Σημειώσεις Στατιστικής

+ εφαρμογή με το **LibreOffice Calc**  και το **R – Project**  **!**

Επαμεινώνδας Διαμαντόπουλος

Νοέμβριος 2012, Ξάνθη.

Επικοινωνία : epdiamantopoulos@yahoo.gr

Ιστοσελίδα : http://users.sch.gr/epdiaman/

Κατάλογος περιεχομένων

Στην οικογένειά μου!

Εισαγωγή

Το βιβλίο αυτό απευθύνεται σε όλους όσους θέλουν να κάνουν μία στατιστική έρευνα. Στόχος του βιβλίου είναι η περιγραφή των σταδίων μίας στατιστικής έρευνας από τη δειγματοληψία έως και τη καταγραφή των αποτελεσμάτων της στατιστικής δοκιμασίας. Το βιβλίο περιέχει ένα μέρος του βιβλίου "Στατιστική με το OpenOffice" που δημοσιεύθηκε το 2008 με άδεια χρήσης GNU GPL, και ένα μέρος από σημειώσεις που είχαν γραφεί για την υποστήριξη του μαθήματος της Στατιστικής στο ΙΕΚ Ξάνθης (διαθέσιμες εδώ). Η περαιτέρω πρωτοτυπία του παρόντος βιβλίου είναι πως για κάθε ένα στατιστικό μέτρο που περιγράφεται ή για κάθε στατιστική μέθοδο που προτείνεται υπάρχει περιγραφή του τρόπου με τον οποίο μπορεί να υπολογίσει το μέτρο ή να υλοποιήσει τη διαδικασία με το λογιστικό φύλλο Calc (OpenOffice.org ή LibreOffice) αλλά και με το R – Project.

Το LibreOffice είναι η πιο σημαντική εξέλιξη του OpenOffice.org. Υποστηρίζεται από το Document Foundation το οποίο δημιουργήθηκε από ένα μέρος της κοινότητας του OpenOffice.org το Σεπτέμβριο του 2010, ύστερα από την εξαγορά της Sun Microsystems από την Oracle (Φεβρουάριος 2010), γεγονός το οποίο αντιμετωπίστηκε αρνητικά από την κοινότητα, λόγω της ασαφούς θέσης της Oracle σχετικά με το μέλλον αυτής της σουίτας γραφείου. Στις 15 Απριλίου 2011 η Oracle ανακοίνωσε πως δεν θα υποστηρίζει πλέον συστηματικά την ανάπτυξη του OpenOffice.org κάτι που τοποθέτησε το Document Foundation στην πρώτη θέση όσον αφορά την ανάπτυξη αυτής της σουίτας γραφείου.Το γεγονός αυτό αναγνωρίστηκε ευρύτερα από την κοινότητα του ελεύθερου λογισμικού καθώς το LibreOffice αποτελεί την προεπιλεγμένη σουίτα γραφείου στις περισσότερες νέες εκδόσεις λειτουργικών Linux (Ubuntu, Mint κλπ). Στις 1 Ιουνίου 2011 η Oracle ανακοίνωσε πως δωρίζει τον κώδικα του OpenOffice.org στο The Apache Software Foundation's Incubator. Το Apache Software Foundation (ASF) είναι μία ιδιαίτερα σημαντική οργάνωση στον κόσμο του ελεύθερου λογισμικού η οποία αναπτύσσει και υποστηρίζει πολλά έργα ανοικτού λογισμικού με γνωστότερο το Apache Server. Από τη στιγμή που ο κώδικας του OpenOffice.org παραχωρήθηκε στο ASF, έγινε προσπάθεια να διαμορφωθεί εκ νέου η κοινότητα του (πλέον Apache OpenOffice.org). Αποτέλεσμα είναι η εμφάνιση μίας ακόμα έκδοσης του OpenOffice.org (έκδοση 3.4, Μάιος 2012) ωστόσο το LibreOffice φαίνεται να έχει την δυναμική και να είναι η κύρια εξέλιξη του OpenOffice.org συγκεντρώνοντας τους περισσότερους και σημαντικότερους υποστηριχτές, γεγονός που εύκολα δικαιολογείται

από την εξέλιξή του. Στη χώρα μας το μεγαλύτερο μέρος (ή ίσως και το σύνολο) της κοινότητας υποστηρικτών του OpenOffice.org ακολούθησε το δρόμο του LibreOffice. Σήμερα, (Νοέμβριος 2012), το Apache OpenOffice.org ευρισκόμενο σε φάση αναδιοργάνωσης προσφέρει δυνατότητα εθελοντικής προσφοράς για οποιον έχει διάθεση!

Το R – Project είναι γλώσσα προγραμματισμού προσανατολισμένη για χρήση στη στατιστική. Η γλώσσα R αναπτύχθηκε έχοντας ως αρχικό πρότυπο τη γλώσσα S ωστόσο πλέον έχει το δικό της κοινό και τη δική της κατεύθυνση ανάπτυξης (28η στη σειρά δημοφιλίας των γλωσσών προγραμματισμού σύμφωνα με το TIOBE). Η R προσφέρει εξαιρετικές δυνατότητες στον ερευνητή. Χαρακτηρίζεται από εύκολη επεκτασιμότητα καθώς ο κάθε ένας χρήστης μπορεί να δημιουργήσει τις δικές του συναρτήσεις και να τις μορφοποιήσει σε μία βιβλιοθήκη η οποία μπορεί να δημοσιοποιηθεί και να χρησιμοποιηθεί αργότερα από κάποιον άλλο χρήστη με τις ίδιες ανάγκες. Μπορεί να παράξει εξαιρετικά πλούσια διαγράμματα απαιτώντας ωστόσο κάποια εξοικείωση με τη γλώσσα. Σε σύγκριση με το εμπορικό λογισμικό SPSS, η R υπολείπεται ενός φιλικού παραθυρικού περιβάλλοντος χρήσης, ωστόσο η χρησιμοποίησή της για την ανάλυση βελτιώνει γρήγορα την ουσιαστική ικανότητα του ερευνητή αλλά και την βαθύτερη αντίληψη της θεωρίας καθώς ο χρήστης αντιλαμβάνεται τον ακριβή τρόπο με τον οποίο επεξεργάζεται τα δεδομένα.

Στο 1ο κεφάλαιο του βιβλίου γίνεται αναφορά στους τρόπους και τις μεθόδους δειγματοληψίας η οποία αποτελεί την αρχή και το πιο σημαντικό βήμα μίας έρευνας που βασίζεται στη συλλογή δεδομένων. Στο 2ο κεφάλαιο περιγράφονται οι μέθοδοι της περιγραφικής στατιστικής. Το 3ο κεφάλαιο αφιερώνεται στη γραμμική παλινδρόμηση, το 4ο κεφάλαιο αποτελεί μία αναφορά στις χρονοσειρές χωρίς όμως ιδιαίτερη εμβάθυνση, το 5ο κεφάλαιο αφιερώνεται στη δοκιμασία $X^{2,}$ και το 6ο κεφάλαιο στην οικογένεια στατιστικών δοκιμασιών Student (t – test) με αναφορά στην ανάλυση διακύμανσης (ANOVA). Τέλος, το 7ο κεφάλαιο αποτελεί μία αναφορά σε κάποια στατιστικά σφάλματα που παρουσιάζονται συνήθως σε έρευνες.

Κεφάλαιο 1 Δειγματοληψία

1.1 Απαραίτητες έννοιες

Μεταβλητές ονομάζονται τα χαρακτηριστικά εκείνα, ως προς τα οποία εξετάζουμε έναν πληθυσμό. Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται τιμές της μεταβλητής. Οι μεταβλητές χωρίζονται σε δύο κατηγορίες, στις ποιοτικές και στις ποσοτικές.

- **Ποιοτικές ή Ονομαστικές (nominal) μεταβλητές** είναι εκείνες που δεν επιδέχονται μέτρηση και οι τιμές τους δεν είναι αριθμοί.
- **Ποσοτικές ή Αριθμητικές μεταβλητές** είναι εκείνες που επιδέχονται μέτρηση και οι τιμές τους είναι αριθμοί.

Οι ποσοτικές μεταβλητές διακρίνονται σε συνεχείς και διακριτές (ασυνεχείς).

- **Συνεχείς (scale)** είναι οι ποσοτικές μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος (α , β).
- **Διακριτές (ordinal)** είναι οι ποσοτικές μεταβλητές που παίρνουν μόνο μεμονωμένες τιμές.

Οι συνεχείς μεταβλητές μπορούν να διαχωριστούν περαιτέρω σε **διαστημικές (interval)** και **αναλογικές (ratio)**, χωρίς ωστόσο να είναι σημαντικός διαχωρισμός υπό την έννοα πως δεν επηρεάζεται η υλοποίηση μίας στατιστικής διαδικασίας από το χαρακτηρισμό μίας συνεχούς μεταβλητής ως διαστημικής ή αναλογικής.

Μία μεταβλητή χαρακτηρίζεται αναλογική αν ο λόγος των τιμών μεταξύ δύο παρατηρήσεων είναι δυνατό να ερμηνευθεί σε όρους της έννοιας που καταγράφει η μεταβλητή. Παράδειγμα αναλογικής μεταβλητής είναι το βάρος. Ένας άνθρωπος που ζυγίζει 80 κιλά έχει *διπλάσιο* βάρος από έναν άλλον άνθρωπο που ζυγίζει 40 κιλά. Η τελευταία σύγκριση δεν είναι δυνατό να συμβεί σε άλλες συνεχείς μεταβλητές όπως για παράδειγμα το Ph που μετρά την οξύτητα ενός υγρού. Το Ph λαμβάνει όλες τις τιμές από 0 έως 14 με την τιμή 7 να αντιστοιχεί σε ουδέτερο υγρό, τις μικρότερες τιμές σε όξινα υγρά και τις μεγαλύτερες σε βασικά υγρά. Ένα υγρό με Ph 8 είναι βασικό ενώ ένα άλλο υγρό με Ph ίσο με 4 είναι όξινο. Οπωσδήποτε, δεν είναι ορθό να δηλώσει κάποιος πως το υγρό με Ph 8 έχει διπλάσια οξύτητα από το άλλο υγρό με Ph 4.

Στατιστικός πληθυσμός ή απλά **πληθυσμός** ονομάζεται κάθε σύνολο, τα στοιχεία του οποίου εξετάζουμε ως προς ένα ή περισσότερα χαρακτηριστικά τους. Τα στοιχεία του πληθυσμού ονομάζονται μονάδες ή άτομα.
Ο πληθυσμός μπορεί να είναι **θεωρητικός** ή **πραγματικός** (μετρήσιμος). Παραδείγματα πραγματικού πληθυσμού είναι :

- Οι πολίτες του Δήμου Ξάνθης (γνωστός ο αριθμός τους κάθε στιγμή)
- Οι φοιτητές ενός τμήματος Α.Ε.Ι.. (ομοίως)
- Οι δημόσιοι υπάλληλοι που απασχολούνται στο Υ.Π.Ε.Π.Θ. (ομοίως)

Παραδείγματα θεωρητικού πληθυσμού είναι :

- Οι καπνιστές που **θα** ασθενήσουν από καρκίνο τον επόμενο μήνα (δεν είναι γνωστός ο αριθμός τους)
- Όσοι **θα** υποβάλλουν αίτηση στο ΑΕΠ για μεταπτυχιακό το 2012.

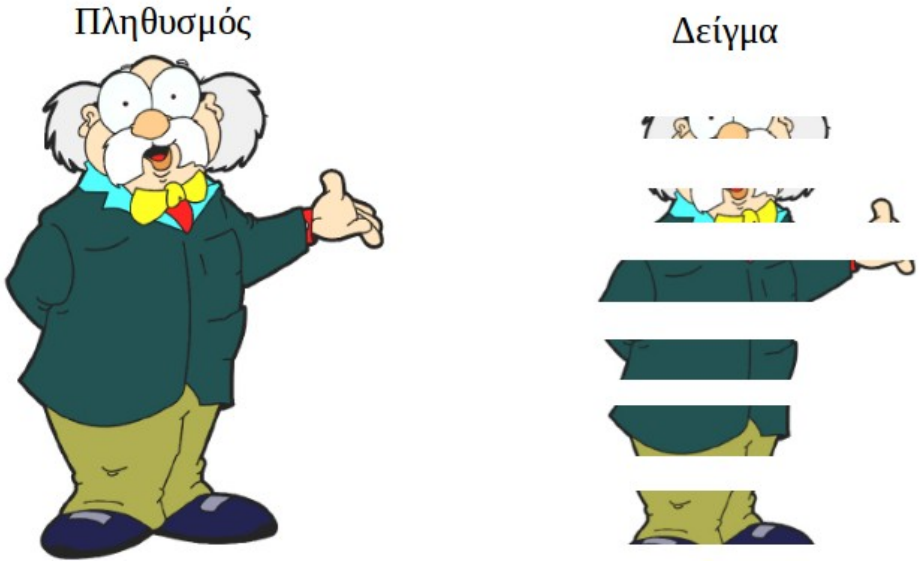
Στην περίπτωση του πραγματικού πληθυσμού πρέπει να είμαστε σε θέση να γνωρίζουμε το μέγεθος του. (ενημέρωση από τα κατάλληλα αρχεία) ενώ στην περίπτωση του θεωρητικού πληθυσμού πρέπει να είμαστε σε θέση να εκτιμήσουμε το μέγεθος του βάσει των παλαιότερων μετρήσεων, αλλά και το ρυθμό μεταβολής του.

Δείγμα ονομάζεται το υποσύνολο του πληθυσμού το οποίο μπορούμε να καταγράψουμε υπό τους περιορισμούς (υλικούς και χρονικούς) της έρευνάς μας.

1.2 Συλλογή στατιστικών δεδομένων

Οι κυριότερες μέθοδοι συλλογής στατιστικών δεδομένων είναι η απογραφή και η δειγματοληψία.

Απογραφή είναι μια μέθοδος συλλογής στατιστικών δεδομένων, που ακολουθούμε για να πάρουμε όλες τις απαραίτητες πληροφορίες, για έναν πληθυσμό εξετάζοντας όλα τα άτομα του πληθυσμού ως προς τα χαρακτηριστικά που μας ενδιαφέρουν.



Εικόνα 1: Δειγματοληψία

Δειγματοληψία ονομάζεται η διαδικασία καταγραφής ενός υποσυνόλου του πληθυσμού. Προχωρούμε σε δειγματοληψία γιατί η απογραφή είναι δύσκολη, οικονομικά και χρονικά ασύμφορη και πολλές φορές αδύνατη. Για αυτόν το λόγο επιλέγουμε μια μικρή ομάδα, δηλαδή ένα υποσύνολο του πληθυσμού το οποίο ονομάζεται δείγμα. Συλλέγουμε τις παρατηρήσεις από το δείγμα και στη συνέχεια γενικεύουμε τα συμπεράσματα για ολόκληρο τον πληθυσμό. Τα συμπεράσματα όμως, που θα προκύψουν από τη μελέτη του δείγματος θα είναι αξιόπιστα, δηλαδή θα ισχύουν με ικανοποιητική προσέγγιση για ολόκληρο τον πληθυσμό, μόνο όταν η επιλογή του δείγματος έχει γίνει με τέτοιο τρόπο, ώστε το δείγμα να είναι αντιπροσωπευτικό. Ένα δείγμα θεωρείται αντιπροσωπευτικό, όταν κάθε άτομο του πληθυσμού έχει δυνατότητα να επιλεγεί και αυτό μπορεί να συμβεί με την ίδια πιθανότητα για όλους.

Οι λόγοι για τον οποίο συμβαίνει μία δειγματοληψία είναι οι οικονομικοί και χρονικοί περιορισμοί που υπάρχουν αλλά και η περιορισμένη πρόσβαση στον πληθυσμό. Οι περιορισμοί αυτοί δεν μειώνουν την αξία της δειγματοληψίας καθώς μπορεί να δώσει ακριβή και αξιόπιστα αποτελέσματα ιδιαίτερα όταν ο πληθυσμός που μελετούμε είναι ομοιογενής ως προς το χαρακτηριστικό που μας ενδιαφέρει. Επίσης, η δειγματοληψία μπορεί να είναι περισσότερο αξιόπιστη από μία απογραφή, όταν η γνώση του ερωτώμενου πως πρόκειται για απογραφή αυξάνει τη μεροληψία της απόκρισης π.χ. οι αποκρίσεις των

αλλοδαπών στην εθνική απογραφή, οι οποίες ίσως να είναι πιο ειλικρινείς σε δειγματοληψία ή ακόμα στην περίπτωση όπου δεν υπάρχουν αξιόπιστοι κατάλογοι του πληθυσμού όπως στις μη αναπτυγμένες χώρες. Τέλος, η δειγματοληψία μειώνει το κόστος της έρευνας σε πραγματικούς πληθυσμούς, δηλαδή σε πληθυσμούς των οποίων το μέγεθος είναι γνωστό κάθε μία χρονική στιγμή.

Η επιλογή του αντιπροσωπευτικού δείγματος είναι “εκ των ων ουκ άνευ”. Αποτελεί πολύ σοβαρή και δύσκολη διαδικασία. Ο κακός σχεδιασμός και η εκτέλεση της στατιστικής έρευνας, η μη αντιπροσωπευτικότητα του δείγματος, ο μη σωστός καθορισμός του μεγέθους του δείγματος αποτελούν μερικά βασικά μειονεκτήματα στη διαδικασία επιλογής ενός δείγματος.

Το **σφάλμα** μίας δειγματοληψίας διαχωρίζεται σε **τυχαίο** και **συστηματικό**. **Τυχαίο σφάλμα δειγματοληψίας** ονομάζεται η διαφορά μεταξύ των μετρήσεων του δείγματος και των πραγματικών μετρήσεων το οποίο θα υπάρχει στην έρευνά μας και δεν μπορούμε να το υπολογίζουμε επακριβώς εκτός αν καταφέρουμε να κάνουμε μία τέλεια εκτελεσμένη απογραφή! Το τυχαίο σφάλμα προκύπτει με φυσικό τρόπο καθώς η μέση τιμή (ή άλλα στατιστικά) του υποσυνόλου του πληθυσμού που επιλέγουμε ως δείγμα είναι πρακτικά αδύνατο να είναι ίση με τη μέση τιμή του πληθυσμού, λόγω των τυχαίων σφαλμάτων της δειγματοληψίας. Αν η δειγματοληψία γίνει με κάποια πιθανοθεωρητική μέθοδο τότε το σφάλμα μπορεί να εκτιμηθεί ενώ αν γίνει με κάποια μη πιθανοθεωρητική μέθοδο (όπως συμβαίνει συχνά στην πράξη) τότε ο υπολογισμός του δεν είναι δυνατός.

Συστηματικό Σφάλμα Δειγματοληψίας ονομάζεται το σφάλμα που εμφανίζεται λόγω των σφαλμάτων σχεδίασης της δειγματοληψίας, όπως π.χ. αν μετράς την ευχαρίστηση από την απόκτηση ενός προϊόντος και έχεις δύο ομάδες που ρωτάνε, με τη μία να έχει μία πολύ όμορφη γυναίκα ως συνεντευξιαστή και την άλλη έναν άνδρα. Το συστηματικό σφάλμα είναι διαφορετικό από το τυχαίο σφάλμα, καθώς οφείλεται αποκλειστικά στον σχεδιασμό της έρευνας.

Η επιλογή του μεγέθους του δείγματος δεν είναι καθόλου εύκολη εργασία. Ένα δείγμα μεγέθους 30 μπορεί να είναι αρκετό αν ο πληθυσμός είναι μεγέθους 100 και ομοιογενής ενώ ένα δείγμα 100.000 κατοίκων του νομού Θεσσαλονίκης ή Αττικής μπορεί να μην είναι.

1.3 Στάδια δειγματοληψίας

Τα στάδια μίας δειγματοληψίας εμφανίζονται παραστατικά στο διάγραμμα 1.

Διάγραμμα 1: Στάδια δειγματοληψίας

Πλαίσιο δειγματοληψίας (Sampling Frame) είναι ο φυσικός περιορισμός που ορίζεται στον πληθυσμό από το χρόνο και τόπο που διεξάγεται η δειγματοληψία Για παράδειγμα αν στόχος είναι η μελέτη της αποδοχή ενός νέου αρώματος στις γυναίκες και η δειγματοληψία πραγματοποιηθεί το Σάββατο πρωί σε μία πλατεία της πόλης τότε πληθυσμός είναι οι γυναίκες της πόλης, ενώ το πλαίσιο της δειγματοληψίας είναι όσες κυκλοφορούν στη συγκεκριμένη πλατεία το Σάββατο το πρωί.

Το μέγεθος του δείγματος ορίζεται είτε από το διαθέσιμο χρόνο και κόστος στην περίπτωση της μη πιθανοθεωρητικής δειγματοληψίας είτε με κατάλληλο υπολογισμό βάσει του επιθυμητού δειγματικού σφάλματος αν η δειγματοληψία πραγματοποιείται με κάποια πιθανοθεωρητική μέθοδο.

1.4 Πιθανοθεωρητική και Μη πιθανοθεωρητική δειγματοληψία.

Πιθανοθεωρητική (propability sampling) ονομάζεται η δειγματοληψία στην οποία κάθε μέλος του πληθυσμού έχει γνωστή πιθανότητα επιλογής πριν την υλοποίηση της δειγματοληψίας, δηλαδή είναι δυνατή η χρήση της θεωρίας Πιθανοτήτων για τον υπολογισμό του τυχαίου σφάλματος της δειγματοληψίας.

Μη πιθανοθεωρητική (nonprobability sampling) ονομάζεται η δειγματοληψία στην οποία η πιθανότητα επιλογής των μελών του πληθυσμού είναι άγνωστη και δεν είναι δυνατή η εκ των προτέρων πιθανότητα επιλογής. Στη μη πιθανοθεωρητική δειγματοληψία το δείγμα επιλέγεται βάσει ορισμένων χαρακτηριστικών για τα οποία δεν γνωρίζουμε την

κατανομή τους αλλά επιθυμούμε να τα έχει το δείγμα. Μειονεκτήματα της μη πιθανοθεωρητικής δειγματοληψίας είναι

- η αδυναμία υπολογισμού τυχαίου σφάλματος.
- η μεγάλη πιθανότητα συστηματικού σφάλματος.
- απαιτείται μεγάλη εμπειρία / επανειλημμένη εφαρμογή.

1.4.1 Είδη πιθανοθεωρητικής δειγματοληψίας.

Απλή τυχαία δειγματοληψία (Simple Random Sampling)

Κάθε μέλος του πληθυσμού έχει ίση πιθανότητα επιλογής στο δείγμα. Στην πράξη η απλή τυχαία δειγματοληψία συμβαίνει όταν υπάρχει η δυνατότητα να τοποθετηθεί ο πληθυσμός στη σειρά 1, 2, ... και μετά επιλέγεται το 10%- 15% με γεννήτρια τυχαίων αριθμών. Απλή τυχαία δειγματοληψία συμβαίνει συνήθως σε τηλεφωνικές δημοσκοπήσεις όπου ο πληθυσμός είναι ταξινομημένος με φυσικό τρόπο αλφαβητικά στον τηλεφωνικό κατάλογο.

Συστηματική δειγματοληψία (Systematic Sampling)

Η συστηματική δειγματοληψία συμβαίνει όταν θέλουμε να επιλέξουμε ένα τυχαίο δείγμα και είναι περισσότερο εύκολο να πάρουμε περιοδικό δείγμα αντί για τυχαίο, όπως για παράδειγμα σε δειγματοληψία μάρκετινγκ στην αγορά. Επιλέγεται μία αρχή με τυχαίο τρόπο και μετά επιλέγεται κάθε ν-οστό μέλος του καταλόγου. Στην πράξη : Τοποθετείται ο πληθυσμός στη σειρά 1, 2, ..., μετά επιλέγεται με τυχαίο τρόπο η πρώτη θέση (π.χ. 10), επιλέγεται το βήμα ανάλογα με το συνολικό μέγεθος του πληθυσμού (π.χ. 5) και μετά επιλέγεται το δείγμα από το 10ο, 15ο, 20ο ... μέλος της σειράς.

Στρωματοποιημένη δειγματοληψία (Stratified Sampling)

Ο ερευνητής ορίζει κάποια χαρακτηριστικά του πληθυσμού για τα οποία επιθυμεί οπωσδήποτε αναλογική εκπροσώπηση στο δείγμα του και επιλέγει απλό τυχαίο δείγμα αναλογικά από κάθε κατηγορία του πληθυσμού.

Αναλογική Στρωματοποιημένη (Proportional Stratified Sample)

Το μέγεθος κάθε δειγματικής μονάδας είναι ανάλογο με το μέγεθος της αντίστοιχης κατηγορίας του πληθυσμού.

Παράδειγμα 1 : Έστω ότι θα επιλέξουμε 100 φοιτητές από ένα τμήμα με 2000 φοιτητές, άρα θα πάρουμε 1 στους 20 από τον πληθυσμό. Ορίζουμε τα στρώματα του πληθυσμού ανά έτος σπουδών και φύλο (άρα 4 x 2 = 8 στρώματα). Αν επιλέξουμε 1 στους 20 = 5% από κάθε ένα στρώμα τότε κάνουμε αναλογική στρωματοποιημένη δειγματοληψία.

Παράδειγμα 2 : Αν γνωρίζεις πως σε μία εταιρεία είναι 60% / 40% άνδρες / γυναίκες και θέλεις να επιλέξεις δείγμα 100 ατόμων για να μετρήσεις την ικανοποίηση θα πάρεις αναλογικά 60 άνδρες και 40 γυναίκες με τυχαία δειγματοληψία σε κάθε υποκατηγορία φύλου.

Μη Αναλογική Στρωματοποιημένη (Disproportional Stratified Sample)
Το μέγεθος του δείγματος είναι διαφορετικό σε κάθε μία κατηγορία και τεκμηριώνεται με θεωρητικά επιχειρήματα.

Παράδειγμα : Αποφασίζεις να συλλέξεις δείγμα 100 παιδιών ηλικίας 5 ετών από κάθε νομό της Ελλάδας και να καταγράψεις τα εμβόλια που έχουν κάνει. Είναι στρωματοποιημένη δειγματοληψία γιατί επιλέγεις ένα υπόδειγμα από τα 52 στρώματα (νομούς) στους οποίους χωρίζεται με φυσικό τρόπο η χώρα αλλά τα 100 παιδιά συνιστούν διαφορετική αναλογία ως προς τον πληθυσμό κάθε νομού.

Δειγματοληψία κατά συστάδες
Στη δειγματοληψία κατά συστάδες η πρωταρχική δειγματική μονάδα δεν είναι ένα μεμονωμένο άτομο αλλά μία ομάδα από άτομα. Οι ομάδες που τελικά συμμετέχουν στη δειγματοληψία επιλέγονται με τυχαίο τρόπο από το σύνολο των ομάδων. Στην πράξη διαχωρίζεται ο πληθυσμός σε ομάδες ομοιογενείς ως προς το χαρακτηριστικό που μελετούνται ενώ μετά με απλή τυχαία δειγματοληψία επιλέγονται κάποιες από αυτές και τις κάνεις δείγμα της έρευνας.
Παράδειγμα : Για να μελετήσεις το ποσοστό των μαθητών που έχουν Η/Υ σε έναν νομό, αποφασίζεις να συλλέξεις δείγμα 300 μαθητών. Πρώτα, με απλή τυχαία δειγματοληψία στα 40 σχολεία του νομού (αναγκαστική υπόθεση εργασίας πως δεν υπάρχουν σημαντικές διαφορές από σχολείο σε σχολείο ως προς το ποσοστό) επιλέγεις 10 σχολεία και λαμβάνεις πάλι με απλή τυχαία 30 μαθητές από κάθε σχολείο.

1.4.2 Είδη μη πιθανοθεωρητικής δειγματοληψίας.

Δειγματοληψία ευκολίας (convenience sampling)

Το δείγμα αποτελείται από τις μονάδες του πληθυσμού που είναι διαθέσιμες εκείνη τη χρονική στιγμή. Είναι η περισσότερο συνηθισμένη πρακτική δειγματοληψίας χωρίς ωστόσο να διασφαλίζεται η αντιπροσωπευτικότητα στον πληθυσμό από τον οποίο προέρχεται το δείγμα. Η κυριότερη πηγή μεροληψίας (σφάλματος) είναι ο τρόπος επιλογής των συμμετεχόντων καθώς το δείγμα αποτελείται από όσα μέλη του πληθυσμού είναι εύκολο να εντοπιστούν ή/και έχουν θετική στάση προς την έρευνα.

Δειγματοληψία σκοπιμότητας (purposive sampling)

Ένας εκπαιδευμένος δειγματολήπτης επιλέγει τις μονάδες του πληθυσμού που θεωρεί πως ανταποκρίνονται σε προκαθορισμένο προφίλ, π.χ. σε ένα εμπορικό κέντρο σταματά και ρωτά μόνο γυναίκες γύρω στα 40 κλπ). Είναι περισσότερο μεθοδική από τη δειγματοληψία ευκολίας, υπό την έννοια πως διασφαλίζεται ορισμένα χαρακτηριστικά του δείγματος χωρίς ωστόσο να τεκμηριώνεται αντιπροσωπευτικότητα στον πληθυσμό.

Δειγματοληψία αναλογίας (quota sampling)

Επιλογή του δείγματος έτσι ώστε να αντανακλάται σε αυτό η δημογραφική δομή του πληθυσμού ως προς ένα ή περισσότερα χαρακτηριστικά. Διαφέρει με τη στρωματοποιημένη δειγματοληψία ως προς τον τρόπο που γίνεται η επιλογή του δείγματος δεν κάθε μία ομάδα : στη δειγματοληψία αναλογίας επιλέγεται δείγμα ευκολίας ή δείγμα σκοπιμότητας ενώ στη στρωματοποιημένη δειγματοληψία επιλέγεται πιθανοθεωρητικό δείγμα.

Δειγματοληψία χιονοστιβάδας (Snowball Sampling)

Αρχική επιλογή ενός δείγματος με πιθανοθεωρητική μέθοδο και σε δεύτερο στάδιο συνέχιση της δειγματοληψίας από φίλο σε φίλο, από γείτονα σε γείτονα κλπ. Συνιστάται μόνο στις περιπτώσεις που είναι επιθυμία του ερευνητή, το δείγμα να έχει κάποια συγκεκριμένα κοινωνικά ή πολιτικά χαρακτηριστικά.

1.4.3 Σύγκριση μεταξύ πιθανοθεωρητικής και μη πιθανοθεωρητικής δειγματοληψίας.

Είναι φανερό πως η περισσότερο επιθυμητή λύση για μία έρευνα είναι η πιθανοθεωρητική

δειγματοληψία. Ωστόσο, ο κόπος, το κόστος και ο χρόνος που απαιτείται για τη σχεδίαση μίας τέτοιας δειγματοληψίας πολλές φορές είναι απαγορευτικό (π.χ. στα πλαίσια μίας πτυχιακής ή διπλωματικής εργασίας) κάτι που στην πράξη σημαίνει πως στις περισσότερες των περιπτώσεων επιλέγεται η μη πιθανοθεωρητική δειγματοληψία και ειδικότερα η δειγματοληψία ευκολίας. Τα πλεονεκτήματα της μη πιθανοθεωρητικής δειγματοληψίας είναι προφανή : απαιτείται μικρός χρόνος συλλογής δεδομένων ενώ έχει ελάχιστο κόστος.

1.4.4 Πολυσταδιακή Γεωγραφική Δειγματοληψία

Η Πολυσταδιακή Γεωγραφική Δειγματοληψία αποτελείται από ένα συνδυασμό δειγματοληπτικών τεχνικών. Η βασική αρχή είναι πως επιλέγονται γεωγραφικές περιοχές με συνεχή εξειδίκευση από βήμα σε βήμα. Ο ερευνητής μπορεί να πραγματοποιήσει όσα βήματα χρειάζονται για να αποκτήσει αντιπροσωπευτικό δείγμα. Στο τελικό στάδιο επιλέγεται η δειγματική μονάδα. Είναι η μέθοδος με την οποία πραγματοποιούνται οι περισσότερες έρευνες σε εθνική εμβέλεια.

1.5 Συνοπτική σύγκριση των δειγματοληπτικών τεχνικών

Πίνακας 1.1 : Σύγκριση Μη Πιθανοθεωρητικών Δειγματοληπτικών Τεχνικών

Τρόπος δειγματοληψίας	Περιγραφή	Πλεονεκτήματα	Μειονεκτήματα
Δειγματοληψία ευκολίας (convenience sampling)	Ο ερευνητής επιλέγει το δείγμα στο οποίο έχει εύκολη πρόσβαση.	Δεν απαιτείται κατάλογος του πληθυσμού.	Δεν μπορεί να τεκμηριωθεί η αντιπροσωπευτικότητα του δείγματος. Δεν είναι ασφαλής η γενίκευση των συμπερασμάτων.
Δειγματοληψία κρίσης (judgment sampling)	Ο ερευνητής επιλέγει τις δειγματικές μονάδες βάσει ορισμένων χαρακτηριστικών όπως π.χ. η ηλικιακή ομάδα ή το φύλο	Χρήσιμο για περιπτώσεις όπου το δείγμα πρέπει να έχει κάποια χαρακτηριστικά.	Υπάρχει μεροληψία του ερευνητή κατά την επιλογή. Δεν είναι ασφαλής η γενίκευση των συμπερασμάτων.
Αναλογική (quota sampling)	Ο ερευνητής διαχωρίζει τον πληθυσμό βάσει ορισμένων χαρακτηριστικών, καθορίζει το ποσοστό που αντιπροσωπεύει κάθε μία υποκατηγορία και επιλέγει με δειγματοληψία	Δεν απαιτείται κατάλογος του πληθυσμού. Αντιπροσώπευση του πληθυσμού ως προς κάποια	Ο διαχωρισμός σε υποομάδες είναι μεροληπτικός (ποιες ομάδες και σε πόσες κατηγορίες;). Δεν είναι δυνατό το σφάλμα δειγματοληψίας. Δεν είναι

	κρίσης το δείγμα από κάθε υποκατηγορία. (το δεύτερο βήμα την καθιστά μη πιθανοθεωρητική)	χαρακτηριστικά.	ασφαλής η γενίκευση των συμπερασμάτων.

Πίνακας 1.2: Σύγκριση Πιθανοθεωρητικών Δειγματοληπτικών Τεχνικών

Τρόπος δειγματοληψίας	Περιγραφή	Πλεονεκτήματα	Μειονεκτήματα
Απλή τυχαία δειγματοληψία (simple random sampling)	Ο ερευνητής ταξινομεί τα μέλη του πληθυσμού σε μία σειρά και μετά επιλέγει με τυχαίο τρόπο το δείγμα. (απαιτείται κατάλογος του πληθυσμού)	Είναι εφικτός ο υπολογισμός του δειγματικού σφάλματος. Δεν απαιτείται προηγούμενη γνώση της δομής του πληθυσμού.	Υπάρχει σφάλμα από τον περιορισμό σε πλαίσιο δειγματοληψίας. Πιθανή μεγάλη γεωγραφική διασπορά του δείγματος. Ενδεχομένως να υπάρξει υποεκτίμηση κάποιων σημαντικών για την έρευνα χαρακτηριστικών.
Συστηματική (systematic sampling)	Ο ερευνητής επιλέγει τις δειγματικές μονάδες βάσει ενός συγκεκριμένου σχεδίου π.χ. Επιλέγει με τυχαίο τρόπο τον πρώτο και μετά επιλέγει κάθε 10 επόμενους. (απαιτείται κατάλογος του πληθυσμού)	Εύκολη λήψη δείγματος.	Το σταθερό διάστημα μεταξύ των επιλογών μπορεί να αυξήσει το δειγματικό σφάλμα στην περίπτωση που θα συμπέσει με κάποια περιοδικότητα του δείγματος.
Στρωματοποιημένη (stratified sampling)	Ο ερευνητής διαχωρίζει τον πληθυσμό σε ομάδες βάσει ορισμένων χαρακτηριστικών, και επιλέγει με τυχαία δειγματοληψία δείγμα από κάθε μία ομάδα.	Αντιπροσωπεύονται όλες οι υποκατηγορίες βάσει επιλεγμένων χαρακτηριστικών.	Απαιτούνται κατάλογοι σε κάθε μία από τις ομάδες που ορίζονται από τα επιλεγμένα χαρακτηριστικά.
Δειγματοληψία διασποράς (cluster sampling)	Ο ερευνητής διαχωρίζει το χώρο σε περιοχές, επιλέγει τυχαία ένα υποσύνολο από αυτές και μετά επιλέγει πιθανοθεωρητικό δείγμα από τις περιοχές αυτές.	Μικρότερο κόστος δειγματοληψίας καθώς περιορίζεται η γεωγραφική κάλυψη που απαιτείται.	Μεγαλύτερο σφάλμα από τις άλλες πιθανοθεωρητικές μεθόδους. Κάθε δειγματική μονάδα πρέπει με μοναδικό τρόπο να μπορεί να συνδεθεί με μία από τις περιοχές.

Πολυσταδιακή (multistage sampling)	Σταδιακή εξειδίκευση περιοχών από τη μεγαλύτερη στη μικρότερη έως ότου επιλεγεί μία δειγματική μονάδα. Ο τρόπος εξειδείκευσης μπορεί να επιλεγεί από τις υπόλοιπες πιθανοθεωρητικές μεθόδους. (συνήθως βρίσκει εφαρμογή σε έρευνες πανεθνικής εμβέλειας)	Ανάλογα με τις μεθόδους που χρησιμοποιούνται για τη συνεχή εξειδίκευση	Ανάλογα με τις μεθόδους που χρησιμοποιούνται για τη συνεχή εξειδίκευση

1.6 Κανονική Κατανομή

Η κανονική κατανομή αφορά τις συνεχείς τυχαίες μεταβλητές, αλλά και τα αρκετά μεγάλα αθροίσματα τυχαίων μεταβλητών οποιασδήποτε κατανομής. Την επινόησε ο Gauss ο οποίος με τον τρόπο αυτό μαθηματικοποίησε τις φυσικές παρατηρήσεις που είχε κάνει σε μεγάλα πλήθη φυσικών παρατηρήσεων οι οποίες επηρεαζόταν από μεγάλο πλήθος σφαλμάτων που δρώντας αθροιστικά στο τελικό αποτέλεσμα διαμόρφωναν μία κατανομή τιμών συμμετρική και κωδωνοειδής, στην οποία η πλειοψηφία των παρατηρήσεων ήταν γύρω από τη μέση τιμή ενώ όλο και λιγότερες υπήρχαν όσο μεγάλωνε η απόσταση από τη μέση τιμή. Η καταλληλότητά της στην ερμηνεία των κατανομών που επηρεάζονται από πολλές πηγές σφαλμάτων εξηγεί άμεσα και τη μεγάλη χρηστικότητά της στην προσέγγιση αθροισμάτων τυχαίων μεταβλητών όλων των κατανομών, λόγος για τον οποίο εμφανίζεται σχεδόν σε κάθε στατιστική διαδικασία!

Αν X είναι μία τυχαία μεταβλητή (τ.μ.) με μέση τιμή μ και τυπική απόκλιση σ τότε λέμε πως η τ.μ. X *ακολουθεί* την **κανονική κατανομή** (γράφουμε **$X \sim N(\mu, \sigma^2)$**) αν η συνάρτηση πυκνότητας πιθανότητας της τ.μ. X είναι η

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

Η μέση τιμή μ ορίζει τον άξονα συμμετρίας της κατανομής ενώ η τυπική απόκλιση σ ορίζει το “πλάτος” της κατανομής υπό την έννοια πως το 99% των τιμών μίας μεταβλητής που ακολουθεί την κανονική κατανομή βρίσκεται σε απόσταση ± 3 τυπικές αποκλίσεις από τη μέση τιμή.

Όπως σε κάθε συνεχής κατανομή έτσι και στην κανονική, η πιθανότητα της τιμής της

τυχαίας μεταβλητής X να βρίσκεται μεταξύ δύο τιμών α και β είναι ίση με το εμβαδόν που βρίσκεται κάτω από την καμπύλη που ορίζεται από τη συνάρτηση πυκνότητας πιθανότητας f, μεταξύ των ευθειών x = α και x = β. Καθώς το εμβαδό αναπαριστά πιθανότητα είναι φανερό πως το συνολικό εμβαδόν της καμπύλης είναι ίσο με 1!

