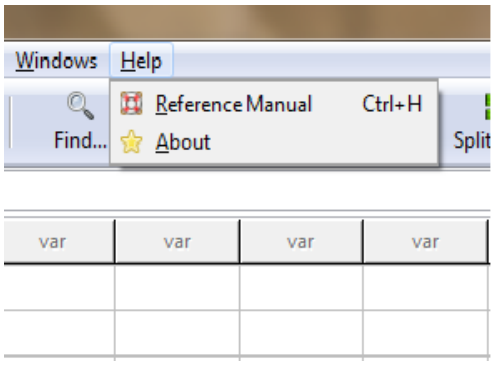
Περιέχει λειτουργίες όπως σμίκρυνση παραθύρου, διαχωρισμό των δεδομένων αρχείου σε υπό-παράθυρα, καθώς και εναλλαγή μεταξύ των ενεργών παραθύρων του αρχείου(Σχήμα 4.26).



Σχήμα 4.26 Η Καρτέλα Windows

4.2.9 Η Καρτέλα Help

Το μενού Help δεν είναι πλήρως λειτουργικό. Θα πρέπει να ανοίξετε το εγχειρίδιο αναφορών (manual), το οποίο συνοδεύεται με την εγκατάσταση του προγράμματος και περιέχει πληροφορίες σχετικά με το πώς να χρησιμοποιούνται οι πολλές δυνατότητες της PSPP. Επίσης μέσω της επιλογής about, εμφανίζεται η τρέχων έκδοση του πακέτου PSPP καθώς και σύνδεση με τον ιστότοπο της εταιρείας παραγωγής (Σχήμα 4.27).



Σχήμα 4.27 Η Καρτέλα Help

5. Επίλυση Μελέτης Περίπτωσης (Case Study)

Ας δούμε τώρα πώς θα εφαρμόσουμε την κωδικοποίηση των δεδομένων και την εισαγωγή τους σε ένα αρχείο, πώς θα προετοιμάσουμε δηλαδή το στατιστικό υλικό για την ανάλυσή του με το στατιστικό πακέτο PSPP, έχοντας δεδομένα από μία έρευνα που διεξήγαγε γνωστή διαφημιστική εταιρία.

Στόχος της έρευνας αυτής ήταν να προσδιοριστεί σε ποιο βαθμό τα στελέχη της εταιρίας είναι ικανοποιημένα από την εργασία τους. Η έρευνα βασίστηκε στη συμπλήρωση ερωτηματολογίων από τα στελέχη της εταιρίας που εργάζονται σε τρία διαφορετικά τμήματα του τμήματος marketing: στην έρευνα αγοράς, στις δημόσιες σχέσεις και στη διαφήμιση. Οι υπάλληλοι απάντησαν σε μία σειρά ερωτήσεων και στη συνέχεια αξιολόγησαν τις συνθήκες εργασίας τους, τις αποδοχές τους και τις δυνατότητες που έχουν για εξέλιξη στην εταιρία, στην κλίμακα 1-100 (με Άριστα=100), όπως φαίνεται από τα πρωτογενή δεδομένα της έρευνας που δίνουμε στον Πίνακα 5.1. Θα παραθέσουμε τώρα τα χαρακτηριστικά - μεταβλητές, που εμείς επιλέξαμε για να δείξουμε τον τρόπο προετοιμασίας τους για την εισαγωγή τους στον Η/Υ.

Την ίδια εφαρμογή θα χρησιμοποιήσουμε και στη συνέχεια για να δείξουμε τη χρήση των βασικών εντολών ανάλυσης που διαθέτει το PSPP και για να εξηγήσουμε τα αποτελέσματα από την εκτέλεση των εντολών αυτών.

Οι Μεταβλητές της έρευνας:

(1) Αύξων Αριθμός (Α/Α): Ο αύξων αριθμός κάθε υπαλλήλου. Δεν αποτελεί μεταβλητή, απλά χρησιμοποιείται προαιρετικά για λόγους ελέγχου των δεδομένων μας.

(2) Τομέας Εργασίας υπαλλήλου: Εδώ έχουμε μία ποιοτική μεταβλητή με τρεις τιμές: "Έρευνα Αγοράς", "Δημόσιες Σχέσεις" και "Διαφήμιση". Στη μεταβλητή θα δώσουμε τον κωδικό όνομα section και οι τιμές θα παρασταθούν αντίστοιχα με τους αριθμούς 1, 2 και 3 (ονομαστική κλίμακα).

(3) Αξιολόγηση Συνθηκών Εργασίας: Είναι ο βαθμός αξιολόγησης των συνθηκών εργασίας και είναι μία ποσοτική συνεχής μεταβλητή. Σύμφωνα μάλιστα με την κλίμακα που δώσαμε, μπορεί να πάρει τιμές σε ένα εύρος 1-100. Θα χρησιμοποιήσουμε τον κωδικό work (αναλογική κλίμακα).

(4) Αξιολόγηση Αποδοχών: Είναι ο βαθμός αξιολόγησης των αποδοχών του υπαλλήλου και είναι όπως και η προηγούμενη μία ποσοτική συνεχής μεταβλητή, που

παίρνει τιμές στο διάστημα 1-100. Θα δώσουμε στη μεταβλητή τον κωδικό pay (αναλογική κλίμακα).

(5) Αξιολόγηση Δυνατοτήτων Εξέλιξης: Είναι ο βαθμός αξιολόγησης των δυνατοτήτων που ο κάθε υπάλληλος εκτιμά πως έχει για περαιτέρω εξέλιξη μέσα στην εταιρία και είναι μία ποσοτική συνεχής μεταβλητή, που παίρνει τιμές στο διάστημα 1-100. Θα δώσουμε στη μεταβλητή τον κωδικό promote (αναλογική κλίμακα).

Παρατηρούμε ότι σε δύο περιπτώσεις έχουμε χαμένες τιμές. Στις περιπτώσεις με Α/Α 29 και 33 λείπει η αξιολόγηση των δυνατοτήτων εξέλιξης και η αξιολόγηση των αποδοχών αντίστοιχα. Στο PSPP μπορούμε να πληκτρολογήσουμε στη θέση των χαμένων αυτών τιμών την παύλα (-).

Πίνακας 5.1: Πρωτογενή δεδομένα της έρευνας

Α/Α	Τομέας εργασίας	Αξιολόγηση συνθηκών εργασίας	Αξιολόγηση αποδοχών	Αξιολόγηση δυνατοτήτων εξέλιξης
1	1	71	49	58
2	3	84	53	63
3	1	84	74	37
4	3	87	66	49
5	3	72	59	79
6	1	72	37	86
7	2	72	57	40
8	1	63	48	78
9	1	72	76	37
10	2	71	25	74
11	3	69	47	16
12	3	90	56	23
13	2	84	28	62
14	1	86	37	59
15	2	70	38	54
16	1	86	72	72
17	1	84	60	29
18	2	90	62	66
19	1	73	56	55
20	1	94	60	52
21	2	84	42	66
22	2	85	56	64
23	3	88	55	52
24	3	74	70	51
25	2	71	45	68
26	2	88	49	42

27	1	90	27	67
28	1	85	89	46
29	3	79	59	
30	3	72	60	45
31	2	88	36	47
32	2	77	60	75
33	2	64		61
34	2	87	51	57
35	2	77	90	51
36	3	71	36	55
37	3	75	53	92
38	2	74	59	82
39	3	76	51	54
40	1	95	66	52
41	2	89	66	62
42	2	85	57	67
43	1	65	42	68
44	1	82	37	54
45	1	82	60	56
46	3	89	80	64
47	2	74	47	63
48	2	82	49	91
49	1	90	76	70
50	1	78	52	72

Μετά από την παραπάνω προετοιμασία, είμαστε έτοιμοι να εισάγουμε τα δεδομένα στον Η/Υ με τη μορφή του Πίνακα 5.2.

Πίνακας 5.2: Κωδικοποιημένα δεδομένα της έρευνας

A/A	section	work	pay	promote
1	1	71	49	58
2	3	84	53	63
3	1	84	74	37
4	3	87	66	49
5	3	72	59	79
6	1	72	37	86
7	2	72	57	40
8	1	63	48	78
9	1	72	76	37
10	2	71	25	74

11	3	69	47	16
12	3	90	56	23
13	2	84	28	62
14	1	86	37	59
15	2	70	38	54
16	1	86	72	72
17	1	84	60	29
18	2	90	62	66
19	1	73	56	55
20	1	94	60	52
21	2	84	42	66
22	2	85	56	64
23	3	88	55	52
24	3	74	70	51
25	2	71	45	68
26	2	88	49	42
27	1	90	27	67
28	1	85	89	46
29	3	79	59	-
30	3	72	60	45
31	2	88	36	47
32	2	77	60	75
33	2	64	-	61
34	2	87	51	57
35	2	77	90	51
36	3	71	36	55
37	3	75	53	92
38	2	74	59	82
39	3	76	51	54
40	1	95	66	52
41	2	89	66	62
42	2	85	57	67
43	1	65	42	68
44	1	82	37	54
45	1	82	60	56
46	3	89	80	64
47	2	74	47	63
48	2	82	49	91
49	1	90	76	70
50	1	78	52	72

➢ Θα εφαρμόσουμε τώρα τις μεθόδους της περιγραφικής στατιστικής. Αρχικά, παρατηρούμε ότι στο δείγμα μας υπάρχει μία ποιοτική μεταβλητή (η section) και τρεις ποσοτικές συνεχείς (work, pay, promote). Η περιγραφική στατιστική θα γίνει με τις διαδικασίες Frequencies για τη μεταβλητή section και τη διαδικασία Descriptives για τις μεταβλητές work, pay και promote.

Αφού ανοίξουμε το παράθυρο διαλόγου της διαδικασίας Frequencies (Analyze->Discriptives Statistics->Frequencies), δηλώνουμε αρχικά την μεταβλητή *section* στη λίστα των μεταβλητών και φροντίζουμε μέσω της επιλογής *Frequency Tables* να είναι τσεκαρισμένη η επιλογή *Always* στο πλαίσιο *Display Frequency Table* έτσι ώστε να εμφανιστεί στα αποτελέσματα πίνακας κατανομής συχνοτήτων. Στη συνέχεια, μέσω της λίστας *Statistics* δηλώνουμε ότι θέλουμε να υπολογιστεί μόνο η επικρατούσα τιμή (*Mode*) αφού οποιοδήποτε άλλο στατιστικό μέτρο δεν έχει απολύτως καμία σημασία για την ποιοτική μεταβλητή *section* και μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Όμοια στο παράθυρο της επιλογής *Charts* δηλώνουμε ότι επιθυμούμε τη δημιουργία ιστογράμματος το οποίο θα έχει στον κατακόρυφο άξονα ποσοστά (*Percentages*).

Τα αποτελέσματα που θα πάρουμε τα βλέπουμε στο Σχήμα 5.1 και περιλαμβάνουν τον πίνακα κατανομής συχνοτήτων της section μαζί με την επικρατούσα τιμή της μεταβλητής που είναι η τιμή 2 (Δημόσιες σχέσεις). Ακόμη, θα εμφανιστεί και το ιστόγραμμα που ζητήσαμε.

```
SAVE

SAVE OUTFILE="C:\Users\User\Desktop\Ptuxiaki\askiseis\askisi_ypodeigma.sav".


FREQUENCIES

FREQUENCIES
    /VARIABLES=section
    /FORMAT=AVALUE TABLE
    /STATISTICS=MODE
    /HISTOGRAM=NONORMAL PERCENT.
```

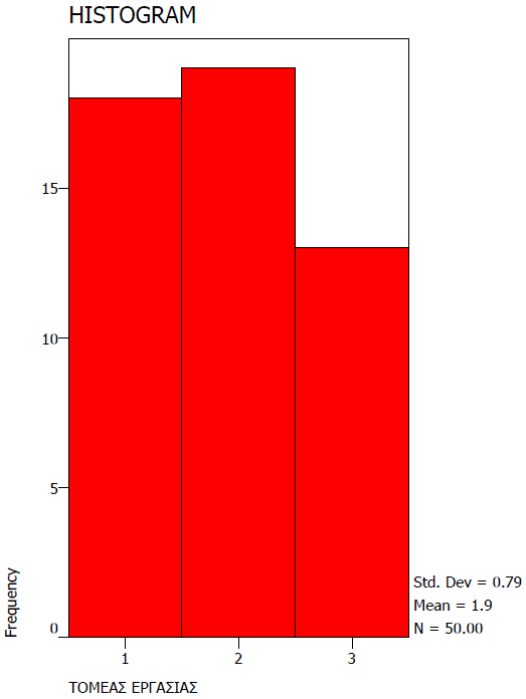
ΤΟΜΕΑΣ ΕΡΓΑΣΙΑΣ

Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum Percent*
ΕΡΕΥΝΑ ΑΓΟΡΑΣ	1	18	36,00	36,00	36,00
ΔΗΜΟΣΙΕΣ ΣΧΕΣΕΙΣ	2	19	38,00	38,00	74,00
ΔΙΑΦΗΜΙΣΗ	3	13	26,00	26,00	100,00
	Total	50	100,0	100,0	

ΤΟΜΕΑΣ ΕΡΓΑΣΙΑΣ

N	*Valid*	50
	Missing	0
Mode		2,00
S.E. Skew		,34

Σχήμα 5.1 Αποτελέσματα της διαδικασίας Frequencies για τη μεταβλητή section



Σχήμα 5.2 Ιστόγραμμα της μεταβλητής sectionαπό τη διαδικασία Frequencies

Στη συνέχεια, ανοίγουμε το παράθυρο της διαδικασίας Descriptives (Analyze->Discriptives Statistics->Discriptives), όπου δηλώνουμε τις ποσοτικές μεταβλητές work, pay και promote στη λίστα Variables. Στη λίστα επιλογών Statistics δηλώνουμε τα στατιστικά μέτρα που θέλουμε (Mean, Standard deviation, Minimum, Maximum, Range). Τα αποτελέσματα που παίρνουμε τα βλέπουμε στο Σχήμα 5.2.

```
SAVE

SAVE OUTFILE="C:\Users\User\Desktop\Ptuxiaki\askiseis\askisi_ypodeigma.sav".


DESCRIPTIVES

DESCRIPTIVES
  /VARIABLES=work pay promote
  /STATISTICS=DEFAULT RANGE.
```

Valid cases = 50; cases with missing value(s) = 2.