Παράδειγμα μεταβλητών που ακολουθούν την κανονική κατανομή είναι τα περισσότερα σωματομετρικά αλλά και ψυχολογικά χαρακτηριστικά στον άνθρωπο όπως το βάρος, το ύψος, η νοημοσύνη, το άγχος, η επιθετικότητα κ.α.. Ο λόγος είναι απλός και εντοπίζεται στο μεγάλο πλήθος παραγόντων που επηρεάζουν τη σωματική και ψυχολογική εξέλιξη του ανθρώπου.

Τυποποιημένη (ή τυπική) κανονική κατανομή ονομάζεται η κατανομή με συνάρτηση πυκνότητας πιθανότητας

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

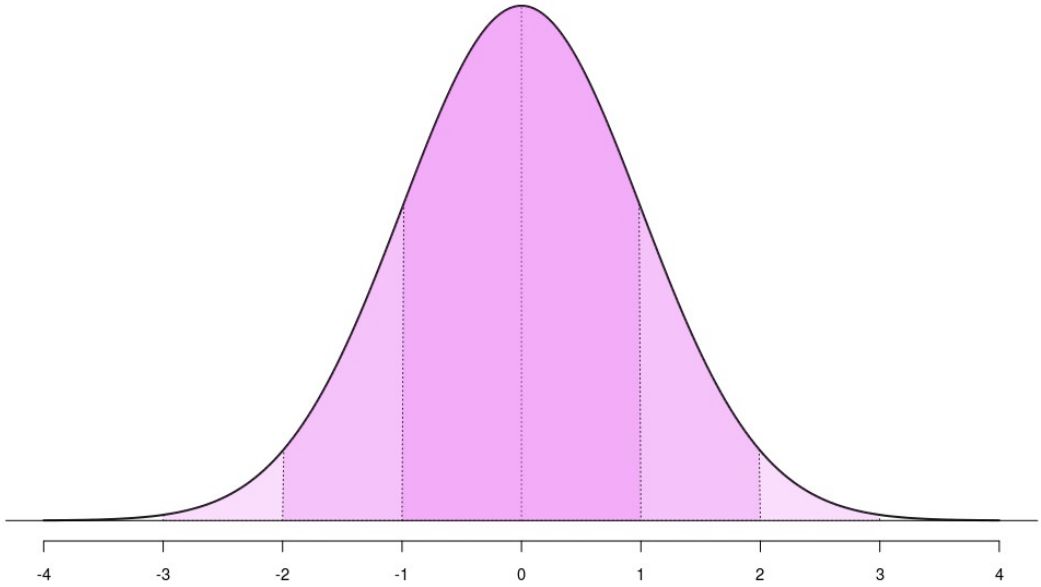
Η γραφική παράσταση της συνάρτησης πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής παρουσιάζεται στο διάγραμμα 2, σελίδα 19. Γράφουμε **Z ~ N(0, 1)** και δεσμεύουμε το γράμμα **Z** να συμβολίζει την τυπική κανονική κατανομή.

Όλες οι κανονικές κατανομές διαφέρουν μεταξύ τους μόνο ως προς το σημείο που βρίσκεται ο άξονας συμμετρίας (ο οποίος ορίζεται από τη μέση τιμή) και στην κυρτότητά τους (η οποία ορίζεται από την τυπική απόκλιση). Για κάθε σημαντικό υπολογισμό χρησιμοποιείται η τυπική κανονική κατανομή αφού πρώτα γίνει η απαραίτητη μεταφορά της οποιασδήποτε κανονικής κατανομής στην τυπική με το επόμενο θεώρημα

:

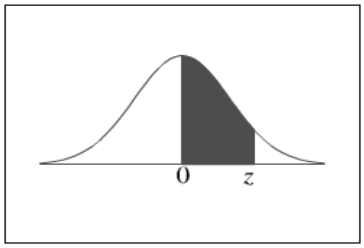
Αν $X \sim N(\mu, \sigma^2)$ τότε $\frac{X-\mu}{\sigma} = Z \sim N(0,1)$

Τύπος 1: Τυποποίηση μίας κανονικής κατανομής



Διάγραμμα 2: Τυπική κανονική κατανομή

Standard Normal Distribution Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177

Πίνακας 1.3: Παράδειγμα πίνακα τιμών τυποποιημένης κανονικής κατανομής

Στη στατιστική είναι αναγκαίος ο υπολογισμός πιθανοτήτων βασισμένων στην τυπική

κανονική κατανομή. Καθώς η πιθανότητα αντιστοιχεί σε κάποιο εμβαδό καταλήγουμε στην αναγκαιότητα ενός υπολογισμού εμβαδού ο οποίος όμως δεν είναι απλός. Παλαιότερα υπήρχαν πίνακες που έδιναν τα εμβαδά για διάφορες τιμές της τ.μ. Z (πίνακας 1.3, σελίδα 22). Πλέον, οι πίνακες αυτοί είναι παρωχημένοι καθώς οι αντίστοιχοι υπολογισμοί γίνονται εύκολα και γρήγορα με έναν υπολογιστή.

Πίνακας 1.4: Υπολογισμός πιθανότητας σε κανονική κατανομή με υπολογιστή

Συνάρτηση **NORMDIST(αριθμός; μ; σ).** Το αποτέλεσμα είναι το εμβαδόν της γραφικής παράστασης της κανονικής κατανομής με μέση τιμή μ, τυπική απόκλιση σ από την ευθεία x = αριθμός και *αριστερά* αυτής.

Για παράδειγμα αν γνωρίζεις πως το βάρος του ενήλικου πληθυσμού ακολουθεί κανονική κατανομή με μέση τιμή 65 κιλά και τυπική απόκλιση 15 κιλά και θέλεις να βρεις την πιθανότητα ένας άνθρωπος να έχει βάρος *μικρότερο* από 72 κιλά τότε αρκεί να υπολογίσεις NORMDIST(72; 65; 15) = 0,6796 ή περίπου 68%. Αντίστοιχα αν θέλεις να βρεις την πιθανότητα ένας άνθρωπος να έχει βάρος *μεγαλύτερο* από 72 κιλά θα υπολογίσεις 1 - NORMDIST(72; 65; 15) = 0,3203 ή περίπου 32%.

Αντίστροφα, αν πρέπει να υπολογιστεί το σημείο του πραγματικού άξονα που αφήνει αριστερά του συγκεκριμένο μέρος του εμβαδού τότε μπορεί να χρησιμοποιηθεί η συνάρτηση **NORMINV(αριθμός; μ; σ)** π.χ. Η NORMINV(0,005;0;1) επιστρέφει -2,5758... Εύκολα καταλαβαίνουμε πως λόγω συμμετρίας μεταξύ -2,5758 και 2,5758 βρίσκεται το 99% του συνολικού εμβαδού της καμπύλης.

Η συνάρτηση **pnorm(c(72), mean=65, sd=15, lower.tail=TRUE)** δίνει το εμβαδόν της γραφικής παράστασης της κανονικής κατανομής με μέση τιμή 65, τυπική απόκλιση 15 από την ευθεία x = 72 και αριστερά αυτής. Αντίστοιχα, η **pnorm(c(72), mean=65, sd=15, lower.tail=FALSE)** δίνει το συμπληρωματικό εμβαδόν.

Η συνάρτηση **qnorm(c(0.005), mean=0, sd=1, lower.tail=TRUE)** δίνει το σημείο του άξονα που περιορίζει το 0.005 = 0.5% του συνολικού εμβαδού της καμπύλης.

Τα αριθμητικά αποτελέσματα είναι ίδια με αυτά του Calc.

Δραστηριότητες

1η. Το παρακάτω ερωτηματολόγιο χρησιμοποιείται για τη μέτρηση του άγχους ως χαρακτηριστικό της προσωπικότητας. Κάθε απάντηση βαθμολογείται με 1 έως 4 μονάδες. Το συνολικό σκορ προκύπτει από το άθροισμα των επιμέρους ερωτήσεων και κυμαίνεται

από 20 έως 80. Όσο μεγαλύτερο είναι το *συνολικό άθροισμα* τόσο μεγαλύτερο είναι και το άγχος του ερωτώμενου. Κάποιες από τις ερωτήσεις πρέπει να βαθμολογηθούν αντίστροφα (δηλαδή το 4 θα γίνει 1, το 3 θα γίνει 2, το 2 θα γίνει 3 και το 1 θα γίνει 4). Α. Βρείτε τις αντίστροφες ερωτήσεις. Β. Συμπληρώστε το ερωτηματολόγιο! Γ. Υπολογίστε το συνολικό σκορ. Δ. Γνωρίζοντας πως η μέση τιμή για την κλίμακα αυτή στον πληθυσμό είναι 42,8 μονάδες με τυπική απόκλιση 10,6 μονάδες (Αναγνωστοπούλου, 2002), βρείτε ποιο ποσοστό του πληθυσμού έχει μικρότερο σκορ στην κλίμακα άγχους από εσάς.

STAI (Spielberger, 1970)

(Απόδοση και προσαρμογή για τον Ελληνικό πληθυσμό: Λιάκος & Γιαννίτση, 1984)

Οδηγίες : Παρακαλούμε διαβάστε προσεκτικά κάθε πρόταση και στη συνέχεια **βάλτε σε κύκλο έναν από τους αριθμούς που αντιστοιχεί στην απάντηση, η οποία θεωρείτε ότι σας αντιπροσωπεύει περισσότερο γενικά στη ζωή σας.**

	ΣΧΕΔΟΝ ΠΟΤΕ	ΜΕΡΙΚΕΣ ΦΟΡΕΣ	ΣΥΧΝΑ	ΣΧΕΔΟΝ ΠΑΝΤΑ
1. Αισθάνομαι ευχάριστα.	1	2	3	4
2. Κουράζομαι εύκολα.	1	2	3	4
3. Βρίσκομαι σε συνεχή αγωνία.	1	2	3	4
4. Εύχομαι να μπορούσα να είμαι τόσο ευτυχισμένος/η, όσο φαίνεται να είναι οι άλλοι.	1	2	3	4
5. Μένω πίσω στις δουλειές μου, γιατί δε μπορώ να αποφασίσω αρκετά γρήγορα.	1	2	3	4
6. Αισθάνομαι αναπαυμένος/η.	1	2	3	4
7. Είμαι ήρεμος/η, ψύχραιμος/η και συγκεντρωμένος/η.	1	2	3	4
8. Αισθάνομαι πως οι δυσκολίες συσσωρεύονται και δε μπορώ να τις ξεπεράσω.	1	2	3	4
9. Ανησυχώ πάρα πολύ για κάτι που στην πραγματικότητα δεν έχει σημασία.	1	2	3	4
10. Βρίσκομαι σε συνεχή υπερένταση.	1	2	3	4
11. Έχω την τάση να βλέπω τα πράγματα δύσκολα.	1	2	3	4
12. Μου λείπει η αυτοπεποίθηση.	1	2	3	4
13. Αισθάνομαι ασφαλής.	1	2	3	4
14. Προσπαθώ να αποφεύγω την αντιμετώπιση μιας κρίσης ή μιας δυσκολίας.	1	2	3	4
15. Βρίσκομαι σε υπερδιέγερση.	1	2	3	4
16. Είμαι ικανοποιημένος/η.	1	2	3	4
17. Κάποια ασήμαντη σκέψη μου περνά από το μυαλό και με ενοχλεί.	1	2	3	4
18. Παίρνω τις απογοητεύσεις τόσο πολύ στα σοβαρά, ώστε δε μπορώ να τις διώξω από τη σκέψη μου.	1	2	3	4
19. Είμαι ένας σταθερός χαρακτήρας.	1	2	3	4
20. Έρχομαι σε μια κατάσταση έντασης ή αναστάτωσης, όταν σκέφτομαι τις τρέχουσες δυσκολίες και τα ενδιαφέροντά μου.	1	2	3	4

2η. Εντοπίστε τον αριθμό α με την ιδιότητα : το εμβαδόν της τυπικής κανονικής κατανομής μεταξύ -α και α είναι ίσο με (i) 0,95 (ii) 0,99 (απάντηση : (i) 1, 96 (ii) 2,57)

Υπόδειξη :

(i) Αναζητούμε το σημείο στον πραγματικό άξονα ο οποίος αφήνει αριστερά του το 0,025 του συνολικού εμβαδού της γραφικής παράστασης (ii) Όμοια αφήνει αριστερά το 0,005.

1.7 Κεντρικό Οριακό Θεώρημα

Ας υποθέσουμε πως από έναν κανονικό πληθυσμό $N(\mu,\sigma^2)$ παίρνουμε ένα μεγάλο πλήθος από δείγματα σταθερού μεγέθους n. Κάθε ένα από αυτά τα δείγματα θα έχει τη δική του μέση τιμή η οποία είναι μάλλον απίθανο να ταυτίζεται με την μέση τιμή του πληθυσμού (το ίδιο ισχύει φυσικά για την τυπική απόκλιση αλλά αυτό τώρα δεν μας ενδιαφέρει).

Είναι λογικό να τεθεί το ερώτημα σχετικά με τη μορφή της κατανομής *των μέσων τιμών* αυτών των δειγμάτων. Η κατανομή αυτή ονομάζεται δειγματική κατανομή (sampling distribution) και τη μορφή της την περιγράφει το κεντρικό οριακό θεώρημα που ακολουθεί.

Αν $X \sim N(\mu,\ \sigma^2)$ τότε $\bar{X} \sim N(\mu,\ \frac{\sigma^2}{n})$

Τύπος 2: Κεντρικό Οριακό Θεώρημα

Με απλά λόγια το Κεντρικό Οριακό Θεώρημα λέει πως αν πάρουμε πολλά δείγματα σταθερού μεγέθους n από έναν κανονικό πληθυσμό, μετρήσουμε τις μέσες τιμές όλων των δειγμάτων και κάνουμε το ιστόγραμμα αυτών των μέσων τιμών αυτό θα ομοιάζει με την κανονική κατανομή.

Το Κεντρικό Οριακό Θεώρημα (ΚΟΘ) είναι ισχυρό : για αρκούντως μεγάλο δείγμα (n > 30) ισχύει ανεξάρτητα από το είδος της κατανομής από την οποία λαμβάνεται το δείγμα.

Το ΚΟΘ γράφεται και $\bar{X} \sim N(\mu,\sigma_{\bar{X}}^2)$ όπου $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ η δειγματική τυπική απόκλιση, δηλαδή η τυπική απόκλιση της κατανομής που προκύπτει από τις μέσες τιμές των

δειγμάτων.

Στην πράξη το ΚΟΘ μας δίνει τη δυνατότητα να υπολογίσουμε την πιθανότητα να εμφανιστεί από δεδομένο κανονικό πληθυσμό ένα δείγμα με μέση τιμή μακριά από την μέση τιμή του πληθυσμού και για το λόγο αυτό το ΚΟΘ είναι η ψυχή όλων των παραμετρικών στατιστικών μεθόδων.

1.8 Διάστημα εμπιστοσύνης

Σε συνδυασμό με τον τύπο 1, σελίδα 21 το ΚΟΘ δίνει το εξής συμπέρασμα

$$\text{Αν} \quad X \sim N(\mu,\ \sigma^2) \quad \text{τότε} \quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z \sim N(0,1)$$

από όπου συμπεραίνουμε άμεσα πως με 95% πιθανότητα (δραστηριότητα 2η)

$$-1{,}96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1{,}96$$

ή λύνοντας ως προς τη μέση τιμή μ του πληθυσμού,

$$\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}} \quad \text{ή} \quad \bar{X} - 1{,}96\,\sigma_{\bar{X}} \le \mu \le \bar{X} + 1{,}96\,\sigma_{\bar{X}}$$

Τύπος 3: 95% διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού

Ο τύπος 3 δίνει στον ερευνητή ένα διάστημα μέσα στο οποίο περιορίζεται με αλγεβρικό τρόπο η μέση τιμή του πληθυσμού. Στην περίπτωση που αυτή είναι άγνωστη (η πιο συνηθισμένη περίπτωση μίας πραγματικής έρευνας) το διάστημα αυτό χρησιμοποιείται για να δώσει ένα εύρος ασάφειας για την άγνωστη μέση τιμή από τα γνωστά μεγέθη του δείγματος. Επίσης καθώς μαζί με τη μέση τιμή του πληθυσμού θα είναι ομοίως άγνωστη και η τυπική απόκλιση, για την εφαρμογή του τύπου 3 χρησιμοποιούνται τα αντίστοιχα δειγματικά μεγέθη.

$$\bar{X} - 1{,}96\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + 1{,}96\frac{s}{\sqrt{n}} \quad \text{ή} \quad \bar{X} - 1{,}96\,s_{\bar{X}} \le \mu \le \bar{X} + 1{,}96\,s_{\bar{X}}$$

Τύπος 4: 95% διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού

Όπου $s_{\bar{X}} = \frac{s}{\sqrt{n}}$ η εκτίμηση της δειγματικής τυπικής απόκλισης. Το διάστημα που ορίζεται από τον τύπο 4 το ονομάζουμε 95% διάστημα εμπιστοσύνης για τη μέση τιμή του

πληθυσμού και αποτελεί την πρώτη εκτίμηση για αυτήν στην περίπτωση που δεν τη γνωρίζουμε και έχουμε κάνει δειγματοληψία για να τη βρούμε.

Το διάστημα εμπιστοσύνης είναι ένα μέτρο της ασάφειας της εκτίμησης που μπορεί να κάνει ο ερευνητής για την άγνωστη μέση τιμή του πληθυσμού από την γνωστή δειγματική μέση τιμή. Ο ερευνητής πρέπει να είναι προσεκτικός ως προς την περιγραφή της ερμηνείας του διαστήματος εμπιστοσύνης. Ο ισχυρισμός πως “Η μέση τιμή του πληθυσμού θα βρίσκεται σε αυτό το διάστημα με 95% πιθανότητα” δεν είναι σωστός. Η ορθή ερμηνεία είναι απλά πως αν η (άγνωστη) μέση τιμή του πληθυσμού είναι μέσα στο διάστημα αυτό τότε το δείγμα θα βρίσκεται σε απόσταση 1,96 δειγματικές τυπικές αποκλίσεις από αυτήν.

Παράδειγμα

130 φοιτητές ρωτήθηκαν για το ποσό που αναλώνουν σε αγορές βιβλίων στη διάρκεια ενός εξαμήνου. Το μέσο ποσό βρέθηκε να είναι 422 ευρώ με τυπική απόκλιση 57 ευρώ. Να υπολογιστεί 95% διάστημα εμπιστοσύνης για το μέσο ποσό που διαθέτει το σύνολο των φοιτητών.

Απάντηση

Υπολογίζουμε $S_{\bar{X}}=\frac{57}{\sqrt{130}}=4{,}9999\simeq 5$ και $422-1{,}96\cdot 5\le\mu\le 422+1{,}96\cdot 5$ ή απλά

$$412{,}2 \le \mu \le 431{,}8$$

1.8.1 Διάστημα εμπιστοσύνης για αναλογία

Λαμβάνουμε δείγμα μεγέθους n από έναν πληθυσμό του οποίου τα μέλη έχουν ένα χαρακτηριστικό με πιθανότητα p (και δεν το έχουν με πιθανότητα 1 – p = q) Αν X είναι η τυχαία μεταβλητή που μετρά το πλήθος των μελών του δείγματος που έχουν το χαρακτηριστικό τότε $X\sim B(n,p)$. Αν το μέγεθος του δείγματος είναι αρκετά μεγάλο τότε $B(n,p)\sim N(np,npq)$ άρα $X\sim N(np,npq)$. Από την τελευταία προσέγγιση και τον τύπο 1, σελίδα 21, παίρνουμε $\frac{X-np}{\sqrt{npq}/\sqrt{n}} = Z \sim N(0,1)$ ή $\frac{X-np}{\sqrt{pq}} = Z \sim N(0,1)$ από όπου συμπεραίνουμε άμεσα πως με 95% πιθανότητα

$$-1{,}96 \le \frac{X-np}{\sqrt{pq}} \le 1{,}96$$

ή λύνοντας ως προς την άγνωστη αναλογία p του πληθυσμού,

$$\frac{X}{n} - 1{,}96\frac{\sqrt{pq}}{\sqrt{n}} \le p \le \frac{X}{n} + 1{,}96\frac{\sqrt{pq}}{\sqrt{n}}$$

ή αν $\hat{p}=\frac{X}{n}$ το ποσοστό εμφάνισης του χαρακτηριστικού στο δείγμα μεγέθους n

$$\hat{p} - 1{,}96\, s_p \le p \le \hat{p} + 1{,}96\, s_p \text{ όπου } s_p = \sqrt{\frac{p\,q}{n}}$$

Τύπος 5: 95% διάστημα εμπιστοσύνης για το ποσοστό ενός χαρακτηριστικού στον πληθυσμό.

Ο τύπος 5 στην πράξη δεν μπορεί να χρησιμοποιηθεί καθώς απαιτείται ο υπολογισμός του τυπικού σφάλματος s_p ο οποίος με τη σειρά του απαιτεί γνώση της πιθανότητας εμφάνισης του χαρακτηριστικού στον πληθυσμό (άρα για ποιο λόγο να υπολογιστεί ένα διάστημα εμπιστοσύνης για αυτήν;) Στην πράξη εκτιμούμε την τιμή του s_p από τα δειγματικά στατιστικά ως $s_{\hat{p}}=\sqrt{\hat{p}\,\hat{q}/n}$ και παίρνουμε πως το διάστημα εμπιστοσύνης για αναλογία δίνεται από

$$\hat{p} - 1{,}96\, s_{\hat{p}} \le p \le \hat{p} + 1{,}96\, s_{\hat{p}} \text{ όπου } s_{\hat{p}} = \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

Τύπος 6: 95% διάστημα εμπιστοσύνης για το ποσοστό ενός χαρακτηριστικού στον πληθυσμό.

Παράδειγμα

Σε μία έρευνα αγοράς 100 ερωτώμενων βρέθηκε πως οι 15 από αυτούς θα αγόραζαν ένα προϊόν. Να βρεθεί 95% διάστημα εμπιστοσύνης για το πραγματικό ποσοστό ενδιαφερομένων για αγορά του προϊόντος στο σύνολο του πληθυσμού.

Απάντηση

Είναι $\hat{p}=\frac{15}{100}=0{,}15$ και $s_{\hat{p}}=\sqrt{\frac{\hat{p}\,\hat{q}}{n}}=\sqrt{\frac{0{,}15\cdot 0{,}85}{100}}=\frac{0{,}357}{10}=0{,}0357.$ Το 95% διάστημα εμπιστοσύνης για το άγνωστο ποσοστό ενδιαφερομένων για την αγορά του προϊόντων σε όλον τον πληθυσμό είναι

$$[0{,}15 - 0{,}0357,\ 0{,}15 + 0{,}0357] = [0.1143,\ 0.4143],$$

δηλαδή με άλλα λόγια στηριζόμενοι στο δείγμα των 100 ατόμων μπορούμε να

ισχυριστούμε σε επίπεδο εμπιστοσύνης 95% πως το άγνωστο ποσοστό ενδιαφερομένων στον πληθυσμό είναι από 11,4% έως 41,4%.

Παρατηρήσεις

1. Ο τύπος 6 για τον υπολογισμό του διαστήματος εμπιστοσύνης βασίζεται στην παραδοχή $B(n,p) \sim N(np, npq)$ η οποία ισχύει για μεγάλα πλήθη παρατηρήσεων n. Εναλλακτικά, έχουν προταθεί και άλλοι τρόποι υπολογισμού για το διάστημα εμπιστοσύνης, ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στο ανάλογο άρθρο της Wikipedia : http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

2. Αν το επιθυμητό επίπεδο εμπιστοσύνης για το διάστημα εμπιστοσύνης είναι 99% τότε οι τύποι 4 και 6 γράφονται

$$\bar{X} - 2{,}57\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 2{,}57\frac{s}{\sqrt{n}}$$

και

$$\hat{p} - 2{,}57\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + 2{,}57\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

αντίστοιχα.

Πίνακας 1.5: Υπολογισμός διαστήματος εμπιστοσύνης με υπολογιστή

Δεν υπάρχει δεσμευμένη συνάρτηση για το διάστημα εμπιστοσύνης. Μπορεί, να γίνει με συνδυασμό των συναρτήσεων AVERAGE(), STDEV() και SQRT().

Το διάστημα εμπιστοσύνης προκύπτει συμπληρωματικά σε κάθε έλεγχο υποθέσεων. Γι παράδειγμα αν x = c(10, 13, 13, 18, 12, 14) και ελέγχουμε την υπόθεση πως το δείγμα προέρχεται από πληθυσμό με μέση τιμή 13 τότε το t – test για ένα δείγμα **t.test(x, alternative='two.sided', mu=13.0, conf.level=.95)** έχει ως αποτέλεσμα

One Sample t-test

data: x

t = 1.4662, df = 9, p-value = 0.1766

alternative hypothesis: true mean is not equal to 13

95 percent confidence interval:

3.662498 56.737502

sample estimates:

mean of x

30.2

Από όπου καταλαβαίνουμε πως το 95% διάστημα εμπιστοσύνης είναι από 3,7 έως 56,7 (παρεμπιπτόντως, η υπόθεση πως το δείγμα προέρχεται από έναν πληθυσμό με μέση τιμή 13 δεν απορρίπτεται! t = 1.466, df = 9, p-value = 0.177)

1.9 Μέγεθος δείγματος

Για τον υπολογισμό του μεγέθους του δείγματος μίας έρευνας απαιτούνται οι εξής στατιστικές πληροφορίες

1. Τυπική απόκλιση του πληθυσμού σ την οποία την προσεγγίζουμε από τη δειγματική τυπική απόκλιση **s**, είτε προχωρούμε σε μία εμπειρική εκτίμηση.
2. Το επίπεδο εμπιστοσύνης το οποίο συνήθως ορίζεται στο 95% ή 99%.
3. Το μέγιστο αποδεκτό σφάλμα **E** που επιθυμούμε να έχει η εκτίμησή της μέσης τιμής του δείγματος από την πραγματική μέση τιμή του πληθυσμού.

Ορίζοντας το επίπεδο εμπιστοσύνης να είναι 95% από τον τύπο 4, σελίδα 26,

$$\bar{X} - 1{,}96\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 1{,}96\frac{s}{\sqrt{n}} \ ,$$

συμπεραίνουμε πως το σφάλμα E της δειγματικής μέσης τιμής από την μέση τιμή του πληθυσμού μ είναι

$$E = 1{,}96\frac{s}{\sqrt{n}} \Leftrightarrow \sqrt{n} = \frac{1{,}96\,s}{E} \Leftrightarrow n = \left(\frac{1{,}96\,s}{E}\right)^2 .$$

Από τον τύπο $n=\left(\frac{1{,}96\,s}{E}\right)^2$ συμπεραίνουμε πως το μέγεθος του δείγματος n είναι ανάλογο της διασποράς s^2 και αντιστρόφως ανάλογο του E^2.

Παράδειγμα

Ένας ερευνητής αναζητά το μέγεθος του ποσού που δίνεται για αγορά lipstick και επιθυμεί 95% διάστημα εμπιστοσύνης και απόλυτο σφάλμα πρόβλεψης (E) μικρότερο από 2€. Γνωρίζουμε από την προκαταρκτική έρευνα πως η διακύμανση του ποσού είναι 29€. (α) Ποιο πρέπει να είναι το μέγεθος του δείγματος; (β) Ποιο θα είναι το μέγεθος του δείγματος αν το επιθυμητό μέγιστο σφάλμα είναι 4;

Λύση

α) $n = \left(\frac{1{,}96\,s}{E}\right)^2 = \left(\frac{1{,}96\cdot 29}{2}\right)^2 = \left(\frac{56{,}8}{2}\right)^2 = 28{,}4^2 \simeq 808$.

β) $n = \left(\frac{1{,}96\,s}{E}\right)^2 = \left(\frac{1{,}96\cdot 29}{4}\right)^2 = \left(\frac{56{,}8}{4}\right)^2 = 14{,}2^2 \simeq 202$.

1.9.1 Μέγεθος δείγματος για αναλογία

Από τον τύπο 5, $\hat{p}-1{,}96\,s_p \le p \le \hat{p}+1{,}96\,s_p$ έχουμε πως

$$E = 1{,}96\sqrt{\frac{p\,q}{n}} \Leftrightarrow \sqrt{n} = \frac{1{,}96\sqrt{p\,q}}{E} \Leftrightarrow n = \left(\frac{1{,}96}{E}\right)^2 pq .$$

Καθώς το p δεν είναι δυνατό να το γνωρίζουμε αντικαθιστούμε το γινόμενο $pq=p(1-p)$ με τη μέγιστη τιμή που μπορεί αυτό να πάρει και είναι ίση με 1/4, παίρνοντας ως τελικό αποτέλεσμα

$$n = \frac{1}{4}\left(\frac{1{,}96}{E}\right)^2 .$$

Παράδειγμα

Αναζητούμε διάστημα εμπιστοσύνης για το ποσοστό των ανδρών σε μία κωμόπολη. (α) Πόσο μεγάλο πρέπει να είναι ένα δείγμα ώστε το 95% διάστημα εμπιστοσύνης που θα υπολογιστεί να μην είναι μεγαλύτερο από 3.5%; (β) Αν το επίπεδο εμπιστοσύνης είναι 99% τότε ποιο είναι το μέγεθος του δείγματος;

Λύση

(α) $n = \frac{1}{4}\left(\frac{1,96}{E}\right)^2 = \frac{1}{4}\left(\frac{1,96}{0,035}\right)^2 = \frac{1}{4}56^2 = \frac{3136}{4} = 784$.

(β) $n = \frac{1}{4}\left(\frac{2,57}{E}\right)^2 = \frac{1}{4}\left(\frac{2,57}{0,035}\right)^2 = \frac{1}{4}73,4^2 = \frac{5391,7}{4} \simeq 1348$.

1.10 Συμπεριφορά του ερευνητή κατά τη δειγματοληψία με ερωτηματολόγιο.

Στις προσωπικές συνεντεύξεις προέχει η περιγραφή της σημαντικότητας της έρευνας και της σπουδαιότητας της συμμετοχής του ερωτώμενου. Στις τηλεφωνικές συνεντεύξεις είναι απαραίτητη η δημιουργία κλίματος εμπιστοσύνης με τον ερωτώμενο και ως πρώτη ενέργεια είναι η πλήρης παρουσίαση των στοιχείων του ερευνητή. Τέλος στις διαδικτυακές έρευνες είναι καλό να έχει προηγηθεί μία αρχική επικοινωνία και η περιγραφή της έρευνας μέσω email.

Η συγκατάθεση του ερωτώμενου μπορεί να επιτευχθεί είτε με σταδιακή αποδοχή όπου πρώτα κερδίζεται η συγκατάθεση του υποψηφίου σε ένα μικρότερο αίτημα και μετά ο ερευνητής προχωρεί στο κύριο μέρος της έρευνας είτε με την “Door-in-the-Face Compliance” τεχνική όπου πρώτα ζητείται κάτι δύσκολο που το πιθανότερο είναι πως θα απορριφθεί και μετά κάτι μικρότερο που φαίνεται περισσότερο λογικό αίτημα από το πρώτο. Ο ερωτώμενος είναι πιθανό να συναινέσει σε αυτό και να συμμετάσχει στην έρευνα.

Οι ερωτήσεις πρέπει να γίνονται ακριβώς όπως είναι γραμμένες στο ερωτηματολόγιο. Η ανάγνωση των ερωτήσεων πρέπει να είναι καθαρή και προσεκτική ενώ οι ερωτήσεις πρέπει να τίθενται με την ίδια σειρά που εμφανίζονται. Αν κάποια ερώτηση δεν γίνεται κατανοητή τότε πρέπει αυτή να επαναλαμβάνεται. Τέλος πρέπει να γίνονται όλες οι ερωτήσεις του ερωτηματολογίου.

Σε περίπτωση μη απόκρισης ο ερευνητής πρέπει να ενθαρρύνει τον ερωτώμενο για περισσότερο πλήρη απάντηση στην ερώτηση. Ως τακτικές που διευκρινίζουν μία μη σαφής

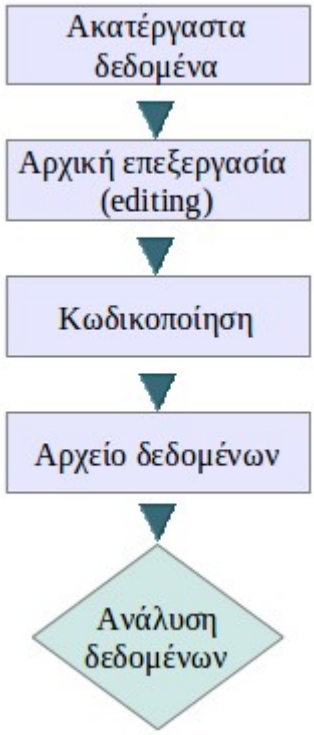
απόκριση μπορεί να χρησιμοποιηθούν οι εξής :

- Η επανάληψη της ερώτησης.
- Η σιωπή, δίνοντας την ευκαιρία για σκέψη στο συνομιλητή.
- Η επανάληψη της απόκρισης που μόλις δόθηκε.
- Η τοποθέτηση μίας “ουδέτερης” πρόσθετης ερώτησης.

Στις ερωτήσεις ανοικτής συμπλήρωσης οι αποκρίσεις πρέπει να καταγράφονται κατά τη διάρκεια της συνέντευξης και να μεταφέρονται αυτολεξεί, χωρίς σύνοψη ή παράφραση των λεγόμενων του ερωτώμενου. Επίσης πρέπει να συμπεριλαμβάνονται και οι ενθαρρύνσεις που έδωσε ο ερευνητής στον ερωτώμενο.

Ο τρόπος τερματισμού της συνέντευξης είναι σημαντικός. Πρέπει πρώτα να έχουν καταγραφεί όλες οι απαιτούμενες πληροφορίες, συμπεριλαμβανομένου και των πρόσθετων σχολίων που έκαναν οι ερωτώμενοι. Ο ερευνητής πρέπει να διευκρινίζει όλες τις απαραίτητες πληροφορίες για τη φύση της έρευνας που τους ζητούνται και να μην αφήνουν απορίες στους ερωτώμενους. Η αποφυγή μίας βιαστικής αναχώρησης είναι βασική ένδειξη ευγένειας εκ μέρους του ερευνητή. Ο ερευνητής, είναι σημαντικό να ευχαριστεί τον ερωτώμενο για το χρόνο που διέθεσε και τη συνεργασία που επέδειξε.

1.11 Ενέργειες πριν τη στατιστική ανάλυση



Τα ακατέργαστα δεδομένα, δηλαδή οι αποκρίσεις των ερωτώμενων, όπως ακριβώς έχουν καταγραφεί από τον ερευνητή ενδεχομένως να κρύβουν κάποια σφάλματα. Πριν από τη Στατιστική Ανάλυση είναι χρήσιμο να γίνει μία αρχική επεξεργασία.

Ως **πρωταρχική επεξεργασία (Editing)** περιγράφονται όλες οι ενέργειες που αποσκοπούν στον έλεγχο για πληρότητα, συνέπεια και ορθότητα των δεδομένων και την προετοιμασία των καταγραφών για πέρασμά τους σε ψηφιακό αρχείο.

Έλεγχος συνέπειας των δεδομένων είναι ο έλεγχος πως η δειγματική μονάδα πράγματι ανήκει στον πληθυσμό που θέλαμε

να περιγράψουμε. Επίσης στην αρχική επεξεργασία πρέπει να γίνει έλεγχος πως το δείγμα καλύπτει το πλαίσιο δειγματοληψίας που είχαμε αρχικά σχεδιάσει. Σε περιπτώσεις προφανούς σφάλματος πρέπει να διορθωθούν οι αποκρίσεις. Ένας Η/Υ μπορεί εύκολα να εντοπίσει παράλογους συνδυασμούς αποκρίσεων στις ερωτήσεις.

Ο *έλεγχος για πληρότητα* πρέπει να εντοπίσει τις αποκρίσεις που έμειναν μη συμπληρωμένες με αποτέλεσμα το ερωτηματολόγιο να μην είναι πλήρως συμπληρωμένο και ενδεχομένως μη αξιοποιήσιμο στα πλαίσια της έρευνας. Στην περίπτωση αυτή ενδεχομένως να πρέπει αν αποφασιστεί αν θα χρησιμοποιηθεί κάποια τιμή αντικατάστασης δηλαδή μία τιμή που αντικαθιστά τις απούσες τιμές. Ο τρόπος υπολογισμού της τιμής αντικατάστασης πρέπει να είναι προκαθορισμένος. Η αντικατάσταση είναι απαραίτητη να συμβεί ώστε να μεγιστοποιηθεί το πλήθος των περιπτώσεων από τις οποίες θα προκύψει το αποτέλεσμα και πραγματοποιείται με προκαθορισμένο στατιστικό τρόπο (π.χ. συμπλήρωση με τη μέση τιμή ή τη διάμεσο των υπολοίπων παρατηρήσεων)

1.12 Κωδικοποίηση των δεδομένων

Κωδικοποίηση (coding) ονομάζεται η αντιστοίχιση διαφορετικών αριθμητικών τιμών σε διαφορετικές τιμές της μεταβλητής. Στόχος είναι η μορφοποίηση των δεδομένων για την εύκολη διαχείρισή τους από το στατιστικό πρόγραμμα. Οι κωδικοί είναι αριθμοί που αντιστοιχούνται σε κάθε διαφορετική τιμή των δεδομένων. Επιπλέον, κατά τη κωδικοποίηση δημιουργούνται (αν είναι απαραίτητο) οι ψευδομεταβλητές (dummy variables) δηλαδή μεταβλητές που έχουν ως μόνο στόχο την ανεξάρτητη στατιστική επεξεργασία τιμών μίας ποιοτικής μεταβλητής. Αν η ποιοτική μεταβλητή έχει k διαφορετικές τιμές τότε απαιτούνται k – 1 ψευδομεταβλητές. Η χρήση ψευδομεταβλητών είναι εκτεταμένη στις μεθόδους παλινδρόμησης.

Είναι φανερό πως οι κωδικοί πρέπει να καλύπτουν όλες τις πιθανές περιπτώσεις και να είναι ανεξάρτητοι μεταξύ τους. Κάθε μία διαφορετική κατηγορία πρέπει να αντιστοιχείται σε μία και μόνο μία διαφορετική τιμή. Για την αποφυγή σφαλμάτων είναι καλό να γίνει μία πρωταρχική ανίχνευση όλων των διαφορετικών αποκρίσεων από το δείγμα και έπειτα να ακολουθήσει αντιστοίχιση με συγκεκριμένους αριθμούς.

Κεφάλαιο 2 Περιγραφή και παρουσίαση των στοιχείων του δείγματος.

Το πρώτο μέλημα ενός ερευνητή είναι να περιγράψει με όσο το δυνατόν περισσότερη ακρίβεια, σαφήνεια και καθαρότητα τα δεδομένα τα οποία συνέλεξε. Ο τρόπος και οι

μέθοδοι που θα χρησιμοποιηθούν για την περιγραφή αυτή εξαρτάται από το είδος των μεταβλητών. Συνοπτικά, στους παρακάτω πίνακες παρουσιάζονται τα βασικά μέτρα και γραφήματα που μπορούν να χρησιμοποιηθούν για την παρουσίαση των τιμών των μεταβλητών.

Πίνακας 2.1: Στατιστικά μέτρα και διαγράμματα για την περιγραφή μίας μεταβλητής

Είδος Μεταβλητής	**Προτεινόμενα Υπολογιστικά Μέτρα**		**Προτεινόμενα Γραφήματα**
Ποιοτική (όπως χρώμα ματιών, φύλο κ.α.)	Πίνακας Συχνοτήτων		Ραβδόγραμμα
			Κυκλικό Διάγραμμα
Ποσοτική (όπως ύψος, βάρος κ.α.)	Μέτρα θέσης	Επικρατούσα Τιμή	Ιστόγραμμα και Πολύγωνο Συχνοτήτων (Για διακριτές ποσοτικές με “λίγες” τιμές είναι αποδεκτό επίσης το ραβδόγραμμα και το κυκλικό διάγραμμα)
		Μέση Τιμή	
		Διάμεση Τιμή	
	Μέτρα διασποράς	Εύρος	
		Διακύμανση	
		Τυπική Απόκλιση	
		Απόλυτη Απόκλιση	

Πίνακας 2.2: Στατιστικά μέτρα και διαγράμματα για την περιγραφή δύο μεταβλητών

Είδος Μεταβλητής	**Προτεινόμενα Υπολογιστικά Μέτρα**	**Προτεινόμενα Γραφήματα**
Ποιοτικές	Διμεταβλητός πίνακας συχνοτήτων, Συντελεστής φ	Ραβδόγραμμα Στοίβας
Ποσοτικές	Συντελεστής συσχέτισης Pearson (Συνεχείς ποσοτικές, π.χ. ύψος και βάρος)	Διάγραμμα Διασποράς (Scatterplot)
	Συντελεστής συσχέτισης Spearman (Διακεκριμένες)	

2.1 Πίνακας Συχνοτήτων

Ένας πίνακας συχνοτήτων χρησιμοποιείται για να παρουσιάσει με συνοπτικό τρόπο

- παρατηρήσεις ποιοτικών μεταβλητών οι οποίες επαναλαμβάνονται.
- παρατηρήσεις ποσοτικών μεταβλητών οι οποίες έχουν ομαδοποιηθεί.

Αποτελείται από τις εξής στήλες :

1. Στήλη τιμών x_i ή κλάσης αριθμών [α, β)
2. Κέντρο της κλάσης αν η μεταβλητή είναι συνεχής ποσοτική.
3. Στήλη συχνότητας v_i
4. Στήλη σχετικής συχνότητας f_i
5. Στήλη αθροιστικής συχνότητας N_i
6. Στήλη αθροιστικής σχετικής συχνότητας F_i

Συχνότητα μίας τιμής x_i ονομάζεται το πλήθος των παρατηρήσεων στο δείγμα μας. Η σχετική συχνότητα f_i είναι απλά το ποσοστό της εμφάνισης της τιμής στο δείγμα μας. Πιο συγκεκριμένα $f_i = \frac{v_i}{N}$. Η αθροιστική συχνότητα προκύπτει από το άθροισμα όλων των προηγούμενων συχνοτήτων, πιο συγκεκριμένα $N_i = v_1 + v_2 + \cdots + v_i$ ενώ η αθροιστική σχετική συχνότητα είναι ο λόγος της αθροιστικής συχνότητας προς το σύνολο των παρατηρήσεων, πιο συγκεκριμένα $F_i = \frac{N_i}{N}$.

Δραστηριότητα

Ρωτήθηκαν 20 γυναίκες για το πλήθος των παιδιών που έχουν και έδωσαν τις παρακάτω αποκρίσεις : 0, 0, 1, 2, 0, 0, 1, 2, 1, 1, 1, 1, 2, 2, 0, 4, 2, 3, 1, 0. Να συμπληρωθεί ο πίνακας συχνοτήτων των παραπάνω παρατηρήσεων.

Πλήθος (x_i)	**Συχνότητα (v_i)**	**Σχετική Συχνότητα (f_i)**	**Αθροιστική Συχνότητα (N_i)**	**Αθροιστική Σχετική Συχνότητα (F_i)**

Πίνακας 2.3: Συμπλήρωση πίνακα συχνοτήτων αριθμητικής μεταβλητής στον υπολογιστή

Υποθέτουμε πως τα αριθμητικά δεδομένα είναι στα κελιά A1:A50

Βήμα 1ο : Ορισμός “με το χέρι” των επιθυμητών ορίων των διαστημάτων σε ένα μέρος του φύλλου εργασίας. Υποθέτουμε πως τοποθετούμε στα κελιά B1:B4 τους αριθμούς 10, 20, 30, 40.

Βήμα 2ο : Σε μία άλλη στήλη επιλέγουμε 5 κελιά (1 περισσότερα από το πλήθος των διαστημάτων) και με τον οδηγό συνάρτησης εισάγουμε τη συνάρτηση – πίνακα FREQUENCY με ορίσματα A1:A50 και B1:B4. Τα διαστήματα είναι της μορφής (-∞, 10], (10, 20], (20, 30], (30, 40] και (40, ∞)

Η συμπλήρωση πίνακα συχνοτήτων ποιοτικής μεταβλητής περιγράφεται στην παράγραφο 2.3, σελίδα 45.

Η συνάρτηση **table(x)** είναι η λύση για την περίπτωση που δεν απαιτείται ομαδοποίηση.

Η συνάρτηση **hist(x, plot=FALSE)** αποδίδει πίνακα συχνοτήτων με αυτόματο προσδιορισμό των διαστημάτων ενώ η **hist(x, breaks = c(0, 10,20, 30), plot=FALSE)** ορίζει τα διαστήματα [0,10], (10, 20], (20, 30] . Προφανώς, αν plot = TRUE ή αν ακόμα δεν εμφανίζεται η διευκρίνηση αυτή τότε δημιουργείται το ιστόγραμμα!

2.2 Μέτρα Θέσης ή Κεντρικής Τάσης

Τα κυριότερα μέτρα κεντρικής τάσης είναι η μέση τιμή, η διάμεση τιμή και η επικρατούσα τιμή. Επιπλέον, η μέση τιμή συνήθως αναφέρεται στον αριθμητικό μέσο.

2.2.1 Επικρατούσα τιμή

Η επικρατούσα τιμή σε ένα σύνολο δεδομένων είναι απλά η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης. Όταν δύο οι περισσότερες τιμές συμπίπτουν στη συχνότητα τότε ονομάζονται όλες επικρατούσες τιμές. Για παράδειγμα η επικρατούσα τιμή του δείγματος 1, 3, 3, 4, 5, 5, 6, 2, 3, 4, 3, 1, 5 είναι το 3.

Δραστηριότητα

Να βρεθεί η επικρατούσα τιμή του δείγματος 12, 12, 7, 0, 0, 3, 5, 4, 4, 3, 0, 12.

Πίνακας 2.4: Υπολογισμός επικρατούσης τιμής στον υπολογιστή

Συνάρτηση **MODE()**. Αν υπάρχουν περισσότερες από μία τότε επιστρέφει τη μικρότερη (σε αντίθεση με το Excel που επιστρέφει τη μεγαλύτερη)

Η συνάρτηση **which.max(table(x))** επιστρέφει την επικρατούσα τιμή ενός διανύσματος x (ή τη μικρότερη αυτών αν είναι περισσότερες από δύο).

Μία πληρέστερη λύση είναι η χρήση της συνάρτησης

```
smode<-function(x){
    xtab<-table(x)
    modes<-xtab[max(xtab)==xtab]
    mag<-as.numeric(modes[1]) #in case mult. modes, this is safer
    themodes<-names(modes)
    mout<-list(themodes=themodes,modeval=mag)
    return(mout)
    }
```

η οποία επιστρέφει όλες τις επικρατούσες και την αντίστοιχη μέγιστη συχνότητα εμφάνισης.

(πηγή : https://stat.ethz.ch/pipermail/r-help/2011-March/273569.html)

2.2.2 Διάμεση τιμή

Η διάμεση τιμή σε ένα σύνολο δεδομένων είναι απλά η μεσαία παρατήρηση αν το πλήθος των στοιχείων είναι μονό ενώ είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων αν το πλήθος των στοιχείων είναι ζυγό. Για να βρούμε τη διάμεσο κάνουμε τα εξής βήματα :

1. Ταξινομούμε τις παρατηρήσεις από τη μικρότερη στη μεγαλύτερη
2. Η μεσαία παρατήρηση βρίσκεται στη θέση (ν + 1) / 2. Αν το (ν+1) / 2 είναι ακέραιος τότε η διάμεσος είναι η παρατήρηση που βρίσκεται στη θέση αυτή , ενώ αν είναι δεκαδικός τότε παίρνουμε το ημιάθροισμα των δύο παρατηρήσεων που βρίσκονται στις γειτονικές θέσεις.

Για παράδειγμα η διάμεσος των 23, 4, 16 είναι το 16 ενώ η διάμεσος των παρατηρήσεων 23, 10, 13, 7 είναι το (10 + 13) / 2 = 11,5.

Πίνακας 2.5: Υπολογισμός διάμεσης τιμής στον υπολογιστή

	Συνάρτηση **MEDIAN()**.
	Συνάρτηση **median(x)**. Αν **x = c(10, 20, 30, 40)** τότε **median(x) = 25**. Μπορεί να χρησιμοποιηθεί και η συνάρτηση **summary(x)**

Δραστηριότητα

Οι απουσίες των 20 σπουδαστών ενός τμήματος είναι 0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 2, 2. Να βρεθούν (α) Η επικρατούσα τιμή (β) Η διάμεση τιμή των απουσιών.

2.2.3 Αριθμητικός Μέσος – Σταθμισμένος αριθμητικός μέσος

Αν $x_{1,}x_{2,}\cdots,x_{v}$ είναι οι παρατηρήσεις του δείγματός μας τότε ο αριθμητικός μέσος είναι ο $\bar{x} = \frac{1}{v}\sum_{i=1}^{v} x_i$ όπου ν το μέγεθος του δείγματος ενώ αν τα στοιχεία $x_{1,}x_{2,}\cdots,x_{v}$ είναι όλος ο πληθυσμός για το οποίο γίνεται η έρευνα τότε γράφουμε $\mu = \frac{1}{v}\sum_{i=1}^{v} x_i$.

Για παράδειγμα αν ένα τμήμα 10 σπουδαστών έχει βαθμολογία στη Στατιστική 12, 15, 10, 18, 17, 19, 15, 20, 13, 15 τότε η μέση βαθμολογία του τμήματος $\bar{x}$ είναι

$$\bar{x} = \frac{12+15+10+18+17+19+15+20+13+15}{10} = 15{,}4 \ .$$

Σημείωση : Αν οι 10 σπουδαστές ήταν όλος ο πληθυσμός της έρευνας μας τότε θα μπορούσαμε να συμβολίσουμε τη μέση τιμή με μ και όχι με $\bar{x}$.

Η παραπάνω μορφή του μέσου ονομάζεται αστάθμητος μέσος αριθμητικός. Εκτός από αυτόν υπάρχει και ο σταθμικός μέσος αριθμητικός ο οποίος χρησιμοποιείται κυρίως στην κατασκευή αριθμοδεικτών. Σε κάθε παρατήρηση x_i αντιστοιχούμε ένα βάρος w_i και ο τύπος του είναι

$$\bar{x}_w = \frac{1}{\sum_{i=1}^{v} w_i}\sum_{i=1}^{v} w_i x_i$$

Για παράδειγμα αν ένας μαθητής γράψει 18 στα Μαθηματικά, 13 στη Βιολογία και 15 στην έκθεση και οι συντελεστές βαρύτητας για τα μαθήματα είναι 1.15, 0,75 και 0.9 αντίστοιχα τότε η βαθμολογία του μαθητή είναι

$$\bar{x} = \frac{18\cdot 1.15+13\cdot 0.75+15\cdot 0.9}{1.15+0.75+0.9}=\frac{43.95}{2.8} = 15{,}7$$

Τέλος αν έχουμε ομαδοποιημένα δεδομένα σε κλάσεις και x_i είναι το κέντρο της i κλάσης τότε $\bar{x} = \frac{1}{v}\sum_{i=1}^{v} f_i x_i$, όπου f_i είναι η συχνότητα στην κλάση i. Εννοείται βέβαια, πως για να έχει σημαντικό νόημα ο παραπάνω υπολογισμός πρέπει να είναι αποδεκτή η παραδοχή πως το κέντρο της κλάσης προσεγγίζει ικανοποιητικά το σύνολο όλων των παρατηρήσεων που ανήκουν στην κλάση αυτή. Για παράδειγμα, από τον πίνακα 2.6 στον οποίο εμφανίζονται οι βαθμολογίες των 20 σπουδαστών ενός τμήματος μπορούμε να υπολογίσουμε τη μέση βαθμολογία του τμήματος να είναι

$$\bar{x} = \frac{3\cdot 2.5+4\cdot 7.5+8\cdot 12.5+5\cdot 17.5}{3+4+8+5}=\frac{225}{20} = 11{,}25 \quad .$$

Ο αριθμητικός μέσος $\bar{x} = \frac{1}{v}\sum_{i=1}^{v} x_i$ των παρατηρήσεων $x_{1,}x_{2,}\cdots,x_v$ έχει την ιδιότητα να επηρεάζεται δυσανάλογα από τις πολύ μεγάλες ή τις πολύ μικρές παρατηρήσεις του δείγματός μας. Πράγματι, η πρόσθεση ενός πολύ μεγάλου αριθμού στον αριθμητή του κλάσματος που ορίζει τη μέση τιμή θα τον αυξήσει δυσανάλογα σε σχέση με την αύξηση στον παρονομαστή η οποία θα είναι μόνο μία μονάδα!

Παρατήρηση : Είναι φανερό πως αν είχαμε πρόσβαση στις μεμονωμένες βαθμολογίες κάθε σπουδαστή θα προτιμούσαμε να υπολογίσουμε τη μέση βαθμολογία με τον απλό αριθμητικό μέσο.

Πίνακας 2.6: Βαθμολογίες τμήματος

Κλάση	Όρια κλάσης	Συχνότητα (f_i)	Κέντρο κλάσης (x_i)	$f_i \cdot x_i$
1	0-5	3	2,5	7,5
2	5-10	4	7,5	30
3	10-15	8	12,5	100
4	15-20	5	17,5	87,5
Άθροισμα		n = 20		225

Πίνακας 2.7: Υπολογισμός αριθμητικού μέσου στον υπολογιστή

	Συνάρτηση **AVERAGE().**
	Συνάρτηση **mean(x)**. Μπορεί να χρησιμοποιηθεί και η συνάρτηση **summary(x)**

Δραστηριότητες

1. Με τα δεδομένα του παραδείγματος 1 της σελίδας 37 να υπολογίσετε το μέσο πλήθος παιδιών του δείγματος των 20 γυναικών : (α) από τα ίδια τα δεδομένα (β) από τον πίνακα συχνοτήτων.
2. Οι μισθοί των εργαζομένων μίας εταιρείας ακολουθούν την παρακάτω κατανομή (ευρώ)

Κλάση	400 έως 600	600 έως 800	800 έως 1000	1000 έως 1200	Σύνολο
Συχνότητα	30	22	25	10	87

Να υπολογιστεί ο μέσος μισθός των εργαζομένων στην εταιρεία.

2.2.4 Αρμονικός μέσος

Ο αρμονικός μέσος των παρατηρήσεων $x_{1,}x_{2,}\cdots,x_{v}$ ορίζεται να είναι η ποσότητα

$$\bar{x}_h = \frac{v}{\sum_{i=1}^{v}\frac{1}{x_i}} .$$

Για παράδειγμα ο αρμονικός μέσος των αριθμών 20 και 10 είναι $\bar{x}_h = \frac{2}{\frac{1}{20}+\frac{1}{10}} \simeq 13{,}3$.

Ο αρμονικός μέσος είναι το κατάλληλο στατιστικό για τον υπολογισμό της “μέσης” τιμής όταν οι παρατηρήσεις εκφράζουν ρυθμούς μεταβολής. Για παράδειγμα, αν ένα όχημα ταξιδεύει μία συγκεκριμένη απόσταση με ταχύτητα α (π.χ. 60 km/h) και μετά την ίδια απόσταση με 40 km/h τότε η μέση του ταχύτητα είναι ο αρμονικός μέσος των 40 και 60 (περίπου 48 km/h) κάτι που σημαίνει πως αν ταξίδευε με αυτήν την ταχύτητα από την αρχή του ταξιδιού μέχρι το τέλος θα έκανε τον ίδιο ακριβώς χρόνο.

Ωστόσο, αν το αυτοκίνητο ταξίδευε για μία ώρα με 60 km/h και μετά για άλλη μία ώρα με

40 km/h τότε η μέση ταχύτητα στη διάρκεια των δύο ωρών θα πρέπει να υπολογιστεί από τον αριθμητικό μέσο και είναι 50 km/h.

Η ίδια αρχή χρησιμοποιείται και για περισσότερα δρομολόγια. Αν κάθε δρομολόγιο καλύπτει την ίδια απόσταση τότε η μέση ταχύτητα είναι ο αρμονικός μέσος των επιμέρους ταχυτήτων, ενώ αν κάθε δρομολόγιο καλύπτει τα ίδια χιλιόμετρα τότε η μέση ταχύτητα είναι ο αριθμητικός μέσος των επιμέρους ταχυτήτων!

Περισσότερα παραδείγματα στη φυσική και στις άλλες επιστήμες μπορούν να βρεθούν στην Wikipedia : http://en.wikipedia.org/wiki/Harmonic_mean

Πίνακας 2.8: Υπολογισμός αρμονικού μέσου στον υπολογιστή

 Συνάρτηση **HARMEAN()**
 Δεν υπάρχει δεσμευμένη συνάρτηση για τον αρμονικό μέσο, ωστόσο μπορεί εύκολα να υπολογιστεί από τη συνάρτηση **1/mean(1/x)**. Αν **x = c(10, 20, 30, 40)** τότε **1/mean(1/x) = 19,2.**

Δραστηριότητες

1. Αν μία αντλία με ρεύμα αντλεί από μία δεξαμενή μεγέθους 10.000 λίτρα, νερό με ρυθμό 30 λίτρα/λεπτό ενώ μία αντλία με μπαταρία αντλεί νερό με ρυθμό 50 λίτρα/λεπτό να βρεθεί σε πόσο χρόνο μπορούν και οι δύο αντλίες να αδειάσουν τη δεξαμενή
2. Αν ένα τρακτέρ Α οργώνει ένα χωράφι σε 5 ημέρες, ένα τρακτέρ Β σε 8 ημέρες και ένα τρακτέρ Γ σε 4 ημέρες να βρεθεί ο χρόνος που απαιτείται για να οργωθεί το χωράφι : (α) από τα τρακτέρ Α και Γ (β) από τα τρακτέρ Α, Β και Γ.

2.2.5 Γεωμετρικός μέσος

Ο γεωμετρικός μέσος των παρατηρήσεων $x_{1,}x_{2,}\cdots,x_{v}$ ορίζεται να είναι η ποσότητα $\bar{x}_g = \sqrt[v]{x_1 \cdot x_2 \cdots x_v}$. Για παράδειγμα ο γεωμετρικός μέσος των αριθμών 3 και 5 είναι

$\bar{x}_g = \sqrt{3 \cdot 5} \simeq 3{,}9$ ενώ ο γεωμετρικός μέσος των αριθμών 3, 5 και 8 είναι

$\bar{x}_g = \sqrt[3]{3 \cdot 5 \cdot 8} \simeq 4{,}9$.

Ο γεωμετρικός μέσος χρησιμοποιείται για τον υπολογισμό της μέσης τιμής ποσοστιαίων

μεταβολών. Για παράδειγμα ας υποθέσουμε πως μία μετοχή που αξίζει 1 ευρώ κερδίζει τον πρώτο χρόνο 10% (δηλαδή το 1 ευρώ γίνεται 1,1), το δεύτερο χρόνο 15% (δηλαδή το 1 ευρώ γίνεται 1,15) και τον τρίτο χρόνο 20% (δηλαδή το 1 ευρώ γίνεται 1,2). Στο παρακάτω σχήμα φαίνεται πως στο τέλος του τρίτου χρόνου το 1 ευρώ θα έχει γίνει 1,518 δηλαδή θα έχουμε συνολική αύξηση 0,518 ευρώ.

$$1 \xrightarrow{\text{επί } 1,1} 1,1 \xrightarrow{\text{επί } 1,15} 1,265 \xrightarrow{\text{επί } 1,2} 1,518$$

Ο υπολογισμός του αριθμητικού μέσου των επιδόσεων είναι 15% ενώ ο υπολογισμός του γεωμετρικού μέσου είναι $\bar{X}_g = \sqrt[3]{1,1 \cdot 1,15 \cdot 1,2} \simeq 1,149$. Δηλαδή, ο αριθμητικός μέσος υπερεκτιμά το τελικό αποτέλεσμα.

Εφαρμογή του 15% για τρία χρόνια :

$$1 \xrightarrow{\text{επί } 1,15} 1,15 \xrightarrow{\text{επί } 1,15} 1,323 \xrightarrow{\text{επί } 1,15} 1,521$$

Εφαρμογή του 14,9% για τρία χρόνια

$$1 \xrightarrow{\text{επί } 1,149} 1,149 \xrightarrow{\text{επί } 1,149} 1,320 \xrightarrow{\text{επί } 1,149} 1,518$$

Πίνακας 2.9: Υπολογισμός γεωμετρικού μέσου στον υπολογιστή

	Συνάρτηση **GEOMEAN()**
	Δεν υπάρχει δεσμευμένη συνάρτηση για το γεωμετρικό μέσο, ωστόσο μπορεί εύκολα να υπολογιστεί από τη συνάρτηση **prod(x)^(1/length(x))**. Αν **x = c(1.1, 1.15, 1.2)** τότε **prod(x)^(1/length(x))= 1,149**

Δραστηριότητες

1. Μία μετοχή κερδίζει τον πρώτο χρόνο 10%, χάνει το δεύτερο χρόνο 5% και κερδίζει τον τρίτο χρόνο 15%. Να βρεθεί η συνολική απόδοση της μετοχής στο διάστημα των τριών ετών.
2. Στους υπαλλήλους μίας επιχείρησης δόθηκε αύξηση μισθού 9% τον πρώτο χρόνο και 5% στον δεύτερο χρόνο. Ποια ήταν η μέση αύξηση του μισθού στο διάστημα των δύο ετών;

15	**Μέτρα Θέσης**			
16	Στατιστικά	Επικρατούσα Τιμή	Διάμεση τιμή	Μέση τιμή
17	**Τιμή**	**12**	**14**	**14,67**
18	**Συνάρτηση**	**MODE(C4:K4)**	**MEDIAN(C4:K4)**	**AVERAGE(C4:K4)**
19				
20	**Μέτρα Διασποράς**			
21	Στατιστικά	Εύρος	Ενδοτεταρτημοριακό Εύρος	Απόλυτη Απόκλιση
22	**Τιμή**	**9**	**5**	**2,52**
23	**Συνάρτηση**	**MAX(C4:K4)-MIN(C4:K4)**	**QUARTILE(C4:K4;3)-QUARTILE(C4:K4;1)**	**AVEDEV(C4:K4)**
24				
25	**Μέτρα Διασποράς (συνέχεια)**			
26	**Στατιστικά**	Διακύμανση	Τυπική Απόκλιση	Διακύμανση πληθυσμού	Τυπική Απόκλιση πληθυσμού
27	**Τιμή**	**9,5**	**3,08**	**8,44**	**2,91**
28	**Συνάρτηση**	**VAR(C4:K4)**	**STDEV(C4:K4)**	**VARP(C4:K4)**	**STDEVP(C4:K4)**

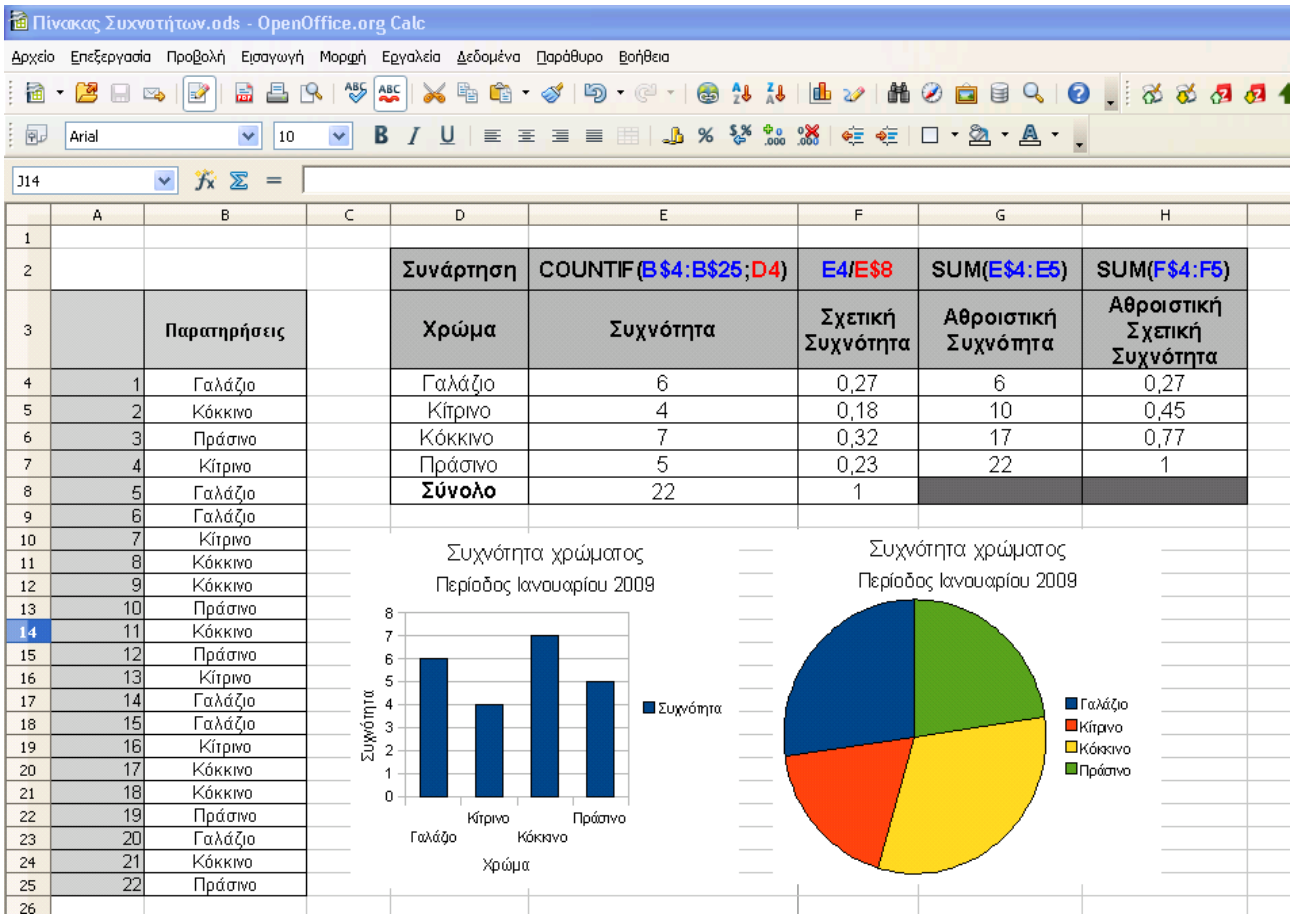
Εικόνα 2: Υπολογισμός περιγραφικών στατιστικών με το Calc

2.3 Συμπλήρωση πίνακα συχνοτήτων ποιοτικής μεταβλητής με το Calc

Ο πίνακας συχνοτήτων είναι ο κατάλληλος πίνακας για την περιγραφή μιας ποιοτικής μεταβλητής ή μιας αριθμητικής διακριτής μεταβλητής. Επιπλέον, είναι απαραίτητο βήμα για τη δημιουργία ραβδογράμματος ή κυκλικού διαγράμματος (απαιτείται ο υπολογισμός της στήλης με τις συχνότητες). Τέλος, είναι δυνατό να χρησιμοποιηθεί και για την περιγραφή των τιμών μιας συνεχής αριθμητικής μεταβλητής μόνο που στην περίπτωση αυτή πρέπει πρώτα να γίνει η ταξινόμηση των τιμών σε κλάσεις.

Η συμπλήρωση ενός πίνακας συχνοτήτων είναι απλή υπόθεση αρκεί να υπάρχουν οι παρατηρήσεις τοποθετημένες σε μία στήλη ή μία γραμμή του Calc και να γνωρίζει ο χρήστης τις συναρτήσεις που πρέπει να χρησιμοποιήσει για τη συμπλήρωση κάθε στήλης του πίνακα συχνοτήτων.

Στην εικόνα 3 φαίνεται ένα χαρακτηριστικό παράδειγμα πίνακα συχνοτήτων των δεδομένων που βρίσκονται στη στήλη **B**. Η σειρά με την οποία εμφανίζονται οι στήλες στον πίνακα συχνοτήτων είναι ενδεικτική. Οι συναρτήσεις του Calc από τις οποίες προήλθαν οι καταχωρήσεις της αντίστοιχης στήλης παρουσιάζονται στην πρώτη γραμμή του πίνακα.



Εικόνα 3: Δημιουργία Πίνακα Συχνοτήτων - Ραβδόγραμμα - Κυκλικό διάγραμμα

Περιγραφή του πίνακα συχνοτήτων : Στην πρώτη στήλη "**Χρώμα**" του πίνακα συχνοτήτων τοποθετούνται οι διαφορετικές τιμές της μεταβλητής. Η ανίχνευση των διαφορετικών τιμών γίνεται με το χέρι! (η αλφαβητική ταξινόμηση στη στήλη **B** με τα δεδομένα βοηθάει πολύ στον εντοπισμό όλων των διαφορετικών τιμών). Επιπλέον, είναι επιβεβλημένη η αλφαβητική τοποθέτηση των τιμών στον πίνακα συχνοτήτων. Αυτό μπορεί να συμβεί χρησιμοποιώντας την επιλογή ταξινόμησης (ή το εικονίδιο  της εργαλειοθήκης) .μετά τον εντοπισμό και την τοποθέτηση όλων των τιμών, επιλέγοντας πρώτα τα κελιά με τις διαφορετικές τιμές (D4:D7 στην περίπτωση του πίνακα της εικόνας 3).

Η στήλη "**Συχνότητα**" περιέχει την συχνότητα κάθε τιμής δηλαδή τις επαναλήψεις κάθε τιμής στο σύνολο της στήλης **B**. Για τον υπολογισμό της συχνότητας κάθε τιμής αρκεί να εφαρμοστεί η συνάρτηση "**=COUNTIF(B$4:B&25;D4)**" στο πρώτο κελί (**E4** στην εικόνα). Μετά η εφαρμογή της συνάρτησης επεκτείνεται αυτόματα στα υπόλοιπα τρία κελιά είτε με

Αντιγραφή – Επικόλληση είτε με αυτόματη μεταφορά “πιάνοντας” το κάτω δεξιά άκρο του κελιού **E4** και σέρνοντας το μέχρι να καλύψει και το **E7**. Προσέξτε τη χρήση του δολαρίου ($) στον ορισμό της περιοχής των δεδομένων στη συνάρτηση COUNTIF() κάτι το οποίο σταθεροποιεί την περιοχή αναζήτησης κατά τη μεταφορά της συνάρτησης από το κελί **E4** στα **E5**, **E6** και **E7**.

2.4 Απλά διαγράμματα

Το ραβδόγραμμα, το κυκλικό διάγραμμα και το ιστόγραμμα είναι τρία απλά στατιστικά διαγράμματα τα οποία μπορούν να γίνουν εύκολα με τη βοήθεια ενός υπολογιστή τσέπης.

2.4.1 Ραβδόγραμμα

Ένα ραβδόγραμμα είναι ένα πολύ απλό διάγραμμα το οποίο αναπαριστά τις συχνότητες των διαφορετικών τιμών μίας ποιοτικής ή ποσοτικής μεταβλητής με σχετικά λίγες τιμές. Σχεδιάζεται σε ορθογώνιο σύστημα αξόνων. Στον οριζόντιο άξονα και σε ίσες αποστάσεις τοποθετούνται οι διαφορετικές τιμές της μεταβλητής. Ορθογώνιες ράβδοι με κέντρο της βάσης τα σημεία αυτά σχεδιάζονται σε κάθε μία από αυτές τις τιμές με ύψος όση και η συχνότητα της αντίστοιχης τιμής. Αν ζητείται ραβδόγραμμα σχετικών συχνοτήτων τότε απλά ορίζουμε ως ύψος των ράβδων τη σχετική συχνότητα της τιμής αντί της απλής συχνότητας.

2.4.2 Κυκλικό διάγραμμα

Ένα κυκλικό διάγραμμα (pie chart) χρησιμοποιείται για τη γραφική αναπαράσταση τόσο των ποιοτικών όσο και των ποσοτικών δεδομένων, όταν οι διαφορετικές τιμές της μεταβλητής είναι σχετικά λίγες. Το κυκλικό διάγραμμα είναι ένας κυκλικός δίσκος ο οποίος χωρίζεται σε κυκλικούς τομείς, με μεγέθη ανάλογα προς τις αντίστοιχες συχνότητες ν_κ των διαφορετικών τιμών x_κ της μεταβλητής. Το μέγεθος α_κ της γωνίας κάθε ενός κυκλικού τομέα υπολογίζεται από τον τύπο

$$\alpha_\kappa = \frac{\nu_\kappa}{\nu} \, 360^o \ .$$

Παράδειγμα

Ρωτήθηκαν 20 γυναίκες για το πλήθος των παιδιών που έχουν και έδωσαν τις παρακάτω

αποκρίσεις : 0, 0, 1, 2, 0, 0, 1, 2, 1, 1, 1, 1, 2, 2, 0, 4, 2, 3, 1, 0. Ο συμπληρωμένος πίνακας συχνοτήτων των παραπάνω παρατηρήσεων ακολουθεί.

Πλήθος (x_i)	**Συχνότητα (ν_i)**	**Σχετική Συχνότητα (f_i)**
0	**6**	**0,3**
1	**7**	**0,35**
2	**5**	**0,25**
3	**1**	**0,05**
4	**1**	**0,05**
Σύνολο	**20**	**1**

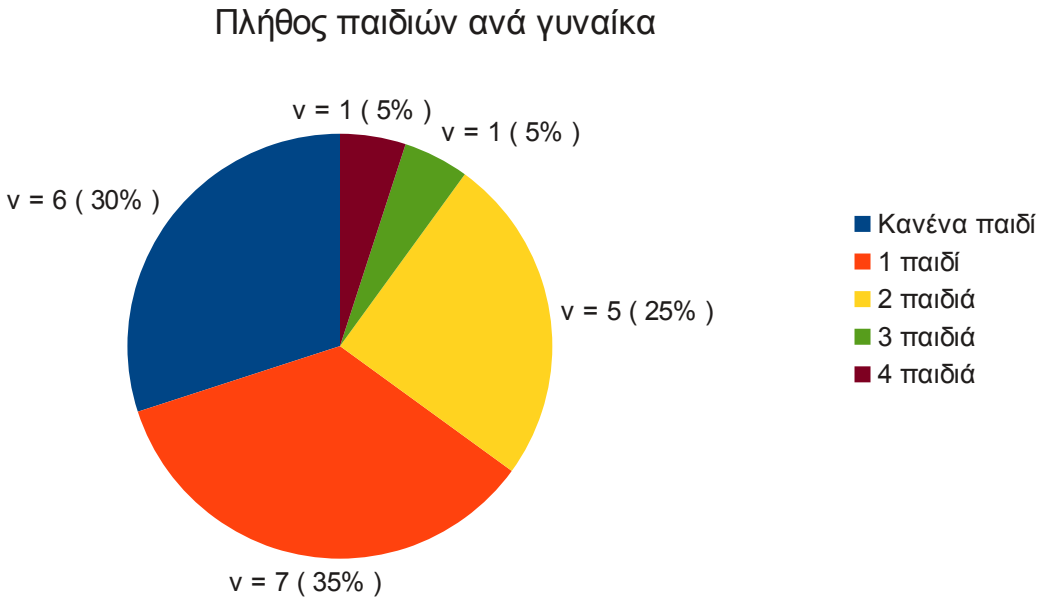
Ο υπολογισμός των γωνιών για κάθε μία διαφορετική κατηγορίας δίνει τα εξής

αποτελέσματα $\alpha_1 = \frac{\nu_1}{\nu} 360^o = \frac{6}{20} 360^o = 0{,}3 \cdot 360^o = 108^o$,

$$\alpha_2 = \frac{\nu_2}{\nu} 360^o = \frac{7}{20} 360^o = 0{,}35 \cdot 360^o = 126^o$$

$\alpha_3 = \frac{\nu_3}{\nu} 360^o = \frac{5}{20} 360^o = 0{,}25 \cdot 360^o = 90^o$, $\alpha_5 = \alpha_4 = \frac{1}{20} 360^o = 0{,}05 \cdot 360^o = 18^o$

Το κυκλικό διάγραμμα είναι το παρακάτω.



2.4.3 Ιστόγραμμα

Το ιστόγραμμα είναι ο κατάλληλος τύπος διαγράμματος για την περιγραφή των τιμών μίας ποσοτικής (αριθμητικής) μεταβλητής. Για τη δημιουργία ενός ιστογράμματος ακολουθούμε τα εξής βήματα :

1ο βήμα : Πρώτα διαχωρίζουμε τις παρατηρήσεις σε ομάδες (ή κλάσεις) ίσου πλάτους. Για το πλήθος των ομάδων δεν υπάρχει κάποιος κοινά αποδεκτός κανόνας. Οι περισσότεροι συγγραφείς προτείνουν τη δημιουργία έως και 10 ομάδων όταν οι παρατηρήσεις μας δεν ξεπερνούν τις 100 και έως 20 ομάδες όταν οι παρατηρήσεις μας είναι περισσότερες από 100. Σε κάποια συγγράμματα εμφανίζεται ο τύπος

Πλήθος ομάδων που πρέπει να διαχωριστούν οι παρατηρήσεις = 1 + 3,3 log(ν)

όπου ν το πλήθος των παρατηρήσεων. Την απόφαση για το πλήθος των ομάδων την παίρνει ο ίδιος ο ερευνητής.

2ο βήμα : Μόλις αποφασιστεί το πλήθος των ομάδων (ή κλάσεων) τότε υπολογίζεται το εύρος κάθε μίας από αυτές με μία απλή διαίρεση του συνολικού εύρους των παρατηρήσεων με το πλήθος των κλάσεων που αποφασίσαμε. Καθώς, το αποτέλεσμα της διαίρεσης είναι απίθανο να είναι κάποιος στρογγυλός αριθμός, συνηθίζεται να τον στρογγυλοποιούμε προς τα πάνω στην πλησιέστερη “βολική” ποσότητα. Οι ομάδες καταγράφονται με τη μορφή διαστημάτων [α, β).

3ο βήμα : Προχωρούμε στη διαλογή των παρατηρήσεων, δηλαδή στην καταμέτρηση των παρατηρήσεων που ανήκουν σε κάθε μία από τις ομάδες που δημιουργήσαμε στα προηγούμενα βήματα. Με τη διαλογή μπορούμε να συμπληρώσουμε έναν πίνακα συχνοτήτων όπως παρακάτω

Κλάση	Συχνότητα	Σχετική συχνότητα
$[\alpha_1 - \beta_1)$	ν_1	f_1
$[\alpha_2 - \beta_2)$	ν_2	f_2
...	...	...
Σύνολο	ν	1

4ο βήμα : Σε ορθοκανονικό σύστημα αξόνων τοποθετούμε τα άκρα κάθε ομάδας (κλάσης)

στον οριζόντιο άξονα και σχεδιάζουμε ορθογώνια παραλληλόγραμμα με βάση την ομάδα και ύψος τη συχνότητα κάθε μίας ομάδας ή τη σχετική συχνότητα της ανάλογα με το αν μας έχει ζητηθεί ιστόγραμμα συχνοτήτων ή σχετικών συχνοτήτων.