Variable	*N*	*Mean*	*Std Dev*	*Range*	*Minimum*	*Maximum*
ΑΞΙΟΛΟΓΗΣΗ ΣΥΝΘΗΚΩΝ ΕΡΓΑΣΙΑΣ	50	79,80	8,29	32,00	63,00	95,00
ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΔΟΧΩΝ	49	54,69	14,81	65,00	25,00	90,00
ΑΞΙΟΛΟΓΗΣΗ ΔΥΝΑΤΟΤΗΤΩΝ ΕΞΕΛΙΞΗΣ	49	58,84	15,96	76,00	16,00	92,00

Σχήμα 5.2 Αποτελέσματα της διαδικασίας Descriptives

➢ Θα εφαρμόσουμε τις εντολές μετασχηματισμού και επιλογής δεδομένων στα δεδομένα μας και συγκεκριμένα:

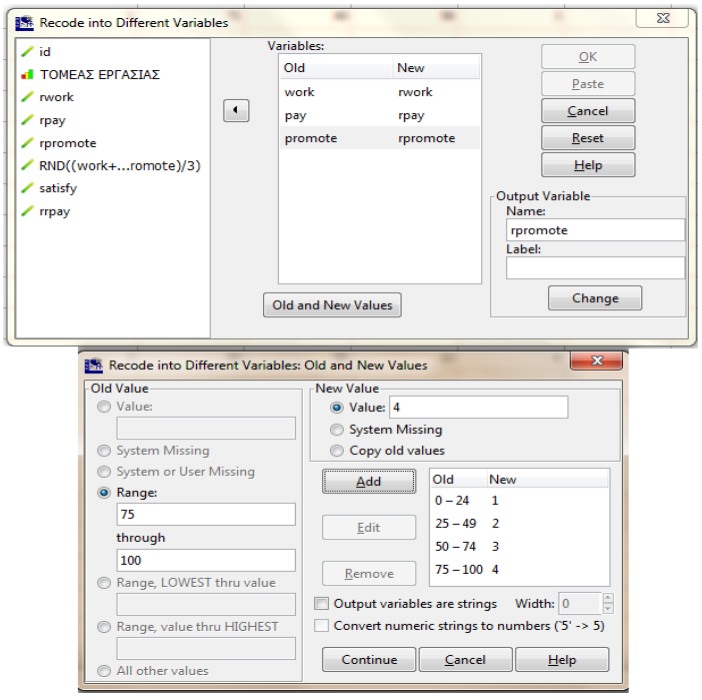
(α) Θα κωδικοποιήσουμε τις μεταβλητές *work, pay* και *promote* σε 4 ομάδες ορίζοντας νέες μεταβλητές rwork, rpay και rpromote οι οποίες θα δεχτούν τις κωδικοποιημένες τιμές.

(β) Θα υπολογίσουμε για τον κάθε εργαζόμενο στο δείγμα ένα μέσο όρο αξιολόγησης της επιχείρησης ο οποίος θα εκφράζεται με ακέραιο αριθμό 0 – 100 και θα καταχωρηθεί σε μία νέα μεταβλητή *avrscore.*

(γ) Θα χαρακτηρίσουμε κάθε εργαζόμενο ικανοποιημένο από την εταιρεία του αν έχει δώσει μέση αξιολόγηση πάνω από 65, δημιουργώντας μία νέα μεταβλητή satisfy με τιμές 1 (ικανοποιημένος) και 0 (μη ικανοποιημένος).

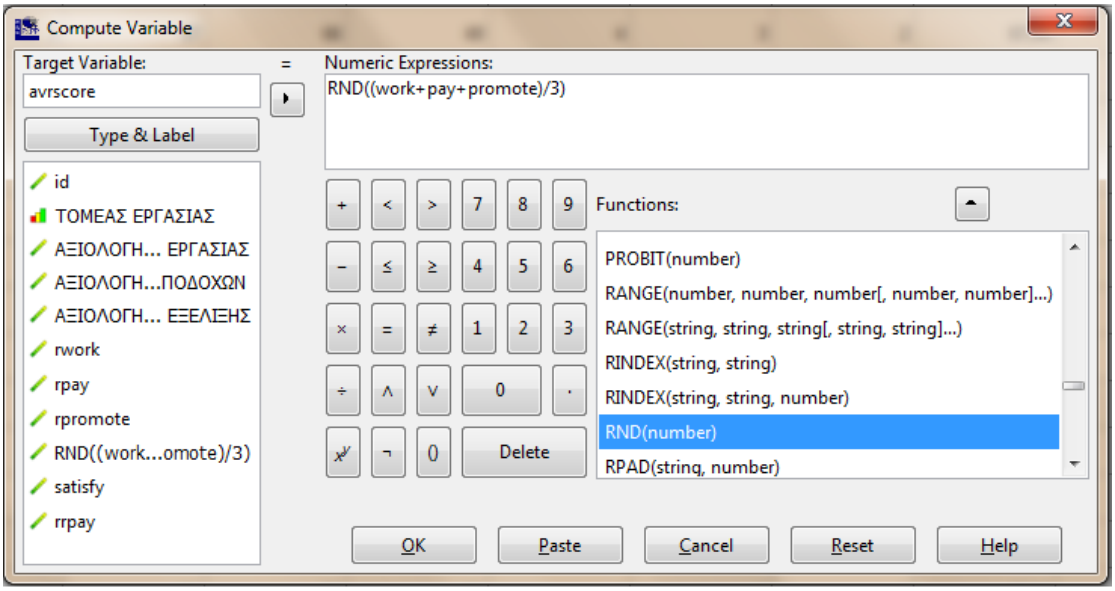
(α) Επιλέγοντας διαδοχικά από το κύριο menu του PSPP, *Transform->Recode Into Different Variables* εμφανίζεται το αρχικό παράθυρο διαλόγου της εντολής (Σχήμα 5.3) όπου δηλώνουμε ποιες μεταβλητές θα κωδικοποιηθούν (work, pay, promote) και ποιες θα δεχτούν τις κωδικοποιημένες τιμές (rwork, rpay, rpromote).

Στο ίδιο παράθυρο αν πατήσουμε το *Old and New Values* θα εμφανιστεί το παράθυρο που βλέπουμε στο Σχήμα 5.3 κάτω, όπου δηλώνουμε τις παλαιές και τις νέες τιμές των τριών μεταβλητών που θα κωδικοποιήσουμε. Η κωδικοποίηση που θα γίνει φαίνεται στο πλαίσιο *Old→New* και σχηματίστηκε δίνοντας τιμές στο πεδίο *Range*: του αριστερού πλαισίου. Μετά από την εκτέλεση της εντολής αυτής θα δημιουργηθούν οι μεταβλητές *rwork, rpay, rpromote* που θα έχουν τιμές 1– 4.



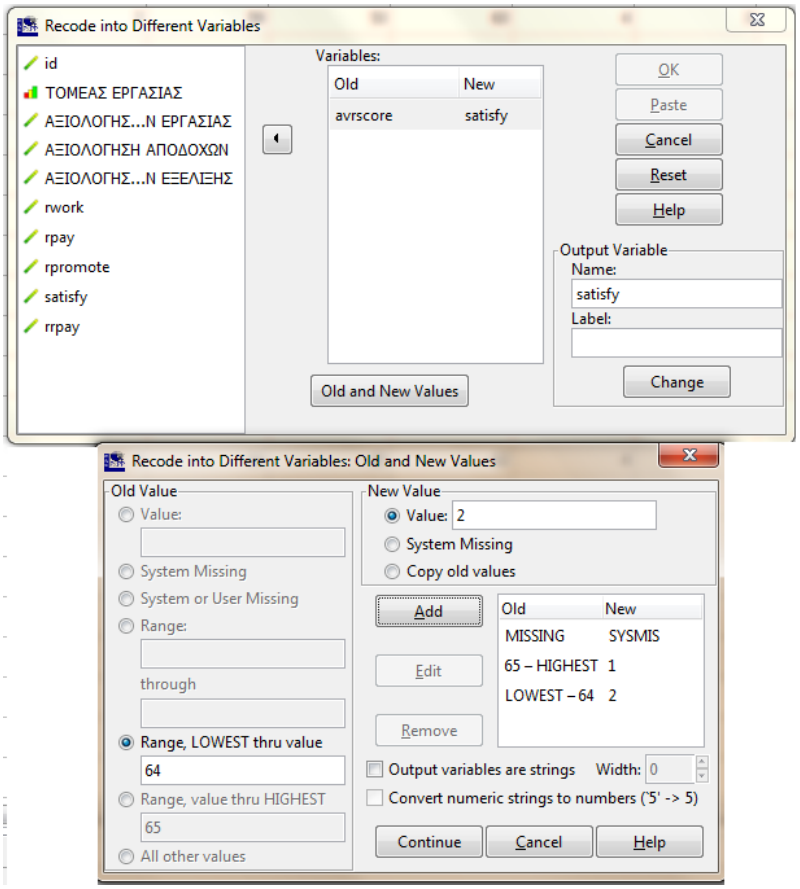
Σχήμα 5.3 Αρχικό παράθυρο διαλόγου Recode Into Different Variables και κάτω η επιλογή Old and New Values.

(β) Στη συνέχεια επιλέγοντας *Transform-> Compute* εμφανίζεται το παράθυρο που βλέπουμε στο Σχήμα 5.4. Στην θέση *Target Variable* γράφουμε το όνομα της νέας μεταβλητής *avrscore* που θα δεχτεί τις τιμές του υπολογισμού εύρεσης της μέσης αξιολόγησης για κάθε υπάλληλο. Ο ακριβής τύπος υπολογισμού της νέας μεταβλητής συντάσσεται στο πλαίσιο *Numeric Expression* με τη βοήθεια του πινακιδίου υπολογισμών και του καταλόγου των συναρτήσεων *(Functions).* Στο συγκεκριμένο παράδειγμα οι τιμές των μεταβλητών work, pay και promote αθροίζονται και το άθροισμα διαιρείται με 3. Στη συνέχεια το αποτέλεσμα στρογγυλοποιείται από τη συνάρτηση *RND().* Η εκτέλεση της εντολής θα έχει σαν αποτέλεσμα τη δημιουργία της νέας μεταβλητής avrscore με τιμές τα αποτελέσματα της πράξης που περιγράψαμε παραπάνω.



Σχήμα 5.4 Παράθυρο υπολογισμού της μεταβλητής avrscore μέσω της διαδικασίας Compute

(γ) Η μεταβλητή *satisfy* θα δημιουργηθεί με την διαδικασία *Recode Into Different.* Το Σχήμα 5.5 δείχνει τον τρόπο με τον οποίο θα δηλώσουμε την νέα μεταβλητή και τις τιμές που θα πάρει αυτή. Μετά την εκτέλεση της εντολής στο αρχείο των δεδομένων μας θα υπάρχει η νέα μεταβλητή *satisfy* που θα έχει τιμές 1 (ικανοποιημένος) και 2 (μη ικανοποιημένος).



Σχήμα 5.5 Αρχικό παράθυρο διαλόγου Recode Into Different Variables και κάτω η επιλογή Old and New Values για τον ορισμό τιμών της satisfy.

Στη συνέχεια δίνουμε κάποια αποτελέσματα περιγραφικής στατιστικής πάνω στις νέες μεταβλητές που έχουμε δημιουργήσει με τις προηγούμενες εντολές *Recode* και *Compute*. Διευκρινίζεται ότι τα αποτελέσματα έχουν εξαχθεί στο σύνολο όλων των δεδομένων και όχι σε επιλεγμένες περιπτώσεις μόνο. Συγκεκριμένα, στο Σχήμα 5.6 βλέπουμε το παράθυρο των αποτελεσμάτων όπου έχουμε υπολογίσει με την εντολή *Descriptives* τα στατιστικά μέτρα της μεταβλητής *avrscore*. Στο Σχήμα 5.7 υπάρχει το ιστόγραμμα της ίδιας μεταβλητής. Στο Σχήμα 5.8 φαίνεται ο πίνακας κατανομής συχνοτήτων της μεταβλητής *satisfy* που υπολογίστηκε με την εντολή *Frequencies.* Τέλος, στο Σχήμα 5.9 έχουμε το ιστόγραμμα και το κυκλικό διάγραμμα στα οποία φαίνεται γραφικά η κατανομή της μεταβλητής *satisfy*.

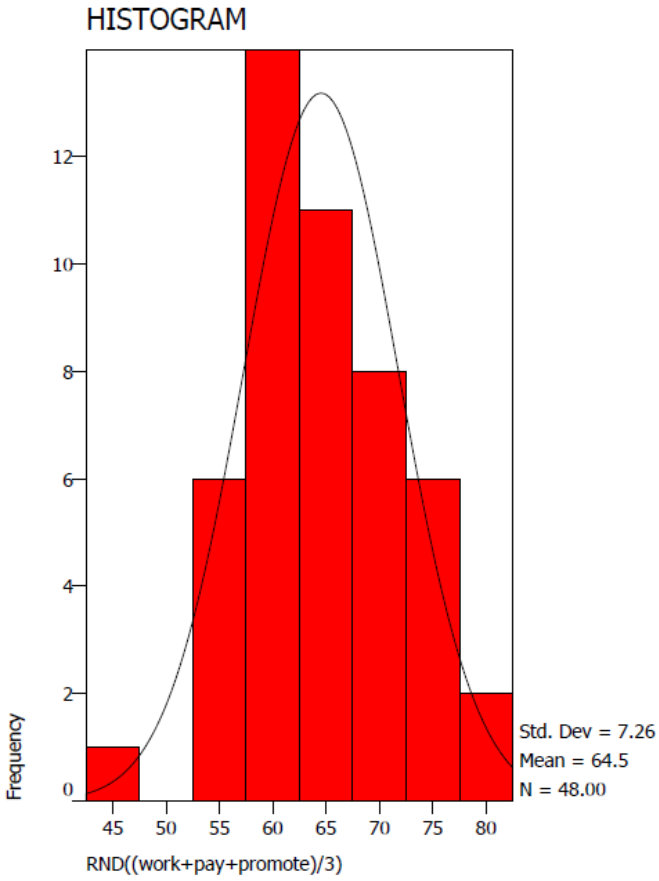
DESCRIPTIVES

```
DESCRIPTIVES
  /VARIABLES=avrscore
  /STATISTICS=DEFAULT RANGE.
```

Valid cases = 50; cases with missing value(s) = 2.