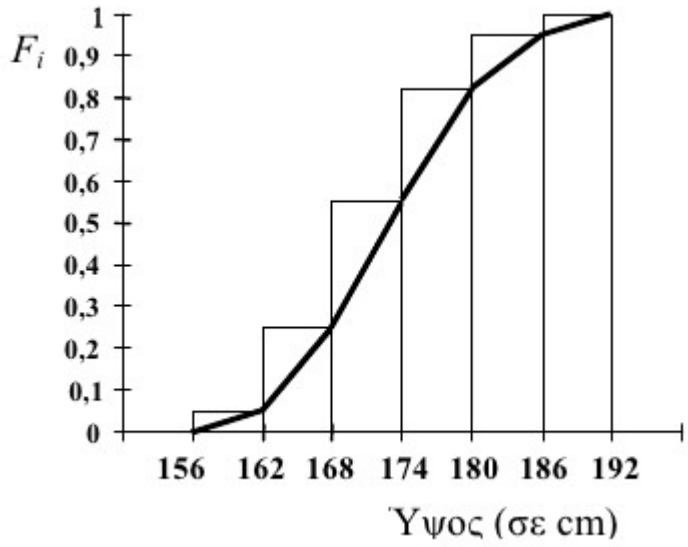
5ο βήμα (πολύγωνο συχνοτήτων) : Αν επιπλέον, έχει ζητηθεί πολύγωνο συχνοτήτων (ή σχετικών συχνοτήτων) τότε ενώνουμε τα μέσα σημεία της πάνω πλευράς του κάθε ορθογωνίου με μία συνεχής γραμμή. Θεωρούμε μία (ιδεατή) κλάση πριν την πρώτη και άλλη μία μετά την τελευταία και ενώνουμε επιπλέον τα μέσα αυτών με τις διπλανές τους (μετά και πριν αντίστοιχα) για να προκύψει μία πολυγωνική γραμμή, η οποία ονομάζεται πολύγωνο συχνοτήτων.(Διάγραμμα 8, σελίδα 59)

2.4.3.1 Ιστόγραμμα αθροιστικών συχνοτήτων

Στην περίπτωση που μας ζητείται ιστόγραμμα αθροιστικών συχνοτήτων ή ιστόγραμμα αθροιστικών σχετικών συχνοτήτων τότε συμπληρώνουμε τον εξής πίνακα

Κλάση	Αθροιστική συχνότητα	Σχετική συχνότητα
$[\alpha_1 - \beta_1)$	$N_1 (= \nu_1)$	$F_1 (= f_1)$
$[\alpha_2 - \beta_2)$	$N_2 (= \nu_1 + \nu_2)$	$F_2 (= f_1 + f_2)$
...	...	...

Το ύψος κάθε ορθογωνίου ορίζεται από την αθροιστική συχνότητα κάθε κλάσης ενώ το πολύγωνο αθροιστικών συχνοτήτων προκύπτει από την ένωση με πολυγωνική γραμμή της κάτω αριστερής με την πάνω δεξιά άκρη κάθε ενός ορθογωνίου (Διάγραμμα 3, σελίδα 51)



Διάγραμμα 3: Ιστόγραμμα και πολύγωνο αθροιστικών συχνοτήτων

Παράδειγμα

Σε κάποια εταιρεία καταγράφηκε η διάρκεια 30 υπεραστικών τηλεφωνημάτων που έγιναν σε μια εβδομάδα. Οι χρόνοι διάρκειας των τηλεφωνημάτων είναι οι ακόλουθοι

11,8 3,6 16,6 13,5 4,8 8,3 8,9 9,1 7,7 2,3 12,1 6,1 10,2 8,0 11,4 6,8 9,6 19,5 15,3 12,3 8,5 15,9 18,7 11,7 6,2 11,2 10,4 7,2 5,5 14,5 Να γίνει (α) το ιστόγραμμα και το πολύγωνο συχνοτήτων αυτών (β) το ιστόγραμμα και το πολύγωνο αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.

Λύση

1ο βήμα : Σύμφωνα με τον τύπο της παραγράφου 2.4.3, σελίδα 49, υπολογίζουμε **1 + 3,3 log(30) = 1 + 3,3* 1,477 = 5,91 ≈ 6**, άρα επιλέγουμε να χωρίσουμε τις παρατηρήσεις σε 6 ομάδες (κλάσεις).

Σημείωση : Η παραπάνω επιλογή του πλήθους κλάσεων δεν είναι δεσμευτική. Θα μπορούσαμε να ορίσουμε 5 ή 7 ή και 8 κλάσεις. Το τελικό ιστόγραμμα θα είχε άλλο σχήμα αλλά ο στόχος της διαδικασίας που είναι η γραφική αναπαράσταση των δεδομένων θα είχε επιτευχθεί!

2ο βήμα : Το εύρος των παρατηρήσεων είναι **R = Max – Min = 19,5 – 2,3 = 17,2** άρα κάθε μία από τις 6 κλάσεις πρέπει να έχει μήκος **17,2 / 6 = 2,86 ≈ 3**. Ξεκινώντας από τον αριθμό

2 τον οποίο επιλέγουμε αυθαίρετα, οι κλάσεις προσδιορίζονται να είναι οι εξής :

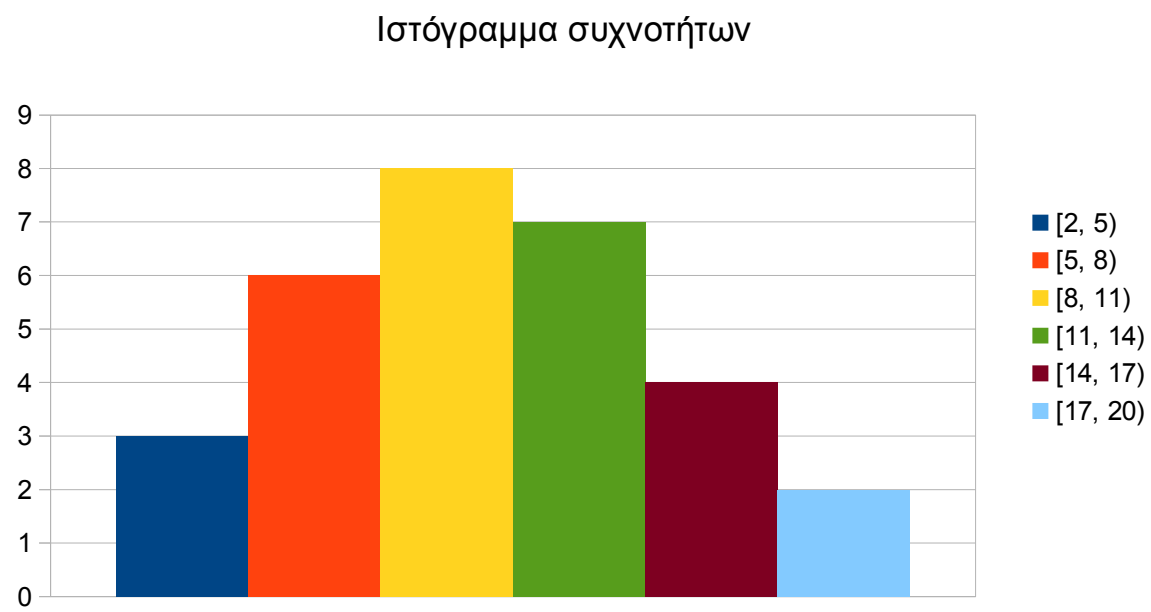
[2,5), [5, 8), [8, 11), [11, 14), [14, 17) και [17, 20).

3ο βήμα : Προχωρούμε στη διαλογή των παρατηρήσεων, και συμπληρώνουμε τον πίνακα συχνοτήτων όπως παρακάτω :

Όρια κλάσεων	Συχνότητα	Σχετική συχνότητα	Αθροιστική Συχνότητα	Αθροιστική Σχετική Συχνότητα
[2, 5)	3	0,1	3	3/30 = 0,1
[5, 8)	6	0,2	9	9/30 = 0,3
[8, 11)	8	0,27	17	17/30 = 0,567
[11, 14)	7	0,23	24	24/30 = 0,8
[14, 17)	4	0,13	28	28/30 = 0,933
[17, 20)	2	0,07	30	30/30 = 1,00
Σύνολο	30	1		

4ο βήμα : Ιστόγραμμα Συχνοτήτων

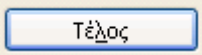
Σχεδιάζουμε πάνω από κάθε μία κλάση, ένα ορθογώνιο με ύψος όσο και η συχνότητα της κλάσης, και προκύπτει το επόμενο διάγραμμα.



Διάγραμμα 4: Ιστόγραμμα συχνοτήτων

2.5 Δημιουργία ραβδογράμματος – κυκλικού διαγράμματος με το Calc

Ένα διάγραμμα δημιουργείται είτε επιλέγοντας ***Εισαγωγή*** → ***Διάγραμμα*** είτε επιλέγοντας

το εικονίδιο  στη γραμμή εργαλείων. Στον οδηγό διαγράμματος που εμφανίζεται στο πρώτο βήμα (**1. Τύπος διαγράμματος**) επιλέγουμε τη δημιουργία κάθετου ραβδογράμματος ( , πρώτη επιλογή και πλέον συνηθισμένη), οριζοντίου ραβδογράμματος ( , δεύτερη επιλογή) ή κυκλικού διαγράμματος ( , τρίτη επιλογή). Σε κάθε περίπτωση περνάμε στο δεύτερο βήμα (**2. Περιοχή δεδομένων**) και επιλέγουμε την περιοχή των δεδομένων μας η οποία πρέπει να περιέχει τα ονόματα των διαφορετικών τιμών και τις συχνότητες τους (δηλαδή τα κελιά D3:E7 στο παράδειγμα της εικόνας 3). Επιλέγουμε την τελευταία επιλογή (**4. Στοιχεία διαγράμματος**) για να εισάγουμε δευτερεύοντα στοιχεία όπως τον τίτλο του διαγράμματος και τις ονομασίες των αξόνων αν υπάρχουν. Επιλέγοντας 

αφήνουμε τον οδηγό και το διάγραμμα είναι έτοιμο και έχει τοποθετηθεί στο φύλλο εργασίας του Calc.

Από τη στιγμή της δημιουργίας του ένα διάγραμμα μπορεί να επεξεργαστεί με διπλό κλικ πάνω του, κάνοντας αλλαγές όπως το χρώμα του φόντου κ.α. Επιπλέον, μπορούμε να δημιουργήσουμε εύκολα και ένα κυκλικό διάγραμμα από το ραβδόγραμμα, δημιουργώντας ένα αντίγραφο όλου του ραβδογράμματος (με αντιγραφή και επικόλληση) το οποίο θα επεξεργαστούμε αλλάζοντας του τον τύπο από ραβδόγραμμα σε κυκλικό διάγραμμα!

Πίνακας 2.10: Δημιουργία ιστογράμματος με υπολογιστή

Για το ιστόγραμμα δεν υπάρχει δεσμευμένη διαδικασία στο Calc. Μπορεί να χρησιμοποιηθεί ένα ραβδόγραμμα που θα αναπαριστά τις συχνότητες των αριθμητικών κλάσεων όπως αυτές έχουν οριστεί στο σχετικό πίνακα συχνοτήτων (Πίνακας 2.3, σελίδα 38) και αποτέλεσμα όπως αυτό του διαγράμματος 4,σελίδα 52.

Ένα ιστόγραμμα προκύπτει εύκολα από τη συνάρτηση **hist(x).** Η συνάρτηση αυτή επιδέχεται μεγάλη παραμετροποίηση. Όλες οι επιλογές μπορούν να εμφανιστούν με **?hist** στο παράθυρο του R – Project. Ενδεικτικά, αναφέρουμε την εντολή

hist(x, labels = TRUE, col = 3, density = 2, breaks = c(5, 10, 15, 20, 25, 30, 35), xlab = "Τιμές", ylab = "Συχνότητα", main = "Ιστόγραμμα συχνοτήτων", ylim = c(0,5), xlim = c(0,30))

η οποία θα δημιουργήσει ένα ιστόγραμμα συχνοτήτων με όρια κλάσεων τα 5, 10, 15, 20, 25, 30, 35, εμφάνιση της συχνότητας πάνω από κάθε ράβδο, γραμμοσκιασμένες ράβδους με πράσινο χρώμα, άξονα x από 0 έως 30, άξονα y από 0 έως 5 και τίτλους όπως έχει οριστεί.

2.6 Ιστόγραμμα και Πολύγωνο Συχνοτήτων με το Calc

	A	B	C	D
1				
2				
3				
4				
5				
6	Πίνακας Συχνοτήτων			
7	α/α	Κλάση (Άνω Όριο)	Κλάση	Συχνότητα
8	1	63	(61,5 έως 63]	0
9	2	65,5	(63 έως 65,5]	1
10	3	68	(65,5 έως 68]	7
11	4	70,5	(68 έως 70,5]	9
12	5	73	(70,5 έως 73]	3
13	6	75,5	(73 έως 75,5]	6
14	7	78	(75,5 έως 78]	4
15			(78 έως 80,5]	0
16				
17		Προκαταρκτικός Πίνακας		
18		Στατιστικά	Τιμή	Συνάρτηση
19		Ελάχιστη Τιμή	63,56	MIN(F2:F31)
20		Μέγιστη Τιμή	77,47	MAX(F2:F31)
21		Εύρος Παρατηρήσεων	13,91	C20-C19
22		Πλήθος Κλάσεων	6	Αυθαίρετος ορισμός
23		Εύρος κάθε κλάσης	2,32	C21/C22
24		Στρογγυλοποίηση	2,5	Αυθαίρετη

Εικόνα 4: Δημιουργία Πίνακα Συχνοτήτων από τον οποίο θα δημιουργηθεί το ιστόγραμμα

Το ιστόγραμμα είναι ο κατάλληλος γραφικός τρόπος παρουσίασης των τιμών μιας συνεχούς μεταβλητής. Δυστυχώς, το Calc δεν περιλαμβάνει το ιστόγραμμα στα διαθέσιμα γραφήματα, ωστόσο είναι εύκολο να το δημιουργήσουμε με λίγη περισσότερη προσπάθεια ως ένα ειδικού τύπου ραβδόγραμμα!

Τα δεδομένα είναι καλό να βρίσκονται σε μία γραμμή ή μία στήλη. Στο παράδειγμα που θα ακολουθήσει θα περιγράψουμε τη δημιουργία Ιστογράμματος για τα δεδομένα του βάρους τριάντα εθελοντών τα οποία εμφανίζονται στην Εικόνα 10.

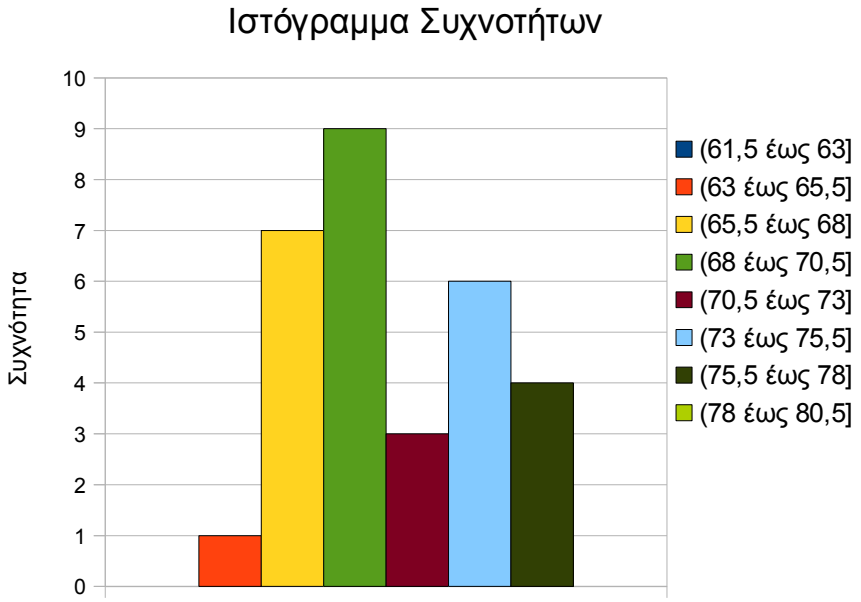
Τα δεδομένα τοποθετήθηκαν στη στήλη **F** του Calc από την γραμμή 2 έως την 31.

Το πρώτο βήμα είναι η συμπλήρωση του απαραίτητου πίνακα συχνοτήτων όπως περιγράφεται στον πίνακα 2.3, σελίδα 38.

	F
1	Δεδομένα
2	63,6
3	65,9
4	66,2
5	67,3
6	67,4
7	67,5
8	67,9
9	68,0
10	68,4
11	68,6
12	68,7
13	68,7
14	69,0
15	69,3
16	69,8
17	69,8
18	70,3
19	70,7
20	72,5
21	72,7
22	73,8
23	74,0
24	74,1
25	74,2
26	75,0
27	75,2
28	75,6
29	76,0
30	76,6
31	77,5

Εικόνα 5: Δεδομένα

Το δεύτερο βήμα είναι να χρησιμοποιήσουμε τη μηχανή γραφικών του Calc για να σχεδιάσουμε το απαραίτητο ιστόγραμμα (Διάγραμμα 5, και 6, σελίδα 57) ή το πολύγωνο συχνοτήτων (Διάγραμμα 7, σελίδα 57)



Διάγραμμα 5: Ιστόγραμμα (Πρώτος τύπος)

Αναλυτική Περιγραφή : Είναι φανερό πως χωρίς πίνακα συχνοτήτων δεν είναι δυνατό να γίνει ιστόγραμμα ή πολύγωνο συχνοτήτων! Η δημιουργία ενός πίνακα συχνοτήτων είναι απλή υπόθεση αρκεί τα απαραίτητα βήματα να γίνουν με προσοχή. Πριν από όλα πρέπει να αποφασίσει ο χρήστης για το πλήθος των κλάσεων στις οποίες θα διαχωριστούν τα δεδομένα. Δεν υπάρχει σωστή και λάθος επιλογή αρκεί να μην είναι πάρα πολλές ή πάρα πολύ λίγες.

Να θυμάστε πως βασικός σκοπός του Ιστογράμματος είναι η άμεση οπτική περιγραφή της κατανομής και η σωστή δημιουργία του επαφίεται κυρίως στην στατιστική αντίληψη του χρήστη! Ωστόσο, κάποιος απλοϊκός κανόνας είναι να χωρίζουμε τα στοιχεία το πολύ σε δέκα κατηγορίες αν τα στοιχεία είναι πλήθους έως εκατό και το πολύ σε είκοσι κατηγορίες στην περίπτωση που είναι περισσότερα.

Η απόφαση που παίρνουμε στην αρχή για το πλήθος των κλάσεων (κελί **C22**) δεν πρέπει

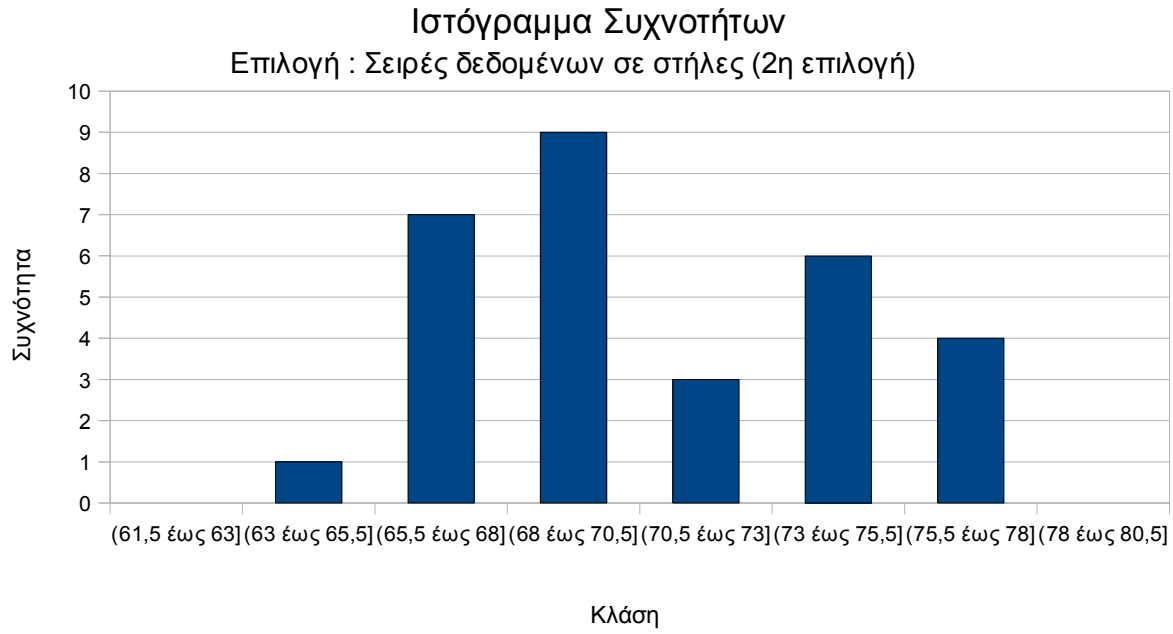
να είναι αυστηρή και ενδεχομένως να αλλάξει αν μειώσουμε ή αυξήσουμε σημαντικά το εύρος κάθε κλάσης στην στρογγυλοποίηση (κελί **C24**) η οποία πολλές φορές είναι απαραίτητη καθώς το ακριβές πλάτος προκύπτει “δύσχρηστος” αριθμός.

Στην στήλη με ετικέτα (**Κλάση (Άνω όριο)**) τοποθετούνται τα άνω όρια κάθε μίας κλάσης τα οποία απαιτούνται για τη συμπλήρωση του πίνακα συχνοτήτων. Το πρώτο άνω όριο (63 στο κελί **B8**) τοποθετείται αυθαίρετα με την μέριμνα να είναι όσο το δυνατόν περισσότερο “στρογγυλός” αριθμός και οπωσδήποτε μικρότερος από τη μικρότερη τιμή των δεδομένων. Οι υπόλοιπες καταχωρήσεις της στήλης αυτής προκύπτουν από την πρώτη με διαδοχική πρόσθεση του πλάτους κλάσης το οποίο στο παράδειγμα το τοποθετήσαμε 2,5 (κελί **B25**), Η συμπλήρωση των υπολοίπων κελιών εύκολα γίνεται στο Calc με την εφαρμογή της συνάρτησης **B8+C$24** στο κελί **B9** και αυτόματη επέκταση στα κελιά **B10:B14**.

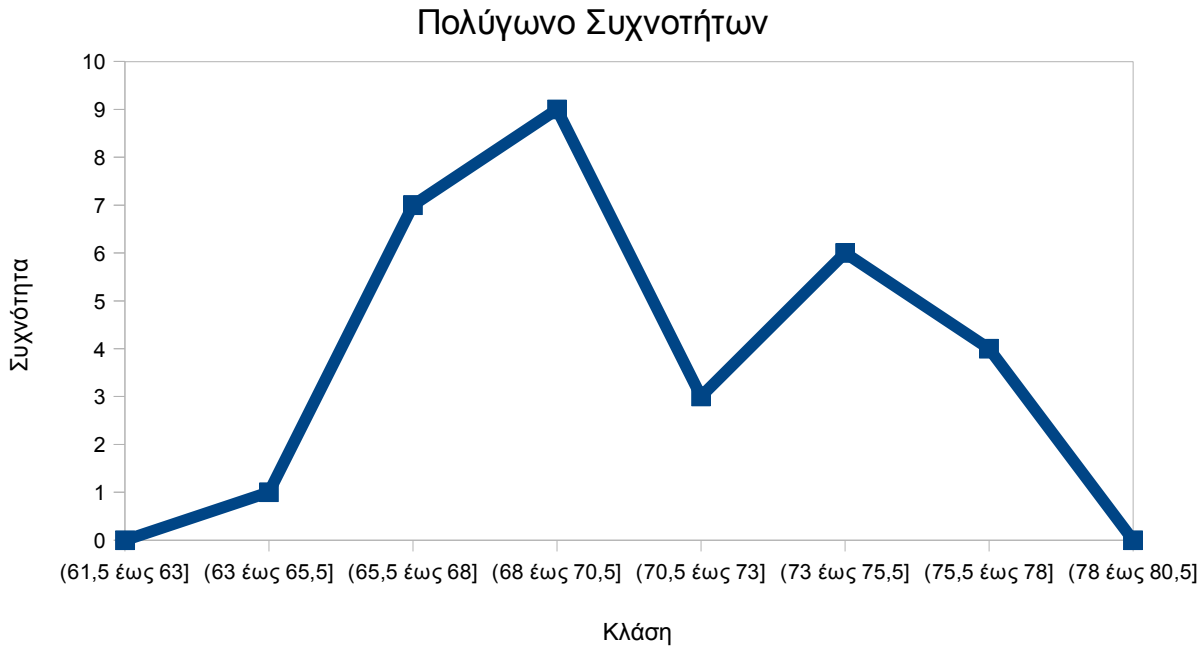
Η στήλη με τα άνω όρια των κλάσεων αρκεί για τη δημιουργία του ιστογράμματος, ωστόσο θα βελτιωθεί ιδιαίτερα η εικόνα του ιστογράμματος αν δημιουργήσουμε τα ανοικτά-κλειστά διαστήματα της στήλης **C** που βρίσκονται κάτω από τον κατανοητό από άνθρωπο τίτλο “**Κλάση**”. Το πρώτο και το τελευταίο διάστημα πρέπει να συμπληρωθούν με το χέρι ωστόσο τα μεσαία διαστήματα εύκολα δημιουργούνται με εφαρμογή της συνάρτησης “**=CONCATENATE("(";B8;" έως ";B9;"]")”** για το **C9** και μεταφορά της ίδιας μέχρι το κελί **C14**.

Τέλος, με εφαρμογή της συνάρτησης **FREQUENCY()** και λίγη προσοχή συμπληρώνεται η στήλη “**Συχνότητα**”. Η συνάρτηση **FREQUENCY()** παίρνει δύο ορίσματα, τα δεδομένα που θα ταξινομηθούν σε κλάσεις (κελιά F2:F31) και τη στήλη με τα επιθυμητά άνω όρια (κελιά B8:B14). Η εφαρμογή της συνάρτησης μπορεί να γίνει σε οποιοδήποτε κελί. Στο παράδειγμα μας εφαρμόστηκε στο κελί **D8** ως “**{=FREQUENCY(F2:F31;B8:B14)}”** (προσέξτε τη χρήση των αγκύλων κάτι που σημαίνει πως το αποτέλεσμα της συνάρτησης είναι πίνακας και όχι ένα στοιχείο) και με αυτήν συμπληρώθηκαν τα κελιά **D8:D15** και ολοκληρώθηκε ο πίνακας συχνοτήτων.

Τέλος, το ιστόγραμμα δημιουργείται ως ένα ραβδόγραμμα, ενώ μπορούμε επιλέγοντας τύπο διαγράμματος “Γραμμή” να δημιουργήσουμε και το πολύγωνο συχνοτήτων.



Διάγραμμα 6: Ιστόγραμμα (Δεύτερος τύπος)

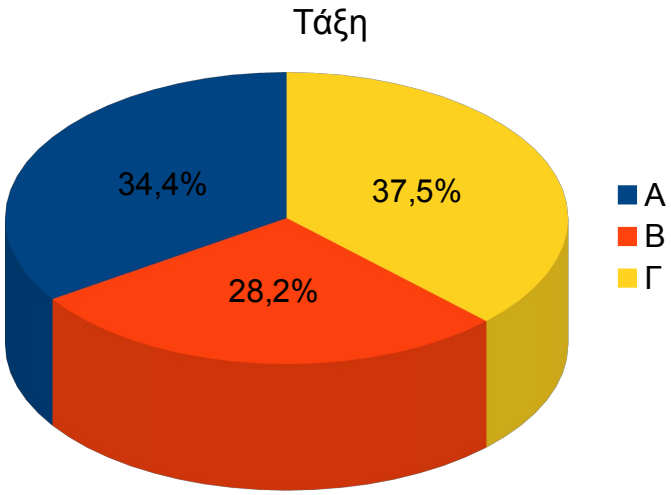


Διάγραμμα 7: Πολύγωνο Συχνοτήτων

Δραστηριότητες

1. Από το 1928 έως το 2011 (Πρωταθλήματα Α΄ Εθνικής) ο Ολυμπιακός έχει κατακτήσει 38 τίτλους, ο Παναθηναϊκός 20, η ΑΕΚ 11, ο Άρης 3, ο ΠΑΟΚ 2 και η Λάρισα 1. Να κατασκευάσετε το ραβδόγραμμα και το κυκλικό διάγραμμα συχνοτήτων και σχετικών συχνοτήτων.

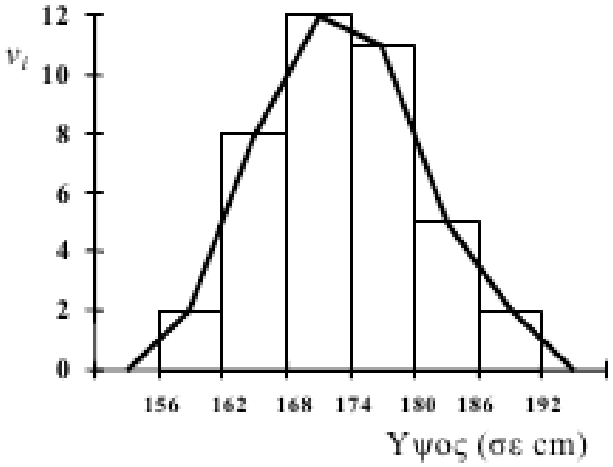
2. Ρωτήθηκαν 291 μαθητές για την τάξη που ανήκουν και από τις αποκρίσεις τους σχεδιάστηκε το παρακάτω κυκλικό διάγραμμα. Να συμπληρωθεί ο αντίστοιχος πίνακας συχνοτήτων.



3. Να μεταφέρεται τα δεδομένα του επόμενου πίνακα (βαθμολογίες σπουδαστών) σε ιστόγραμμα συχνοτήτων και σχετικών συχνοτήτων και να σχεδιάσετε τα αντίστοιχα πολύγωνα συχνοτήτων.

Κλάση	Συχνότητα	Σχετική συχνότητα
[10, 12)	2	2/20=0,10
[12, 14)	5	5/20=0,25
[14, 16)	5	5/20=0,25
[16, 18)	7	7/20=0,35
[18, 20)	1	1/20=0,05
Σύνολο	20	20/20=1,00

4. Μετρήθηκε το ύψος 40 μαθητών και τα δεδομένα παραστάθηκαν στο παρακάτω ιστόγραμμα συχνοτήτων.



Διάγραμμα 8: Ιστόγραμμα και πολύγωνο συχνοτήτων

(α) Πόσοι είναι οι μαθητές με ύψος μεταξύ 174 και 180; Ποιο είναι το ποσοστό αυτών των μαθητών;

(β) Πόσοι είναι οι μαθητές με ύψος μεταξύ 162 και 180; Ποιο είναι το ποσοστό αυτών των μαθητών;

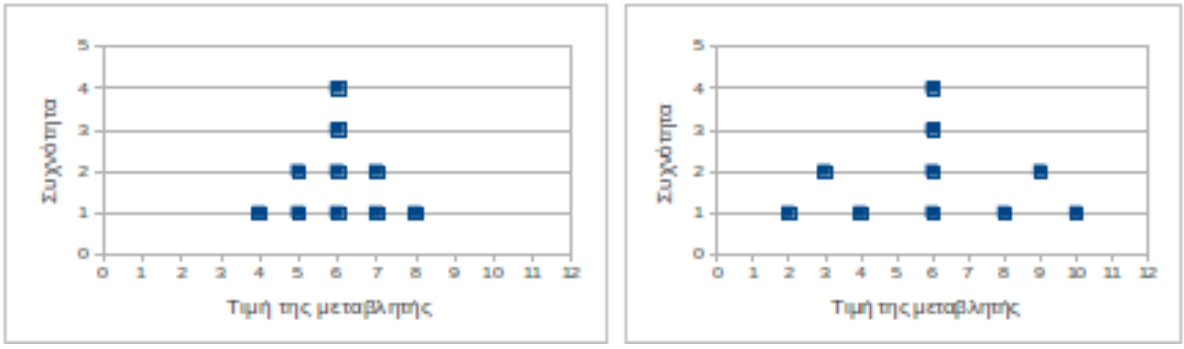
(γ) Να συμπληρωθεί ο πίνακας συχνοτήτων από τον οποίο δημιουργήθηκε το παραπάνω ιστόγραμμα.

(δ) Να υπολογιστεί το μέσο ύψος και η τυπική απόκλιση αυτού.

5. Δίνεται ο μηνιαίος μισθός (σε ευρώ) των 20 υψηλόμισθων υπαλλήλων μιας μεγάλης εταιρείας : 1800, 2000, 2700, 3600, 4000, 1100, 3000, 2700, 3600, 4500, 1700, 2000, 3000, 2800, 3800, 1900, 3000, 3300, 3600, 3800. Να γίνει το ιστόγραμμα συχνοτήτων και το αντίστοιχο πολύγωνο συχνοτήτων.

2.7 Μέτρα Διασποράς

Με τη γενική ονομασία “Μέτρα Διασποράς” περιγράφουμε όλα τα στατιστικά που αποσκοπούν στην περιγραφή της διασποράς των παρατηρήσεων. Η περιγραφή της διασποράς μίας ομάδας παρατηρήσεων είναι απαραίτητη καθώς η μέση τιμή δεν δίνει πλήρη εικόνα για τη φύση της κατανομής. Χαρακτηριστικά, στο διάγραμμα 9 εμφανίζονται δύο κατανομές με ίδιο “κέντρο” αλλά διαφορετική διασπορά.



Διάγραμμα 9: Κατανομές με ίδια μέση τιμή και διαφορετική διασπορά

2.7.1 Εύρος

Το εύρος ενός δείγματος είναι απλά η διαφορά της μέγιστης από την ελάχιστη τιμή του. Για παράδειγμα το εύρος των παρατηρήσεων 1, 5, 4, 9, 11 είναι 11 – 1 = 10. Το εύρος συνήθως συμβολίζεται με R από την αγγλική λέξη Range.

Δραστηριότητα

Οι μέγιστες θερμοκρασίες 10 ημερών του χειμώνα μετρήθηκαν 5, 8, 10, 7, 3, -1, -3 , 0, 2, 7. Να βρεθούν : (α) Η μέγιστη (β) Η ελάχιστη θερμοκρασία (γ) Το εύρος της θερμοκρασίας (δ) Η μέση θερμοκρασία (ε) Η διάμεση θερμοκρασία

Πίνακας 2.11: Υπολογισμός του εύρους στον υπολογιστή

	Δεν υπάρχει ενσωματωμένη συνάρτηση, ωστόσο μπορεί εύκολα να υπολογιστεί ως **MAX() - MIN()**
	Η συνάρτηση **range(x)** επιστρέφει διάνυσμα με την ελάχιστη και τη μέγιστη τιμή. π.χ. Αν x = c(10, 20, 30, 40) τότε range(x) = [10, 40] άρα Έυρος = range(x)[2] - range(x)[1]

2.7.2 Ενδοτεταρτημοριακό Εύρος

Δύο σημαντικά σημεία σε μία κατανομή είναι αυτά που βρίσκονται στο 25ο και στο 75ο σημείο της κατανομής, τα οποία ονομάζονται πρώτο και τρίτο τεταρτημόριο και συμβολίζονται Q_1 και Q_3 αντίστοιχα.

Το πρώτο τεταρτημόριο Q_1 βρίσκεται στην $\frac{v+1}{4}$ θέση των ταξινομημένων παρατηρήσεων και αν ο αριθμός αυτός δεν είναι ακέραιος τότε υπολογίζουμε το ημιάθροισμα των στοιχείων που βρίσκονται στις δύο γειτονικές θέσεις. Ανάλογα, το τρίτο τεταρτημόριο Q_3

βρίσκεται στην $3\frac{v+1}{4}$ θέση και υπολογίζεται με τον ίδιο τρόπο. Από τον ορισμό των Q_1, Q_3 είναι φανερό πως : Μεταξύ του πρώτου και του τρίτου τεταρτημορίου βρίσκονται οι μισές παρατηρήσεις.

Ως **ενδοτεταρτημοριακό εύρος** ορίζεται η διαφορά

$$\mathbf{Q = Q_3 - Q_1}.$$

Παράδειγμα

Να βρεθεί το πρώτο, το τρίτο τεταρτημόριο και το ενδοτεταρτημοριακό εύρος των παρατηρήσεων

0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 2, 2.

Απάντηση

Πρώτα ταξινομούμε τις παρατηρήσεις από τη μικρότερη στη μεγαλύτερη :

0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4 .

Οι παρατηρήσεις είναι 20 άρα ν = 20.

Είναι $\frac{v+1}{4} = \frac{21}{4} = 5{,}25$ άρα $Q_1 = \frac{1+1}{2} = 1$ και $3\frac{v+1}{4} = 3\frac{21}{4} = 15{,}75$ άρα $Q_3 = \frac{2+2}{2} = 2$. Το ενδοτεταρτημοριακό εύρος είναι $Q = Q_3 - Q_1 = 2 - 1 = 1$.

Δραστηριότητα

Για το δείγμα θερμοκρασιών 5, 8, 10, 7, 3, -1, -3 , 0, 2, 7 να υπολογιστεί (α) το πρώτο τεταρτημόριο (β) το τρίτο τεταρτημόριο (γ) Το ενδοτεταρτημοριακό εύρος.

Πίνακας 2.12: Υπολογισμός του ενδοτεταρτημοριακού εύρους στον υπολογιστή

	Συνάρτηση **QUARTILE(δεδομένα; α)**. Αν α = 0 τότε επιστρέφει την ελάχιστη τιμή ενώ αν α = 4 επιστρέφει τη μέγιστη. Για α = 1 επιστρέφει το Q_1, για α = 2 επιστρέφει τη διάμεσο και για α = 3 επιστρέφει το Q_3.
	Η συνάρτηση **summary(x)** επιστρέφει διάνυσμα με τα βασικά στατιστικά μεταξύ των οποίων είναι και τα τεταρτημόρια. Εναλλακτικά μπορεί να χρησιμοποιηθεί η συνάρτηση **fivenum(x)**. Οι δύο διαδικασίες δεν δίνουν πάντα τα ίδια αποτελέσματα καθώς υπολογίζουν τα τεταρτημόρια με διαφορετικές μεθόδους.

2.7.3 Μέση απόκλιση, διακύμανση και τυπική απόκλιση

Μέση (απόλυτη) απόκλιση των παρατηρήσεων $x_{1,}x_{2,}\ldots,x_{v}$ ονομάζεται η ποσότητα

$MAD = \frac{1}{v}\sum_{k=1}^{v}\left|x_k-\bar{x}\right|$, όπου $\bar{x}$ είναι η μέση τιμή των παρατηρήσεων.

Διακύμανση ή διασπορά των παρατηρήσεων $x_{1,}x_{2,}\ldots,x_{v}$ ονομάζεται η ποσότητα $\sigma^2 = \frac{1}{v}\sum_{k=1}^{v}(x_k-\mu)^2$, όπου μ είναι η μέση τιμή των παρατηρήσεων. Αν οι παρατηρήσεις $x_{1,}x_{2,}\ldots,x_{v}$ αποτελούν δείγμα από ένα πληθυσμό του οποίου δεν γνωρίζουμε τη μέση τιμή τότε η τυπική απόκλιση πρέπει να υπολογιστεί από τον τύπο $s^2 = \frac{1}{v-1}\sum_{k=1}^{v}(x_k-\bar{x})^2$

Τυπική απόκλιση ονομάζεται η τετραγωνική ρίζα της διακύμανσης ή πιο απλά $\sigma=\sqrt{\sigma^2}$ (ή $s=\sqrt{s^2}$ για δείγμα). Επιπλέον, ένας ισοδύναμος τύπος για τον υπολογισμό της διακύμανσης είναι ο $\sigma^2 = \frac{1}{v}\left(\sum_{k=1}^{v}x_k^2 - \frac{\left(\sum_{k=1}^{v}x_k\right)^2}{v}\right)$ ο οποίος είναι καλύτερη επιλογή όταν ο υπολογισμός πρέπει να γίνει με το χέρι. Όταν οι παρατηρήσεις επαναλαμβάνονται οι ίδιες πολλές φορές ή όταν έχουμε παρατηρήσεις χωρισμένες σε διαστήματα τότε μπορούμε να χρησιμοποιήσουμε τους τύπους $\sigma^2 = \frac{1}{v}\sum_{k=1}^{\rho}(x_k-\mu)^2 v_k$ (αντ. $s^2 = \frac{1}{v-1}\sum_{k=1}^{\rho}(x_k-\bar{x})^2 v_k$)ή $\sigma^2 = \frac{1}{v}\left(\sum_{k=1}^{\rho}x_k^2 v_k - \frac{\left(\sum_{k=1}^{\rho}x_k v_k\right)^2}{v}\right)$ όπου ρ είναι το πλήθος των ομάδων και v_k είναι η συχνότητα εμφάνισης της παρατήρησης x_k.

Παρατήρηση. Η δειγματική τυπική απόκλιση s προκύπτει από τον πολλαπλασιασμό της διακύμανσης του πληθυσμού σ επί τον παράγοντα ν/(ν-1) που γενικότερα αναφέρεται ως η διόρθωση Bessel. Ο λόγος που απαιτείται μία διόρθωση είναι απλός : η δειγματική μέση τιμή δεν συμπίπτει με τη μέση τιμή του πληθυσμού άρα το άθροισμα τετραγωνικών

αποκλίσεων των παρατηρήσεων του δείγματος από τη μέση τιμή είναι μικρότερο από αυτό που θα υπολογιζόταν αν στη θέση της δειγματικής μέσης τιμής ήταν η μέση τιμή του πληθυσμού και ως εκ τούτου εφαρμόζεται η αύξηση που προκύπτει με την διόρθωση Bessel. Το μειονέκτημα είναι η αύξηση του τυπικού σφάλματος Στα περισσότερα εξειδικευμένα προγράμματα (SPSS, R - Project), θεωρείται πως οι παρατηρήσεις είναι ένα δείγμα από το οποίο πρέπει να προσεγγιστούν τα στατιστικά του πληθυσμού άρα όταν αναφέρεται η διακύμανση εννοείται η δειγματική διακύμανση.

2.7.3.1 Παραδείγματα υπολογισμού

Α' τρόπος : Αντικατάσταση και πράξεις (χρήση του 1ου τύπου διακύμανσης).

Παράδειγμα.

Να βρεθεί η μέση απόκλιση, η διακύμανση και η τυπική απόκλιση των παρατηρήσεων 11, 13, 14, 15, 17.

Λύση Υπολογίζουμε τη μέση τιμή $\bar{x} = \frac{11+13+14+15+17}{5} = 14$. Η μέση απόλυτη απόκλιση είναι

$$MAD = \frac{1}{5}[|11-14|+|13-14|+|14-14|+|15-14|+|17-14|] =$$
$$= \frac{1}{5}[|-3|+|-1|+|0|+|1|+|3|] = \frac{1}{5}(3+1+0+1+3) = \frac{7}{5} = 1{,}4$$

Η διακύμανση είναι

$$s^2 = \frac{1}{4}[(11-14)^2+(13-14)^2+(14-14)^2+(15-14)^2+(17-14)^2] =$$
$$= \frac{1}{4}[(-3)^2+(-1)^2+0^2+1^2+3^2] = \frac{1}{4}(9+1+0+1+9) = \frac{20}{4} = 5$$

ενώ η τυπική απόκλιση υπολογίζεται να είναι $s = \sqrt{s^2} = \sqrt{5} = 2{,}23$.

Β' τρόπος : Με τη βοήθεια ενός συνοπτικού πίνακα (χρήση του 2ου τύπου διακύμανσης).

Παράδειγμα.

Να βρεθεί η διακύμανση και η τυπική απόκλιση των 20 παρατηρήσεων ενός πληθυσμού,

0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 2, 2.

Λύση

Συμπληρώνουμε τον παρακάτω πίνακα

Τιμή (x_k)	Συχνότητα (ν_k)	x_k^2	$x_k\,\nu_k$	$x_k^2\,\nu_k$
0	3	0	0	0
1	6	1	6	6
2	7	4	14	28
3	3	9	9	27
4	1	16	4	16
Σύνολο	**20**	**30**	**33**	**77**

Είναι ν = 20, ρ = 5 (πλήθος ομάδων), $\sum_{k=1}^{5} x_k^2 \nu_k = 77$ και $\sum_{k=1}^{5} x_k \nu_k = 33$. Υπολογίζουμε

$$\sigma^2 = \frac{1}{20}\left(\sum_{k=1}^{5} x_k^2 \nu_k - \frac{\left(\sum_{k=1}^{5} x_k \nu_k\right)^2}{20}\right) = \frac{1}{20}\left(77 - \frac{33^2}{20}\right) = 1{,}1275$$ και

$\sigma = \sqrt{1{,}1275} = 1{,}062$.

Πίνακας 2.13: Υπολογισμός των μέτρων διασποράς στον υπολογιστή

Οι συναρτήσεις **VAR** και **VARP** υπολογίζουν τη δειγματική διασπορά και τη πληθυσμιακή διασπορά των παρατηρήσεων ενώ οι συναρτήσεις **STDEV** και **STDEVP** υπολογίζουν αντίστοιχα τη διακύμανση (P = Population). Η διαφορά τους είναι ο παρονομαστής με τον οποίο διαιρείται το άθροισμα τετραγωνικών αποκλίσεων : στη VAR είναι ο (n – 1) ενώ στη VARP είναι ο n (n = πλήθος παρατηρήσεων).

var(x) και **sd(x)** για τον υπολογισμό της δειγματικής διασποράς και τυπικής απόκλισης αντίστοιχα. Αντιστοιχούν στις συναρτήσεις STDEV και VAR του Calc.

2.8 Συντελεστής Μεταβολής ή Ομοιογένειας (CV)

Ο συντελεστής μεταβολής ή ομοιογένειας, ορίζεται για κάθε δείγμα ποσοτικών (αριθμητικών) παρατηρήσεων να είναι

$$CV = \frac{\text{τυπική απόκλιση}}{\text{μέση τιμή}} = \frac{s}{\bar{x}} \quad .$$

Υπολογίζεται και χρησιμοποιείται για δύο λόγους

(α) για τη σύγκριση της ομοιογένειας δύο δείγματος που μετρούνται σε διαφορετικές μονάδες είτε για δύο δείγματα που έχουν σημαντικά διαφορετικές μέσες τιμές. Το δείγμα

που έχει το μικρότερο συντελεστή μεταβολής χαρακτηρίζεται περισσότερο ομοιογενές από το άλλο.

(β) για το χαρακτηρισμό ενός δείγματος ως ομοιογενές ή μη ομοιογενές : Ένα δείγμα χαρακτηρίζεται ομοιογενές όταν ο συντελεστής μεταβολής είναι μικρότερος ή ίσος από 0,1 ή 10%.

Παράδειγμα

Ένα δείγμα είκοσι μαθητών της Α΄ Γυμνασίου έχει μέσο βάρος $\bar{x}_A = 40\,kgr$ και τυπική απόκλιση s_A = 6 kgr, ενώ ένα δεύτερο δείγμα τριάντα μαθητών της Γ΄ Λυκείου βρήκαμε μέσο βάρος $\bar{x}_B = 75\,kgr$ και τυπική απόκλιση s_B = 6 kgr. (α) Ποια τάξη είναι περισσότερο ομοιογενής ως προς το βάρος των μαθητών. (β) Ποια τάξη μπορεί να χαρακτηριστεί ομοιογενής ως προς το βάρος ων μαθητών;

Λύση

Υπολογίζουμε $CV_A = \frac{s_A}{\bar{x}_A} = \frac{6}{40} = 0{,}15 = 15\%$ και $CV_B = \frac{s_B}{\bar{x}_B} = \frac{6}{75} = 0{,}08 = 8\%$

(α) Είναι $CV_B < CV_A$ άρα η τάξη Β είναι περισσότερο ομοιογενής από την τάξη Α.

(β) Είναι $CV_B = 8\% < 10\%$ άρα η τάξη Β θεωρείται ομοιογενής. Καθώς $CV_A = 15\% > 10\%$, η τάξη Α δεν θεωρείται ομοιογενής.

Πίνακας 2.14: Υπολογισμός του συντελεστή ομοιογένειας στον υπολογιστή

	Δεν υπάρχει αντίστοιχη συνάρτηση αλλά μπορεί εύκολα να υπολογιστεί ως **STDEVP() / AVERAGE()**
	Δεν υπάρχει συνάρτηση, υπολογίζεται ως **sd(x)/mean(x)**

Δραστηριότητες

1. Ο παρακάτω πίνακας δίνει τον αριθμό των επισκέψεων 40 μαθητών σε διάφορα μουσεία της χώρας κατά τη διάρκεια ενός έτους

Επισκέψεις	**0 έως 2**	**2 έως 4**	**4 έως 6**	**6 έως 8**	**8 έως 10**
Συχνότητα	**8**	**12**	**10**	**6**	**4**

(α) Να συμπληρωθεί ο παρακάτω πίνακας

Επισκέψεις	**Κέντρο κλάσης (x_k)**	**Συχνότητα (ν_k)**	x_k^2	$x_k\,\nu_k$	$x_k^2\,\nu_k$
0 έως 2		8			
2 έως 4		12			
4 έως 6		10			
6 έως 8		6			
8 έως 10		4			
Σύνολο					

(β) Να υπολογιστεί το μέσο πλήθος επισκέψεων των 40 μαθητών.

(γ) Να υπολογιστεί η διακύμανση και η τυπική απόκλιση του πλήθους των επισκέψεων.

2. Η βαθμολογία δέκα μαθητών σε ένα διαγώνισμα ήταν: 7, 11, 10, 13, 15, 3, 12, 11, 4, 14. Να υπολογίσετε:

   α) τη μέση τιμή, την επικρατούσα τιμή και τη διάμεσο,

   β) τα Q_1, Q_3 και το ενδοτεταρτημοριακό εύρος.

   γ) το εύρος, τη διακύμανση, την τυπική απόκλιση και το συντελεστή ομοιογένειας.

2.9 Γεωμετρική ερμηνεία μέσης τιμής και τυπικής απόκλισης

Η μέση τιμή σε συνδυασμό με την τυπική απόκλιση δίνουν μια απλή και σύντομη εκτίμηση της κατανομής όλων των τιμών της μεταβλητής. Η εκτίμηση αυτή βασίζεται στην εμπειρική παρατήρηση πως οι τιμές μιας μεταβλητής θα βρίσκονται γύρω από τη μέση τιμή, σε απόσταση τριών τυπικών αποκλίσεων πριν και μετά τη μέση τιμή. Ποιο αναλυτικά :

α) Σε απόσταση μίας τυπικής απόκλισης από τη μέση τιμή αναμένουμε να βρίσκεται το 65% των παρατηρήσεων.

β) Σε απόσταση δύο τυπικών αποκλίσεων από τη μέση τιμή αναμένουμε να βρίσκεται το 95% των παρατηρήσεων.

γ) Σε απόσταση τριών τυπικών αποκλίσεων από τη μέση τιμή αναμένουμε να βρίσκεται το 99% των παρατηρήσεων.

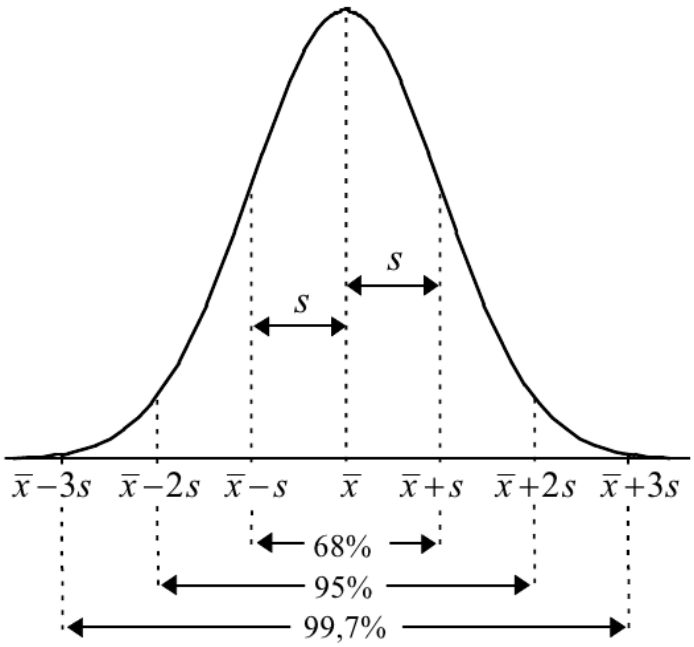
Οι παραπάνω παρατηρήσεις καταδεικνύουν την ουσία της στατιστικής επιστήμης η οποία είναι η μετάδοση όσο το δυνατόν περισσότερων πληροφοριών σχετικά με ένα σύνολο

αριθμών με τον περισσότερο σύντομο και εύληπτο τρόπο.

Έτσι, το λιγότερο που μπορούμε να κάνουμε για να περιγράψουμε σύντομα την κατανομή ενός συνόλου αριθμών είναι να υπολογίσουμε τη μέση τιμή και την τυπική απόκλιση των αριθμών αυτών. Κάθε τρίτος που γνωρίζει τους κανόνες α), β) και γ) είναι σε θέση να αναπαράξει σε γενικές γραμμές την κατανομή.

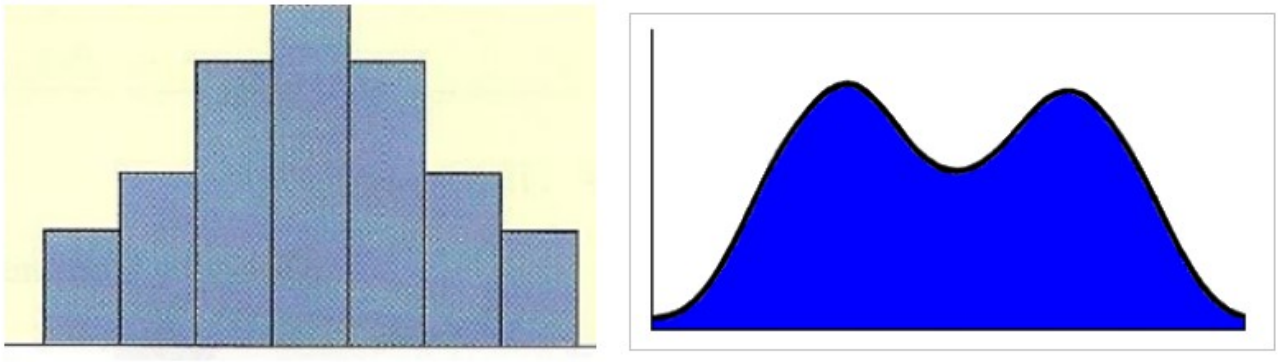
Δραστηριότητα

Σε δείγμα 30 σπουδαστών ενός ΙΕΚ βρέθηκε μέσο ύψος 168,3 και τυπική απόκλιση 4,2 μονάδες. (α) Αν το δείγμα αυτό είναι αντιπροσωπευτικό του συνόλου των σπουδαστών του ΙΕΚ τότε να βρεθεί διάστημα συμμετρικό γύρω από τη μέση τιμή στο οποίο θα ανήκει το ύψος του 68% περίπου των σπουδαστών. (β) Αν το ΙΕΚ έχει 150 σπουδαστές τότε πόσοι από αυτούς περιμένουμε να έχουν ύψος που θα ανήκει στο παραπάνω διάστημα;



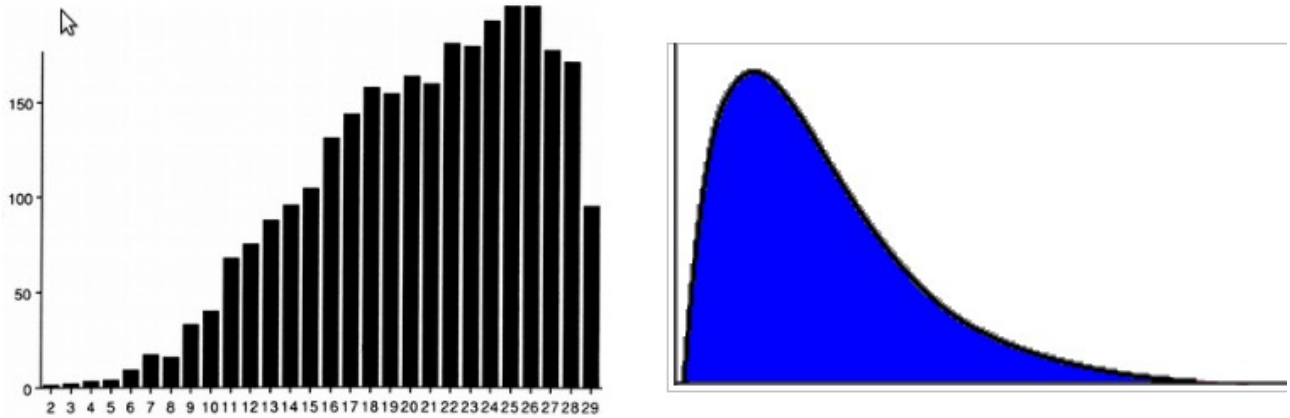
Διάγραμμα 10: Περιγραφή της κατανομής από τη μέση τιμή και την τυπική απόκλιση

2.10 Ασυμμετρία μίας κατανομής (Skewness).



Εικόνα 6: Παραδείγματα συμμετρικών κατανομών. Είναι φανερό πως και στις δύο περιπτώσεις υπάρχει μία κατακόρυφη γραμμή (η οποία ταυτίζεται με τη μέση τιμή αλλά και τη διάμεσο) η οποία χωρίζει σε δύο συμμετρικά μέρη την κατανομή.

Μία κατανομή συχνοτήτων (ή σχετικών συχνοτήτων) ονομάζεται συμμετρική όταν είναι φανερό πως υπάρχει ένας κατακόρυφος άξονας ο οποίος λειτουργεί ως καθρέπτης της μισής κατανομής στην άλλη μισή. Χαρακτηριστικό παράδειγμα είναι η κανονική κατανομή χωρίς αυτό να σημαίνει πως δεν μπορεί να είναι συμμετρική μία περισσότερο “ανώμαλη” κατανομή. Στο διάγραμμα 8, (σελίδα 97) παρουσιάζονται κάποια παραδείγματα συμμετρικών κατανομών.



Εικόνα 7: Παραδείγματα ασύμμετρων κατανομών.

Σημείωση για την εικόνα 10: Στην αριστερή εικόνα υπάρχουν κάποιες δυσανάλογα μικρές παρατηρήσεις, για τις οποίες η απόσταση από τη μέση τιμή είναι δυσανάλογα μεγάλη και αρνητική, κατά συνέπεια $\gamma < 0$ και λέμε ότι η κατανομή είναι λοξή προς τα αριστερά ή αρνητικά ασύμμετρη. Αντίθετα, στη δεξιά εικόνα η “ουρά” προς τα δεξιά δίνει

θετικό πρόσημο στο συντελεστή ασυμμετρίας (γ > 0) και λέμε ότι η κατανομή είναι λοξή προς τα δεξιά ή θετικά ασύμμετρη.

2.10.1 Αριθμητική εκτίμηση της ασυμμετρίας μίας κατανομής

Ο συντελεστής ασυμμετρίας ορίζεται ως :

$$\gamma = \frac{\frac{1}{v}\sum_{i=1}^{v}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{v}\sum_{i=1}^{v}(x_i - \bar{x})^2}\right)^3} = \frac{1}{s^3 v}\sum_{i=1}^{v}(x_i - \bar{x})^3$$

Τύπος 7: Συντελεστής ασυμμετρίας

Ο τύπος που δίνει το συντελεστή ασυμμετρίας φαίνεται ιδιαίτερα περίπλοκος, ωστόσο οι παρακάτω παρατηρήσεις κάνουν πιο καθαρό το τοπίο :

1. Ο συντελεστής ασυμμετρίας πρέπει να “ποσοτικοποιεί” το γεγονός ότι κάποιες από παρατηρήσεις απέχουν υπερβολικά προς τη μία μεριά σε σχέση με μέση τιμή της κατανομής.

2. Οι αποστάσεις των παρατηρήσεων από τη μέση τιμή δεν είναι παρά οι διαφορές $x_1-\bar{x}, x_2-\bar{x}, \cdots, x_v-\bar{x}$.

3. Το άθροισμα $\sum_{i=1}^{v}(x_i-\bar{x})$ είναι πάντα ίσο με το 0 άρα δεν μπορεί να χρησιμοποιηθεί για την ανίχνευση πιθανής ασυμμετρίας.

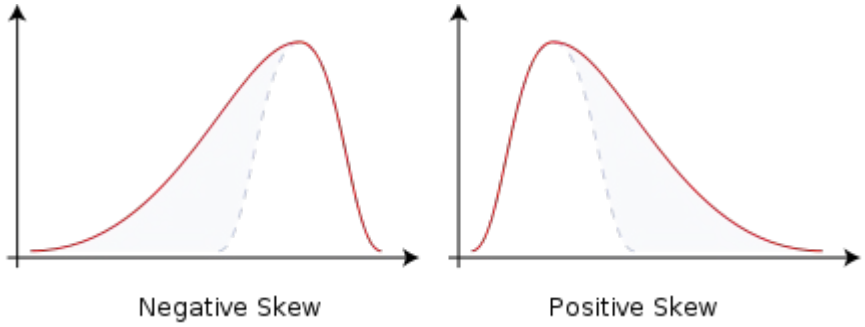
4. Το άθροισμα $\sum_{i=1}^{v}(x_i-\bar{x})^2$ ναι μεν εκφράζει το συνολικό μέγεθος της απόστασης των παρατηρήσεων από τη μέση τιμή αλλά δεν δίνει κάποια πληροφορία σχετικά με την συμμετρία της κατανομής καθώς όλοι οι όροι είναι θετικοί.

5. Το άθροισμα $\sum_{i=1}^{v}(x_i-\bar{x})^3$ διατηρεί το πρόσημο των διαφορών ενώ συγκεντρώνει και το μέγεθος της ασυμμετρίας σε μία αριθμητική ποσότητα. Η κανονικοποίηση του δείκτη ασυμμετρίας με τη διαίρεση με το πλήθος v και την τυπική απόκλιση s ακολουθεί με φυσικό τρόπο καθώς ο δείκτης πρέπει να μπορεί να έχει κοινή ερμηνεία σε όλες τις κατανομές ανεξαρτήτως διακύμανσης. Ο λόγος για τον οποίο

το στατιστικό 7 αποτελεί το κατάλληλο μέτρο ασυμμετρίας αποδεικνύεται και περισσότερο αυστηρά αλλά η απόδειξη ξεφεύγει από τους σκοπούς αυτών των σημειώσεων.

6. Στα περισσότερα στατιστικά προγράμματα ο υπολογισμός του συντελεστή ασυμμετρίας γίνεται με τον προσαρμοσμένο τυποποιημένο συντελεστή ασυμμετρίας Fisher-Pearson $\gamma = \frac{\nu}{s^3(\nu-1)(\nu-2)}\sum_{i=1}^{\nu}(x_i - \bar{x})^3$.

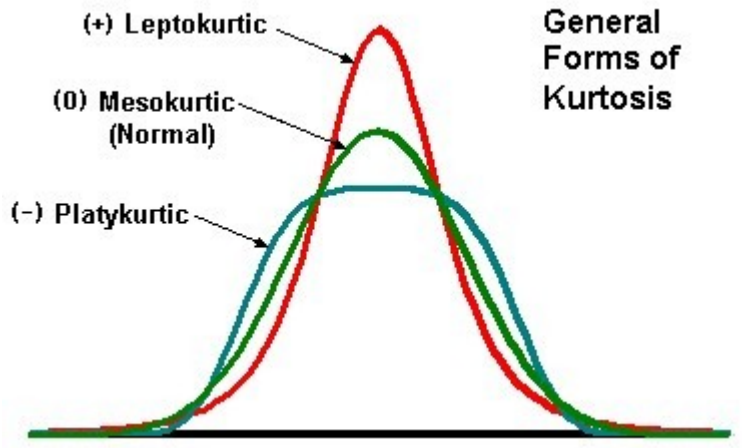
Ο συντελεστής ασυμμετρίας γ ερμηνεύεται με δύο τρόπους : Ο πρώτος είναι από το πρόσημό του και ο άλλος από το μέγεθος της απόλυτης τιμής του. Όταν γ > 0 τότε η κατανομή λέγεται λοξή προς τα δεξιά, ενώ όταν γ < 0 τότε η κατανομή λέγεται λοξή προς τα αριστερά (Διάγραμμα 13, σελίδα 90)



Διάγραμμα 11: Παραδείγματα με γ < ο (αρνητική συμμετρία) και γ > 0 (θετική ασυμμετρία) αντίστοιχα

2.11 Κυρτότητα μίας κατανομής (Kurtosis).

Η κανονική κατανομή είναι το μέτρο και για την κυρτότητα μίας κατανομής. Αν κάποια κατανομή έχει περισσότερο “οξεία” κορυφή από αυτή της κανονικής κατανομής τότε ονομάζεται Λεπτόκυρτη, ενώ όταν έχει περισσότερο “πλατιά” κορυφή τότε ονομάζεται Πλατύκυρτη (Διάγραμμα 16, σελίδα 96). Η κυρτότητα μίας κατανομής υπολογίζεται με το συντελεστή κυρτότητας του Pearson ο οποίος δίνεται από τον εξής τύπο :



Διάγραμμα 12: Οι τρεις εκδοχές κυρτότητας. Στη λεπτόκυρτη κατανομή είναι α < 3, στη μεσόκυρτη είναι α = 3, ενώ στην πλατύκυρτη είναι α > 3.

$$\alpha = \frac{\frac{1}{v}\sum_{i=1}^{v}(x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{v}\sum_{i=1}^{v}(x_i - \bar{x})^2}\right)^4} = \frac{1}{s^4 v}\sum_{i=1}^{v}(x_i - \bar{x})^4$$

Τύπος 8: Συντελεστής κυρτότητας.

Αποδεικνύεται πως στην κανονική κατανομή ο παραπάνω συντελεστής είναι ίσος με τον αριθμό 3. Στην περίπτωση της πλατύκυρτης κατανομής από τον τρόπο που ορίζεται ο συντελεστής κυρτότητας μπορούμε να καταλάβουμε πως θα παίρνει τιμές μεγαλύτερες του 3 (σκεφθείτε το γιατί!) ενώ στην περίπτωση της λεπτόκυρτης κατανομής θα παίρνει τιμές μικρότερες του 3.

Παρατήρησεις

1. Σε πολλά στατιστικά λογισμικά, πριν την εμφάνιση της τιμής της παραμέτρου γίνεται η αφαίρεση με το 3 και ως εκ τούτου η ερμηνεία της τιμής πρέπει να γίνει με ανάλογο τρόπο (α = 0 για την κανονική, α > 0 αν είναι πλατύκυρτη και α < 0 αν είναι λεπτόκυρτη). Ο χρήστης ενός στατιστικού προγράμματος είναι καλό να γνωρίζει τον τρόπο υπολογισμού για να είναι σε θέση να εκτιμήσει και το αποτέλεσμα.

2. Ο προσεκτικός αναγνώστης ίσως πρόσεξε πως ο συντελεστής ασυμμετρίας μπορεί να γραφεί

$$\gamma = \frac{1}{v}\sum_{i=1}^{v}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{1}{v}\sum_{i=1}^{v} z_i^3 \ ,$$

ενώ ο συντελεστής κυρτότητας

$$\alpha = \frac{1}{v}\sum_{i=1}^{v}\left(\frac{x_i - \bar{x}}{s}\right)^4 = \frac{1}{v}\sum_{i=1}^{v} z_i^4 \ ,$$

όπου $z_i = \frac{x_i - \bar{x}}{s}$ οι τυποποιημένες τιμές της μεταβλητής.

3. Αν τα δεδομένα προέρχονται από ένα δείγμα το οποίο θα χρησιμοποιηθεί για την εκτίμηση παραμέτρων του πληθυσμού στους τύπους 7 και 8 (σελίδα 69 και 71 αντίστοιχα), ο λόγος $\frac{1}{v}$ πρέπει να αντικατασταθεί από το $\frac{1}{v-1}$ καθώς αποδεικνύεται πως οι συντελεστές που θα προκύψουν με το νέο παρονομαστή θα αποτελούν καλύτεροι εκτιμητές για τις αντίστοιχες παραμέτρους του πληθυσμού. Η παρατήρηση αυτή δεν αναιρεί την παρατήρηση 6 της προηγούμενης παράγράφου : Για περισσότερες πληροφορίες : http://en.wikipedia.org/wiki/Kurtosis. Για λόγους απλοποίησης όπου εμφανίζονται ανάλογοι τύπου θα εμφανίζεται ο παρονομαστής ν.

Πίνακας 2.15: Υπολογισμός των συντελεστών ασυμμετρίας και κυρτότητας σε υπολογιστή

SKEW() και **KURT()** Για τον υπολογισμό της KURT απαιτούνται τουλάχιστον 4 παρατηρήσεις.

Υπολογίζουμε το συντελεστή ασυμμετρίας με **skewness(x)** και το συντελεστή κυρτότητας με **kurtosis(x)**. Οι συναρτήσεις βρίσκονται (και) στο πακέτο **moments**. Αν δεν είναι ήδη εγκατεστημένο τότε το εγκαθιστούμε με την εντολή **install.packages("moments", dependencies = TRUE)** και ακολουθεί η εντολή **library(moments)** για να είμαστε σε θέση να χρησιμοποιήσουμε τις συναρτήσεις.

Προσοχή : Από τον συντελεστή κυρτότητας όπως υπολογίζεται με τη συνάρτηση kurtosis δεν έχει αφαιρεθεί το 3 που αντιστοιχεί στην κανονική κατανομή.

2.12 Πότε η συμμετρία και η κυρτότητα της κατανομής διαφέρει σημαντικά από την κανονική;

Ο συντελεστής ασυμμετρίας γ είναι ίσος με 0 όταν οι παρατηρήσεις είναι συμμετρικά τοποθετημένες γύρω από τη μέση τιμή, γεγονός ασυνήθιστο στις πραγματικές έρευνες. Καθώς, ο ερευνητής σε περίπτωση που σκοπεύει να εφαρμόσει παραμετρικές δοκιμασίες πρέπει να είναι σε θέση να πιστοποιήσει τη συμμετρία της κατανομής, προκύπτει το

ερώτημα : πόσο μεγάλος πρέπει να είναι ο συντελεστής ασυμμετρίας για να θεωρήσουμε την κατανομή μη κανονική;

Μία απάντηση που μπορεί να χρησιμοποιηθεί στην πράξη είναι η εξής : Αν η απόλυτη τιμή του συντελεστή ασυμμετρίας είναι μικρότερη από το διπλάσιο του τυπικού σφάλματος της ασυμμετρίας (Standard Error of Skewness) τότε θεωρείται πως η συμμετρία της κατανομής δεν διαφέρει σημαντικά από την κανονική. Επιπλέον, υπάρχει η δοκιμασία D'Agostino με την οποία ελέγχεται η στατιστική υπόθεση πως η ασυμμετρία της δειγματικής κατανομής δεν διαφέρει σημαντικά από αυτή της κανονικής (δηλαδή από το 0), διαθέσιμη στο R – Project στη βιβλιοθήκη moments (διαθέσιμη με την εντολή **library(moments)**) εκτελώντας

agostino.test(x, alternative = "two.sided") . Αν μας απασχολεί μόνο η μία από τις ανισότητες τότε μπορεί να οριστεί alternative = "less", ή alternative = "greater"

Ανάλογος πρακτικός κανόνας ισχύει και για την κυρτότητα μίας κατανομής : Αν η απόλυτη τιμή του συντελεστή κυρτότητας απέχει από τον αριθμό 3 το πολύ το διπλάσιο του τυπικού σφάλματος της κυρτότητας (Standard Error of Kurtosis) τότε θεωρείται πως η κυρτότητα της κατανομής δεν διαφέρει σημαντικά από την κανονική. Επιπλέον, υπάρχει η δοκιμασία Anscombe-Glynn με την οποία ελέγχεται η στατιστική υπόθεση πως η κυρτότητα της δειγματικής κατανομής δεν διαφέρει σημαντικά από αυτή της κανονικής (δηλαδή από το 3), διαθέσιμη στο R – Project στη βιβλιοθήκη moments (**library(moments)**) με την εντολή

anscombe.test(x, alternative = "two.sided"). Αν μας απασχολεί μόνο η μία από τις ανισότητες τότε μπορεί να οριστεί alternative = "less", ή alternative = "greater"

Στην πράξη, οι συντελεστές ασυμμετρίας και κυρτότητας χρησιμοποιούνται για μία πρώτη επισκόπηση της κατανομής. Αν, από την ανάγνωση των συντελεστών δεν προκύπτει καθαρό συμπέρασμα τότε ο ερευνητής πρέπει να προχωρήσει σε άλλες μεθόδους όπως οι δοκιμασίες Kolmogorov – Smirnof και Shapiro – Wilk με τις οποίες ελέγχεται η απόκλιση των τιμών της κατανομής από τις αντίστοιχες τιμές μίας κανονικής κατανομής.

Πίνακας 2.16: Δοκιμασίες κανονικότητας κατανομής με υπολογιστή

Όπως αναφέρθηκε η εντολή **agostino.test(x, alternative = “two.sided”)** ελέγχει την απόκλιση της ασυμμετρίας της δειγματικής κατανομής από την κανονική ενώ η **anscombe.test(x, alternative = "two.sided")** κάνει το ίδιο για την κυρτότητα. Για τη δοκιμασία K-S αρκεί η συνάρτηση **ks.test(x, pnorm, mean(x), sd(x)).** π.χ. Η δοκιμή στο διάνυσμα **x = rnorm(1000)** κάνει δεκτή την μηδενική υπόθεση της κανονικότητας ενώ αν εφαρμοστεί στο διάνυσμα **x = runif(1000)** απορρίπτεται.

2.13 Οι συντελεστές ασυμμετρίας και κυρτότητας ως στιγμές μίας τυχαίας μεταβλητής.

Η αναμενόμενη τιμή EX (expected value) μίας διακριτής τυχαίας μεταβλητής X ορίζεται να είναι το άθροισμα όλων των πιθανών τιμών που παίρνει η μεταβλητή πολλαπλασιασμένο με την εκάστοτε πιθανότητα να ληφθεί η τιμή αυτή.

Στην πράξη, αν X είναι τυχαία μεταβλητή και $x_{1,}x_{2,}\cdots,x_{v}$ οι διαφορετικές τιμές της, τότε η αναμενόμενη τιμή μίας συνάρτησης του X υπολογίζεται ως η μέση τιμή των τιμών της συνάρτησης αυτής στις διαφορετικές τιμές που λαμβάνει η X. Για παράδειγμα, σύμφωνα με τα παραπάνω,

$$EX \;=\; \frac{x_1+x_2+\cdots+x_v}{v} \;=\; \bar{x}$$

Η EX ονομάζεται πρώτη στιγμή της τυχαίας μεταβλητής X. Γενικότερα, η κ στιγμή (k-th moment) της τυχαίας μεταβλητής X ορίζεται να είναι

$$\kappa_\kappa \;=\; E\,X^\kappa$$

ενώ αν μ = EX είναι η αναμενόμενη (μέση) τιμή τότε η κ-κεντρική στιγμή (k-th central moment) της τυχαίας μεταβλητής X ορίζεται να είναι

$$\mu_\kappa \;=\; E[(X-EX)^\kappa] \;=\; E[(X-\mu)^\kappa]$$

Για παράδειγμα, αν X είναι τυχαία μεταβλητή και $x_{1,}x_{2,}\cdots,x_{v}$ οι διαφορετικές τιμές της τότε

$$E(X-\mu)^2 \;=\; \frac{1}{v}\sum_{i=1}^{v}(x_i-\bar{x})^2 \;=\; \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_v-\bar{x})^2}{v} \;=\; s^2$$

Ο αναγνώστης μπορεί εύκολα να επαληθεύσει πως με την ορολογία αυτή, ο τύπος 7, σελίδα 69 του συντελεστή ασυμμετρίας γράφεται

$$\gamma = \frac{1}{s^3 v}\sum_{i=1}^{v}(x_i - \bar{x})^3 = \frac{\mu_3}{s^3}$$

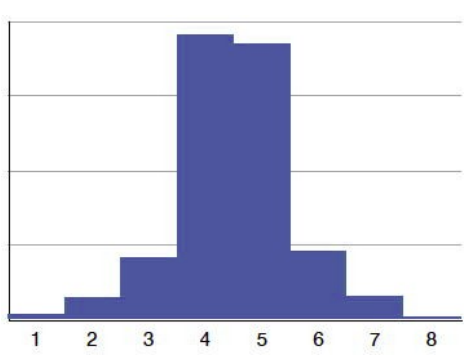
ενώ ο τύπος 8, σελίδα 71 του συντελεστή κυρτότητας γράφεται

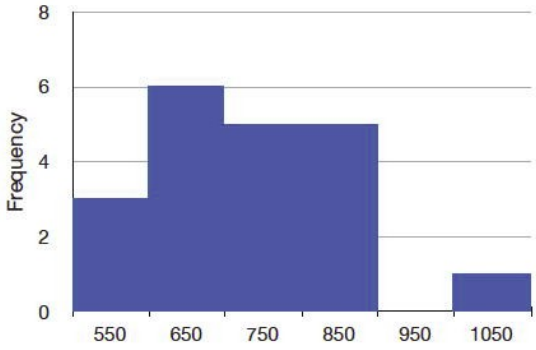
$$\alpha = \frac{1}{s^4 v}\sum_{i=1}^{v}(x_i - \bar{x})^4 = \frac{\mu_4}{s^4}$$

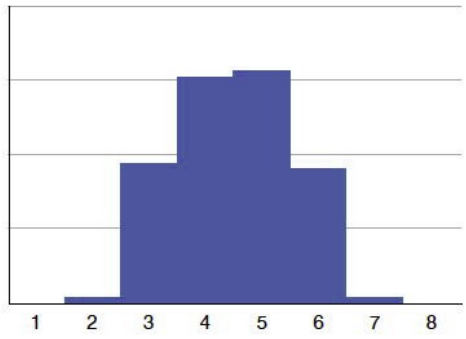
Στη βιβλιογραφία ορίζονται και άλλες παραλλαγές των στιγμών μίας τυχαίας μεταβλητής. Αναφέρονται χαρακτηριστικά οι ν-οστές αντίστροφες στιγμές $E(X^{-v})$ και οι ν-οστές λογαριθμικές στιγμές ως $E(\ln^v X)$.

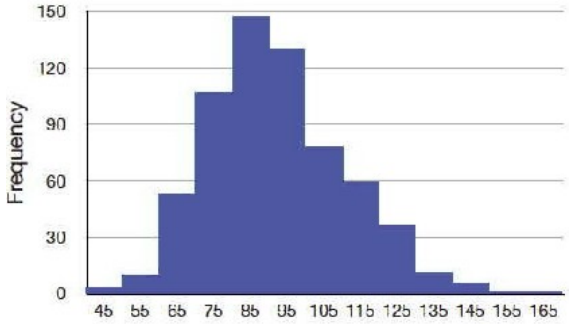
Δραστηριότητες

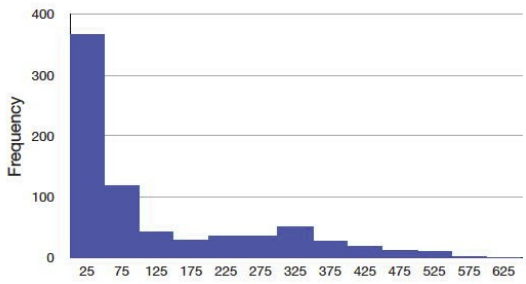
1. Να χαρακτηρίσετε τις παρακάτω κατανομές ως προς την κυρτότητα και τη συμμετρία τους.

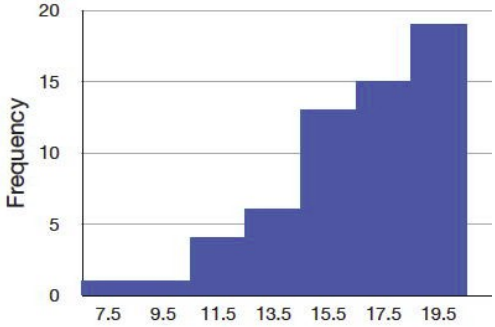












2. Δίνεται η τυχαία μεταβλητή X η οποία παίρνει τις ισοπίθανες τιμές $x_1 = 1$, $x_2 = 6$, $x_3 = 8$ και $x_4 = 17$, δηλαδή $P(X = x_1) = P(X = x_2) = P(X = x_3) = P(X = x_4) = 1/4$. Να υπολογιστούν :

(α) η αναμενόμενη τιμή της μεταβλητής X. (β) η $2^η$, η $3^η$ και η $4^η$ κεντρική στιγμή της μεταβλητής X.

(γ) ο συντελεστής ασυμμετρίας της κατανομής (δ) ο συντελεστής κυρτότητας της κατανομής.

(ε) ποιο το συμπέρασμα για την κατανομή των τιμών;

2.14 Χρήση των συναρτήσεων βάσης δεδομένων του Calc

Στον πίνακα 2.17 που ακολουθεί παρουσιάζονται ορισμένες βασικές στατιστικές συναρτήσεις που υποστηρίζονται από το LibreOffice Calc.