Variable	*N*	*Mean*	*Std Dev*	*Range*	*Minimum*	*Maximum*
RND((work+pay+promote)/3)	48	64,52	7,26	35,00	44,00	79,00

Σχήμα 5.6 Στατιστικά μέτρα της avrscore



Σχήμα 5.7 Ιστόγραμμα της μεταβλητής avrscore

```
FREQUENCIES

FREQUENCIES
    /VARIABLES=satisfy
    /FORMAT=AVALUE TABLE
    /STATISTICS=MEAN
    /MISSING=INCLUDE
    /HISTOGRAM=NONORMAL PERCENT
    /PIECHART=.
```

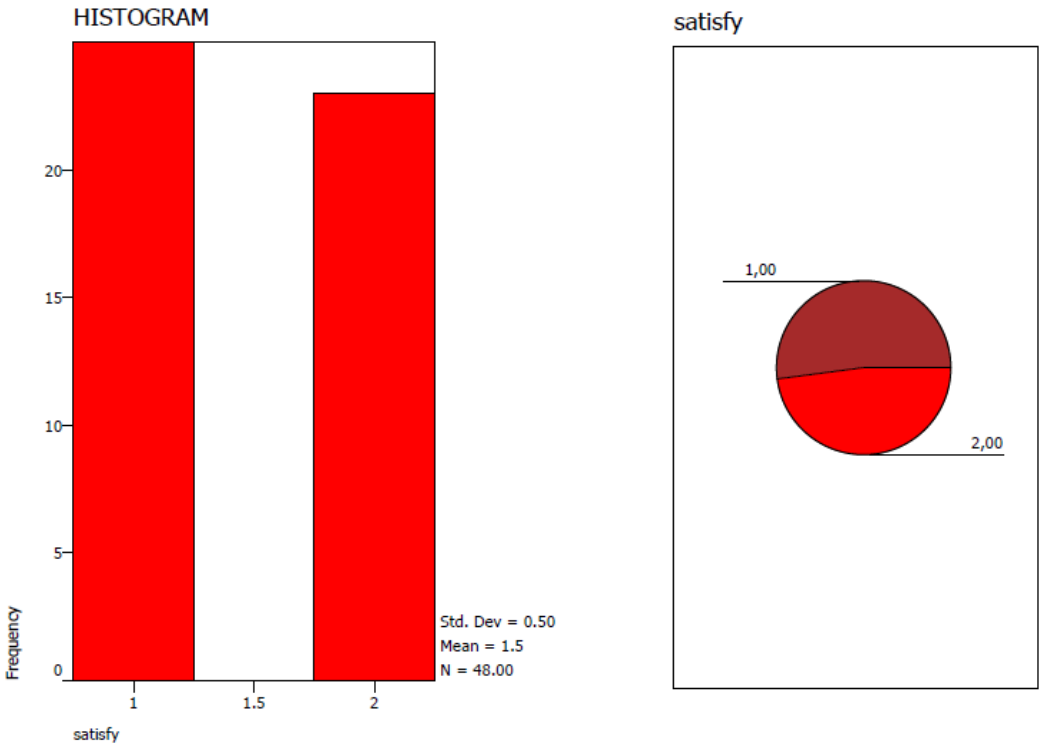
satisfy

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	1,00	25	50,00	52,08	52,08
	2,00	23	46,00	47,92	100,00
	.	2	4,00	Missing	
	Total	50	100,0	100,0	

satisfy

N	Valid	48
	Missing	2
Mean		1,48

Σχήμα 5.8 Πίνακας κατανομής συχνοτήτων της μεταβλητής satisfy



Σχήμα 5.9 Ιστόγραμμα και κυκλικό διάγραμμα της μεταβλητής satisfy

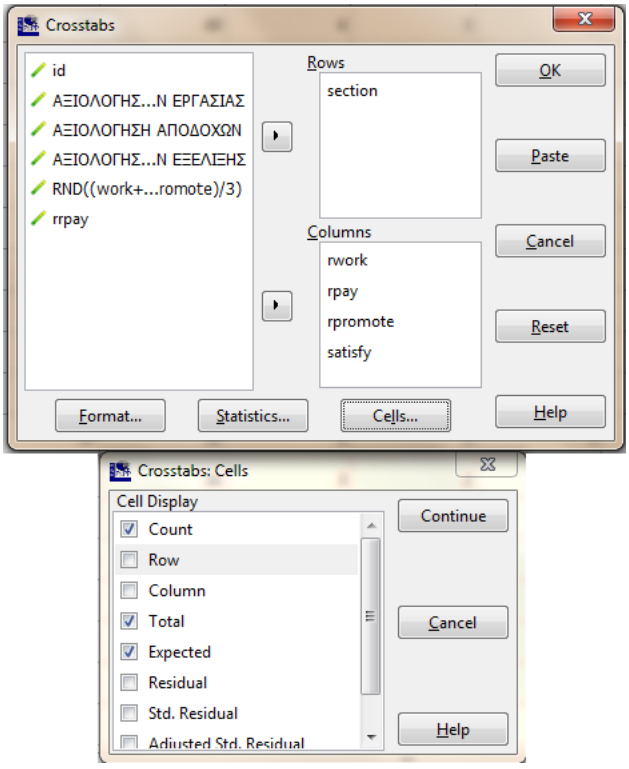
➢ Θα χρησιμοποιήσουμε τώρα στο παράδειγμά μας, την διαδικασία Crosstabs, για την κατασκευή πινάκων συνάφειας και για ελέγχους ανεξαρτησίας. Για να το πετύχουμε αυτό χρησιμοποιούμε τις νέες κατηγοριοποιημένες μεταβλητές που δημιουργήσαμε, κωδικοποιώντας τις ποσοτικές που ήδη υπάρχουν *(rwork, rpay, rpromote, satisfy).*

Αφού ανοίξουμε το παράθυρο διαλόγου της διαδικασίας *Crosstabs (Analyze->Descriptive Statistics->Crosstabs)*, όπως φαίνεται και στο Σχήμα 5.10, δηλώνουμε αρχικά τη μεταβλητή *section* στη λίστα *Row(s)* και τις μεταβλητές *rwork, rpay, rpromote* και *satisfy* στη λίστα *Column(s).* Με τον τρόπο αυτό εκφράζουμε την επιθυμία μας να δημιουργηθούν τέσσερις πίνακες συνάφειας, όπου ο καθένας θα έχει τις τιμές της μεταβλητής *section* στις γραμμές του και τις τιμές των υπόλοιπων μεταβλητών στις στήλες.

Για να ελέγξουμε τις υποθέσεις ανεξαρτησίας της *section* με τις άλλες μεταβλητές, επιλέγουμε το *Statistics* και στο παράθυρο που ανοίγει σημειώνουμε μόνο το *Chi-square*.

Από το αρχικό παράθυρο, για να καθορίσουμε το περιεχόμενο των κελιών, επιλέγουμε το *Cells* και σημειώνουμε τα τετραγωνίδια *Count* , *Expected* και *Total (Percentages),* έτσι ώστε σε κάθε κελί να εμφανίζεται η αναμενόμενη συχνότητά του και το ποσοστό (σχετική συχνότητα) ως προς το σύνολο των παρατηρήσεων. Αυτά φαίνονται επίσης στο Σχήμα 5.10.

Το *Format* το αφήνουμε να έχει τις εξορισμού επιλογές του, δηλαδή να εμφανίζονται οι πίνακες και οι γραμμές να είναι σε αύξουσα σειρά.



Σχήμα 5.10 Δήλωση μεταβλητών στο αρχικό παράθυρο της διαδικασίας Crosstabs και κάτω οι επιλογές του Cells

Μετά την εκτέλεση της διαδικασίας θα πάρουμε στο παράθυρο αποτελεσμάτων (Σχήμα 5.11) τέσσερις πίνακες συνάφειας διαμορφωμένους όπως καθορίσαμε, που θα συνοδεύονται από το στατιστικό x^2. Εδώ πρέπει να σημειώσουμε ότι στα αποτελέσματα τυπώνονται 3 διαφορετικές εκδοχές του x^2 *(Pearson's, Likelihood Ratio, Linear-by-Linear Association).* Θα λάβουμε υπόψη μόνο αυτό του *Pearson*, αγνοώντας τα υπόλοιπα. Παρακάτω δίνουμε τα αποτελέσματα του *Crosstabs* μόνο για ένα από τα 4 ζευγάρια, τη μεταβλητή *section* με την *rpay*. Βλέπουμε ότι το *Significance* που μας ενδιαφέρει είναι 0.23 > 0.05 που σημαίνει ότι δεν υπάρχει εξάρτηση ανάμεσα στον τομέα εργασίας και στην αξιολόγηση των αποδοχών. Το αποτέλεσμα αυτό όμως, δεν είναι αξιόπιστο αφού το ποσοστό των κελιών με αναμενόμενη συχνότητα <5 (4 Cells with Expected Frequency <5) είναι 44.4%.

ΤΟΜΕΑΣ ΕΡΓΑΣΙΑΣ * rpay [count, total %, expected].

ΤΟΜΕΑΣ ΕΡΓΑΣΙΑΣ	*rpay* 2	3	4	Total
ΕΡΕΥΝΑ ΑΓΟΡΑΣ	7,0	8,0	3,0	18,0
	6,6	9,6	1,8	,0
	14,3%	16,3%	6,1%	36,7%
ΔΗΜΟΣΙΕΣ ΣΧΕΣΕΙΣ	9,0	8,0	1,0	18,0
	6,6	9,6	1,8	,0
	18,4%	16,3%	2,0%	36,7%
ΔΙΑΦΗΜΙΣΗ	2,0	10,0	1,0	13,0
	4,8	6,9	1,3	,0
	4,1%	20,4%	2,0%	26,5%
Total	18,0	26,0	5,0	49,0
	36,7%	53,1%	10,2%	100,0%

Chi-square tests.

Statistic	*Value*	*df*	*Asymp. Sig. (2-tailed)*
Pearson Chi-Square	5,60	4	,23
Likelihood Ratio	5,79	4	,22
Linear-by-Linear Association	,23	1	,63
N of Valid Cases	49		

Σχήμα 5.11 Πίνακας αποτελεσμάτων section*rpay

Λόγω της μη εγκυρότητας του ελέγχου, είμαστε υποχρεωμένοι να συμπτύξουμε κατηγορίες σε μία τουλάχιστον από τις μεταβλητές. Για το λόγο αυτό κωδικοποιούμε την αρχική μεταβλητή *pay*, χρησιμοποιώντας την *Recode*, σε μία νέα ποιοτική μεταβλητή, την *rrpay* που παίρνει δύο τιμές, 1 αν η αξιολόγηση είναι 0-49 και 2 αν είναι 50-100. Με αυτόν

τον τρόπο επιτυγχάνουμε σύμπτυξη του πίνακα και μείωση του ποσοστού των κελιών με αναμενόμενη συχνότητα < 5. Στα αποτελέσματα της εκτέλεσης της διαδικασίας *Crosstabs* για τις μεταβλητές *section* και *rrpay* που ακολουθούν (Σχήμα 5.12), παρατηρούμε ότι *Significance = 0.14*, δηλαδή το συμπέρασμα της ανεξαρτησίας παραμένει ουσιαστικά το ίδιο, αλλά τα αποτελέσματά μας είναι πλέον αξιόπιστα αφού το ποσοστό των κελιών με αναμενόμενη συχνότητα <5 είναι τώρα 16.7%.

ΤΟΜΕΑΣ ΕΡΓΑΣΙΑΣ * rrpay [count, total %, expected].

	rrpay		
ΤΟΜΕΑΣ ΕΡΓΑΣΙΑΣ	1,00	2,00	Total
ΕΡΕΥΝΑ ΑΓΟΡΑΣ	7,0	11,0	18,0
	6,6	11,4	,0
	14,3%	22,4%	36,7%
ΔΗΜΟΣΙΕΣ ΣΧΕΣΕΙΣ	9,0	9,0	18,0
	6,6	11,4	,0
	18,4%	18,4%	36,7%
ΔΙΑΦΗΜΙΣΗ	2,0	11,0	13,0
	4,8	8,2	,0
	4,1%	22,4%	26,5%
Total	18,0	31,0	49,0
	36,7%	63,3%	100,0%

Chi-square tests.