Πίνακας 2.17: Βασικές Στατιστικές Συναρτήσεις του LibreOffice Calc

Μέτρα Θέσης		
Στατιστικό	**Συνάρτηση**	**Σχόλιο**
Επικρατούσα Τιμή	MODE()	
1ο Τεταρτημόριο	QUARTILE(κελιά, 1)	
Διάμεση Τιμή	MEDIAN()	ή QUARTILE(κελιά, 2)
3ο Τεταρτημόριο	QUARTILE(κελιά, 3)	
Αριθμητικός Μέσος	AVERAGE()	
Αριθμητικός Μέσος	DAVERAGE()	Υπολογισμός μέσου κελιών από βάση δεδομένων.
Μέτρα Διασποράς		
Εύρος	RANGE()	
Ενδοτεταρτημοριακό Εύρος	QUARTILE(κελιά, 3) - QUARTILE(κελιά, 1)	
Απόλυτη Απόκλιση	AVEDEV()	
Διακύμανση	VAR()	Διαίρεση με (πλήθος - 1)
Τυπική Απόκλιση	STDEV()	Διαίρεση με (πλήθος - 1)
Διακύμανση	DVAR()	Υπολογισμός διακύμανσης από κελιά βάσης δεδομένων
Τυπική Απόκλιση	DSTDEV()	Υπολογισμός τυπικής απόκλισης από κελιά βάσης

		δεδομένων
Διακύμανση Πληθυσμού	VARP()	Διαίρεση με (πλήθος)
Τυπική Απόκλιση Πληθυσμού	STDEVP()	Διαίρεση με (πλήθος)
	Διάφορα	
Άθροισμα	SUM()	

Τα στατιστικά του πίνακα 2.17 είναι ικανά να καλύψουν τις περισσότερες ανάγκες ενός μέσου χρήστη, για την περιγραφή ενός συνόλου δεδομένων.

Παράδειγμα

Οι Hoaglin, Mosteller και Tukey (1983) παρουσιάζουν δεδομένα από μια έρευνα που πραγματοποίησαν μελετώντας την επίδραση του άγχους στα επίπεδα της β-ενδορφίνης στο αίμα. Μέτρησαν το επίπεδο της β-ενδορφίνης σε 19 ασθενείς 12 ώρες πριν από την εγχείριση και 10 λεπτά πριν από την εγχείριση. Τα δεδομένα παρουσιάζονται στον πίνακα που ακολουθεί σε fmol/ml καταχωρημένα σε φύλλο Calc:

	A	B	C	D
1	**Ασθενής**	**Φύλο**	**12 ώρες πριν**	**10 λεπτά πριν**
2	1	Άνδρας	10	6,5
3	2	Γυναίκα	6,5	14
4	3	Άνδρας	8	13,5
5	4	Άνδρας	12	18
6	5	Γυναίκα	5	14,5
7	6	Γυναίκα	11,5	9
8	7	Άνδρας	5	18
9	8	Άνδρας	3,5	42
10	9	Γυναίκα	7,5	7,5
11	10	Γυναίκα	5,8	6
12	11	Γυναίκα	4,7	25
13	12	Άνδρας	8	12
14	13	Άνδρας	7	52
15	14	Γυναίκα	17	20
16	15	Άνδρας	8,8	16
17	16	Άνδρας	17	15
18	17	Γυναίκα	15	11,5
19	18	Άνδρας	4,4	2,5
20	19	Άνδρας	2	2

Πίνακας 2.18: Δεδομένα παραδείγματος

Να συμπληρωθούν οι παρακάτω πίνακες

Πίνακας 1

	Μέση Τιμή (μ)		Διακύμανση (σ^2)		Τυπική Απόκλιση (σ)	
	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν
Σύνολο						
Άνδρες						
Γυναίκες						

Πίνακας 2

Υπολογισμοί 12 ώρες πριν		Πλήθος ασθενών
μ – σ	μ + σ	
μ – 2σ	μ + 2σ	
μ – 3σ	μ + 3σ	

Πίνακας 3

Υπολογισμοί 10 λεπτά πριν		Πλήθος ασθενών
μ – σ	μ + σ	
μ – 2σ	μ + 2σ	
μ – 3σ	μ + 3σ	

Διαδικασία Επίλυσης

Για το σύνολο των δεδομένων, τα στατιστικά της μέσης τιμής, της διακύμανσης και της τυπικής απόκλισης υπολογίζονται άμεσα από τις στατιστικές συναρτήσεις AVERAGE(), VAR() και STDEV() αντίστοιχα. Ωστόσο, για τον υπολογισμό των ίδιων στατιστικών για άνδρες και γυναίκες πρέπει να εφαρμόσουμε τις συναρτήσεις DAVERAGE(), DVAR() και DSTDEV() αντίστοιχα.

Για παράδειγμα, ο υπολογισμός της μέσης τιμής 12 ώρες πριν για τους άνδρες συμμετέχοντες στην έρευνα μπορεί να γίνει με τη συνάρτηση

=DAVERAGE(A1:D20; C1; B23:B24)

όπου στα κελιά B23:B24 έχει καταχωρηθεί ο πίνακας

Φύλο
Άνδρας

ο οποίος χρησιμοποιείται για την επιλογή των δεδομένων.

Περισσότερο αναλυτικά : Οι συναρτήσεις που εφαρμόζονται σε βάση δεδομένων έχουν ως πρώτο γράμμα το **D** (όπως Database – Βάση δεδομένων). Μία βάση δεδομένων στο λογιστικό φύλλο LibreOffice Calc αντιστοιχεί σε ένα σύνολο από κελιά στην πρώτη γραμμή των οποίων βρίσκονται οι ετικέτες κάθε μίας στήλης. Στα δεδομένα του πίνακα 2.18, σελίδα 77 η βάση δεδομένων είναι τα κελιά **A1:D20** ενώ στην γραμμή 1 είναι οι ετικέτες κάθε μίας στήλης.

Το πρώτο όρισμα κάθε συνάρτησης αυτού του τύπου όπως η DAVERAGE() είναι τα κελιά με όλα τα στοιχεία που απαιτούνται για να γίνει ο επιθυμητός υπολογισμός, στην περίπτωσή μας τα κελιά A1:D20. Το δεύτερο όρισμα είναι η επικεφαλίδα της στήλης από την οποία θα επιλεγούν τα στοιχεία, “12 ώρες πριν” στην δική μας περίπτωση ή απλά το κελί C1. Τέλος, το τρίτο όρισμα είναι τα κριτήρια επιλογής, τα οποία δίνονται στη μορφή (επικεφαλίδα – τιμή) σε δύο γραμμές. Με ανάλογη χρήση των συναρτήσεων DVAR() και DSTDEV() συμπληρώνεται ο πίνακας όπως παρακάτω :

Πίνακας 1

	Μέση Τιμή (μ)		Διακύμανση (σ^2)		Τυπική Απόκλιση (σ)	
	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν	12 ώρες πριν	10 λεπτά πριν
Σύνολο	8,4	16,1	19,3	156,5	4,4	12,5
Άνδρες	7,8	18,0	18,0	243,0	4,2	15,6
Γυναίκες	9,1	13,4	22,8	41,7	4,8	6,5

Για τη συμπλήρωση των ποσοτήτων μ ± σ, μ ± 2σ, και μ ± 3σ στους πίνακες 2 και 3 του παραδείγματος αρκεί ο απευθείας υπολογισμός με εισαγωγή τύπων. Η καταμέτρηση ωστόσο των περιπτώσεων που βρίσκονται μεταξύ των δύο αριθμών απαιτεί τη χρήση της συνάρτησης SUMPRODUCT η οποία λαμβάνει όσα ορίσματα χρειάζονται στη μορφή Πίνακας – Συνθήκη για την καταμέτρηση των κελιών. Για παράδειγμα αν θέλουμε να βρούμε το πλήθος των μετρήσεων 12 ώρες πριν που είναι μεταξύ 5 και 8 μπορούμε να γράψουμε :

=SUMPRODUCT(C2:C20>5;C2:C20<8)

Το τελικό αποτέλεσμα θα είναι το παρακάτω :

Πίνακας 2				
	Υπολογισμοί 12 ώρες πριν			Πλήθος ασθενών
μ – σ	4,0	μ + σ	12,7	14
μ – 2σ	-0,4	μ + 2σ	17,1	19
μ – 3σ	-4,8	μ + 3σ	21,5	19

Πίνακας 3				
	Υπολογισμοί 10 λεπτά πριν			Πλήθος ασθενών
μ – σ	3,5	μ + σ	28,6	15
μ – 2σ	-9,0	μ + 2σ	41,1	17
μ – 3σ	-21,5	μ + 3σ	53,6	19

Σημείωση : Οι υπολογισμοί που παρουσιάζονται παραπάνω αρχικά εμφανίστηκαν με πολλά δεκαδικά ψηφία και έπειτα επιλέχθηκε η προβολή μόνο ενός με την επιλογή **Μορφή -> Κελιά** -> *Κατηγορία : Αριθμός* και *Επιλογές : Δεκαδικά ψηφία : 1*.

Στον παρακάτω πίνακα παρουσιάζονται μερικές ακόμα συναρτήσεις υπολογισμού στοιχείων από βάση δεδομένων.

Συνάρτηση	Περιγραφή
DCOUNT	Μετρά τις αριθμητικές καταχωρήσεις σε μία στήλη μίας βάσης δεδομένων που ικανοποιούν τα κριτήρια που ζητούνται.
DMAX	Βρίσκει τη μέγιστη τιμή σε μία στήλη μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DMIN	Βρίσκει τη ελάχιστη τιμή σε μία στήλη μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DAVERAGE	Υπολογίζει τον αριθμητικό μέσο των τιμών μίας στήλης μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DPRODUCT	Υπολογίζει το γινόμενο των τιμών μίας στήλης μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.
DSUM	Υπολογίζει το άθροισμα των τιμών μίας στήλης μίας βάσης δεδομένων που ικανοποιεί τα κριτήρια που ζητούνται.

2.15 Τυποποιημένες τιμές

Έστω $x_1, x_2, \ldots, x_k$ αριθμητικές παρατηρήσεις με μέση τιμή $\bar{x}$ και τυπική απόκλιση s. Ονομάζουμε τυποποιημένες ή τυπικές τιμές των παρατηρήσεων αυτών τους αριθμούς :

$$z_1 = \frac{x_1 - \bar{x}}{s},\ z_2 = \frac{x_2 - \bar{x}}{s},\ \ldots,\ z_k = \frac{x_k - \bar{x}}{s}\ .$$

Παράδειγμα

Αν $x_1=10$, $x_2=15$ και $x_3=35$ τότε

- $\bar{x} = \frac{10 + 15 + 35}{3} = 20$
- $s^2 = \frac{[(10-20)^2 + (15-20)^2 + (35-20)^2]}{3} = \frac{(100 + 25 + 225)}{3} = \frac{350}{3} \simeq 116{,}67$
- $s = \sqrt{s^2} \simeq 10{,}8$

και οι τυποποιημένες (ή τυπικές) τιμές των x_1 , x_2 , x_3 είναι αντίστοιχα :

$$z_1=\frac{x_1-\bar{x}}{s}=\frac{10-20}{10{,}8}=-0{,}926,\ z_2=\frac{x_2-\bar{x}}{s}=\frac{15-20}{10{,}8}=-0{,}463 \text{ και } z_3=\frac{x_3-\bar{x}}{s}=\frac{35-20}{10{,}8}=1{,}389$$

Προσέξτε πως αν και οι αρχικές τιμές ήταν διψήφιοι αριθμοί οι αριθμοί που προκύπτουν ως τυποποιημένες τιμές είναι σχετικά “μικροί”. Το γεγονός αυτό δεν είναι τυχαίο καθώς εύκολα αποδεικνύεται ότι :

Αν οι παρατηρήσεις $x_1, x_2, \ldots, x_k$ έχουν μέση τιμή $\bar{x}$ και τυπική απόκλιση s τότε οι αντίστοιχες τυποποιημένες τιμές έχουν μέση τιμή 0 και τυπική απόκλιση 1.

Η παραπάνω παρατήρηση πρακτικά σημαίνει (αν θυμηθούμε και τον κανόνα των τριών τυπικών αποκλίσεων που αναφέρθηκε στην παράγραφο 2.9, σελίδα 66, πως πρακτικά οι τυποποιημένες τιμές πρέπει να βρίσκονται μεταξύ -3 και +3.

Οι τυποποιημένες τιμές είναι χρήσιμες όταν θέλουμε να συγκρίνουμε την σχετική απόσταση δύο παρατηρήσεων από διαφορετικά δείγματα με διαφορετικές μέσες τιμές και διαφορετικές τυπικές αποκλίσεις. Απλά υπολογίζουμε τις τυποποιημένες αντίστοιχες τιμές και η παρατήρηση με τη μεγαλύτερη τυποποιημένη τιμή είναι περισσότερο ακραία στο δικό της δείγμα από την άλλη στο άλλο δείγμα.

Πίνακας 2.19: Υπολογισμός τυποποιημένων τιμών από υπολογιστή

Συνάρτηση **STANDARDIZE()** η οποία παίρνει 3 ορίσματα, την τιμή προς τυποποίηση, την μέση τιμή του δείγματος και την τυπική απόκλιση αυτού.

Στο βασικό πακέτο δεν υπάρχει ανάλογη συνάρτηση ωστόσο μπορεί εύκολα να υπολογιστούν οι τυποποιημένες τιμές με τη συνάρτηση **(x – mean(x))/sd(x).** Αν **x = c(10, 20, 30)** και **y = (x – mean(x))/sd(x)** τότε y = [-1, 0, 1].

Παράδειγμα

Στην χώρα Α το μέσο εισόδημα είναι 6.000 ευρώ με τυπική απόκλιση 2.000 ενώ στη χώρα Β το μέσο εισόδημα είναι 10.000 με τυπική απόκλιση 3.000. Ρωτάμε έναν κάτοικο από τη χώρα Α και μας λέει πως έχει εισόδημα 8.000 και έναν κάτοικο από τη χώρα Β και αποκρίνεται πως έχει εισόδημα 11.000. Ποιος είναι περισσότερο πλούσιος σχετικά με τους υπόλοιπους κατοίκους της χώρας του;

Λύση

Αρκεί να υπολογίσουμε τις τυποποιημένες μέσες τιμές. Για τον κάτοικο της χώρας Α υπολογίζουμε

$$z_A = \frac{8.000 - 5.000}{2.000} = 1.5$$

ενώ για τον κάτοικο της χώρας Β είναι

$$z_B = \frac{11.000 - 10.000}{3.000} = 0.333$$

Καθώς $z_A > z_B$, συμπεραίνουμε πως ο κάτοικος της χώρας Α αν και έχει μικρότερο εισόδημα από αυτόν της χώρας Β, είναι περισσότερο πλούσιος συγκριτικά με το περιβάλλον του σε σχέση με τον κάτοικο της χώρας Β.

2.16 Μέση Διαφορά του Gini

Ένα ακόμη και όχι τόσο δημοφιλές μέτρο διασποράς είναι η μέση διαφορά του Gini (Gini's mean difference), γνωστή και ως απόλυτη μέση διαφορά (absolute mean difference). Αν $x_1, x_2, \ldots, x_k$ είναι οι παρατηρήσεις, η μέση διαφορά του Gini ορίζεται να είναι :

$$MD = \frac{1}{\kappa(\kappa - 1)} \sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} |x_i - x_j|$$

Παράδειγμα

Αν $x_1=10$, $x_2=15$ και $x_3=35$ τότε κ = 3 και

$$\begin{aligned} MD &= \frac{1}{3(3-1)}\sum_{i=1}^{3}\sum_{j=1}^{3}|x_i-x_j| \\ &= \frac{1}{6}(|10-10|+|15-10|+|35-10|+|10-15|+|15-15|+|35-15|+|10-35|+|15-35|+|35-35|) \\ &= \frac{1}{6}(0+5+25+5+0+20+25+20+0) \\ &= \frac{1}{6}\cdot 100=16{,}67 \end{aligned}$$

2.16.1 Σύγκριση της μέσης διαφοράς του Gini με την τυπική απόκλιση.

Τόσο η μέση διαφορά του Gini όσο και η τυπική απόκλιση αποτελούν μέτρα διασποράς, δηλαδή είναι στατιστικά που αποσκοπούν στην περιγραφή της διασποράς των παρατηρήσεων. Η μέση διαφορά του Gini δεν χρειάζεται κάποιο προηγούμενο στατιστικό υπολογισμό όπως η μέση τιμή που απαιτείται για την τυπική απόκλιση. Καθώς, για τον υπολογισμό της τυπικής απόκλισης υπολογίζονται τα τετράγωνα των αποστάσεων κάθε παρατήρησης από τη μέση τιμή, υπάρχει υπερεκτίμηση των μεγάλων παρατηρήσεων και υποεκτίμηση των μικρών παρατηρήσεων σε αντίθεση με τη μέση διαφορά του Gini.

2.17 Καμπύλη Lorenz και δείκτης Gini

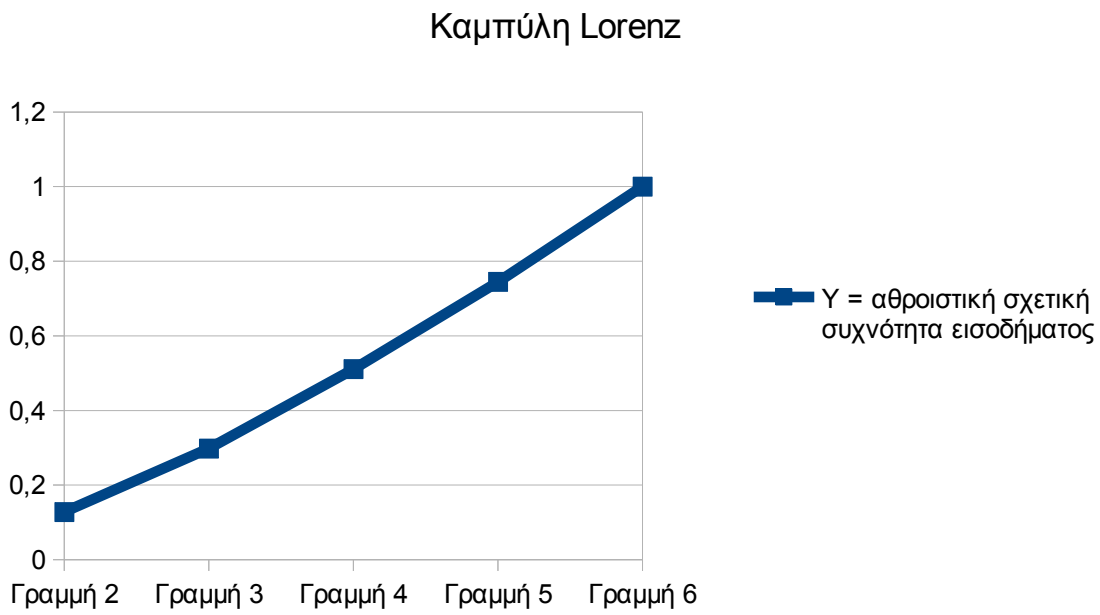
Η καμπύλη του Lorenz (ή Lorenz curve) είναι ένα γράφημα με το οποίο αναπαριστάται γραφικά τυχόν δυσαρμονία στις τιμές μίας μεταβλητής. Υπολογίζεται κυρίως όταν η μεταβλητή που μελετούμε σχετίζεται με κάποιο χρηματικό μέγεθος. Η καμπύλη αποτελείται από τόσα σημεία όσες και οι παρατηρήσεις. Αρχικά, ταξινομούνται σε αύξουσα σειρά οι παρατηρήσεις και μετά ακολουθούν οι απαραίτητοι υπολογισμοί. Κάθε σημείο της καμπύλης έχει ως τετμημένη την αθροιστική σχετική συχνότητα της παρατήρησης ενώ ως τεταγμένη λαμβάνεται η αθροιστική σχετική συχνότητα της ίδιας της παρατήρησης σε σχέση με το συνολικό άθροισμα όλων των παρατηρήσεων.

Παράδειγμα

Ρωτήθηκαν έξι άτομα για τον μηνιαίο μισθό τους σε ευρώ και δόθηκαν οι παρακάτω αποκρίσεις $x_1=600$, $x_2=1200$, $x_3=1000$, $x_4=800$ και $x_5=1100$. Το άθροισμα όλων των παρατηρήσεων είναι $\sum_{i=1}^{6} x_i=4700$. Συμπληρώνουμε τον παρακάτω πίνακα :

Άτομο (φθίνουσα σειρά μισθού)	Μισθός	Αθροιστικός μισθός	X (αθροιστική σχετική συχνότητα)	Y = αθροιστική σχετική συχνότητα εισοδήματος	Ιδεατή τιμή του Y
1	600	600	0,2	0,13	0,2
4	800	1400	0,4	0,3	0,4
3	1000	2400	0,6	0,51	0,6
5	1100	3500	0.8	0,74	0.8
2	1200	4700	1.0	1	1.0

Και η καμπύλη Lorenz είναι η εξής:



Παρατηρούμε πως η καμπύλη Lorenz έχει μία μικρή κοιλιά προς τα κάτω, κάτι που σημαίνει πως υπάρχει μία μικρή δυσαναλογία στην κατανομή των μισθών : Αυτός που έχει μισθό 1200 ευρώ παίρνει δυσανάλογα περισσότερα από ότι παίρνει αυτός με μισθό 600 ευρώ.

2.17.1 Δείκτης Gini

Ο δείκτης Gini είναι ένα μέτρο της ανισότητας μίας κατανομής. Αναπτύχθηκε από τον Ιταλό Στατιστικό Corrado Gini και δημοσιεύθηκε στην εργασία του “Variabilite e mutabilite” το 1912. Χρησιμοποιείται κυρίως για να εκφράζει την ανισότητα των εισοδημάτων αλλά μπορεί να χρησιμοποιηθεί για τη μέτρηση οποιαδήποτε “άδικης” κατανομής.

Μιλώντας απλά, ο δείκτης Gini είναι ο λόγος του εμβαδού E_1 που ορίζεται από την ευθεία y

= x και της καμπύλης Lorenz και του εμβαδού E_2 του ορθογώνιου τριγώνου με βάση το πλήθος των παρατηρήσεων και ύψος μονάδα. Ο δείκτης Gini λαμβάνει τιμές μεταξύ 0 και 1. Η τιμή 0 αντιστοιχεί σε μία απολύτως “ισορροπημένη” κατάσταση όπου οι μισθοί κατανέμονται σε αναλογικά μέρη του πληθυσμού. Αντίθετα η τιμή 1 λαμβάνεται όταν υπάρχει “ολοκληρωτισμός” : ένας κατέχει το σύνολο του πλούτου ενώ οι υπόλοιποι δεν κατέχουν οτιδήποτε. Στην πράξη, στις χώρες του κόσμου, ο δείκτης Gini λαμβάνει τιμές από 0,25 έως 0,7. Στον παρακάτω πίνακα παρουσιάζονται οι 10 χώρες με την καλύτερη και οι 10 με την χειρότερη βαθμολογία στο δείκτη Gini. Ο πλήρης κατάλογος μπορεί να βρεθεί στην ιστοσελίδα : https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html#top

10 Καλύτερες τιμές δείκτη Gini			**10 Χειρότερες τιμές δείκτη Gini**		
1	Finland	26.8	140	Namibia	70.7
2	Kazakhstan	26.7	139	Seychelles	65.8
3	Slovakia	26.0	138	South Africa	65.0
4	Luxembourg	26.0	137	Lesotho	63.2
5	Malta	26.0	136	Botswana	63.0
6	Austria	26.0	135	Sierra Leone	62.9
7	Czech Republic	26.0	134	Central African Republic	61.3
8	Norway	25.0	133	Haiti	59.2
9	Hungary	24.7	132	Colombia	58.5
10	Sweden	23.0	131	Bolivia	58.2

2.18 Καταγραφή περιγραφικών στατιστικών

Ως προς τους αριθμούς που εμφανίζονται σε μία δημοσίευση ή διατριβή αρκεί να παρουσιάζονται όσα δεκαδικά ψηφία είναι απαραίτητα για την σωστή ερμηνεία των αποτελεσμάτων. Αν ωστόσο ο συγγραφέας δεν είναι σίγουρος για την επιλογή του τότε μπορεί να συμβουλευθεί τον παρακάτω πίνακα :

Πίνακας 2.20: Παρουσίαση αριθμητικών δεδομένων			
Είδος αριθμού	**Δεκαδικά ψηφία που πρέπει να εμφανιστούν**	**Παράδειγμα**	
		Υπολογισμός	**Παρουσίαση**
Μεγαλύτερος του 100	Κανένα	μέση τιμή 1322,58	M = 1323
Μεταξύ 10 και 100	1	μέση τιμή 34,133	M = 34,1
Μεταξύ 0,1 και 10	2	6,5555	M = 6,56
Μικρότερος από 0,1	Όσα πρέπει ώστε να μην εμφανίζεται μηδέν.	Μέση τιμή 0,003278	M = 0,003

Ως προς τα περιγραφικά στατιστικά υπάρχουν κάποιες συμβάσεις στην καταγραφή τους που υιοθετούνται από τα περισσότερα περιοδικά. Πιο συγκεκριμένα :

- M : Μέση τιμή (mean)
- SD : Τυπική απόκλιση (standard deviation)
- SE : Τυπικό σφάλμα (standard error)

Η μέση τιμή και η τυπική απόκλιση μπορούν να παρουσιάζονται στο κείμενο είτε σε έναν πίνακα αλλά δεν συνιστάται και με τους δύο τρόπους. Παραδείγματα είναι τα εξής :

- Η μέση ηλικία των ερωτώμενων του δείγματος ήταν 25,5 χρόνια (SD = 7.94).
- Η ηλικία των ερωτώμενων ήταν από 18 έως 70 έτη (M = 25.5, SD = 7.94). Η ηλικία δεν ήταν κανονικά κατανεμημένη καθώς ο συντελεστής ασυμμετρίας υπολογίστηκε 1.87 (SE = 0.05) και ο συντελεστής κυρτότητας 3.93 (SE = 0.10)
- Οι συμμετέχοντες ήταν 98 άνδρες και 132 γυναίκες ηλικίας 17 έως 25 ετών (άνδρες : M = 19.2, SD = 2.32, γυναίκες : M = 19.6, SD = 2.54).

Δραστηριότητες

1. Τι από τα παρακάτω είναι σωστό;
   i) Η τυπική απόκλιση είναι το μισό της διακύμανσης
   ii) Η τυπική απόκλιση είναι το τετράγωνο της διακύμανσης.
   Iii) Η διακύμανση είναι το μισό της τυπικής απόκλισης.
   iv) Η διακύμανση είναι το τετράγωνο της τυπικής απόκλισης.
2. Η μέση τιμή του συνόλου 1, 2, 4, 5 και 13 είναι... i) 1, ii) 4, iii) 4,8, iv) 5, v) 12

3. Η διάμεσος του συνόλου 1, 2, 4, 5 και 13 είναι... i) 1, ii) 4, iii) 4,8, iv) 5, iv) 12
4. Η διακύμανση του συνόλου 1, 2, 4, 5 και 13 είναι... i) 4, ii) 15, iii)18, iv) 22, iv) 34
5. Η τυπική απόκλιση του συνόλου 1, 2, 4, 5 και 13 είναι

   i) 2, ii) $3,9 \simeq \sqrt{15}$, iii) $4,2 \simeq \sqrt{18}$, iv) $4,7 \simeq \sqrt{22}$, iv) $5,8 \simeq \sqrt{34}$

Δίνεται ο παρακάτω πίνακας συχνοτήτων του πλήθους των παιδιών που υπάρχουν σε 50 οικογένειες :

Πλήθος παιδιών	**Συχνότητα**	**Σχετική Συχνότητα**	**Αθροιστική Συχνότητα**	**Αθροιστική Σχετική Συχνότητα**
0	19	f_1	N_1	F_1
1	ν_2	f_2	30	F_2
2	ν_3	0,3	N_3	F_3
3	ν_4	f_4	N_4	F_4
Σύνολο	50	1		

6. Η συχνότητα ν_2 είναι ίση με : i) 5, ii) 11, iii) 15, iv) 30
7. Η συχνότητα ν_3 είναι ίση με : i) 5, ii) 11, iii) 15, iv) 30
8. Η συχνότητα ν_4 είναι ίση με : i) 5, ii) 11, iii) 15, iv) 30
9. Η σχετική συχνότητα f_1 είναι ίση με : i) 0,1, ii) 0,22, iii) 0,3, iv) 0,38
10. Η σχετική συχνότητα f_2 είναι ίση με : i) 0,1, ii) 0,22, iii) 0,3, iv) 0,38
11. Η σχετική συχνότητα f_4 είναι ίση με : i) 0,1, ii) 0,22, iii) 0,3, iv) 0,38
12. Η αθροιστική συχνότητα N_1 είναι ίση με : i) 19, ii) 30, iii) 45, iv) 50
13. Η αθροιστική συχνότητα N_3 είναι ίση με : i) 19, ii) 30, iii) 45, iv) 50
14. Η αθροιστική συχνότητα N_4 είναι ίση με : i) 19, ii) 30, iii) 45, iv) 50
15. Η αθροιστική σχετική συχνότητα F_1 είναι ίση με : i) 0,38, ii)0,6, iii)0,9, iv)1
16. Η αθροιστική σχετική συχνότητα F_2 είναι ίση με : i) 0,38, ii) 0,6, iii) 0,9, iv) 1
17. Η αθροιστική σχετική συχνότητα F_3 είναι ίση με : i) 0,38, ii) 0,6, iii) 0,9, iv) 1
18. Η αθροιστική σχετική συχνότητα F_4 είναι ίση με : i) 0,38, ii) 0,6, iii) 0,9, iv) 1
19. Το μέσο πλήθος παιδιών είναι : i) 1, ii) 1,02, iii) 1,12, iv) 1,22
20. Το διάμεσο πλήθος παιδιών είναι : i) 0, ii) 1, iii) 2, iv) 3

Απαντήσεις :

Ερώτηση	1	2	3	4	5	6	7	8	9	10
Απάντηση	iv	iv	ii	iii	iii	ii	iii	i	iv	ii
Ερώτηση	11	12	13	14	15	16	17	18	19	20
Απάντηση	i	i	iii	iv	i	ii	iii	iv	iii	ii

Κεφάλαιο 3 Παλινδρόμηση

Με τον όρο “ανάλυση παλινδρόμησης” περιγράφονται όλες οι στατιστικές διαδικασίες που έχουν ως στόχο την ανίχνευση της σχέσης δύο ή περισσοτέρων μεταβλητών αλλά και την καταγραφή μοντέλου πρόβλεψης. Η ανάλυση παλινδρόμησης προσπαθεί να καταγράψει τον τρόπο με τον οποίο μεταβάλλεται η αναμενόμενη (μέση) τιμή μίας εξαρτημένης μεταβλητής όταν αλλάξει η τιμή μίας ή περισσοτέρων ανεξάρτητων μεταβλητών. Το αποτέλεσμα μίας τέτοιας διαδικασίας καταγράφεται ως μία συνάρτηση

$$\mathbf{Y = f(X_1, X_2, ...,X_\kappa).}$$

Η συνάρτηση f μπορεί να έχει οποιαδήποτε μορφή ωστόσο ο ερευνητής πρέπει να προσπαθεί να εντοπίζει το απλούστερο δυνατό μοντέλο, δηλαδή ένα της μορφής

$$\mathbf{Y = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_\kappa X_\kappa}$$

και να το εγκαταλείπει μόνο όταν αυτό δεν είναι ικανοποιητικό στην πρόβλεψή του.

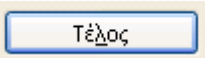
Στις επόμενες παραγράφους θα μελετήσουμε ένα παράδειγμα δημιουργίας γραμμικού μοντέλου, χρησιμοποιώντας δεδομένα προσδόκιμου ζωής ανδρών και γυναικών του έτους 2005 σε 30 τυχαία επιλεγμένες χώρες. Η διαδικασία ξεκινά από την οπτική αναγνώριση

μίας γραμμικής ή μη γραμμικής σχέσης (αν υπάρχει κάποια) και αυτό μπορεί να συμβεί με το διάγραμμα διασποράς (scatterplot).

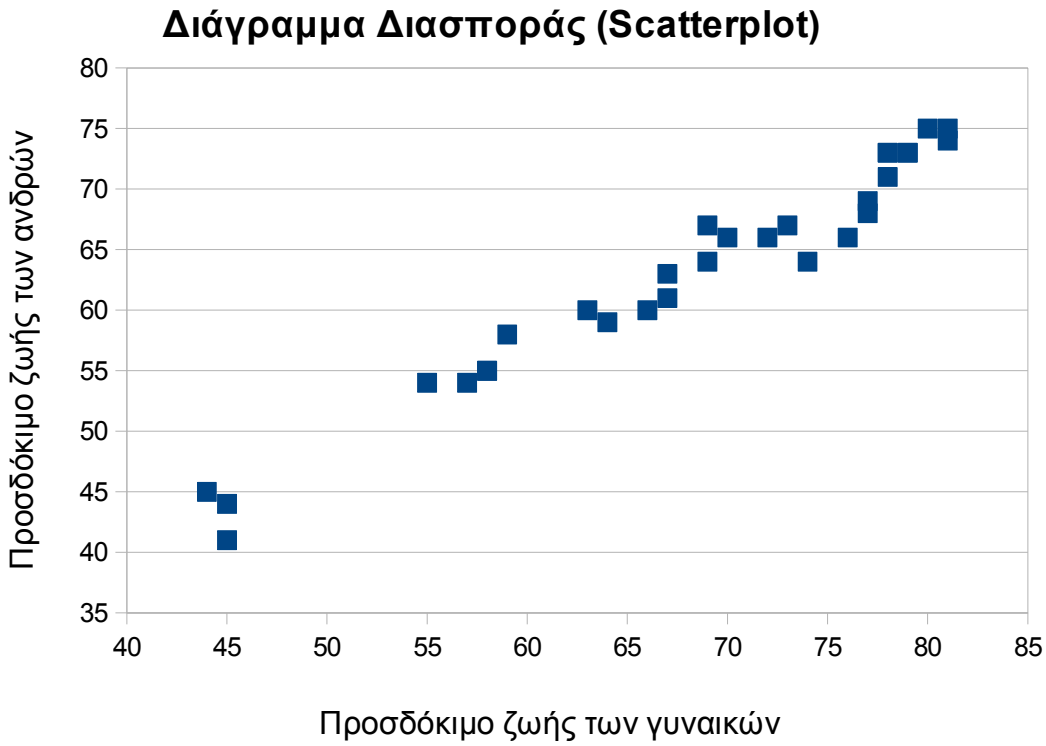
3.1 Διάγραμμα διασποράς (Scatterplot)

Το διάγραμμα διασποράς είναι το κατάλληλο γράφημα που δημιουργούμε ως πρώτο βήμα για να μελετήσουμε τη σχέση που υπάρχει μεταξύ δύο συνεχών αριθμητικών μεταβλητών, ιδιαίτερα αν αποσκοπούμε στη δημιουργία μοντέλου πρόγνωσης των τιμών της μίας μεταβλητής από την άλλη. Με το διάγραμμα διασποράς και μια έμπειρη στατιστική ματιά ανιχνεύεται η σχέση που ενδεχομένως να υπάρχει μεταξύ των δύο μεταβλητών.

3.1.1 Διάγραμμα διασποράς με το Calc

Το διάγραμμα διασποράς δημιουργείται στο LibreOffice Calc, όπως όλα τα γραφήματα επιλέγοντας ***Εισαγωγή*** → ***Διάγραμμα*** είτε επιλέγοντας το εικονίδιο  στη γραμμή εργαλείων. Στον οδηγό διαγράμματος που εμφανίζεται στο πρώτο βήμα (**1. Τύπος διαγράμματος**) επιλέγουμε τη δημιουργία διαγράμματος τύπου “**XY (Διασπορά)**” Στο δεύτερο βήμα (**2. Περιοχή δεδομένων**) επιλέγουμε την περιοχή των δεδομένων μας (**B3:C32** στην περίπτωση του παραδείγματος) και προσέχουμε να αποεπιλέξουμε τις επιλογές “**Πρώτη γραμμή σαν ετικέτα**” και “**Πρώτη στήλη σαν ετικέτα**” (Οι ονομασίες των αξόνων τοποθετούνται σε επόμενο βήμα). Προχωρούμε στην τελευταία επιλογή (**4. Στοιχεία διαγράμματος**), αποεπιλέγουμε την επιλογή “**Προβολή υπομνήματος**” και εισάγουμε τον τίτλο του διαγράμματος και τις ονομασίες των αξόνων έχοντας στο μυαλό μας πως το Calc θεωρεί την πρώτη στήλη πάντα σαν τετμημένη για κάθε σημείο (δηλαδή X) και τη δεύτερη σαν τεταγμένη (δηλαδή Y). Έτσι, στην περίπτωση των δεδομένων του πίνακα 3.2, σελίδα 98, ο άξονας X ονομάζεται “Προσδόκιμο ζωής των γυναικών” ενώ ο άξονας Y “Προσδόκιμο ζωής των ανδρών”. Επιλέγοντας 

αφήνουμε τον οδηγό και το διάγραμμα είναι έτοιμο και έχει τοποθετηθεί στο φύλλο εργασίας του Calc. (Διάγραμμα 13).



Διάγραμμα 13: Διάγραμμα Διασποράς

Στην περίπτωση όπου το διάγραμμα εμφανίζεται αρκετά “κενό” μπορούμε να αλλάξουμε το κατώτερο όριο εμφάνισης τιμών στον έναν ή και στους δύο άξονες κάνοντας διπλό κλικ πάνω στο γράφημα και ξανά διπλό κλικ πάνω στον άξονα (Στο διάγραμμα διασποράς 13 τοποθετήθηκε ως κατώτερη τιμή του άξονα Υ ο αριθμός 35). Επιπλέον, με περιήγηση πάνω στα μικρά τετράγωνα εμφανίζεται ο αύξων αριθμός και το ζεύγος τιμών από τις οποίες προήλθε.

3.1.2 Διάγραμμα διασποράς με το R – Project

Καταχωρούμε τα δεδομένα του πίνακα 3.2, σελίδα 98, όπως παρακάτω :

x = c(63, 79, 44, 79, 64, 70, 69, 80, 45, 59, 73, 58, 81, 69, 78, 76, 77, 72, 66, 78, 57, 67, 81, 67, 74, 58, 55, 45, 77, 78)

y = c(60, 73, 45, 73, 59, 66, 64, 75, 44, 58, 67, 55, 74, 67, 73, 66, 68, 66, 60, 73, 54, 61, 75, 63, 64, 55, 54, 41, 69, 71)

Αρκεί τώρα να δώσουμε την εντολή **plot(x,y)** για να πάρουμε το διάγραμμα διασποράς της y πάνω στη x. Με προσοχή στην παραμετροποίηση θα πάρουμε ένα δημοσιεύσιμο αποτέλεσμα (Διάγραμμα 14)

plot(x,y, xlab = "Προσδόκιμο ζωής γυναικών", ylab= "Προσδόκιμο ζωής ανδρών", main = "Διάγραμμα διασποράς (Scatterplot)", type = "p", pch = 3)

Μία άλλη εκδοχή είναι η εντολή **scatterplot(x, y, xlab = "Προσδόκιμο ζωής γυναικών", ylab= "Προσδόκιμο ζωής ανδρών", main = "Διάγραμμα διασποράς (Scatterplot)")**

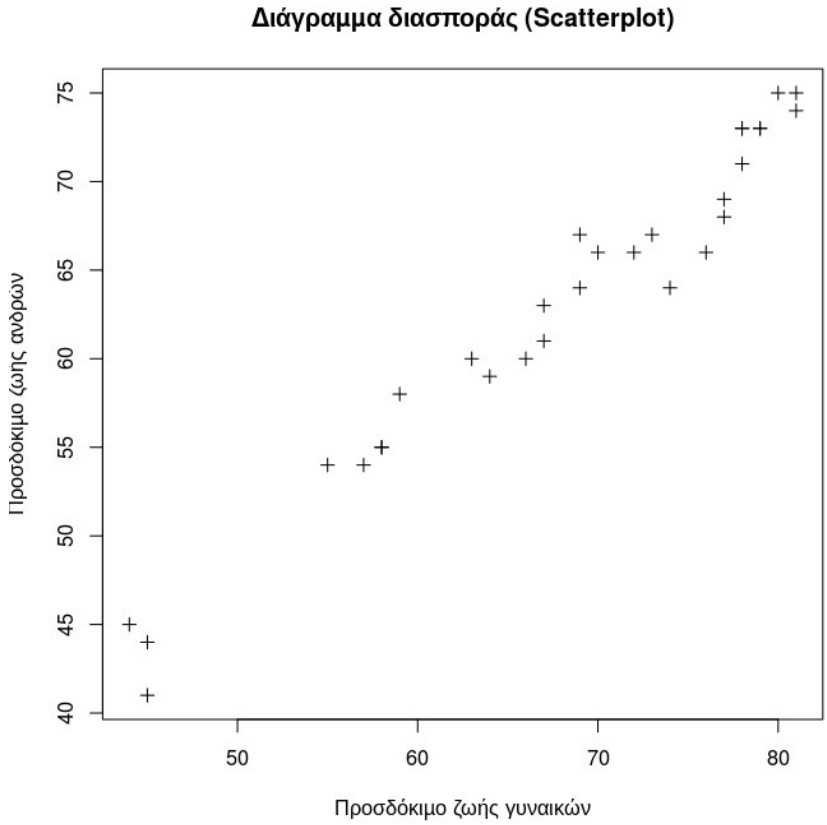
που δίνει το περισσότερο πλούσιο αποτέλεσμα του διαγράμματος 15 το οποίο εκτός από τα σημεία περιέχει και την ευθεία γραμμικής παλινδρόμησης, τις ευθείες που περικλείουν το 95% διάστημα εμπιστοσύνης αλλά και τα θηκογράμματα των δύο μεταβλητών.

Αν επιπλέον, ορίσουμε

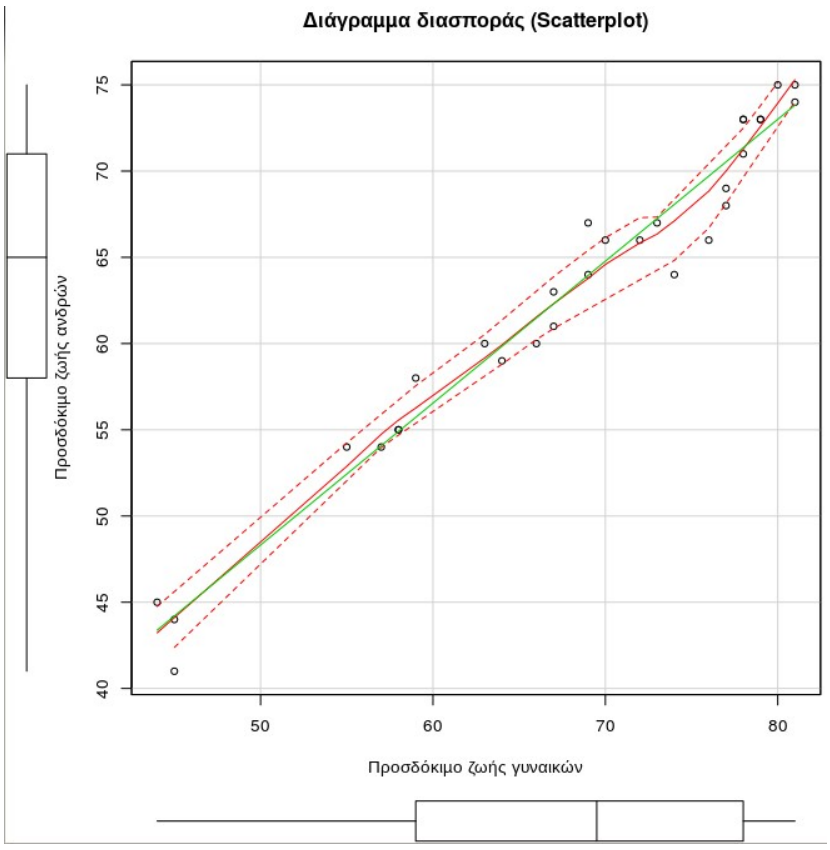
countries = c('Αίγυπτος', 'Αυστρία', 'Αφγανιστάν', 'Βέλγιο', 'Βολιβία', 'Δομινικανή Δημοκρατία', 'Ελ Σαλβαδόρ', 'Ελλάδα', 'Ζάμπια', 'Ινδία', 'Ισημερινός', 'Καμερούν', 'Καναδάς', 'Κίνα', 'Κουβέιτ', 'Λευκορωσία', 'Λιθουανία', 'Μαλαισία', 'Μποτσουάνα', 'Νησιά Μπαρμπάντος', 'Νιγηρία', 'Νικαράγουα', 'Ολλανδία', 'Περού', 'Ρωσία', 'Σενεγάλη', 'Σομαλία', 'Τανζανία', 'Τσεχία', 'Χιλή')

με την πρόσθετη παραμετροποίηση **scatterplot(x, y, xlab = "Προσδόκιμο ζωής γυναικών", ylab= "Προσδόκιμο ζωής ανδρών", main = "Διάγραμμα διασποράς (Scatterplot)", labels = countries)**

εντοπίζονται εύκολα οι ακραίες παρατηρήσεις.



Διάγραμμα 14: Διάγραμμα διασποράς στο R - Project



Διάγραμμα 15: Διάγραμμα διασποράς στο R – Project (2)

Παρατηρήσεις

1. Από την παρατήρηση του διαγράμματος 13 φαίνεται πως τα σημεία του διαγράμματος είναι κοντά σε μία ευθεία. Αυτό σημαίνει πως η εξίσωση παλινδρόμησης πρέπει να είναι γραμμική!

2. Το διάγραμμα διασποράς μας δίνει την πληροφορία πως οι δύο μεταβλητές συνδέονται με γραμμικό τρόπο αλλά δεν εξηγεί ποια θα έχει το ρόλο της εξαρτημένης και ποια το ρόλο της ανεξάρτητης. Αυτό το αποφασίζει μόνος του ο ερευνητής.

3.2 Συνδιακύμανση

Η συνδιακύμανση (covariance) είναι μία στατιστική ποσότητα η οποία ποσοτικοποιεί το είδος των μεταβολών που εμφανίζονται στις τιμές μίας συνεχούς τυχαίας μεταβλητής όταν μία άλλη μεταβάλλεται.

Αν X και Y είναι δύο τυχαίες μεταβλητές τότε η συνδιακύμανση ορίζεται ως

$$Cov(X,Y) = E[(X - EX)(Y - EY)] = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - EX \cdot EY$$

Στην περισσότερο συνηθισμένη περίπτωση όπου δεν γνωρίζουμε τις θεωρητικές κατανομές συχνοτήτων αλλά έχουμε ένα δείγμα τιμών των τυχαίων μεταβλητών τότε υπολογίζουμε τη δειγματική συνδιακύμανση ως

$$s^2_{XY} = \frac{1}{\nu}\sum_{i=1}^{\nu}(x_i - \bar{x})(y_i - \bar{y})$$

Πίνακας 3.1: Υπολογισμός συνδιακύμανσης από υπολογιστή

	Συνάρτηση **COVAR(δεδομένα1;δεδομένα2)**
	Συνάρτηση **cov(x, y)** Σημείωση : Οι δύο συναρτήσεις δίνουν διαφορετικά αποτελέσματα γιατί η διαίρεση γίνεται με το ν στο Calc ενώ με το ν – 1 στο R – Project (εφαρμόζεται η διόρθωση Bessel).

Παράδειγμα

Δίνεται ο εξής πίνακας με τις παρατηρήσεις των τ.μ. X και Y.

X	1	2	3
Y	3	5	10

Υπολογίζουμε $\bar{x} = \frac{1+2+3}{3} = 2$ και $\bar{y} = \frac{3+5+10}{3} = 6$ και

$$s_{XY} = \frac{1}{3}[(1-2)(3-6)+(2-2)(5-6)+(3-2)(10-6)] = \frac{1}{3}(3+0+4) = \frac{7}{3} = 2{,}33$$

3.2.1 Ερμηνεία της συνδιακύμανσης

Το πρόσημο της τιμής της συνδιακύμανσης δείχνει την τάση στη γραμμική σχέση των δύο μεταβλητών. Πιο συγκεκριμένα, αν οι μεγαλύτερες τιμές της μίας αντιστοιχούν στις μεγαλύτερες τιμές της άλλης και το ίδιο συμβαίνει για τις μικρότερες τιμές, δηλαδή όταν οι δύο μεταβλητές εμφανίζουν παρόμοια συμπεριφορά τότε η συνδιακύμανση είναι θετική. Στην αντίθετη περίπτωση όπου οι μεγαλύτερες τιμές της μίας μεταβλητής αντιστοιχούν στις μικρότερες της άλλης τότε η συνδιακύμανση είναι αρνητική. Εκτός από το πρόσημο, η συνδιακύμανση μπορεί να ερμηνευθεί και ως προς την απόλυτη τιμή της χωρίς όμως αυτό να είναι εύκολο καθώς το μέγεθος της εξαρτάται από το είδος των μονάδων που χρησιμοποιούνται αλλά και από το μέγεθος του δείγματος. Περισσότερο καθαρά συμπεράσματα για το είδος και την ισχύ της γραμμικής σχέσης προκύπτουν από την χρήση της κανονικοποιημένης εκδοχής της συνδιακύμανσης που είναι ο συντελεστής συσχέτισης του Pearson.

3.3 Συντελεστής συσχέτισης Pearson

Ο συντελεστής συσχέτισης του Pearson είναι το κατάλληλο στατιστικό για την ανίχνευση της γραμμικής σχέσης δύο ποσοτικών μεταβλητών, για συνεχείς ή αριθμητικές διακριτές μεταβλητές . Συμπληρώνει το διάγραμμα διασποράς υπό την έννοια ότι αντιστοιχεί μια συγκεκριμένη αριθμητική τιμή στην οπτική παρατήρηση της προσαρμογής του διαγράμματος διασποράς σε μία ευθεία.

Αν X και Y είναι δύο τυχαίες μεταβλητές τότε ο συντελεστής συσχέτισης ορίζεται ως

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Όταν δεν γνωρίζουμε τις θεωρητικές κατανομές συχνοτήτων αλλά έχουμε ένα δείγμα τιμών των τυχαίων μεταβλητών τότε υπολογίζουμε το συντελεστή συσχέτισης ως

$$r_{XY} = \frac{\sum_{i=1}^{v}(x_i-\bar{x})(y_i-\bar{y})}{v\, s_X, s_Y} = \frac{\sum_{i=1}^{v}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{v}(x_i-\bar{x})^2 \sum_{i=1}^{v}(y_i-\bar{y})}}$$

Παράδειγμα

Με τα δεδομένα της παραγράφου 3.2, σελίδα 93 υπολογίζουμε

$$\sum_{i=1}^{3}(x_i-\bar{x})(y_i-\bar{y}) = (1-2)(3-6)+(2-2)(5-6)+(3-2)(10-6) = 7 \ ,$$

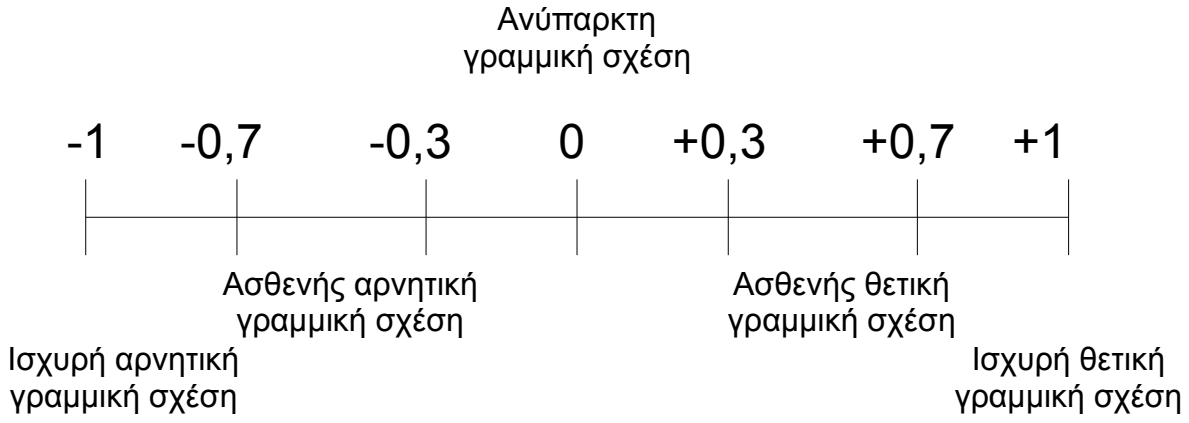
$$\sum_{i=1}^{3}(x_i-\bar{x})^2 = (1-2)^2+(2-2)^2+(3-2)^2 = 2 \quad \text{και} \quad \sum_{i=1}^{3}(y_i-\bar{y})^2 = (3-6)^2+(5-6)^2+(10-6)^2 = 26$$

άρα ο συντελεστής συσχέτισης του Pearson είναι

$$r_{XY} = \frac{7}{\sqrt{2\cdot 26}} = 0{,}97$$

3.4 Προϋποθέσεις υπολογισμού

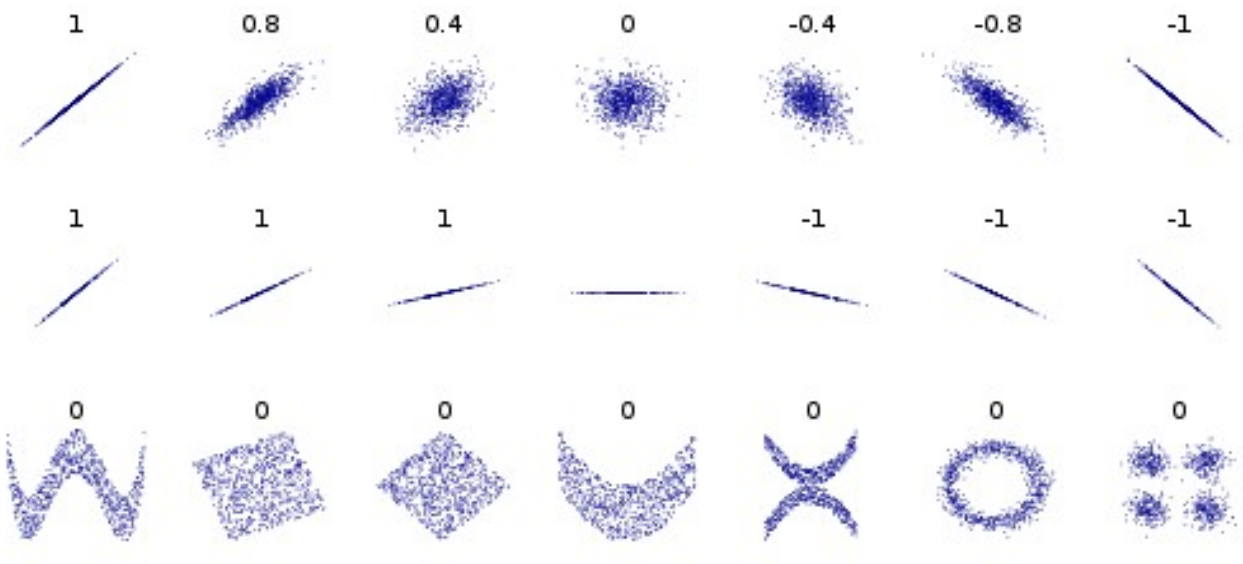
Ο συντελεστής Pearson είναι παραμετρικός, δηλαδή προϋποθέτει πως οι τιμές των δύο μεταβλητών προέρχονται από κανονικούς πληθυσμούς. Η κανονικότητα των πληθυσμών τεκμηριώνεται χρησιμοποιώντας το ιστόγραμμα ή κάποιο από τα διαγράμματα p – p και q – q. Αν υπάρχουν σοβαρές αμφιβολίες για την υπόθεση αυτή ή οι μεταβλητές είναι διακριτές με σχετικά λίγες τιμές τότε καλύτερα να υπολογιστεί ο συντελεστής Spearman ο οποίος είναι το μη παραμετρικό ανάλογο του συντελεστή Pearson.



Σχέδιο 1: Αξιολόγηση του συντελεστή συσχέτισης

3.4.1 Αξιολόγηση του συντελεστή Pearson

Ο συντελεστής Pearson, ο οποίος συμβολίζεται συνήθως **r** ή **R**, παίρνει πάντα τιμές μεταξύ -1 και +1 και ανάλογα με την τιμή συνάγουμε το είδος και την ισχύ της γραμμικής σχέσης μεταξύ των μεταβλητών. Στο Σχέδιο 1 παρουσιάζεται συνοπτικά η αξιολόγηση του συντελεστή.



Διάγραμμα 16: Ενδεικτικά διαγράμματα διασποράς και ο αντίστοιχος συντελεστής Pearson (πηγή : Wikipedia)

Παράδειγμα

Με τα δεδομένα της παραγράφου 3.2, σελίδα 93 βρήκαμε $r_{XY}=0,97$ κάτι που μας οδηγεί να συμπεράνουμε πως οι μεταβλητές X και Y είναι ισχυρά θετικά γραμμικά συσχετισμένες.

3.4.2 Υπολογισμός συντελεστή συσχέτισης με το Calc

Οι τιμές των μεταβλητών πρέπει να είναι τοποθετημένες ανά ζεύγη σε δύο, όχι κατ΄ ανάγκη συνεχόμενες, στήλες του Calc. Για παράδειγμα στον πίνακα της επόμενης σελίδας (Πίνακας 3.2, σελίδα 98) παρουσιάζονται δεδομένα για το προσδόκιμο ζωής ανδρών και γυναικών σε τριάντα χώρες του κόσμου το έτος 1995. Αυτές τοποθετήθηκαν στις στήλες A, B, C ενός φύλλου του Calc από την γραμμή 1 έως την 32 (Τα δεδομένα βρίσκονται στις στήλες 3 έως 32).

Η ποιοτική ανάλυση των μεταβλητών φανερώνει πως οι μεταβλητές “Προσδόκιμο ζωής ανδρών” και “Προσδόκιμο ζωής γυναικών” είναι θετικά συσχετισμένες. Πράγματι, όσο μεγαλύτερο είναι το προσδόκιμο ζωής για το ένα από τα δύο φύλα σε μία χώρα ανάλογα μεγάλο περιμένουμε να είναι και για το άλλο φύλο. Υπολογίζοντας τον συντελεστή Pearson θα αποκτήσουμε μια ποσοτική μέτρηση του γεγονότος αυτού ενώ η επιβεβαίωση της ισχυρής γραμμικής σχέσης θα ανοίξει τον δρόμο για την αξιόπιστη πρόβλεψη της τιμής της μίας μεταβλητής από την άλλη χρησιμοποιώντας την εξίσωση της ευθείας γραμμικής

παλινδρόμησης.

Ο συντελεστής γραμμικής συσχέτισης υπολογίζεται με τη συνάρτηση **PEARSON()** η οποία χρειάζεται δύο ορίσματα, το πρώτο από το οποίο θα είναι οι τιμές της μίας μεταβλητής και το δεύτερο οι τιμές της άλλης (Εικόνα 8). Είναι φανερό πως πρέπει οι δύο ομάδες δεδομένων να είναι ίσες στο πλήθος! Αν τοποθετήσουμε ομάδες διαφορετικού πλήθους τότε το αποτέλεσμα θα είναι **Σφάλμα: 502.**

Συντελεστής Γραμμικής Συσχέτισης Pearson	Συνάρτηση
0,98	PEARSON(B3:B32;C3:C32)

Εικόνα 8: Συντελεστής συσχέτισης Pearson

3.5 Συντελεστής συσχέτισης Spearman

Ο συντελεστής συσχέτισης Spearman είναι το μη παραμετρικό ανάλογο του συντελεστή Pearson υπό την έννοια πως υπολογίζεται χρησιμοποιώντας την τάξη κάθε στοιχείου δηλαδή τη σειρά κατάταξης του στα ταξινομημένα συνολικά δεδομένα. Συμβολίζεται συνήθως με το γράμμα ρ. Λόγω του τρόπου ορισμού του είναι καλύτερος εκτιμητής της συσχέτισης δύο διατακτικών μεταβλητών δηλαδή μεταβλητών που παίρνουν τιμές ακέραιους αριθμούς όπως συμβαίνει στις μεταβλητές με κλίμακα Liker.

Δυστυχώς το Calc δεν έχει συνάρτηση άμεσου υπολογισμού του συντελεστή Spearman αλλά εύκολα μπορούμε να το υπολογίσουμε χρησιμοποιώντας τον τύπο ορισμού του

$$\rho = 1 - \frac{6\sum d^2}{n\left(n^2-1\right)} \quad .$$

Δεδομένα		
Χώρα	**Προσδόκιμο ζωής γυναικών**	**Προσδόκιμο ζωής ανδρών**
Αίγυπτος	63	60
Αυστρία	79	73
Αφγανιστάν	44	45
Βέλγιο	79	73
Βολιβία	64	59
Δομινικανή Δημοκρατία	70	66
Ελ Σαλβαδόρ	69	64
Ελλάδα	80	75
Ζάμπια	45	44
Ινδία	59	58
Ισημερινός	73	67
Καμερούν	58	55
Καναδάς	81	74
Κίνα	69	67
Κουβέιτ	78	73
Λευκορωσία	76	66
Λιθουανία	77	68
Μαλαισία	72	66
Μποτσουάνα	66	60
Νησιά Μπαρμπάντος	78	73
Νιγηρία	57	54
Νικαράγουα	67	61
Ολλανδία	81	75
Περού	67	63
Ρωσία	74	64
Σενεγάλη	58	55
Σομαλία	55	54
Τανζανία	45	41
Τσεχία	77	69
Χιλή	78	71

Πίνακας 3.2: Δεδομένα προσδόκιμου ζωής έτους 1995

Στον τύπο υπολογισμού του συντελεστή Spearman το n είναι το πλήθος των ζευγών παρατηρήσεων, ενώ το d είναι η διαφορά της τάξης της μίας παρατήρησης ενός ζεύγους

από την άλλη, αριθμοί οι οποίοι υψώνονται στο τετράγωνο και αθροίζονται για να μας δώσουν τον αριθμητή του κλάσματος.

Συντελεστής Συσχέτισης Spearman	Συνάρτηση
0,98	1-6*I9/(I10*(I10^2-1))

		Συνάρτηση
Άθροισμα d^2	91	SUM(F3:F32)
Πλήθος Παρατηρήσεων	30	COUNT(E3:E32)

Εικόνα 9: Υπολογισμός συντελεστή συσχέτισης Spearman

RANK(B3;B$3:B$32)	RANK(C3;C$3:C$32)	(D3-E3)^2	(D3-E3)^2
Τάξη Προσδόκιμου Ζωής Γυναικών	Τάξη Προσδόκιμου Ζωής Ανδρών	Διαφορά	Διαφορά^2
22	20	2	4
4	4	0	0
30	28	2	4
4	4	0	0
21	22	-1	1
15	13	2	4
16	16	0	0
3	1	2	4
28	29	-1	1
23	23	0	0
13	11	2	4
24	24	0	0
1	3	-2	4
16	11	5	25
6	4	2	4
11	13	-2	4
9	10	-1	1
14	13	1	1
20	20	0	0
6	4	2	4
26	26	0	0
18	19	-1	1
1	1	0	0
18	18	0	0
12	16	-4	16
24	24	0	0
27	26	1	1
28	30	-2	4
9	9	0	0

Εικόνα 10: Υπολογισμός της τάξης των παρατηρήσεων τα οποία τοποθετήθηκαν στις στήλες D,E και F

Για κάθε μία από τις δύο στήλες των δεδομένων δημιουργούμε μια στήλη στην οποία αποθηκεύονται οι τάξεις των παρατηρήσεων χρησιμοποιώντας τη συνάρτηση **RANK()**. Οι υπόλοιπες ενέργειες είναι απλές!

Ο συντελεστής Spearman αξιολογείται με τον ίδιο τρόπο με αυτόν του Pearson (Σχέδιο 1). Απόλυτη τιμή του συντελεστή μεγαλύτερη από 0,7 συνήθως αξιολογείται ως ισχυρή γραμμική σχέση, μεταξύ 0,3 και 0,7 ως ασθενής γραμμική σχέση ενώ μεταξύ 0 και 0,3 ως μη γραμμική σχέση.

Πίνακας 3.3: Υπολογισμός συντελεστή συσχέτισης από υπολογιστή

Ο συντελεστής συσχέτισης του Pearson μπορεί να υπολογιστεί από τη συνάρτηση **PEARSON(δεδομένα1;δεδομένα2).** Ο συντελεστής του Spearman ή του Kendall δεν μπορεί να υπολογιστεί άμεσα από το Calc.

Συνάρτηση **cor(x, y, method = "pearson")** ή πιο απλά **cor(x, y)** για το συντελεστή συσχέτισης του Pearson, **cor(x, y, method = "spearman")** και **cor(x, y, method = "kendall")** για τους συντελεστές Spearman και Kendall αντίστοιχα.

3.6 Γραμμική παλινδρόμηση.

Το πρώτο βήμα για τη μελέτη της σχέσης δύο συνεχών μεταβλητών είναι ο υπολογισμός του συντελεστή γραμμικής συσχέτισης Pearson. Το επόμενο βήμα είναι η εύρεση της ευθείας γραμμικής παλινδρόμησης της μίας μεταβλητής πάνω στην άλλη, ιδιαίτερα αν μεταξύ των μεταβλητών υπάρχει ισχυρή γραμμική συσχέτιση (συντελεστής μεγαλύτερος κατά απόλυτη τιμή από το 0,7).

Η ευθεία γραμμικής παλινδρόμησης είναι η ευθεία η οποία βρίσκεται όσο το δυνατόν “κοντύτερα” στα σημεία του διαγράμματος διασποράς. Το “κοντύτερα” είναι υποκειμενικό και μπορεί να οριστεί με διάφορους τρόπους. Η πλέον συνηθισμένη επιλογή είναι ο εντοπισμός της ευθείας για την οποία το άθροισμα των κάθετων αποστάσεων όλων των σημείων από αυτήν είναι ελάχιστο Η ευθεία αυτή ονομάζεται και ευθεία ελαχίστων τετραγώνων.

Η εξίσωσή της ευθείας ελαχίστων τετραγώνων, όπως και η εξίσωση κάθε ευθείας έχει τη μορφή

$$Y = \alpha * X + \beta.$$

Στην παραπάνω εξίσωση το X ονομάζεται ανεξάρτητη μεταβλητή (independent variable) ενώ το Y εξαρτημένη (dependent variable) διότι οι τιμές του λαμβάνονται από τις τιμές του

X.

Οι συντελεστές του γραμμικού μοντέλου υπολογίζονται από τα ζεύγη τιμών των μεταβλητών X και Y ως εξής :

$$\alpha = \frac{\sum_{i=1}^{v}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{v}(x_i - \bar{x})^2} \quad \text{και} \quad \beta = \bar{y} - \alpha \cdot \bar{x}$$

Παράδειγμα

1. Με τα δεδομένα της παραγράφου 3.2, σελίδα 93 είναι $\bar{x}=2$, $\bar{y}=6$, $\sum_{i=1}^{3}(x_i-\bar{x})^2=2$ και $\sum_{i=1}^{3}(x_i-\bar{x})(y_i-\bar{y})=7$ και υπολογίζουμε $\alpha=\frac{7}{2}=3{,}5$ ενώ $\beta=6-3{,}5\cdot 2=-1$ άρα το γραμμικό μοντέλο είναι

Y = 3,5 * X -1.

Η απόφαση για το ρόλο που θα αναλάβει κάθε μεταβλητή στην εξίσωση της ευθείας γραμμικής παλινδρόμησης είναι αποκλειστικά του ερευνητή (δεν υπάρχει σωστή και λάθος επιλογή) και καθορίζεται από την εσωτερική σχέση που έχουν οι δύο μεταβλητές, δηλαδή την συμπεριφορά της μίας ως “αιτίας” και της άλλης ως “αιτιατό” στο συγκεκριμένο εννοιολογικό πλαίσιο που αυτές ορίζονται.

2. Το παράδειγμα αφορά τα δεδομένα της εικόνας 10, σελίδα 99 και θα το λύσουμε με τις συναρτήσεις που προσφέρει το Calc. Θεωρούμε ως εξαρτημένη μεταβλητή **Y** το προσδόκιμο ζωής των γυναικών (τοποθετημένα στη στήλη **B** στα δεδομένα της εικόνας 10) και ως ανεξάρτητη μεταβλητή **X** το προσδόκιμο ζωής των ανδρών (τοποθετημένα στη στήλη **C** στα δεδομένα της εικόνας 10). Επιζητούμε την εξίσωση της ευθείας

Προσδόκιμο ζωής γυναικών = α * Προσδόκιμο ζωής ανδρών + β

Είναι φανερό πως χρησιμοποιώντας την τελευταία εξίσωση είναι δυνατή η πρόβλεψη του προσδόκιμου ζωής των γυναικών σε μία χώρα για την οποία δεν θα το γνωρίζουμε αλλά θα γνωρίζουμε το αντίστοιχο προσδόκιμο των ανδρών.

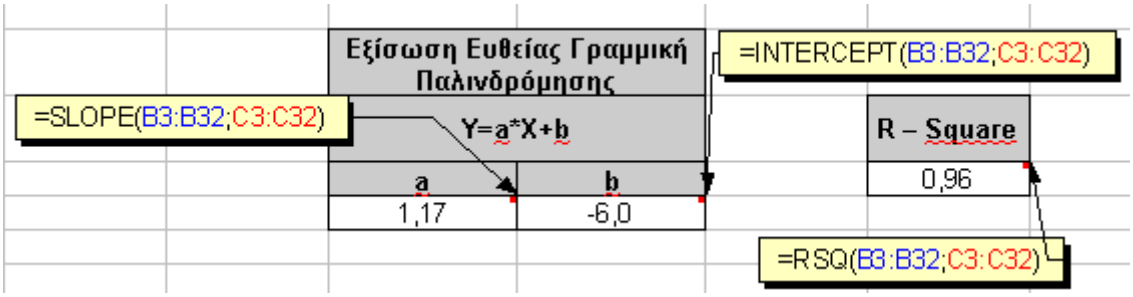
3.6.1 Γραμμική παλινδρόμηση με το Calc

Το Calc προσφέρει έτοιμες συναρτήσεις για τον υπολογισμό των συντελεστών **α** και **β** της εξίσωσης της ευθείας. Η εφαρμογή των συναρτήσεων αυτών στα δεδομένα του προσδόκιμου ζωής (Εικόνα 10) παρουσιάζεται στην εικόνα 11.

Προσέξτε πως και στις δύο συναρτήσεις **SLOPE()** και **INTERCEPT()** πρώτα τοποθετούνται τα δεδομένα της εξαρτημένης μεταβλητής (Y) και μετά τα δεδομένα της ανεξάρτητης μεταβλητής (X). Αντίστροφη τοποθέτηση θα οδηγούσε στον υπολογισμό των συντελεστών της εξίσωσης

Προσδόκιμο ζωής ανδρών = a * Προσδόκιμο ζωής γυναικών + b

το οποίο έχει ενδιαφέρον αλλά δεν είναι ο σκοπός της υποτιθέμενης έρευνας μας.



Εικόνα 11: Υπολογισμός συντελεστών της ευθείας γραμμικής παλινδρόμησης

Από τα εξαγόμενα των συναρτήσεων προκύπτει πως η ζητούμενη εξίσωση είναι η

Προσδόκιμο ζωής γυναικών = 1,17 * Προσδόκιμο ζωής ανδρών – 6

Εφαρμογή : Να βρεθεί το προσδόκιμο ζωής των γυναικών σε μία χώρα στην οποία οι άνδρες έχουν προσδόκιμο ζωής τα 65 χρόνια.

Υλοποίηση : Απλά αντικαθιστούμε όπου **X (Προσδόκιμο ζωής ανδρών) = 65** στην εξίσωση της ευθείας γραμμικής παλινδρόμησης και υπολογίζουμε

Αναμενόμενο προσδόκιμο ζωής γυναικών = 1,17 * 70 – 6 = 75,9 έτη

Προσέξτε πως το παραπάνω αποτέλεσμα δεν σημαίνει πως σε όλες τις χώρες όπου οι άνδρες θα ζούνε 70 χρόνια κατά μέσο όρο, οι γυναίκες θα ζούνε 75,9. Ο αριθμός που υπολογίσαμε είναι το αναμενόμενο προσδόκιμο ζωής των γυναικών, δηλαδή το μέσο προσδόκιμο ζωής των γυναικών σε όλες τις χώρες στις οποίες οι άνδρες ζούνε 70 χρόνια.

3.6.2 Γραμμική παλινδρόμηση με το R – Project

Η πιο απλή εντολή είναι **lm(y~x).** Για τα δεδομένα του προσδόκιμου ζωής του πίνακα 3.2

women = c(63, 79, 44, 79, 64, 70, 69, 80, 45, 59, 73, 58, 81, 69, 78, 76, 77, 72, 66, 78, 57, 67, 81, 67, 74, 58, 55, 45, 77, 78)

man = c(60, 73, 45, 73, 59, 66, 64, 75, 44, 58, 67, 55, 74, 67, 73, 66, 68, 66, 60, 73, 54, 61, 75, 63, 64, 55, 54, 41, 69, 71)

το αποτέλεσμα της συνάρτησης **lm(women~man)** είναι

Call:

lm(formula = women ~ men)

Coefficients:

(Intercept)	y
-6.000	1.172

Από το αποτέλεσμα συμπεραίνουμε άμεσα το γραμμικό μοντέλο

Women = 1.17 * Men – 6.0

το οποίο ταυτίζεται με αυτό που βρέθηκε στην 3.6.1 με χρήση συναρτήσεων του Calc.

3.6.3 Αξιολόγηση μοντέλου γραμμικής παλινδρόμησης

Η ευθεία γραμμικής παλινδρόμησης δίνει πρόβλεψη της άγνωστης τιμής της εξαρτημένης μεταβλητής. Η ποιότητα της πρόβλεψης αυτής ελέγχεται από την τιμή του συντελεστή προσδιοριστίας (Coefficient of determination) **R^2 (R – Square)** το οποίο πρέπει να υπολογίζεται μαζί με του συντελεστές της εξίσωσης (Εικόνα 11). και το οποίο ερμηνεύεται ως το ποσοστό της μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής που προσδιορίζεται από τις τιμές της ανεξάρτητης μεταβλητής, και είναι καλό να είναι μεγάλο δηλαδή κοντά στη μονάδα.

Πιο αναλυτικά, είναι

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

όπου SS_{err} είναι το άθροισμα τετραγώνων των σφαλμάτων του μοντέλου στις

παρατηρήσεις της εξαρτημένης μεταβλητής

$$SS_{err} = \sum (y_i - \hat{y}_i)^2$$

ενώ SS_{tot} είναι το συνολικό άθροισμα τετραγώνων των αποστάσεων των παρατηρήσεων της εξαρτημένης μεταβλητής από τη μέση της τιμή

$$SS_{tot} = \sum (y_i - \bar{y})^2$$

Ερμηνεία R^2 : Με μεθόδους διαφορικού λογισμού Γ' Λυκείου αποδεικνύεται πως η μέγιστη τιμή του SS_{err} είναι ακριβώς το SS_{tot} άρα ο λόγος SS_{err} / SS_{tot} είναι μεγαλύτερος από 0 (καθώς τα αθροίσματα είναι θετικά) και μικρότερος του 1. Καθώς, τα αθροίσματα τετραγωνικών αποκλίσεων ερμηνεύονται ως η “μεταβλητότητα” των παρατηρήσεων από τα οποία υπολογίζονται, το R^2 ερμηνεύεται ως το *ποσοστό της μεταβλητότητας* της εξαρτημένης μεταβλητής που εξηγείται από τις τιμές της ανεξάρτητης.

Ιδανική περίπτωση είναι η $SS_{err}=0$, περίπτωση κατά την οποία το μοντέλο προβλέπει επακριβώς τις τιμές της εξαρτημένης μεταβλητής από τις οποίες υπολογίστηκε. Σπάνια συμβαίνει αυτό, στην πράξη ένας συντελεστής μεγαλύτερος από 80% θεωρείται ικανοποιητικός ενώ αν είναι μεγαλύτερος από 90% τότε έχουμε καλό λόγο για να θεωρήσουμε πως η πρόβλεψη των τιμών του μοντέλου θα είναι αξιόπιστη για κάθε πρακτική χρήση. Αν παρ' ελπίδα η τιμή του R^2 είναι μικρή τότε πρέπει να αναζητήσουμε περισσότερο περίπλοκο μοντέλο (μη γραμμικό ή γραμμικό με περισσότερες μεταβλητές)

Παραδείγματα στο Calc

1. Με τα δεδομένα της παραγράφου 3.2, σελίδα 93 είναι $\bar{x}=2$, $\bar{y}=6$, και το μοντέλο πρόβλεψης της Y από τη X είναι **Y = 3,5 * X -1** (σελίδα 100). Υπολογίζουμε :

X	1	2	3
Y	3	5	10
$Y-\bar{Y}$	-3	-1	4
$(Y-\bar{Y})^2$	9	1	16
$\hat{Y}$	2,5	6	9,5
$Y-\hat{Y}$	0.5	-1	0.5
$(Y-\hat{Y})^2$	0.25	1	0.25

Από τον πίνακα έχουμε $SS_{err} = \sum (y_i - \hat{y}_i)^2 = 0.25+1+0.25 = 1.5$ και

$SS_{tot} = \sum (y_i - \bar{y})^2 = 9+1+16 = 26$. Κατά συνέπεια

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{1.5}{26} = 0.942 = 94{,}2\% \ .$$

2. Το παράδειγμα αφορά τα δεδομένα της εικόνας 10, σελίδα 99 και θα το λύσουμε με τη συνάρτηση **RSQ()** που προσφέρει το Calc Στην περίπτωση του παραδείγματος μας είναι 0,96 ή 96% (Εικόνα 11). Αυτό σημαίνει πως το 96% της μεταβλητότητας του προσδόκιμου ζωής των γυναικών Y ορίζεται από το προσδόκιμο ζωής των ανδρών X ενώ το υπόλοιπο 4% (=100%-96%) της μεταβολής της Y ορίζεται από άλλους παράγοντες εκτός από την μεταβλητή X κάτι που στην πράξη σημαίνει πως το μοντέλο μας μάλλον κάνει καλά την δουλειά του!

Μία επιπλέον χρήσιμη παρατήρηση είναι πως η αξιοπιστία της πρόβλεψης που θα προκύψει από την εξίσωση της ευθείας γραμμικής παλινδρόμησης είναι ανάλογη της αξιοπιστίας του υπολογισμού των συντελεστών α και β. Καθώς η μεταβλητότητα των δεδομένων επηρεάζει το μέγεθος των συντελεστών αυτών είναι καλό μετά τον υπολογισμό των α και β να προχωρούμε και σε στατιστικούς ελέγχους οι οποίοι θα διαβεβαιώνουν πως τα α και β δεν είναι μηδέν!

Πιο συγκεκριμένα, για το συντελεστή α,

Ερευνητική Υπόθεση H_1: Ο συντελεστής α στην παραπάνω εξίσωση είναι διάφορος από το μηδέν.

Στατιστική Υπόθεση H_0: Ο συντελεστής α στην παραπάνω εξίσωση είναι ίσος με το μηδέν.

και ανάλογες υποθέσεις για το συντελεστή β.

Είναι επιθυμητή η απόρριψη των στατιστικών υποθέσεων και στους δύο ελέγχους καθώς η αποδοχή κάποιας από τις δύο θα σημαίνει πως υπάρχει τόσο μεγάλη μεταβλητότητα στις τιμές που δεν μπορούμε να υπολογίσουμε με ασφάλεια τον αντίστοιχο συντελεστή. Δυστυχώς, οι έλεγχοι αυτοί ενώ είναι δεδομένοι σε περισσότερο ειδικά λογισμικά (R-Project ή SPAW) στο Calc δεν υπάρχουν ανάλογες διαδικασίες!