Statistic	*Value*	*df*	*Asymp. Sig. (2-tailed)*
Pearson Chi-Square	3,95	2	,14
Likelihood Ratio	4,27	2	,12
Linear-by-Linear Association	1,38	1	,24
N of Valid Cases	49		

Σχήμα 5.12 Πίνακας αποτελεσμάτων μετά τον μετασχηματισμό section*rrpay

➢ Ας υποθέσουμε ότι θέλουμε να απαντήσουμε στις παρακάτω ερωτήσεις που αναφέρονται στις ποσοτικές μεταβλητές:

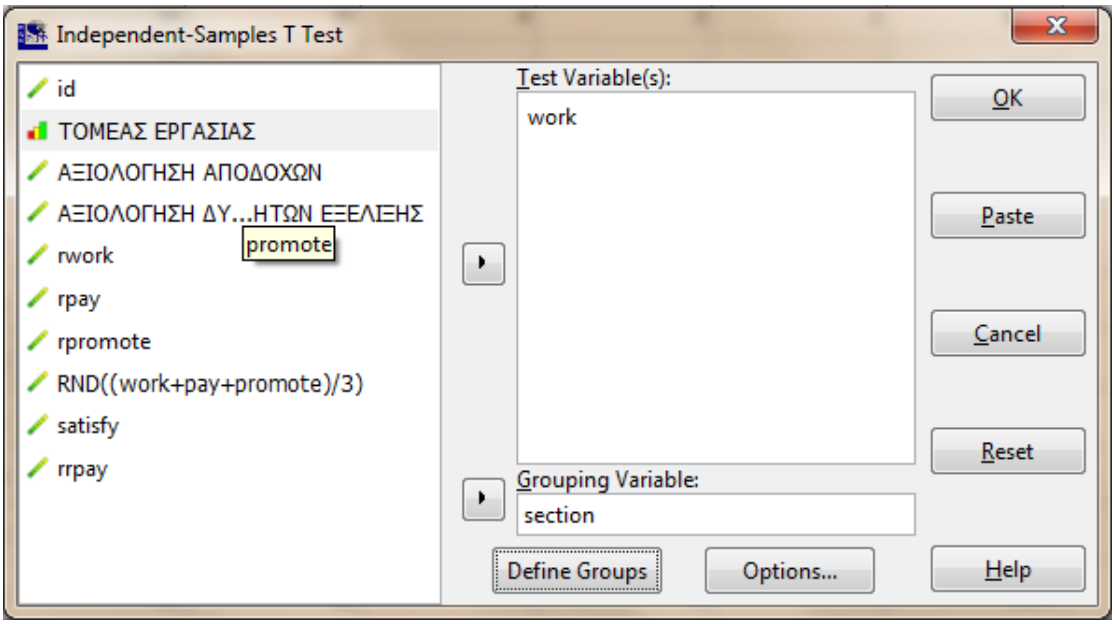
1. Οι αξιολογήσεις των υπαλλήλων, καθώς και οι μέσες αξιολογήσεις αυτών, διαφέρουν κατά μέση τιμή μεταξύ των εργαζομένων στον τομέα έρευνας αγοράς και των εργαζομένων στον τομέα της διαφήμισης;

2. Οι αξιολογήσεις των εργαζομένων για τις συνθήκες εργασίες και τις αποδοχές διαφέρουν κατά μέση τιμή ;

3. Η μέση τιμή αξιολόγησης των αποδοχών ισούται με 65 μονάδες ;

Οι απαντήσεις στα 3 αυτά ερωτήματα θα δοθούν μετά από εφαρμογή των κατάλληλων διαδικασιών t-tests. Αρχικά θα εκτελέσουμε τους ελέγχους στο PSPP, συνοδεύοντας τις περιγραφές με αποσπάσματα από τις εκτυπώσεις των αποτελεσμάτων.

1. Αφού ανοίξουμε το παράθυρο της διαδικασίας *Independent Samples T Test (Analyze->Compare Means->Independent Samples T Test)*, δηλώνουμε τις μεταβλητές που θέλουμε να ελεγχθούν στο πλαίσιο *Test Variable(s)*. Στο πλαίσιο *Grouping Variable* δηλώνουμε την μεταβλητή ομαδοποίησης και στο *Define Groups* τις κατηγορίες ανάμεσα στις οποίες θα γίνει ο έλεγχος(Σχήμα 5.13). Θα ελεγχθούν οι μέσες τιμές των μεταβλητών *work, pay, promote* και *avrscore* ως προς τους δύο τομείς εργασίας (1 και 3) που ορίζονται από την μεταβλητή ομαδοποίησης *section*. Απόσπασμα από τα αποτελέσματα, μόνο για τη μεταβλητή *work*, βλέπουμε στη συνέχεια (Σχήμα 5.14).



Σχήμα 5.13 Δήλωση μεταβλητών στο Independent-Samples T Test

Παρατηρούμε ότι αφού υπολογιστούν κάποια στατιστικά μέτρα για τις δύο ανεξάρτητες ομάδες της μεταβλητής work, γίνεται έλεγχος για τη σύγκριση των διασπορών *(Levene's Test for Equality of Variances)* όπου από τη σημαντικότητα 0.46 >0.05 συμπεραίνουμε ότι δεν υπάρχει σημαντική διαφορά στις διασπορές και επομένως μπορούμε να τις θεωρήσουμε ίσες. Από τα δύο t-test που ακολουθούν θα επιλέξουμε τη γραμμή *Equal variances assumed που* δίνει σημαντικότητα ελέγχου 0.59 > 0.05 και επομένως οι δύο μέσες τιμές δεν εμφανίζουν στατιστικά σημαντική διαφορά (οι εργαζόμενοι στους τομείς 1=Έρευνα Αγοράς και 3=Διαφήμιση αξιολογούν κατά μέσο όρο με τον ίδιο τρόπο την εργασία τους).

```
T-TEST

T-TEST /VARIABLES=work
    /GROUPS=section(1,3)     /MISSING=ANALYSIS
    /CRITERIA=CIN(0.95).
```

Group Statistics

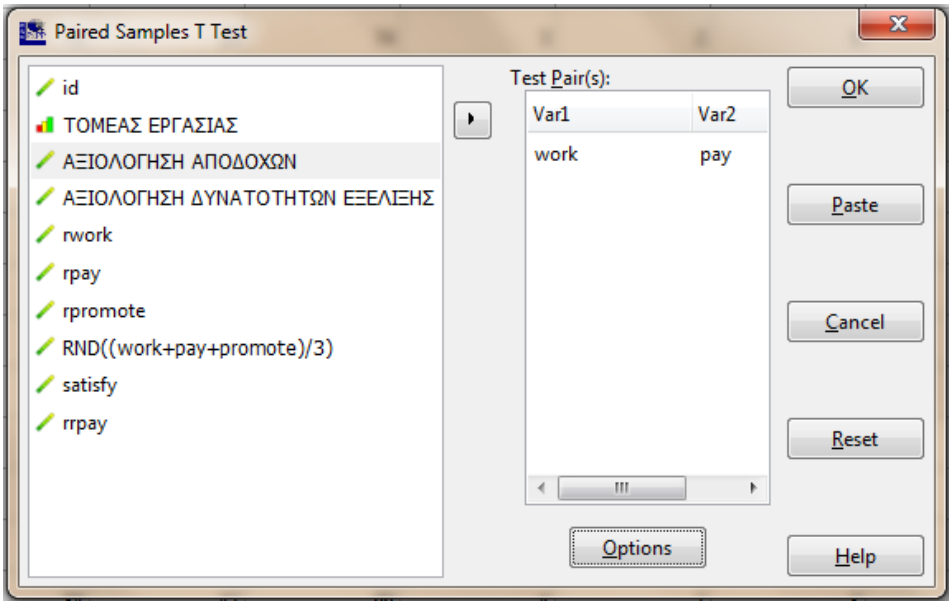
	ΤΟΜΕΑΣ ΕΡΓΑΣΙΑΣ	N	Mean	Std. Deviation	S.E. Mean
ΑΞΙΟΛΟΓΗΣΗ ΣΥΝΘΗΚΩΝ ΕΡΓΑΣΙΑΣ	ΕΡΕΥΝΑ ΑΓΟΡΑΣ	18	80,67	9,47	2,23
	ΔΙΑΦΗΜΙΣΗ	13	78,92	7,65	2,12

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
ΑΞΙΟΛΟΓΗΣΗ ΣΥΝΘΗΚΩΝ ΕΡΓΑΣΙΑΣ	Equal variances assumed	,55	,46	,55	29,00	,59	1,74	3,08	-4,56	8,05
	Equal variances not assumed			,57	28,56	,58	1,74	3,08	-4,56	8,05

Σχήμα 5.14 Αποτελέσματα του Independent Samples T Test για τη μεταβλητή work

2. Στο Σχήμα 5.15 βλέπουμε τον τρόπο που δηλώνουμε σε ζευγάρια τις μεταβλητές των οποίων τις μέσες τιμές θέλουμε να συγκρίνουμε στο παράθυρο *Paired SamplesT Test (Analyze->CompareMeans->PairedSamplesTTest).* Το ζευγάρι που μας ενδιαφέρει να ελέγξουμε είναι αυτό των μεταβλητών work και *pay*.



Σχήμα 5.15 Δήλωση μεταβλητών στο Paired Samples T Test

Στα αποτελέσματα του ελέγχου (Σχήμα 5.16), παρατηρούμε ότι ο συντελεστής συσχέτισης *Correlation= 0.24* είναι θετικός και επομένως ο έλεγχος σε ζευγάρια είναι δικαιολογημένος. Η σημαντικότητα του ελέγχου (Sig.) είναι 0.00 < 0.05 και επομένως φαίνεται ότι υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στις δύο αξιολογήσεις (εργασίας και αποδοχών).

T-TEST

```
T-TEST
  PAIRS = work WITH pay (PAIRED)
  /MISSING=ANALYSIS
  /CRITERIA=CIN(0.95).
```

Paired Sample Statistics

		Mean	N	Std. Deviation	S.E. Mean
Pair 0	ΑΞΙΟΛΟΓΗΣΗ ΣΥΝΘΗΚΩΝ ΕΡΓΑΣΙΑΣ	80,12	49	8,05	1,15
	ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΔΟΧΩΝ	54,69	49	14,81	2,12

Paired Samples Correlations

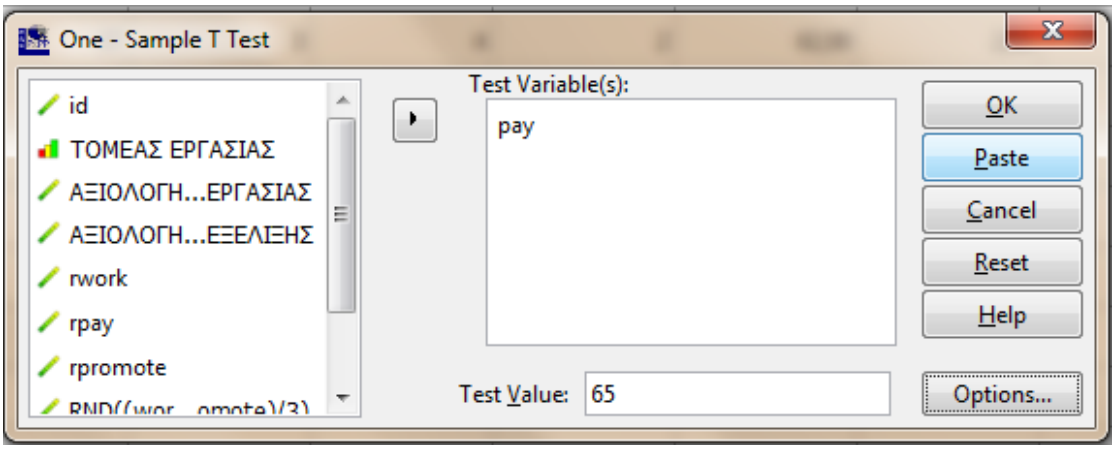
		N	Correlation	Sig.
Pair 0	ΑΞΙΟΛΟΓΗΣΗ ΣΥΝΘΗΚΩΝ ΕΡΓΑΣΙΑΣ & ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΔΟΧΩΝ	49	,24	,10

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 0	ΑΞΙΟΛΟΓΗΣΗ ΣΥΝΘΗΚΩΝ ΕΡΓΑΣΙΑΣ - ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΔΟΧΩΝ	25,43	15,07	2,15	21,10	29,76	11,81	48	,00

Σχήμα 5.16 Αποτελέσματα του Paired Samples T Test

3. Στο Σχήμα 5.17 βλέπουμε τον τρόπο με τον οποίο δηλώνουμε μεταβλητές που θα ελεγχθούν για το αν η μέση τιμή τους διαφέρει σημαντικά από την δεδομένη τιμή 65 *(Analyze->CompareMeans->One-SampleTTest)*. Στα αποτελέσματα που βλέπουμε παρακάτω (Σχήμα 5.18), μόνο για την μεταβλητή pay, παρατηρούμε ότι η σημαντικότητα είναι 0.00 < 0.05 και επομένως φαίνεται ότι υπάρχει σημαντική διαφορά της μέσης τιμής από την τιμή 65.



Σχήμα 5.17 Δήλωση μεταβλητών στο One-Sample T Test

T-TEST

```
T-TEST /TESTVAL=65
    /VARIABLES=pay /MISSING=ANALYSIS
    /CRITERIA=CIN(0.95).
```

One-Sample Statistics

	N	Mean	Std. Deviation	S.E. Mean
ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΔΟΧΩΝ	49	54,69	14,81	2,12

One-Sample Test

	Test Value = 65.000000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΔΟΧΩΝ	-4,87	48	,00	-10,31	-14,56	-6,05

Σχήμα 5.18 Αποτελέσματα του One-Sample T Test για την μεταβλητή pay

6. Ασκήσεις

6.1 Εκφωνήσεις Ασκήσεων

Σημείωση: Το σύνολο των εκφωνήσεων των ασκήσεων καθώς και τα αρχεία δεδομένων PSPP στα οποία γίνεται αναφορά, είναι διαθέσιμα στην ιστοσελίδα του κ.Βασίλη Κώστογλου (http://aetos.it.teithe.gr/~vkostogl/) στην καρτέλα Μαθήματα – Εργαστήριο Στατιστικής.

✓ Άσκηση 1

Οι υπεύθυνοι του εργαστηρίου «Θεωρία Πιθανοτήτων & Στατιστική» θέλοντας να αξιολογήσουν τον τρόπο διεξαγωγής του μαθήματος αποφάσισαν την διενέργεια μιας έρευνας. Στην έρευνα αυτή συμμετείχαν οι φοιτητές ενός τυχαία επιλεγμένου τμήματος από τα τμήματα, τα οποία απαρτίζουν το συγκεκριμένο εργαστήριο. Η έρευνα διενεργήθηκε με τη χρήση ερωτηματολογίου, το οποίο περιλάμβανε μεταξύ άλλων και τις παρακάτω ερωτήσεις:

1. Ποια είναι η ηλικία σας;

2. Πόσες φορές παρακολουθήσατε στο παρελθόν το συγκεκριμένο εργαστήριο;

3. Η ύλη που διδάσκεται είναι καλά οργανωμένη;
   [Καθόλου - Λίγο - Μέτρια - Πολύ - Πάρα πολύ]

4. Ο στόχος σας στο εργαστήριο είναι:
   [Να το περάσω - Να το περάσω με μεγάλο βαθμό - Να το περάσω και να αποκτήσω γνώσεις]

5. Πόσο ικανοποιητική είναι η καθαριότητα των εργαστηρίων;
   [Βαθμολογήστε στην κλίμακα 0 (καθόλου) – 100 (πάρα πολύ)]

6. Ποια από τα παρακάτω θεωρείτε ως τα στοιχεία εκείνα που απαιτούν βελτίωση σε ό,τι αφορά τον διδάσκοντα του εργαστηρίου (μπορείτε να επιλέξετε περισσότερα του ενός);
   [Μεταδοτικότητα - Προετοιμασία - Φιλικότητα - Κανένα]

Συζήτηση για τις παρακάτω εισαγωγικές έννοιες της Στατιστικής:

- Πληθυσμός & δείγμα
- Μεταβλητές (ποιοτικές- ποσοτικές)
- Ορισμός πεδίων μεταβλητής
- Περιγραφική Στατιστική (Descriptive Statistics)
- Συχνότητες (Frequencies)
- Διαγράμματα (Charts)

Περιγραφική στατιστική ανάλυση των 20 συμπληρωμένων ερωτηματολογίων.
Τα πρωτογενή δεδομένα της έρευνας είναι τα ακόλουθα:

α/α φοιτητή	**Ερ. 1**	**Ερ. 2**	**Ερ. 3**	**Ερ. 4**	**Ερ. 5**	**Ερ. 6**
1	18	0	Λίγο	Να το περάσω	65	Μεταδοτικότητα Φιλικότητα
2	19	0	Μέτρια		78	Μεταδοτικότητα
3	18	0	Μέτρια	Να το περάσω με μεγάλο βαθμό	68	
4	20	1	Πολύ	Να το περάσω	82	Μεταδοτικότητα Φιλικότητα
5	18	0	Λίγο	Να το περάσω με μεγάλο βαθμό	56	Μεταδοτικότητα Προετοιμασία Φιλικότητα
6	24	5	Μέτρια		74	
7	19,5	2	Πολύ	Να το περάσω	63,75	Μεταδοτικότητα Φιλικότητα
8	22	3	Λίγο	Να το περάσω	55	Μεταδοτικότητα
9	19	0	Πολύ	Να το περάσω με μεγάλο βαθμό	50	
10	18	0	Πολύ	Να το περάσω	30	Κανένα
11	19,75	1	Πάρα πολύ	Να το περάσω με μεγάλο βαθμό	90	Κανένα
12	20	1	Πολύ	Να το περάσω	75	Μεταδοτικότητα Φιλικότητα
13	21	0	Λίγο	Να το περάσω και να αποκτήσω γνώσεις	78,5	Μεταδοτικότητα
14	21	2	Πολύ	Να το περάσω	65	Κανένα
15	19	0	Μέτρια	Να το περάσω και να αποκτήσω γνώσεις	70	Κανένα

16	19	1	Πολύ	Να το περάσω και να αποκτήσω γνώσεις	80	Μεταδοτικότητα Φιλικότητα
17	19	0	Μέτρια	Να το περάσω	80	Κανένα
18	20	1	Μέτρια	Να το περάσω	60	Μεταδοτικότητα
19	21	2	Μέτρια	Να το περάσω και να αποκτήσω γνώσεις	68	Μεταδοτικότητα
20	18	0	Μέτρια	Να το περάσω	75	

✓ Άσκηση 2

Τα Συστήματα Ανίχνευσης Εισβολής (Intrusion Detection Systems) αποτελούν ένα απαραίτητο συστατικό κάθε ολοκληρωμένης αρχιτεκτονικής ασφάλειας και έχουν ως κύρια αποστολή την έγκαιρη ανίχνευση αλλά και αποτροπή των διάφορων τύπων επιθέσεων που απειλούν τα δίκτυα υπολογιστών.

Το MIT Lincoln Labs υλοποίησε το πρόγραμμα DARPA (1998), το οποίο είχε ως στόχο την αξιολόγηση της έρευνας στο πεδίο της Ανίχνευσης Εισβολών. Για το σκοπό αυτό προσομοίωσε ένα πλήθος διαφορετικού τύπου εισβολών σ' ένα δικτυακό περιβάλλον και κατέγραψε τις τιμές διάφορων χαρακτηριστικών κατά την διάρκεια της διαδικασίας αυτής. Η καταγραφή των χαρακτηριστικών αυτών πραγματοποιήθηκε για κάθε μία "σύνδεση". Μία "σύνδεση" προσδιορίζεται από μία σειρά TCP πακέτων, τα οποία ξεκινούν και ολοκληρώνονται σε καθορισμένες χρονικές στιγμές κατά τις οποίες δεδομένα ρέουν μεταξύ δύο IP διευθύνσεων. Για κάθε "σύνδεση" είναι γνωστό εάν είναι "κανονική" ή αποτελεί μία "εισβολή" στο δίκτυο.

Η συλλογή των δεδομένων αυτών έχει ως στόχο την περιγραφή των συγκεκριμένων χαρακτηριστικών κατά τη διάρκεια μίας "σύνδεσης", έτσι ώστε να διερευνηθεί η σχέση τους με τον τύπο της "σύνδεσης", δηλαδή με το αν η συγκεκριμένη "σύνδεση" αποτελεί μία εισβολή ή όχι. Αναλυτικότερες πληροφορίες για τη μελέτη αυτής της περίπτωσης, μπορείτε να αντλήσετε από την παρακάτω ιστοσελίδα:

http://archive.ics.uci.edu/ml/databases/kddcup99/kddcup99.html

Στην επόμενη σελίδα παρουσιάζονται δεδομένα τα οποία έχουν καταγραφεί για 25 "συνδέσεις" και τα οποία αποτελούν ένα μέρος των πραγματικών δεδομένων.

Τα χαρακτηριστικά των οποίων οι τιμές παρουσιάζονται είναι τα ακόλουθα:

- protocol_type: type of the protocol (tcp, udp)
- scr_bytes: number of data bytes from source to destination
- dst_bytes: number of data bytes from destination to source
- num_failed_logins: number of failed logins
- attack_type: type of the attack or normal

protocol_type	scr_bytes	dst_bytes	num_failed_logins	attack_type
tcp	241	259	0	normal
tcp	212	4433	0	normal
tcp	252	11627	0	normal
tcp	185	1263	0	normal
icm	1032	0	0	smurf
tcp	209	566	0	normal
udp	44	78	0	normal
tcp	428	7512	0	normal
tcp	305	5357	0	normal
tcp	253	11696	0	normal
tcp	238	11680	0	normal
tcp	212	487	0	normal
tcp	54540	8314	1	back
tcp	0	0	0	normal
icm	1032	0	2	smurf
icm	1032	0	1	smurf
icm	8	0	1	ipsweep
tcp	1116	326	0	normal
tcp	0	0	1	neptune
tcp	0	0	1	neptune
udp	44	113	0	normal
tcp	325	1565	0	normal
tcp	219	2678	0	normal
tcp	11492	1437	0	normal
tcp	308	415	0	normal

1. Να δημιουργηθεί στο πρόγραμμα PSPP κατάλληλο αρχείο, στο οποίο να καταχωρηθούν τα δεδομένα του παραπάνω πίνακα.
2. Να περιγραφούν/αναλυθούν τα παραπάνω δεδομένα εφαρμόζοντας κατάλληλες μεθόδους της Περιγραφικής Στατιστικής.

✓ Άσκηση 3

Τα παρακάτω ερωτήματα αφορούν το συνημμένο αρχείο δεδομένων με τίτλο:

ERG-STAT-Askisi_3_World.sav

1. Να κατανοηθούν τα δεδομένα του προβλήματος μέσα από τις διάφορες παραμέτρους των μεταβλητών. Ποιες μεταβλητές είναι ποιοτικές και ποιες ποσοτικές; Ποιο είναι το μέγεθος του δείγματος;
2. Πόσες είναι οι μουσουλμανικές χώρες και τι ποσοστό αποτελούν επί του συνόλου του δείγματος;
3. Να παρουσιαστεί η κατανομή των χωρών ως προς την περιοχή, στην οποία βρίσκονται και τα αντίστοιχα ποσοστά.
4. Ποια είναι η μέση αναμενόμενη διάρκεια ζωής των γυναικών και ποια η αντίστοιχη τυπική απόκλιση τους; Ποια η μικρότερη και ποια η μεγαλύτερη διάρκεια ζωής;
5. Να παρουσιαστεί διαγραμματικά η κατανομή των χωρών ως προς το θρήσκευμά τους.
6. Να παρουσιαστεί διαγραμματικά η κατανομή των χωρών σε σχέση με την αναμενόμενη διάρκεια ζωής των ανδρών.
7. Ποια είναι η μέση βρεφική θνησιμότητα και η αντίστοιχη τυπική απόκλιση;
8. Να κατασκευαστεί το ιστόγραμμα για το ποσοστό των κατοίκων που γνωρίζουν ανάγνωση. Ποιο είναι το πλάτος των διαστημάτων που έχει προσδιορίσει το πρόγραμμα; Ποιο είναι το πρώτο και ποιο το τελευταίο διάστημα; Τι συμπέρασμα προκύπτει από τη μορφή του ιστογράμματος;
9. Να υπολογιστεί η έκταση κάθε χώρας. Ποια είναι η μικρότερη, η μεγαλύτερη και η μέση έκταση;
10. Να ομαδοποιηθούν οι χώρες σε 4 κατηγορίες ως προς τον πληθυσμό τους: (i) πληθυσμός < 10, (ii) 10 <= πληθυσμός < 50, (iii) 50 <= πληθυσμός < 100, (iv) πληθυσμός >= 100 (οι αριθμοί εκφράζονται σε εκατομμύρια).
Να κατασκευαστεί ο πίνακας συχνοτήτων των χωρών ως προς τον πληθυσμό τους (σύμφωνα με τη νέα μεταβλητή).
11. Ποιο ποσοστό των χωρών παρουσιάζει βρεφική θνησιμότητα μικρότερη από 50 θανάτους ανά 1000 γεννήσεις;
12. Πάνω από πόσους θανάτους ανά 1000 γεννήσεις εμφανίζει το 10% (προσεγγιστικά) των χωρών με τη μεγαλύτερη βρεφική θνησιμότητα;
13. Να περιγραφούν οι χώρες αυτές στατιστικά ως προς τον πληθυσμό και το αντίστοιχο κλίμα τους. Να γραφεί σύντομη αναφορά.
14. Να υπολογιστεί ο μέσος πληθυσμός των χωρών της Λατινικής Αμερικής;
15. Από τις χώρες εκείνες που παρουσιάζουν αναμενόμενη διάρκεια ζωής μεγαλύτερη από 60 έτη για τις γυναίκες και τους άντρες, ποιο είναι το ποσοστό εκείνων που βρίσκονται στη Λατινική Αμερική;

16. Για τις χώρες εκείνες που βρίσκονται στην Αφρική και στην Ασία, να υπολογιστεί η τυπική απόκλιση για τη βρεφική θνησιμότητα.

✓ Άσκηση 4

1η Περίπτωση

Μια φαρμακευτική εταιρία υποστηρίζει ότι η μέση χοληστερόλη αίματος σε καρδιοπαθείς μετά από χορήγηση ενός νέου φαρμάκου της για 15 ημέρες είναι 230 mg/100ml. Ο Εθνικός Οργανισμός Φαρμάκων προκειμένου να διαπιστώσει εάν είναι αλήθεια αυτό που υποστηρίζει η εταιρία, πήρε ένα τυχαίο δείγμα από 30 καρδιοπαθείς και μετά από χορήγηση 15 ημερών του νέου φαρμάκου μέτρησαν τη χοληστερόλη του αίματος. Οι μετρήσεις αυτές παρουσιάζονται παρακάτω.

Μετρήσεις χοληστερόλης αίματος μετά από χορήγηση για 15 ημέρες του φαρμάκου:

260, 270, 260, 250, 240, 250, 230, 225, 220, 250, 230, 230, 225, 220, 210, 230, 200, 250, 230, 210, 220, 220, 230, 235, 240, 245, 230, 240, 220, 210

Με την υπόθεση ότι τα αποτελέσματα αυτά ακολουθούν την κανονική κατανομή, μπορούμε να ισχυριστούμε με κίνδυνο σφάλματος 5% ότι η φαρμακευτική εταιρία έχει δίκιο σε αυτά που υποστηρίζει για το νέο της φάρμακο;

2η Περίπτωση

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα που είχε η χορήγηση δύο φαρμάκων που χορηγήθηκαν σε καρδιοπαθείς ασθενείς (σε κάθε ασθενή χορηγήθηκε ένα από τα δύο φάρμακα) και αναφέρονται σε μετρήσεις της χοληστερόλης στο αίμα (mg/100ml) ανά ασθενή.

Φάρμακο Α: 250 259 350 325 300 275 215 290 290 258 260 270 300
Φάρμακο Β: 230 200 240 280 290 257 217 230 260 250 270

Με την υπόθεση ότι τα αποτελέσματα αυτά ακολουθούν την κανονική κατανομή, μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 5% ότι τα δύο φάρμακα είναι ισοδύναμα;

3η Περίπτωση

Στον παρακάτω πίνακα παρουσιάζονται οι μετρήσεις χοληστερόλης στο αίμα (mg/100ml) 14 ασθενών πριν και μετά από χορήγηση σε αυτούς μιας θεραπευτικής αγωγής με ένα νέο δοκιμασμένο φάρμακο.

Πριν:	280	290	300	285	270	260	290	290	280	300	250	280	290	300
Μετά:	260	280	280	260	245	250	250	280	280	280	250	270	270	280

Αν η διαφορά δ ακολουθεί την κανονική κατανομή, μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 5% ότι τα επίπεδα χοληστερόλης στο αίμα είναι ισοδύναμα πριν από τη χορήγηση του φαρμάκου και μετά από αυτήν;

Βοηθητικές οδηγίες:

ΜΕΡΙΚΟΙ ΧΡΗΣΙΜΟΙ ΤΥΠΟΙ

	Πληθυσμός	**Δείγμα**
Μέση Τιμή (mean value)	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Διακύμανση (variance)	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Τυπική Απόκλιση (standard deviation)	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
Τυπικό Σφάλμα του Μέσου (standard error of mean)	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, όταν το σ είναι άγνωστο, τότε μπορεί να προσεγγιστεί με το s	

ΕΦΑΡΜΟΓΗ 1. – ΣΤΑΤΙΣΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

$\bar{x} = 232,667$ s = 16,595 n = 30

S.E. = 3,02986

t = 0,880 df = 29 sig. = 0,386

Έλεγχος υποθέσεων για τη μέση τιμή ενός πληθυσμού

Βήμα 1ο: Προσδιορισμός υποθέσεων.

(μηδενική) $H_0 : \mu = 230$

(εναλλακτική) $H_a : \mu \neq 230$

Βήμα 2ο: Προσδιορισμός επιπέδου σημαντικότητας

a = 0,05 σφάλμα τύπου I

Βήμα 3ο: Υπολογισμός του στατιστικού της υπόθεσης

$$t = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{232,7 - 230}{16,6 / \sqrt{30}} = 0,880$$

Βήμα 4ο: Προσδιορισμός του κρίσιμου πεδίου

(Απορριπτική περιοχή για την H_0)

$|t| > t_{a/2}$ **0,880 < 1,96**

Βήμα 4ο (PSPP): Προσδιορισμός του κρίσιμου πεδίου

(Απορριπτική περιοχή για την H_0)

significance level < 0,05

significance level = 0,386 > 0,05

Βήμα 5ο: Συμπέρασμα

Αποδεχόμαστε την H_0, δηλαδή δεν έχουμε επαρκή στοιχεία για να απορρίψουμε τον ισχυρισμό της εταιρίας.

ΣΦΑΛΜΑΤΑ ΕΛΕΓΧΟΥ ΥΠΟΘΕΣΕΩΝ

	H_0 Αληθής	**H_0 Ψευδής**
Απόρριψη της H_0	**Σφάλμα τύπου I (α)**	**Σωστή Απόφαση**
Αποδοχή της H_0	**Σωστή Απόφαση**	**Σφάλμα Τύπου II (β)**

Η τιμή **significance** ή **p-value** (**sig.**) ενός ελέγχου υποθέσεων εκφράζει το μικρότερο επίπεδο σημαντικότητας (α) για το οποίο θα μπορούσαμε να απορρίψουμε την H_0.

significance level < 0,05 απόρριψη της H_0

significance level >0,05 αποδοχή της H_0

✓ Άσκηση 5

1η Περίπτωση

Σε μία μονάδα κατασκευής επίπλων έχουν καταγραφεί 309 ελαττωματικά τεμάχια και έχουν ταξινομηθεί ανάλογα με το είδος του ελαττώματος και τη βάρδια, στην οποία κατασκευάστηκαν. Τα σχετικά αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα.

	Ελάττωμα				
Βάρδια	Α	Β	Γ	Δ	Σύνολο
1η	15	21	45	13	94
2η	26	31	34	5	96
3η	33	17	49	20	119
Σύνολο:	74	69	128	38	309

Μπορούμε να ισχυριστούμε ότι το είδος του ελαττώματος είναι ανεξάρτητο από την βάρδια στην οποία έχει κατασκευαστεί το αντίστοιχο έπιπλο;

2η Περίπτωση

Στον πίνακα που ακολουθεί, παραθέτουμε τα στοιχεία που κατατάσσουν τους 9960 κατοίκους μιας πόλης, ηλικίας άνω των 30 ετών, ως προς το βαθμό έκθεσής τους σε ατμοσφαιρικούς ρύπους και ως προς τη βαρύτητα του αναπνευστικού προβλήματος που παρουσιάζουν.

Αναπνευστικά προβλήματα	Βαθμός έκθεσης σε ατμοσφαιρικούς ρύπους			
	Υψηλός	Μέτριος	Περιορισμένος	Καθόλου
Βήχας	200	180	30	15
Δύσπνοια	900	980	200	55
Άσθμα	3000	3500	700	200

Μπορούν τα στοιχεία αυτά να υποστηρίξουν την υπόθεση ότι η ένταση του αναπνευστικού προβλήματος και ο βαθμός έκθεσης σε ατμοσφαιρικούς ρύπους έχουν στατιστικά σημαντική σχέση μεταξύ τους;

✓ Άσκηση 6

Στον παρακάτω πίνακα παρουσιάζεται η μέση διάρκεια ζωής των κατοίκων που έχει καταγραφεί σε 20 χώρες δύο ηπείρων.

Χώρες με πληθυσμό μικρότερο των 50 εκατομμυρίων

Ήπειρος / Διάρκεια Ζωής	Αφρική									Νότια Αμερική		
Άνδρες	47	41	60	48	55	54	41	46	55	53	62	61
Γυναίκες	50	44	66	52	58	57	43	50	58	53	67	67

Χώρες με πληθυσμό μεγαλύτερο των 50 εκατομμυρίων

Ήπειρος / Διάρκεια Ζωής	Αφρική		Νότια Αμερική					
Άνδρες	52	44	61	50	66	59	63	71
Γυναίκες	58	45	65	52	72	64	67	78

Ζητείται:

1. Να περιγραφεί το αρχείο δεδομένων του PSPP που πρέπει να δημιουργηθεί για την εισαγωγή των παραπάνω δεδομένων.
2. Να ελεγχθεί αν διαφέρει στατιστικά σημαντικά η μέση ηλικία των ανδρών από τη μέση ηλικία των γυναικών στις χώρες αυτές.
3. Μπορούμε να ισχυριστούμε ότι η μέση ηλικία των αντρών είναι 58 έτη;
4. Να ελεγχθεί αν η μέση ηλικία των γυναικών διαφέρει στατιστικά μεταξύ χωρών από διαφορετική ήπειρο.
5. Να ελεγχθεί αν υπάρχει στατιστικά σημαντική σχέση μεταξύ του πληθυσμού μιας χώρας και της ηπείρου στην οποία βρίσκεται.

- ✓ Άσκηση 7

Να μελετηθεί το αρχείο ERG-STAT-Askisi_7_data και να διερευνηθεί με χρήση των κατάλληλων στατιστικών ελέγχων η σχέση που έχει η πιστωτική κατάσταση (credit_rating) των πελατών με τα διάφορα χαρακτηριστικά τους:

i. Ηλικία
ii. Εισόδημα
iii. Αριθμός πιστωτικών καρτών
iv. Εκπαίδευση
v. Αριθμός δανείων αυτοκινήτων

✓ Άσκηση 8

Η διεξαγωγή μίας έρευνας για τον προσδιορισμό του ρόλου της τηλεόρασης στη ζωή των συνταξιούχων έχει ως σκοπό να παρέχει οδηγίες για το σχεδιασμό και τον προγραμματισμό τηλεοπτικών προγραμμάτων που θα ικανοποιούν επαρκώς τις ανάγκες αυτού του κοινού.

Μετά από έρευνα που έγινε σε ένα δείγμα 25 κατοίκων της Θεσσαλονίκης, συνταξιούχων ηλικίας μεγαλύτερης των 60 ετών, συγκεντρώθηκαν τα παρακάτω στοιχεία:

- ο μέσος αριθμός ωρών / μέρα, που ένας συνταξιούχος παρακολουθεί τηλεόραση
- η οικογενειακή κατάσταση (1 ο συνταξιούχος ζει με την σύζυγό του, 0 αλλιώς)
- η ηλικία (σε χρόνια)
- το επίπεδο μόρφωσης (σε έτη εκπαίδευσης)

Ο σκοπός της έρευνας είναι να ελεγχθεί η σχέση μεταξύ του μέσου αριθμού ωρών που ένας ηλικιωμένος παρακολουθεί τηλεόραση και των δημογραφικών του στοιχείων.

1. Να υπολογιστεί και να ερμηνευτεί ο συντελεστής γραμμικής συσχέτισης του Pearson μεταξύ του μέσου χρόνου που παρακολουθεί τηλεόραση ένας συνταξιούχος και κάθε ενός από τα υπόλοιπα δημογραφικά του στοιχεία.
   Ποιοι από τους συντελεστές αυτούς κρίνονται ως στατιστικά σημαντικοί;

2. Να εκτιμηθεί η εξίσωση της γραμμής (πολλαπλής) παλινδρόμησης, η οποία περιγράφει τη σχέση που έχει ο μέσος χρόνος που ένας συνταξιούχος παρακολουθεί τηλεόραση με όλα τα υπόλοιπα δημογραφικά στοιχεία.

3. Ποιος είναι ο αναμενόμενος μέσος αριθμός ωρών που παρακολουθεί ημερησίως τηλεόραση ένας ηλικιωμένος 70 ετών, με 12 έτη εκπαίδευσης, ο οποίος δεν ζει με τη σύζυγό του;
   [Υπόδειξη: Να χρησιμοποιηθεί η εξίσωση της γραμμής παλινδρόμησης που προσδιορίσατε στο προηγούμενο ερώτημα]

4. Να ερμηνευτούν οι συντελεστές παλινδρόμησης.
   Ποιες μεταβλητές συνεισφέρουν στατιστικά σημαντικά στο μοντέλο αυτό ($\alpha = 0.05$);

5. Με βάση την γραμμή παλινδρόμησης του ζητήματος 2, να εκτιμηθεί ο μέσος αριθμός ωρών/μέρα που παρακολουθεί τηλεόραση κάθε ένας από τους συνταξιούχους που πήραν μέρος στην έρευνα.
   Να υπολογιστεί ο μέσος όρος αυτών των χρόνων, όπως επίσης και οι μέσοι όροι των υπόλοιπων μεταβλητών που σχετίζονται με τα δημογραφικά στοιχεία.
   Να επιβεβαιωθεί ότι αυτοί οι μέσοι όροι ικανοποιούν την εξίσωση της γραμμής παλινδρόμησης (δηλαδή ότι, $\bar{y} = \hat{a} + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_3 \bar{x}_3$). Η αντικατάσταση στον τύπο και οι πράξεις να πραγματοποιηθούν διατηρώντας 3 δεκαδικά ψηφία.

6. Να αξιολογηθεί η γραμμή παλινδρόμησης και να προταθεί εναλλακτικά ένα "ανταγωνιστικό" μοντέλο.

ΔΕΔΟΜΕΝΑ

ΑΤΟΜΟ	ΩΡΕΣ	ΟΙΚ.ΚΑΤ.	ΗΛΙΚΙΑ	ΜΟΡΦΩΣΗ
1	0.5	1	73	14
2	0.5	1	66	16
3	0.7	0	65	15
4	0.8	0	65	16
5	0.8	1	68	9
6	0.9	1	69	10
7	1.1	1	82	12
8	1.6	1	83	12
9	1.6	1	81	12
10	2.0	0	72	10
11	2.5	1	69	8
12	2.8	0	71	16
13	2.8	0	71	12
14	3.0	0	80	9
15	3.0	0	73	6
16	3.0	0	75	6
17	3.2	0	76	10
18	3.2	0	78	6

19	3.3	1	79	6
20	3.3	0	79	4
21	3.4	1	78	6
22	3.5	0	76	9
23	3.6	0	65	12
24	3.7	0	72	12
25	3.7	0	80	6

✓ Άσκηση 9

Το αρχείο δεδομένων ERG-STAT-Askisi_9_data περιέχει δεδομένα σε κωδικοποιημένη μορφή σχετικά με μια έρευνα που έγινε για μπύρες στις ΗΠΑ. Συγκεκριμένα, και με τη σειρά των στηλών που καταλαμβάνουν στο αρχείο, οι μεταβλητές για τις οποίες υπάρχουν δεδομένα είναι :

1	RATING	Ποιότητα 1=Πολύ καλή, 2=Καλή, 3=Μέτρια
2	ORIGIN	Προέλευση 1=ΗΠΑ, 2=Καναδάς, 3=Γαλλία, 4=Ολλανδία, 5=Μεξικό, 6=Γερμανία, 7=Ιαπωνία
3	AVAIL	Διάθεση 1=Εθνικό επίπεδο, 2=Τοπικό επίπεδο
4	PRICE	Τιμή πώλησης
5	COST	Τιμή κόστους
6	CALORIES	Θερμίδες
7	SODIUM	Σόδα
8	ALCOHOL	Οινόπνευμα
9	CLASS	Κατηγορία τιμής 0=Δεν εκτιμήθηκε, 1=Πολύ υψηλή, 2=Υψηλή, 3=Μέτρια
10	LIGHT	Ελαφριά ή όχι 1=Ελαφριά, 0=Όχι ελαφριά

1. Πόσες είναι οι πολύ καλές μπύρες και τι ποσοστό αποτελούν επί του συνόλου των μπυρών;
2. Ποια είναι η μέση τιμή κόστους των ελαφριών μπυρών;
   Ποια είναι η αντίστοιχη τυπική απόκλιση;
3. Να υπολογιστεί το κέρδος για κάθε μία μπύρα.
   Ποιο είναι το μέσο κέρδος και η τυπική απόκλιση των Γερμανικών μπυρών;
4. Να ομαδοποιηθεί η ποσότητα των θερμίδων σε δύο κατηγορίες:
   i) θερμίδες λιγότερες από 138 ii) θερμίδες περισσότερες ή ίσες με 138.

Να ελεγχθεί αν η ποιότητα είναι ανεξάρτητη της ποσότητας θερμίδων, σύμφωνα με τη νέα μεταβλητή για τις θερμίδες.

5. Να ελεγχθεί για τις μπύρες από τις ΗΠΑ, αν η μέση ποσότητα οινοπνεύματος διαφέρει στατιστικά σημαντικά μεταξύ των «πολύ καλών» και των «μέτριων» μπυρών;

6. Μπορούμε να ισχυριστούμε ότι η μέση περιεκτικότητα σε σόδα των μπυρών που διατίθενται σε εθνικό επίπεδο δεν διαφέρει στατιστικά σημαντικά από τις 12 μονάδες;

Παράρτημα Α: Συνοπτικά Αποτελέσματα Ασκήσεων-Υποδείξεις

✓ Άσκηση 1:

- Ως *δείγμα* της έρευνας θεωρούμε τους φοιτητές ενός τμήματος οι οποίοι απαρτίζουν το συγκεκριμένο εργαστήριο. Συγκεκριμένα ο αριθμός των ερωτηθέντων είναι 20. Ο *συνολικός πληθυσμός* της μελέτης είναι όλοι οι μαθητές, από όλα τα εργαστήρια που παρακολουθούν το συγκεκριμένο μάθημα.
- Οι *ποσοτικές* μεταβλητές είναι: η ηλικία του φοιτητή (συνεχής), οι φορές παρακολούθησης του εργαστηρίου (διακριτή) και πόσο ικανοποιητική είναι η καθαριότητα του εργαστηρίου

  Οι *ποιοτικές* μεταβλητές είναι: η μεταβλητή που εκφράζει το κατά πόσο η ύλη που διδάσκεται είναι καλά οργανωμένη, η μεταβλητή που αναφέρεται στον στόχο επίτευξης του εκάστοτε φοιτητή με το πέρας του εργαστηρίου και τα στοιχεία που απαιτούν βελτίωση σε ότι αφορά τον διδάσκοντα.

- Η μεταβλητή *ηλικία* παίρνει τιμές από 18 μέχρι 24.
  Η μεταβλητή *φορές παρακολούθησης* του εργαστηρίου παίρνει τιμές από 0 μέχρι 5.
  Η μεταβλητή *οργάνωση της ύλης* χωρίζεται στις εξής κατηγορίες :
  1 όταν παίρνει τον χαρακτηρισμό Καθόλου
  2 όταν παίρνει τον χαρακτηρισμό Λίγο
  3 όταν παίρνει τον χαρακτηρισμό Μέτρια
  4 όταν παίρνει τον χαρακτηρισμό Πολύ
  5 όταν παίρνει τον χαρακτηρισμό Πάρα πολύ

  Η μεταβλητή *στόχος* χωρίζεται στις εξής κατηγορίες :
  1 όταν κάποιος μαθητής δήλωσε πως απλώς θέλει να περάσει το μάθημα.
  2 όταν κάποιος μαθητής δήλωσε πως θέλει να περάσει το μάθημα με βαθμό
  3 κάποιος μαθητής δήλωσε πως θέλει να περάσει το μάθημα και να αποκτήσει γνώσεις.

  Η μεταβλητή *καθαριότητα* έχει κλίμακα από 0(καθόλου) μέχρι 100(πάρα πολύ)

Η μεταβλητή *μεταδοτικότητα* είναι δίτιμη. Παίρνει την τιμή 1 όταν η απάντηση του φοιτητή ήταν πως χρειάζεται να βελτιωθεί και την τιμή 0 στην αντίθετη περίπτωση.

Η μεταβλητή *προετοιμασία* είναι δίτιμη. Παίρνει την τιμή 1 όταν η απάντηση του φοιτητή ήταν πως χρειάζεται να βελτιωθεί και την τιμή 0 στην αντίθετη περίπτωση.

Η μεταβλητή *Φιλικότητα* είναι δίτιμη. Παίρνει την τιμή 1 όταν η απάντηση του φοιτητή ήταν πως χρειάζεται να βελτιωθεί και την τιμή 0 στην αντίθετη περίπτωση.

Η μεταβλητή *Κανένα* είναι δίτιμη. Παίρνει την τιμή 1 όταν η απάντηση του φοιτητή ήταν πως χρειάζεται να βελτιωθεί και την τιμή 0 στην αντίθετη περίπτωση.

- Μέσω των διαδικασιών frequencies και descriptives για τις ποσοτικές και ποιοτικές μεταβλητές παίρνουμε την περιγραφική ανάλυση και τα διαγράμματα που επιθυμούμε.

Με την χρήση της διαδικασίας *descriptives* για τις ποσοτικές μεταβλητές συμπεραίνουμε: Όλοι οι φοιτητές που ρωτήθηκαν (20) είναι ηλικίας από 18 μέχρι 24 ετών με μέσο όρο 19,66 και διακύμανση 2,4 που σημαίνει ότι δεν έχουμε έκτροπες παρατηρήσεις. Δήλωσαν πως παρακολούθησαν κατά μέσο όρο το μάθημα 1 φορά. Η άποψη τους για την καθαριότητα του εργαστηρίου είναι ικανοποιητική εφόσον κατά μέσω όρο βαθμολόγησαν με 68,16 που είναι παραπάνω από το 50 που ορίζεται σαν βάση της κλίμακας.

Με την χρήση της διαδικασίας *frequencies* για τις ποιοτικές μεταβλητές συμπεραίνουμε:
α) οργάνωση ύλης
Παρατηρήθηκε πως για την οργάνωση της ύλης απάντησαν και οι 20 φοιτητές. Το 30% (6 φοιτητές) απάντησε πως η ύλη είναι λίγο οργανωμένη, το 40% (8 φοιτητές) πως είναι μέτρια οργανωμένη, το 25%(5 φοιτητές) πως είναι καλά οργανωμένη και μόλις το 5%(1 φοιτητής) πως είναι πάρα πολύ καλά οργανωμένη. Στην τελευταία στήλη υπολογίζεται η σχετική αθροιστική συχνότητα επί τις 100. Παρατηρείται πως οι περισσότεροι από τους ερωτηθέντες θεωρούν πως είναι μέτρια οργανωμένη, επομένως μια καλύτερη οργάνωση του μαθήματος θα βελτιώσει την διεξαγωγή του μαθήματος.

β) στόχος φοιτητή

Παρατηρήθηκε πως για τον στόχο που έθεσαν οι φοιτητές με το πέρας του εργαστηρίου απάντησαν και οι 18 φοιτητές. Το 55,56% (10 φοιτητές) απάντησε πως απλά θέλει να περάσει το μάθημα, το 22,22% (4 φοιτητές) πως θέλει να το περάσει με βαθμό, το 22,22% (4 φοιτητές) πως θέλει να το περάσει και να αποκτήσει γνώσεις. Οι υπόλοιποι (2 φοιτητές) δεν απάντησαν στην ερώτηση. Το μεγαλύτερο ποσοστό των μαθητών επιθυμεί απλά να περάσει το μάθημα.

γ) μεταδοτικότητα καθηγητών

Το 62,50% των φοιτητών (10 άτομα) απάντησε πως θα έπρεπε να βελτιωθεί η μεταδοτικότητα των καθηγητών. Το 37,50 % είπε πως είναι ικανοποιημένο με την μεταδοτικότητα των καθηγητών.

δ) προετοιμασία καθηγητών

Το 87,50% των φοιτητών (14 άτομα) απάντησε πως είναι ικανοποιημένοι από την προετοιμασία των καθηγητών. Το 12,50% (2 άτομα) είπε πως δεν είναι ικανοποιητική η προετοιμασία των καθηγητών

ε) φιλικότητα καθηγητών

Το 68,75% των φοιτητών (10 άτομα) απάντησε πως είναι ικανοποιημένοι από την φιλική προσέγγιση των καθηγητών. Το 25% (5 άτομα) είπε πως η φιλικότητα των καθηγητών ενδέχεται βελτίωση.

δ) κανένα

Μόλις το 37,50% των φοιτητών (6 άτομα) απάντησε πως είναι ικανοποιημένοι από την διεξαγωγή του μαθήματος

✓ Άσκηση 2

α) Οι ποσοτικές μεταβλητές είναι: ο αριθμός bytes από πηγή σε προορισμό (scr_bytes), ο αριθμός bytes από προορισμό σε πηγή (dst_bytes) και ο αριθμός αποτυχημένων εισόδων (num_failed_logins).
Οι ποιοτικές μεταβλητές είναι: ο τύπος πρωτοκόλλου (protocol_type) με τιμές 1=’icm’, 2=’tcp’, 3=’udp’ και ο τύπος επίθεσης (attack_type) με τιμές 1=’back’, 2=’ipsweep’, 3=’neptune’, 4=’normal’, 5=’smurf’ .

Αντιστοίχως γίνεται και η εισαγωγή των τιμών από τον πίνακα δεδομένων στο αρχείο PSPP.

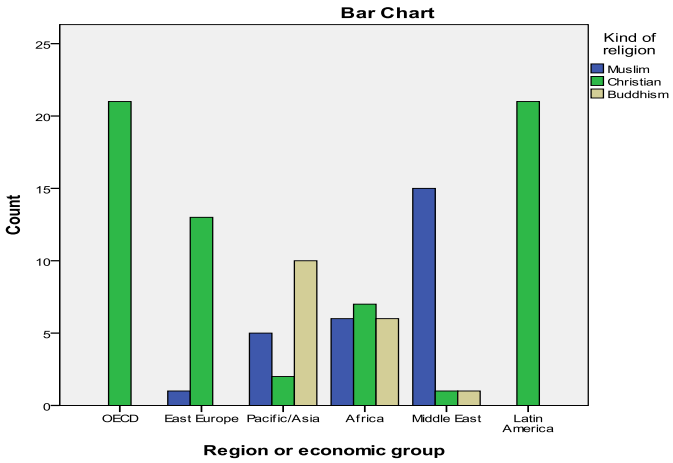
β) Μέσω της διαδικασίας Descriptives για τις ποσοτικές μεταβλητές παίρνουμε τα αντίστοιχα στατιστικά αποτελέσματα όπως το πλήθος των περιπτώσεων(25), την διακύμανση, την μέγιστη και ελάχιστη τιμή καθώς και το μέσο όρο της κάθε μεταβλητής.

Ανάλογα, μέσω της διαδικασίας Frequencies για τις ποιοτικές μεταβλητές έχουμε τα εξής αποτελέσματα:

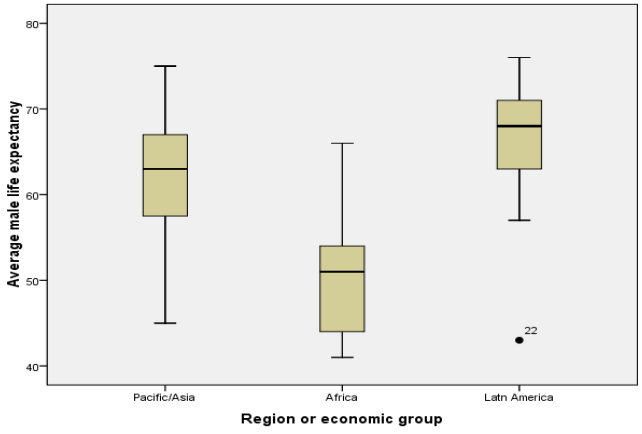
Ο τύπος πρωτοκόλλου με την μεγαλύτερη συχνότητα είναι το tcp με 19 (76,00%) εμφανίσεις, ακολουθεί το icm με 4 (16,00%) και τέλος το udp με 2 (8,00%).
Επίσης οι αντίστοιχες συχνότητες των τύπων επιθέσεων είναι η normal με 18 (72,00 %) εμφανίσεις, η smurf με 3 (12,00%), η neptune με 2 (8,00%) και ακολουθούν οι ipsweep και back με 1 (4,00%).

✓ Άσκηση 3

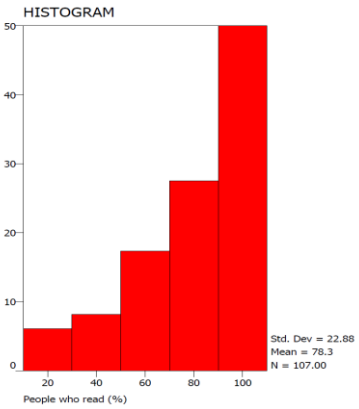
1. Οι ποιοτικές μεταβλητές είναι: η region, η climate και η religion. Όλες οι άλλες μεταβλητές είναι ποσοτικές.
   Το δείγμα αποτελείται από 109 παρατηρήσεις.
2. Οι μουσουλμανικές χώρες είναι 27 και αποτελούν το 24,77% επί του συνόλου του δείγματος.
3. Τα ποσοστά της κατανομής των χωρών είναι: OECD με 19.27%, East Europe με 12.84%, Pacific / Asia με 15.60%, Africa με 17.43%, Middle East με 15.60% και Latin America με 19.27%.
4. Η μέση αναμενόμενη διάρκεια ζωής των γυναικών είναι: 70.16 έτη
   Η τυπική απόκλιση είναι: 10.57
   Η μικρότερη διάρκεια ζωής είναι: 43 έτη ενώ η μεγαλύτερη είναι: 82 έτη
5. Διάγραμμα κατανομής χωρών ως προς το θρήσκευμα τους:



6. Διάγραμμα κατανομής χωρών ως προς την αναμενόμενη διάρκεια ζωής των ανδρών:



7. Η μέση βρεφική θνησιμότητα είναι: 42.31 και η αντίστοιχη τυπική απόκλιση είναι: 38.08
8. Ιστόγραμμα για το ποσοστό των κατοίκων που γνωρίζουν ανάγνωση:



Το πλάτος των διαστημάτων του προγράμματος είναι: 10

Το πρώτο διάστημα είναι από 15 έως 25 και το τελευταίο από 95 έως 105

Συμπεραίνουμε ότι το δείγμα δεν ακολουθεί κανονική κατανομή.

9. Μικρότερη έκταση έχει η χώρα με 423.14

   Μεγαλύτερη έκταση η χώρα με 16954545

   Η μέση έκταση των χωρών είναι 1008657
10. Συχνότητα εμφάνισης των χωρών ανά κατηγορία:
    i) πληθυσμός<10 έχουν 49 χώρες,
    ii) 10<=πληθυσμός<50 έχουν 37 χώρες,
    iii) 50<=πληθυσμός<100 έχουν 14 χώρες,
    iv) πληθυσμός>=100 έχουν 9 χώρες.
11. Το ποσοστό των χωρών που παρουσιάζει παιδική βρεφική θνησιμότητα μικρότερη από 50 θανάτους ανά 1000 γεννήσεις είναι το 65,14%.
12. Ο αριθμός των θανάτων ανά 1000 γεννήσεις είναι 110
13. Αναφορά αποτελεσμάτων
14. Ο μέσος πληθυσμός των χωρών της Λ. Αμερικής είναι : 21928.86 κάτοικοι
15. Το ποσοστό ανδρών και γυναικών με διάρκεια ζωής >60 έτη της Λ.Αμερικης είναι 21.69%
16. Η τυπική απόκλιση για την βρεφική θνησιμότητα στην Αφρική και την Ασία είναι: 42.71 θάνατοι ανά 1000 γεννήσεις

✓ Άσκηση 4

1η Περίπτωση:

H_0: μ=230

H_1: μ≠230

Από τα αποτελέσματα του One Sample T-Test συμπεραίνουμε: εφόσον το sig=0,39 το οποίο είναι μεγαλύτερο από 5% που έχουμε ορίσει ως επίπεδο σημαντικότητας τότε δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Επομένως έχει δίκιο η εταιρεία στον ισχυρισμό της.

Μέσο επίπεδο χοληστερόλης = 232,67 / Τυπική απόκλιση = 16,6

2η Περίπτωση:

H_0: τα δυο φάρμακα είναι ισοδύναμα

H_1: τα δυο φάρμακα δεν είναι ισοδύναμα

Από τα αποτελέσματα του Independent Samples T-Test συμπεραίνουμε: εφόσον το sig=0,53 στο τεστ του Levene (equal variances) το οποίο είναι μεγαλύτερο από 5% που έχουμε ορίσει ως επίπεδο σημαντικότητας μπορούμε να πούμε ότι τα δύο δείγματα έχουν ίδιες διακυμάνσεις.

Εν συνεχεία παρατηρήθηκε ότι το sig του ελέγχου είναι 0,02 (2-tailed). Η τιμή είναι μικρότερη από το 5% που έχει οριστεί ως επίπεδο σημαντικότητας επομένως απορρίπτουμε την μηδενική υπόθεση ότι τα δύο φάρμακα είναι ισοδύναμα. (t = 2.51, df = 22, sig. = 0.02)

Το φάρμακο Β φαίνεται να είναι πιο αποτελεσματικό από το φάρμακο Α

3η Περίπτωση:

H_0: η διαφορά των δύο δειγμάτων είναι μηδέν

H_1: υπάρχει διαφορά μεταξύ των δύο δειγμάτων

Από τα αποτελέσματα του Paired Samples T-Test συμπεραίνουμε: εφόσον το sig=0 είναι μικρότερο από 5% που έχουμε ορίσει ως επίπεδο σημαντικότητας, απορρίπτουμε την μηδενική υπόθεση. Επομένως, υπάρχει στατιστικά σημαντική διαφορά στο επίπεδο της χοληστερόλης προ της χορήγησης του φαρμάκου και μετά τη χορήγηση του φαρμάκου.

(t = 5.78, df = 13, sig. = 0.00)

Πριν την χορήγηση του φαρμάκου το επίπεδο ήταν υψηλότερο (283.21 ± 14.89) από το αντίστοιχο επίπεδο χοληστερόλης μετά τη χορήγηση του φαρμάκου (266.79 ± 13.81)

✓ Άσκηση 5

1η Περίπτωση:

H_0: ελάττωμα ανεξάρτητο από τη βάρδια

H_1: ελάττωμα δεν είναι ανεξάρτητο από τη βάρδια

Από τα αποτελέσματα συμπεραίνουμε: εφόσον το sig=0.00 είναι μικρότερο από 5% που έχουμε ορίσει ως επίπεδο σημαντικότητας, απορρίπτουμε την μηδενική υπόθεση.

Άρα, υπάρχει στατιστικά σημαντική σχέση μεταξύ βάρδιας και είδους ελαττώματος

(x^2 = 19.18, df = 6, sig. = 0.00)

2η Περίπτωση:

H_0: η ένταση του αναπνευστικού προβλήματος και ο βαθμός έκθεσης σε ατμοσφαιρικούς ρύπους έχουν στατιστικά σημαντική σχέση μεταξύ τους

H_1: η ένταση του αναπνευστικού προβλήματος και ο βαθμός έκθεσης σε ατμοσφαιρικούς ρύπους δεν έχουν στατιστικά σημαντική σχέση μεταξύ τους

Από τα αποτελέσματα του Test Chi-square συμπεραίνουμε: εφόσον το sig=0.09 είναι μεγαλύτερο από 0,05 (τιμή p) τότε δεν υπάρχει στατιστικά σημαντική σχέση μεταξύ αναπνευστικού προβλήματος και βαθμού έκθεσης σε ατμοσφαιρικούς ρύπους

(x^2 = 11.3, df = 6, sig. = 0.09)

✓ Άσκηση 6

1. Έχουμε ένα σύνολο δεδομένων με τέσσερα χαρακτηριστικά που είναι και οι μεταβλητές που θα κωδικοποιήσουμε: η μέση διάρκεια ζωής, η Ήπειρος διαμονής, το φύλο του ατόμου και αν ο πληθυσμός της χώρας υπερβαίνει τα 50 εκατομμύρια ή όχι. Έχουμε τις εξής μεταβλητές:
   i. Η μέση διάρκεια ζωής: είναι ποσοτική (αριθμητική) μεταβλητή
   ii. Η Ήπειρος διαμονής: είναι μια ποιοτική μεταβλητή που παίρνει τιμές 1 για την Αφρική και 2 για τη Νότιο Αμερική
   iii. Το φύλο του ατόμου: είναι μια ποιοτική μεταβλητή που παίρνει τιμές 1 για τους άνδρες και 2 για τις γυναίκες
   iv. Πληθυσμός άνω των 50 εκατομμυρίων: είναι μια ποιοτική μεταβλητή που παίρνει τιμές 1 για τις χώρες με άνω των 50 εκατομμυρίων κατοίκων και 2 για τις χώρες που έχουν λιγότερους κατοίκους από 50 εκατομμύρια
2. H_0: διαφέρει στατιστικά σημαντικά η μέση ηλικία των ανδρών από τη μέση ηλικία των γυναικών

   H_1: δεν διαφέρει στατιστικά σημαντικά η μέση ηλικία των ανδρών από τη μέση ηλικία των γυναικών

   Από το Independent Samples T-Test συμπεραίνουμε: εφόσον το sig=0.00 είναι μικρότερο από 5% που έχουμε ορίσει ως επίπεδο σημαντικότητας, απορρίπτουμε την μηδενική υπόθεση. Άρα, δεν διαφέρει στατιστικά σημαντικά η μέση ηλικία των ανδρών από τη μέση ηλικία των γυναικών. (t = -9.34, df = 19, sig. = 0.00)
3. H_0: η μέση ηλικία των αντρών είναι 58 έτη

   H_1: η μέση ηλικία των αντρών δεν είναι 58 έτη

   Από το One Sample T-Test συμπεραίνουμε: εφόσον το sig=0.07 είναι μεγαλύτερο από 5% που έχουμε ορίσει ως επίπεδο σημαντικότητας, αποδεχόμαστε την μηδενική υπόθεση. Άρα, η τιμή των 58 ετών μπορεί να θεωρηθεί ως έγκυρη μέση ηλικία.

   (t = -1.89, df = 19, sig. = 0.07)
4. H_0: η μέση ηλικία των γυναικών διαφέρει στατιστικά μεταξύ χωρών από διαφορετική ήπειρο.

   H_1: η μέση ηλικία των γυναικών δεν διαφέρει στατιστικά μεταξύ χωρών από διαφορετική ήπειρο.

Από τα αποτελέσματα του Independent Samples T-Test συμπεραίνουμε: εφόσον το sig=0,92 στο τεστ του Levene (equal variances) το οποίο είναι μεγαλύτερο από 5% που έχουμε ορίσει ως επίπεδο σημαντικότητας μπορούμε να πούμε ότι τα δύο δείγματα έχουν ίδιες διακυμάνσεις.

Εν συνεχεία παρατηρήθηκε ότι το sig του ελέγχου είναι 0,00 (2-tailed). Η τιμή είναι μικρότερη από το 5% που έχει οριστεί ως επίπεδο σημαντικότητας επομένως απορρίπτουμε την μηδενική υπόθεση. (t = -3.52, df = 18, sig. = 0.00)

Άρα, υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μέσων τιμών των ηλικιών των γυναικών των δύο ηπείρων.

Οι γυναίκες της Λατινικής Αμερικής έχουν μέση ηλικία 65 ± 8.25, ενώ οι γυναίκες της Αφρικής 58.82 ± 7.24.

5. H_0: υπάρχει στατιστικά σημαντική σχέση μεταξύ του πληθυσμού μιας χώρας και της ηπείρου στην οποία βρίσκεται

   H_1: δεν υπάρχει στατιστικά σημαντική σχέση μεταξύ του πληθυσμού μιας χώρας και της ηπείρου στην οποία βρίσκεται

   Από τα αποτελέσματα του Test Chi-square συμπεραίνουμε: εφόσον το sig=0.53 (και sig= 0.06) είναι μεγαλύτερο από 0,05 (τιμή p) τότε δεν υπάρχει στατιστικά σημαντική σχέση μεταξύ του πληθυσμού μιας χώρας και της ηπείρου στην οποία βρίσκεται

✓ <u>Άσκηση 7</u>

Με την διαδικασία Crosstabs για την μελέτη της πιστωτικής κατάστασης σε σχέση με τα διάφορα χαρακτηριστικά των πελατών συμπεραίνουμε ότι:

- Σε σχέση με το εισόδημα (ποιοτική): Υπάρχει στατιστικά σημαντική σχέση καθώς το sig.=0.00<0.05 (x^2 = 662.46, df = 2, sig. = 0.00)
- Σε σχέση με τον αριθμό των πιστωτικών καρτών (ποιοτική): Υπάρχει στατιστικά σημαντική σχέση καθώς το sig.=0.00<0.05 (x^2 = 415.97, df = 1, sig. = 0.00)
- Σε σχέση με τον αριθμό των δανείων αυτοκινήτων (ποιοτική): Υπάρχει στατιστικά σημαντική σχέση καθώς το sig.=0.00<0.05 (x^2 = 265.96, df = 1, sig. = 0.00)
- Σε σχέση με την εκπαίδευση (ποιοτική): Δεν υπάρχει στατιστικά σημαντική σχέση καθώς το sig.=0.75>0.05 (x^2 = 0.10, df = 1, sig. = 0.75).
- Σε σχέση με την ηλικία (ποσοτική): Υπάρχει στατιστικά σημαντική διαφορά μεταξύ των διακυμάνσεων (F = 35.13, sig. = 0.00<0.05) καθώς και μεταξύ των μέσων τιμών (t = -20.16, df = 2379.64, sig. = 0.00<0.05)

✓ <u>Άσκηση 8</u>

α) Υπολογίζουμε το συντελεστή γραμμικής συσχέτισης του Pearson μέσω της διαδικασίας Bivariate Correlations και συμπεραίνουμε ότι:

- Οι ώρες παρακολούθησης με την οικογενειακή κατάσταση έχουν μέτρια έως ικανοποιητική αρνητική συσχέτιση (r = -0.52, sig.= 0.01)
- Οι ώρες παρακολούθησης με την ηλικία δεν έχουν σημαντική συσχέτιση (r = 0.36, sig. = 0.08)
- Οι ώρες παρακολούθησης με το επίπεδο μόρφωσης έχουν μέτρια έως ικανοποιητική αρνητική συσχέτιση (r = -0.61, sig. = 0.00)

β) Η εξίσωση της γραμμής (πολλαπλής) παλινδρόμησης είναι:
ώρες = 1,50 – 1,18*(οικογ. κατάσταση) + 0,04*(ηλικία) – 0,15*(μόρφωση)

γ) Ο αναμενόμενος μέσος αριθμός ωρών είναι: 2,5 ώρες/ ημέρα

δ) Οι μεταβλητές οι οποίες συνεισφέρουν σημαντικά είναι αυτές όσων οι συντελεστές παλινδρόμησης έχουν sig. < 0,05. Επομένως συμπεραίνουμε:

- Οικογενειακή κατάσταση: t = -3,73 sig = 0,00 < 0,05 → συνεισφέρει στατιστικά σημαντικά
- Ηλικία: t = 1,21 sig = 0,24 > 0,05 → δεν συνεισφέρει στατιστικά σημαντικά
- Μόρφωση: t = -3,04 sig = 0,01 < 0,05 → συνεισφέρει στατιστικά σημαντικά

ε) Με χρήση της εντολής Compute δημιουργούμε μία νέα μεταβλητή και υπολογίζουμε ότι ο μέσος όρος είναι: 2,46 ώρες / μέρα

στ) Ερμηνεία του R-square (R2 = 0.63): Εφόσον το μοντέλο εξηγεί το 63% της μεταβλητότητας θεωρείται πολύ καλό και αξιόπιστο.
Εξέταση διαφορετικών συνδυασμών ανεξάρτητων μεταβλητών
Εναλλακτικό απλούστερο μοντέλο:
ώρες = 4,65 – 1,1*(οικογ. κατάσταση) – 0.18*(μόρφωση) με R2 = 0.60

✓ <u>Άσκηση 9</u>

1. Στατιστικά σημαντικές συσχετίσεις υπάρχουν μεταξύ των μεταβλητών:
    - head_ic και chest_decel
    - chest_decel και r_leg
    - l_leg και r_leg

2. Η γραμμή παλινδρόμησης προσδιορίζεται από την εξίσωση:
   head_ic = 421.04 + 0.17 * weight
   Το εκτιμώμενο μοντέλο αξιολογείται στατιστικά σημαντικό (ANOVA: sig.<0.05)
   Ο συντελεστής παλινδρόμησης του βάρους του αυτοκινήτου ερμηνεύεται ως:
   R-square = 0.05 → ιδιαίτερα χαμηλό
3. Με βάση τα αποτελέσματα των tests για τις μεταβλητότητες (F=5.98, sig.=0.01) και για τις μέσες τιμές (t=6.85, df=334.84, sig.=0.00) συμπεραίνουμε ότι ο βαθμός τραυματισμού στο στήθος διαφέρει στατιστικά σημαντικά μεταξύ του οδηγού και του συνοδηγού.
   Συγκεκριμένα, για τον οδηγό: 51,65 ± 9,60 για τον συνοδηγό: 44,99 ± 8,33
4. Με την χρήση των εντολών Compute, Select και Frequencies καταλήγουμε ότι η μέση επιβάρυνση που δέχεται ο οδηγός στο κεφάλι είναι 866.41 και η αντίστοιχη τυπική απόκλιση είναι 343.73
5. Με βάση το Chi-square test συμπεραίνουμε ότι δεν υπάρχει στατιστικά σημαντική σχέση με το εάν είναι δίθυρο ή τετράθυρο καθώς sig.= 0.15>0.05 →απόρριψη μηδενικής υπόθεσης (sig.= 0.15, x^2=6.73, df=4, sig.= 0.15)
6. Υπάρχει στατιστικά σημαντική διαφορά καθώς sig. = 0.00<0.05 →αποδοχή μηδενικής υπόθεσης (t=9.02, df=333, sig. = 0.00)
7. Με χρήση της εντολής Recode δημιουργούμε μια νέα μεταβλητή και υπολογίζουμε ότι το ποσοστό των βαρέων (heavy) αυτοκινήτων είναι 22.73%
8. Με χρήση της εντολής Compute δημιουργούμε έναν γενικό δείκτη τραυματισμού και υπολογίζουμε ότι η μέση τιμή του είναι 683.95 και η τυπική απόκλιση 226.58