Παράδειγμα στο R – Project

Με τα δεδομένα του προσδόκιμου ζωής ανδρών και γυναικών (3.6.2) δίνουμε την εντολή

mymodel = lm(women~man)

και συνεχίζουμε δίνοντας

summary(mymodel)

Το αποτέλεσμα της εντολής εμφανίζεται παρακάτω

Call:

lm(formula = women ~ men)

Residuals:

Min	1Q	Median	3Q	Max
-3.5383	-1.3578	-0.4718	0.8228	4.9783

Coefficients:

	Estimate	Std Error	t value	Pr(>|t|)
(Intercept)	-5.99976	2.70505	-2.218	0.0348 *
men	1.17221	0.04242	27.630	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.128 on 28 degrees of freedom

Multiple R-squared: 0.9646, Adjusted R-squared: 0.9634

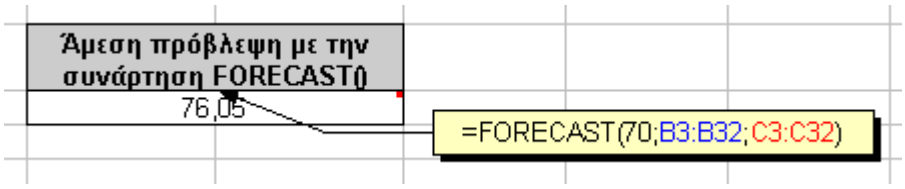
F-statistic: 763.4 on 1 and 28 DF, p-value: < 2.2e-16

Ερμηνεία των στατιστικών που προκύπτουν

Αρχικά παρατηρούμε πως το γραμμικό μοντέλο στο σύνολό του είναι στατιστικά σημαντικό ($F(1, 28) = 763,4$, $p < 0,001$) ενώ τόσο η σταθερά του μοντέλου όσο και ο συντελεστής του προσδόκιμου ζωής των ανδρών υπολογίστηκαν με στατιστικά σημαντικό τρόπο ($p = 0,03$ και $p < 0,001$ αντίστοιχα). Το γραμμικό μοντέλο που προκύπτει εξηγεί το 96,5% της μεταβλητότητας του προσδόκιμου ζωής των γυναικών γεγονός που του αποδίδει ιδιαίτερη αξιοπιστία στην πρόβλεψη.

3.6.4 Άμεση πρόβλεψη με γραμμικό μοντέλο με συνάρτηση του Calc

Αν δεν επιθυμούμε να βρούμε την εξίσωση της ευθείας αλλά θέλουμε απευθείας να την εφαρμόσουμε μπορούμε να χρησιμοποιήσουμε την συνάρτηση **FORECAST()** η οποία υπολογίζει άμεσα την πρόβλεψη βάσει του γραμμικού μοντέλου χωρίς να το εμφανίζει!



Εικόνα 12: Άμεση πρόβλεψη

Στην εικόνα 12 παρουσιάζεται η εφαρμογή της συνάρτησης. Είναι χαρακτηριστικό πως η τιμή που προκύπτει από την εφαρμογή της συνάρτησης **FORECAST()** είναι λίγο διαφορετική από την τιμή που υπολογίσαμε με το γραμμικό μας μοντέλο της παραγράφου 3.6.1. Η διαφορά οφείλεται στο γεγονός της αναγκαστικής στρογγυλοποίησης που εφαρμόσαμε στους συντελεστές a και b της εξίσωσης την οποία δεν πράττει στον ίδιο βαθμό η συνάρτηση **FORECAST()**.

3.7 Πολλαπλή γραμμική παλινδρόμηση

Με το R – Project είναι εύκολη η δημιουργία μοντέλου πολλαπλής γραμμική παλινδρόμησης.

Για παράδειγμα αν για το προσδόκιμο ζωής γυναικών και ανδρών έχουμε τα δεδομένα

women = c(63, 79, 44, 79, 64, 70, 69, 80, 45, 59, 73, 58, 81, 69, 78, 76, 77, 72, 66, 78, 57, 67, 81, 67, 74, 58, 55, 45, 77, 78)

man = c(60, 73, 45, 73, 59, 66, 64, 75, 44, 58, 67, 55, 74, 67, 73, 66, 68, 66, 60, 73, 54, 61, 75, 63, 64, 55, 54, 41, 69, 71)

ενώ επιπλέον γνωρίζουμε πως το ποσοστό ανθρώπων που γνωρίζουν γραφή και ανάγνωση είναι

literacy = c(30, 90, 20, 95, 70, 80, 75, 96, 33, 60, 67, 40, 93, 83, 94, 90, 87, 80, 60, 85, 67, 67, 96, 77, 76, 70, 70, 40, 80, 92)

Για να πάρουμε γραμμικό μοντέλο της μορφής

women = a* man + b * literacy + c

αρκεί να δώσουμε τις εντολές

data = data.frame(w = women, m = man, l = literacy)

attach(data)

mylinearmodel = lm(w ~ m + l)

summary(mylinearmodel)

Το αποτέλεσμα είναι

Call:

lm(formula = w ~ m + l)

Residuals:

Min	1Q	Median	3Q	Max
-3.6927	-1.2278	-0.2872	0.9421	4.8769

Coefficients:

	Estimate	Std Error	t value	Pr(>|t|)
(Intercept)	-3.53983	3.30975	-1.070	0.294
m	1.07888	0.08495	12.700	6.71e-13 ***
l	0.04756	0.03764	1.264	0.217

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.106 on 27 degrees of freedom

Multiple R-squared: 0.9666, Adjusted R-squared: 0.9641

F-statistic: 390.7 on 2 and 27 DF, p-value: < 2.2e-16

Από τα παραπάνω συμπεραίνουμε πως από τους συντελεστές του γραμμικού μοντέλου ο συντελεστής του ποσοστού γνώσης γραφής και ανάγνωσης δεν είναι στατιστικά σημαντικός.

Παρατήρηση : Το R – Project στη στήλη Estimate εμφανίζει τους συντελεστές του γραμμικού μοντέλου οι οποίοι ερμηνεύονται ως η απόλυτη μεταβολή του προσδόκιμου ζωής των γυναικών που αναμένεται να επιφέρει μεταβολή μίας μονάδας στην εκάστοτε ανεξάρτητη μεταβλητή. Αυτός ο τρόπος εμφάνισης δεν κρίνεται ως ο προσφορότερος για τα περισσότερα επιστημονικά περιοδικά αλλά προτιμάται η εμφάνιση των τυποποιημένων συντελεστών, οι οποίοι εμφανίζονται αν εκτελεστεί ξανά η διαδικασία με τις τυποποιημένες τιμές των μεταβλητών αντί των κανονικών. Για να συμβεί αυτό αρκεί να δοθεί η εντολή

mylinearmodelstd = lm(scale(w) ~ scale(m) + scale(l))

και

summary(mylinearmodelstd)

και παίρνουμε ως εξαγόμενο

Call:

lm(formula = scale(women) ~ scale(man) + scale(literacy))

Residuals:

Min	1Q	Median	3Q	Max
-0.33212	-0.11043	-0.02583	0.08473	0.43863

Coefficients:

	Estimate	Std Error	t value	Pr(>|t|)
(Intercept)	-4.716e-17	3.458e-02	0.000	1.000
scale(m)	9.040e-01	7.118e-02	12.700	6.71e-13 ***
scale(l)	8.994e-02	7.118e-02	1.264	0.217

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1894 on 27 degrees of freedom

Multiple R-squared: 0.9666, Adjusted R-squared: 0.9641

F-statistic: 390.7 on 2 and 27 DF, p-value: < 2.2e-16

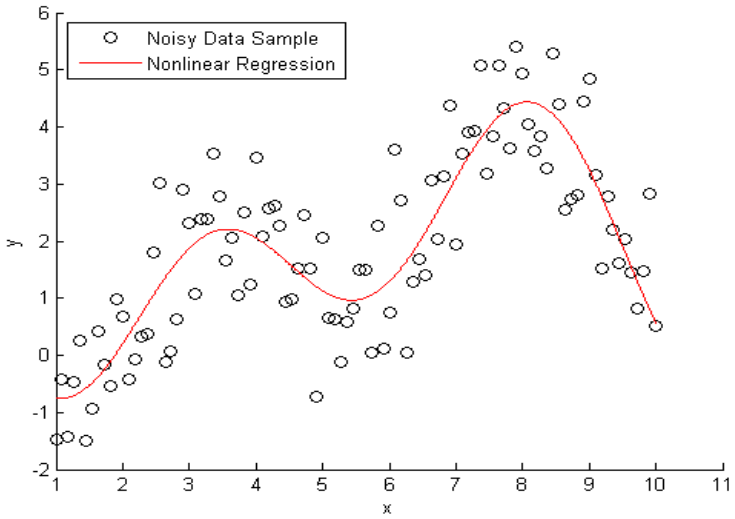
Από όπου συνάγουμε πως ο τυποποιημένος συντελεστής του προσδόκιμου ζωής των ανδρών (B ή Beta) είναι ίσος με 0,9 και είναι στατιστικά σημαντικός ($p < .001$) ενώ ο αντίστοιχος συντελεστής για το ποσοστό των κατοίκων που γνωρίζει γραφή και ανάγνωση είναι 0,09 και δεν είναι στατιστικά σημαντικός ($p = .217$).

Σημείωση : Αν στη θέση της μεταβλητής literacy (l) ήταν μία ποιοτική μεταβλητή (όπως για παράδειγμα η επικρατούσα θρησκεία) τότε θα έπρεπε να δοθεί η εντολή :

mylinearmodelstd = lm(scale(w) ~ scale(m) + factor(l))

3.8 Μη γραμμική παλινδρόμηση

Πολλές φορές, ιδιαίτερα στις βιολογικές επιστήμες το διάγραμμα διασποράς συνηγορεί σε μία μη γραμμική σχέση.

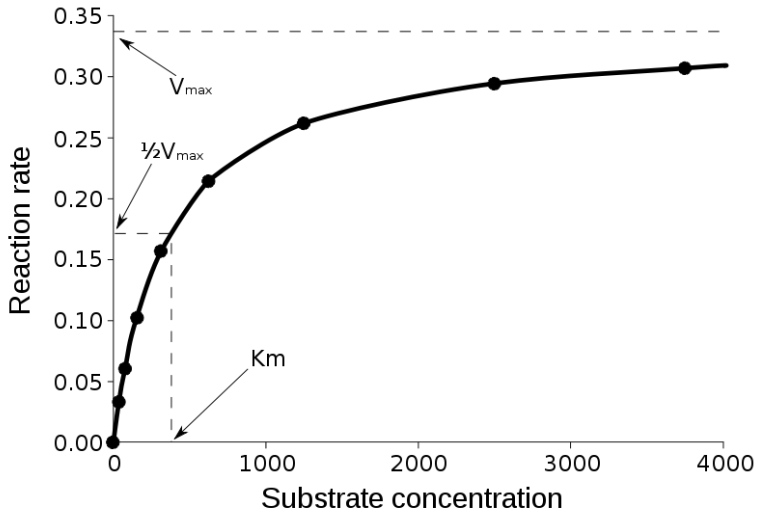


Διάγραμμα 17: Μη γραμμικό μοντέλο (1ο παράδειγμα)

Στην περίπτωση του διαγράμματος 17 το βέλτιστο μοντέλο προκύπτει να είναι

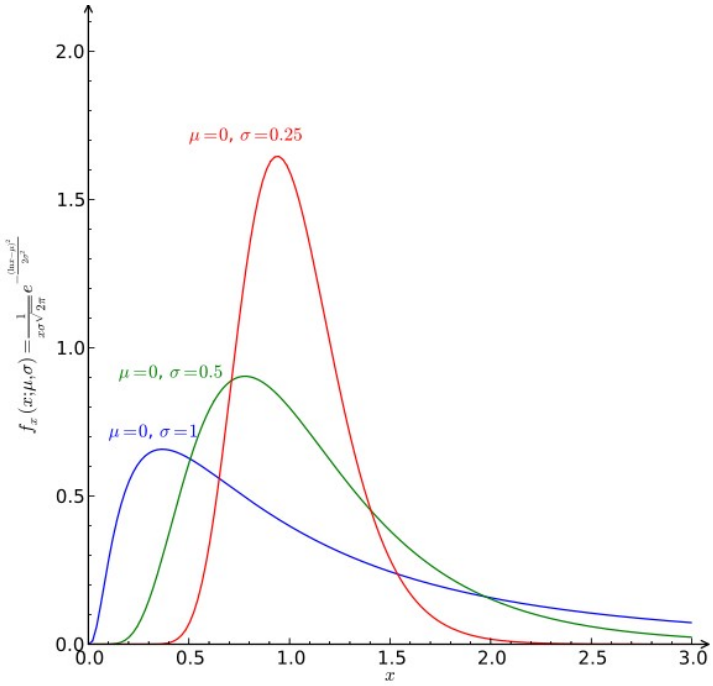
$$Y = 1{,}73 - 0{,}2\,\sigma\upsilon\nu\left(0{,}63\,X\right) - 1{,}41\,\eta\mu\left(0{,}63\,X\right) - 0{,}77\,\sigma\upsilon\nu\left(2\cdot 0{,}63\,X\right) - 1{,}31\,\eta\mu\left(2\cdot 063\,X\right)$$

το οποίο ένα έμπειρο μάτι ίσως τα αναγνωρίζει ως τους πρώτους 4 όρους μίας σειράς Fourier!



Διάγραμμα 18: Μη γραμμικό μοντέλο (2ο παράδειγμα)

Στην περίπτωση του διαγράμματος 18 το βέλτιστο μοντέλο είναι $\upsilon = \frac{V_{max} X}{K_m + X}$. Υπάρχουν απεριόριστες περιπτώσεις για τα μη γραμμικά μοντέλα.



Διάγραμμα 19: Λογαριθμοκανονική κατανομή

Ένας τρόπος για να επεξεργαστούμε μη γραμμικά μοντέλα είναι η κατάλληλη αλλαγή μεταβλητής και η έκφρασή τους γραμμικά. Η διαδικασία αυτή βρίσκει εφαρμογή σε βιολογικά μοντέλα όπου η εξέλιξη τους είναι συνήθως πολλαπλασιαστική και εμφανίζεται λογαριθμοκανονική κατανομή (διάγραμμα 19, σελίδα 111). Άλλη εκδοχή είναι η χρήση ειδικών διαδικασιών όπως η **nlreg** του R – Project (πίνακας 3.4).

Πίνακας 3.4: Μη γραμμική παλινδρόμηση με υπολογιστή

library(nlreg) [ή install.packages("nlreg", dependencies = TRUE)]

mydata = data.frame(w = women, m = man, l = literacy)

Αναζήτηση μοντέλου της μορφής : $\text{women} = b_0 + b_1 * \text{men}^3 + c$

mymodel <- nlreg(w~b0 + b1*m^3, start = c(b0=4, b1=0.1), data = mydata)

Αναζήτηση μοντέλου της μορφής : $\text{women} = b_0 + b_1 \sqrt{\text{men}} + c$

mymodel <- nlreg(w~b0 + b1*sqrt(m), start = c(b0=4, b1=0.1), data = mydata)

3.9 Παρουσίαση των αποτελεσμάτων της παλινδρόμησης

Πρώτα από όλα πρέπει να παρουσιάζονται οι συντελεστές γραμμικής συσχέτισης Pearson είτε στο κείμενο είτε σε πίνακα αν το μοντέλο αφορά περισσότερες από δύο μεταβλητές. Συνήθως, ο συντελεστής r και η στατιστική του σημαντικότητα p γράφεται σε πλάγια γραφή και δεν αναφέρεται το αρχικό μηδενικό. Παράδειγμα : “Το προσδόκιμο ζωής μεταξύ ανδρών και γυναικών βρέθηκε ισχυρά θετικά συσχετισμένο, Pearson’s $r(29) = .95$, $p < .001$.”

Στη συνέχεια, αν το γραμμικό μοντέλο είναι με μία ανεξάρτητη μεταβλητή, αρκεί να δοθεί η εξίσωση και ο συντελεστής προσδιορισμού R^2. Για παράδειγμα

“*Το προσδόκιμο ζωής των γυναικών (women) μπορεί να προβλεφθεί από αυτό των ανδρών (men) με την εξίσωση : women = 1.17 * men – 6, $R^2 = .96$*”

Στην περίπτωση που το γραμμικό μοντέλο έχει παραπάνω από μία ανεξάρτητη μεταβλητή τότε αρκεί να αναφερθούν οι συντελεστές του μοντέλου σε τυποποιημένη μορφή και η στατιστική σημαντικότητα για κάθε έναν από αυτούς. Αν στα πλαίσια της δημοσίευσης είναι χρήσιμη η εμφάνιση της εξίσωσης τότε ο ερευνητής μπορεί να την εμφανίσει είτε ως εξίσωση είτε σε μορφή πίνακα όπου θα αναφέρονται οι απόλυτες και οι τυποποιημένες τιμές των συντελεστών. Για παράδειγμα αναφορικά με τα αποτελέσματα της παραγράφου 3.7 :

“*Ως προς την πρόβλεψη του προσδόκιμου ζωής των γυναικών βρέθηκε πως το προσδόκιμο ζωής των ανδρών (Beta = 0.9, p <.001) είναι στατιστικά σημαντικός παράγοντας. Το ποσοστό του πληθυσμού που γνωρίζει γραφή και ανάγνωση δεν προέκυψε στατιστικά σημαντικός παράγοντας (Beta = 0.09, μ.σ.). Ο συντελεστής προσδιοριστίας του γραμμικού μοντέλου ήταν $R^2 = .96$*”.

Δραστηριότητες

1. Δίνονται οι εξής παρατηρήσεις των μεταβλητών X και Y

X	0	2	5	6
Y	8	8	5	3

(α) Να γίνει το διάγραμμα διασποράς των μεταβλητών X και Y.

(β) Να βρεθεί η συνδιακύμανση των μεταβλητών X και Y. Πως ερμηνεύεται η τιμή που βρέθηκε;

(γ) Να βρεθεί ο συντελεστής συσχέτισης Pearson. Πως ερμηνεύεται η τιμή που βρέθηκε;

(δ) Να βρεθεί γραμμικό μοντέλο πρόβλεψης των τιμών της μεταβλητής Y από τη μεταβλητή X.

(ε) Αν X = 4 τότε ποια περιμένουμε να είναι η τιμή του Y;

(στ) Να βρεθεί ο συντελεστής προσδιοριστίας του γραμμικού μοντέλου που βρέθηκε.

2. Επισκεφθείτε την ιστοσελίδα της ελληνικής στατιστικής αρχής http://www.statistics.gr/

   και επιλέξτε ένα κατάλληλο σύνολο δεδομένων και δημιουργήστε ένα γραμμικό μοντέλο πρόβλεψης. Πιο συγκεκριμένα πρέπει να κάνετε τα εξής :

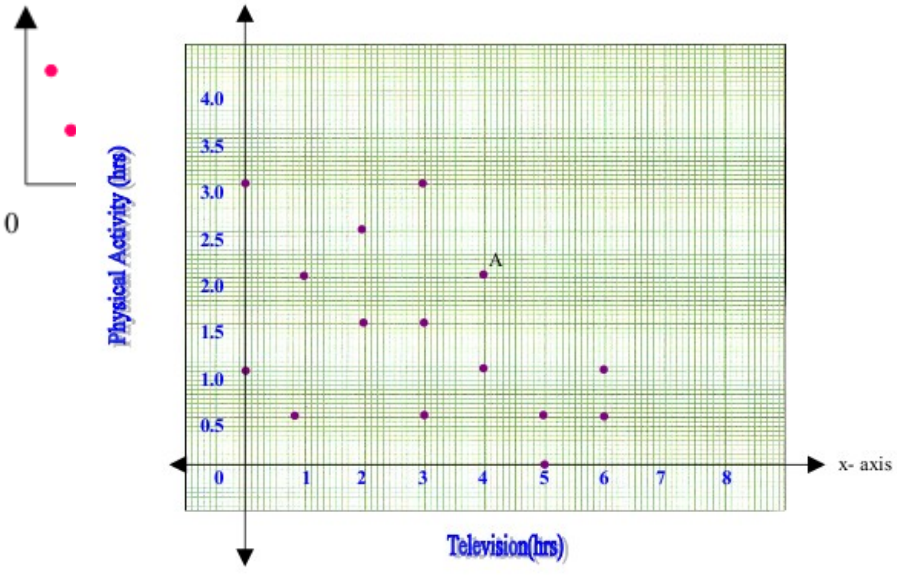
   (α) Επιλέξτε ένα σύνολο δεδομένων με τουλάχιστον 2 μεταβλητές.

   (β) Αποφασίστε ποια θα είναι η εξαρτημένη και ποια η ανεξάρτητη μεταβλητή.

   (γ) Δημιουργήστε γραμμικό μοντέλο πρόβλεψης (με το Calc!)

   (δ) Προχωρήστε σε πρόβλεψη τιμής της Y από τη X και συγκρίνεται με τα πραγματικά δεδομένα.

3. Βρείτε το πρόσημο και το μέγεθος του συντελεστή συσχέτισης του Pearson που αντιστοιχεί στα παρακάτω διαγράμματα

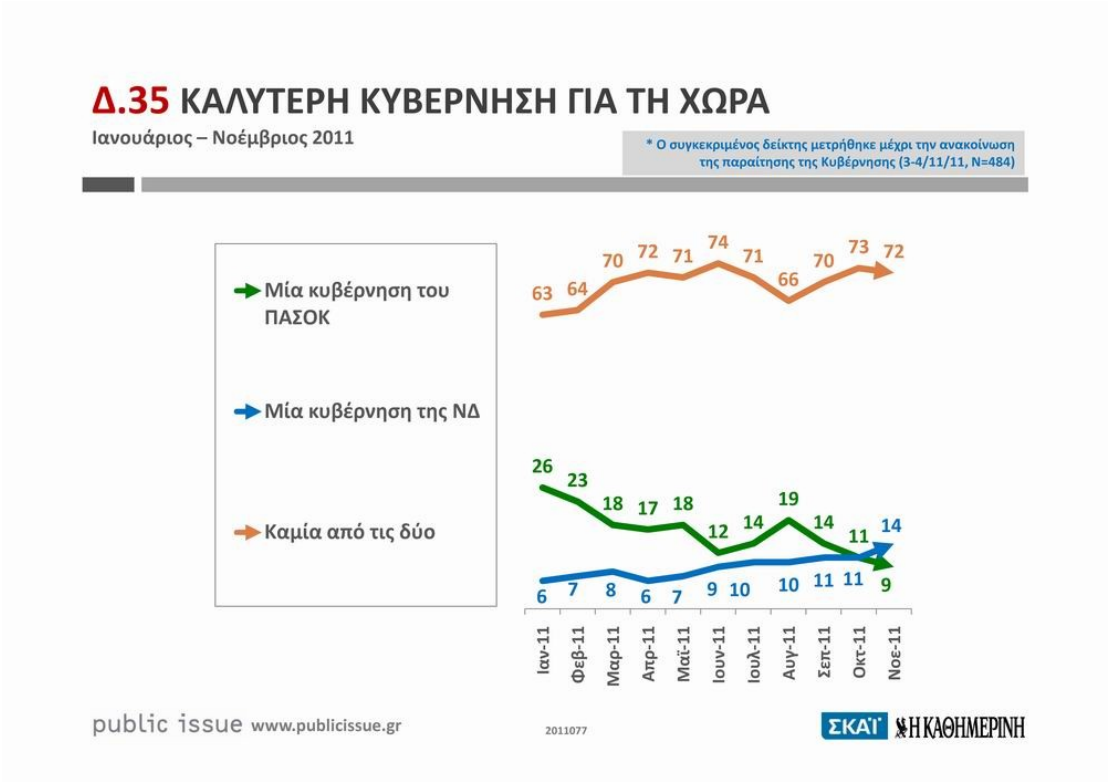


4. Από το διάγραμμα διασποράς που ακολουθεί να ανακτήσετε τα δεδομένα από τα οποία προήλθε

Κεφάλαιο 4 Χρονοσειρές

Στη στατιστική επιστήμη στην οικονομετρία και τα οικονομικά μαθηματικά, μια χρονολογική σειρά είναι μια ακολουθία χρονολογικών δεδομένων, η οποία μετράται συνήθως σε διαδοχικές χρονικές στιγμές οι οποίες απέχουν μεταξύ τους κατά ίσα χρονικά διαστήματα. Παραδείγματα χρονοσειρών είναι η ημερήσια τιμή κλεισίματος του γενικού δείκτη του χρηματιστηρίου ή οι ημερήσιες πωλήσεις ενός προϊόντος στη διάρκεια του τελευταίου έτους.

Η ανάλυση των χρονολογικών σειρών αποσκοπεί στην εύρεση χρήσιμων χαρακτηριστικών της χρονοσειράς που θα επιστρέψει στον ερευνητή να κατανοήσεις καλύτερα το φαινόμενο που παρατηρεί. Επιπλέον, ιδιαίτερα επιθυμητή είναι συνήθως η δημιουργία μοντέλων πρόβλεψης της μελλοντικής συμπεριφοράς της χρονοσειράς με βάση τις τιμές που παρατηρήθηκαν στο παρελθόν.



Διάγραμμα 20: http://www.publicissue.gr/1925/varometro-analysis-nov-2011/

Οι χρονοσειρές συνήθως περιγράφονται γραφικά με τα χρονοδιαγράμματα, στα οποία ο χρόνος τοποθετείται στον οριζόντιο άξονα σε κατάλληλες μονάδες ενώ η τιμή του χαρακτηριστικού που παρατηρείται τοποθετείται στον κάθετο άξονα σε κατάλληλη κλίμακα. Στο διάγραμμα 20 παρουσιάζεται ένα χρονοδιάγραμμα από μία πρόσφατη δημοσκόπηση.

4.1 Τρόποι στατιστικής ανάλυσης χρονοσειρών

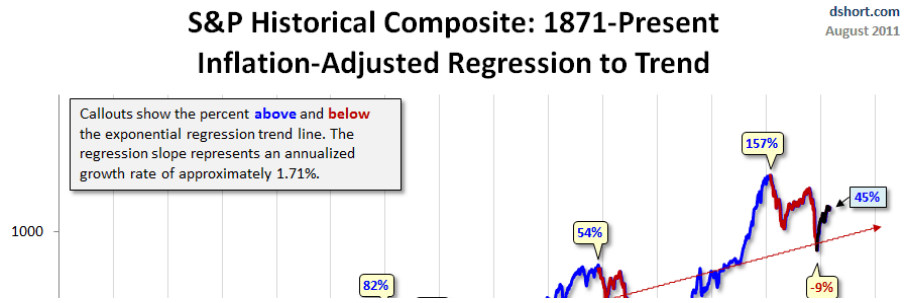
Οι βασικότεροι τρόποι με τους επεξεργαζόμαστε μία χρονοσειρά είναι οι εξής :

1. Ανάλυση της χρονοσειράς στις κυριότερες συνιστώσες
2. Έλεγχος για αυτοσυσχέτιση – ανίχνευση γραμμικής εξάρτησης συνεχόμενων τιμών
3. Φασματική ανάλυση

4.2 Ανάλυση της χρονοσειράς στις κυριότερες συνιστώσες

Η ανάλυση μίας χρονοσειράς στις κυριότερες συνιστώσες της είναι η βασικότερη διαδικασία επεξεργασίας μίας χρονοσειράς. Στόχος είναι η χρησιμοποίηση των διαθέσιμων δεδομένων για την καταγραφή όλων των “επιρροών” που διαμορφώνουν την τελική τιμή της χρονοσειράς. Συνήθως, η ανάλυση γίνεται στις εξής συνιστώσες :

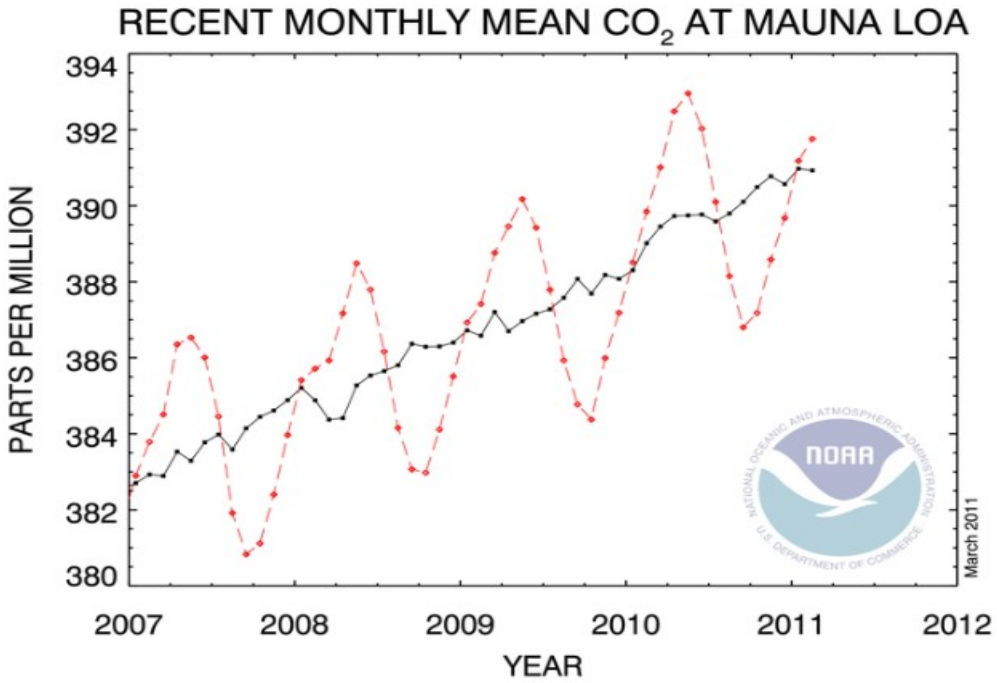
- Τη συνιστώσα της κύριας τάσης T_t (secular trend) η οποία περιγράφει τη μακροχρόνια εξέλιξη της χρονοσειράς (σχεδιάζεται ως μία ευθεία, π.χ. Διάγραμμα 21 *Παρατήρηση :* Δεν είναι απαραίτητο για μία χρονοσειρά να έχει κάποια κύρια τάση αύξησης ή μείωσης των τιμών της.
- Τη συνιστώσα της κυκλικής εναλλαγής C_t (cyclical fluctuation) η οποία περιγράφει επαναλαμβανόμενες μη περιοδικές μεταβολές στη χρονοσειρά. (π.χ διάγραμμα 22) *Παρατήρηση :* Σημειώνεται πως για να χαρακτηριστεί μία εναλλαγή γύρω από τη γραμμή τάσης ως κυκλική εναλλαγή δεν πρέπει να είναι αυτή περιοδική. Δηλαδή, στην πράξη πρέπει τα σημεία της χρονοσειράς να βρίσκονται κάτω από τη γραμμή τάσης για μια σειρά ετών και στη συνέχεια πάνω από τη γραμμή τάσης για την επόμενη σειρά ετών πάνω από τη γραμμή τάσης. Τα δύο χρονικά διαστήματα δεν πρέπει να είναι σταθερά, καθώς αν είναι σταθερά τότε η εναλλαγή ίσως αρμόζει να ονομαστεί περιοδική (ή εποχιακή).
- Την εποχιακή συνιστώσα (seasonal variation) η οποία ορίζεται από τις εποχιακές (περιοδικές) μεταβολές (π.χ. Διάγραμμα 23)



Διάγραμμα 22:Σύγκριση γάμων ανά 1.000 άτομα πληθυσμού και ποσοστού ανεργίας ενηλίκων. Μέθοδος : κινούμενος μέσος 24 μηνών.

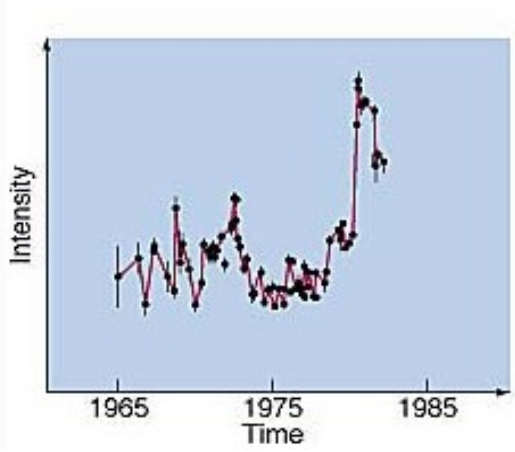
Πηγή : http://stateofourunions.org/2009/marriage_and_the_recession.php

- Τη συνιστώσα του τυχαίου σφάλματος της χρονοσειράς (irregular variations) η οποία περιέχει όλες τις μεταβολές που δεν ανήκουν στις προηγούμενες συνιστώσες. (π.χ. Διάγραμμα 24)



Διάγραμμα 23: Εποχιακές διακυμάνσεις της παγκόσμιας έκλυσης CO2. Η εποχιακότητα ανά τρίμηνο του έτους είναι εμφανής.

Πηγη : http://joannenova.com.au/2011/04/is-man-made-co2-different-1000-years-try-4-years/



Διάγραμμα 24: Ακανόνιστες μεταβολές της φωτεινότητας ενός γαλαξία Shefert σε περίοδο 20 ετών.

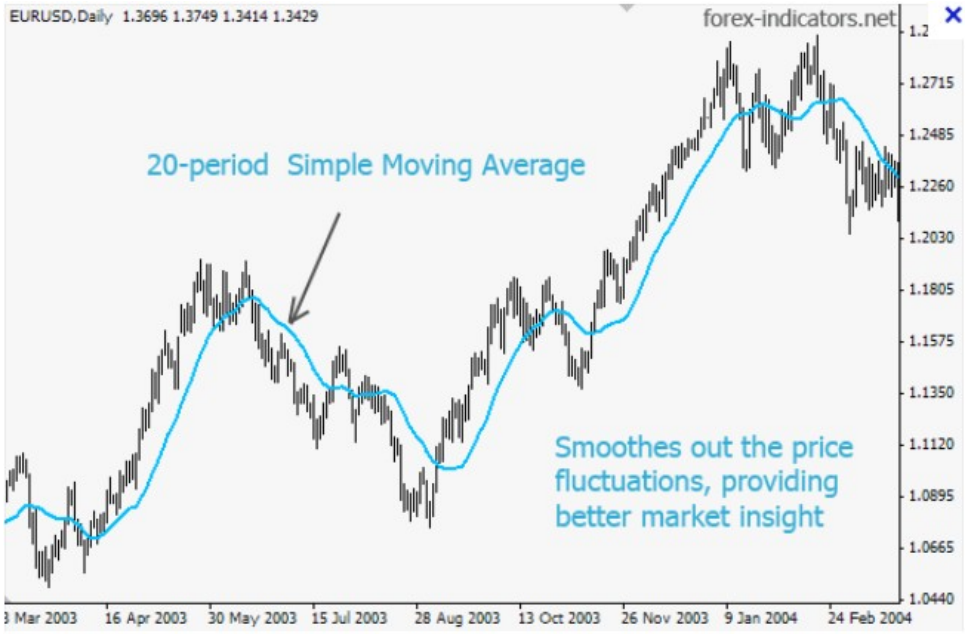
Πηγή : http://obsn3.on.br/~jlkm/astron2e/AT_MEDIA/CH25/CHAP25AT.HTM

4.3 Μέθοδοι προσδιορισμού της κύριας τάσης

Ο προσδιορισμός της κύριας τάσης (αν αυτή υπάρχει) είναι το πρώτο βήμα στη μελέτη μίας χρονοσειράς.

4.3.1 Μέθοδος του κινητού μέσου

Για τον εντοπισμό της κύριας τάσης υπάρχουν πολλές μέθοδοι, απλές και περισσότερο σύνθετες. Ως απλή μέθοδος θεωρείται η μέθοδος του κινητού μέσου κατά την οποία απλά υπολογίζεται η μέση τιμή των τιμών της χρονοσειράς σε ορισμένο χρονικό διάστημα. Το διάγραμμα του κινητού μέσου πολλές φορές φανερώνει τη γενική εξέλιξη της χρονοσειράς.



Διάγραμμα 25: Παράδειγμα υπολογισμού κινούμενου μέσου όρου. Πηγή : http://forex-indicators.net/trend-indicators

Παράδειγμα

Δίνεται η χρονοσειρά :

Χρόνος	1	2	3	4	5	6	7	8	9	10
Τιμή	3	8	25	20	22	16	14	6	12	15

Να βρεθεί ο κινούμενος μέσος όρος 3 σημείων.

Λύση

Ονομάζουμε x_i τις αρχικές παρατηρήσεις και y_i τις παρατηρήσεις του κινούμενου χρονικού μέσου. Θέλουμε τον κινούμενο μέσο όρο 3 σημείων άρα πρέπει για κάθε μία χρονική στιγμή να υπολογίσουμε τη μέση τιμή των τριών τελευταίων παρατηρήσεων. Είναι κατανοητό πως δεν μπορεί να υπολογιστεί τέτοιος μέσος όρος για την πρώτη και τη

τελευταία παρατήρηση άρα οι υπολογισμοί μας ξεκινούν από την τρίτη παρατήρηση. Υπολογίζουμε

$$y_3 = \frac{x_1+x_2+x_3}{3} = \frac{3+8+25}{3} = \frac{36}{3} = 12 \quad \text{και} \quad y_4 = \frac{x_2+x_3+x_4}{3} = \frac{8+25+20}{3} = \frac{53}{3} = 17{,}7$$

$$y_5 = \frac{x_3+x_4+x_5}{3} = \frac{25+20+22}{3} = \frac{67}{3} = 22{,}3 \ , \quad y_6 = \frac{x_4+x_5+x_6}{3} = \frac{20+22+16}{3} = \frac{58}{3} = 19{,}3 \ ,$$

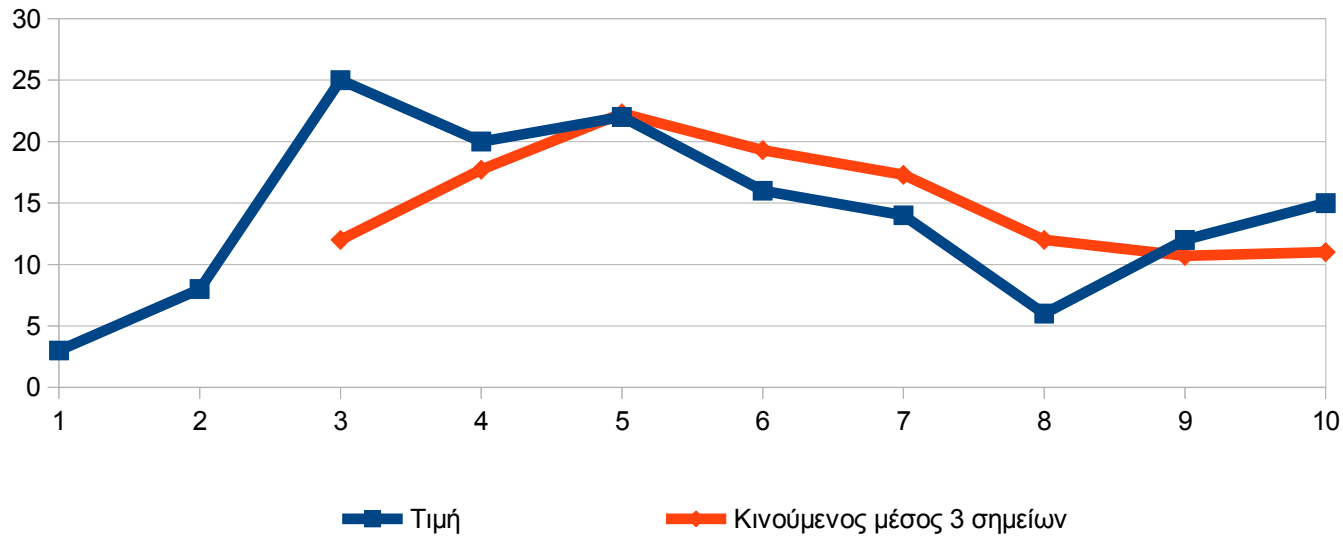
$$y_7 = \frac{x_5+x_6+x_7}{3} = \frac{22+16+14}{3} = \frac{52}{3} = 17{,}3 \ , \quad y_8 = \frac{x_6+x_7+x_8}{3} = \frac{16+14+6}{3} = \frac{36}{3} = 12 \ ,$$

$$y_9 = \frac{x_7+x_8+x_9}{3} = \frac{14+6+12}{3} = \frac{32}{3} = 10{,}7 \ , \quad y_{10} = \frac{x_8+x_9+x_{10}}{3} = \frac{6+12+15}{3} = \frac{33}{3} = 11 \ .$$

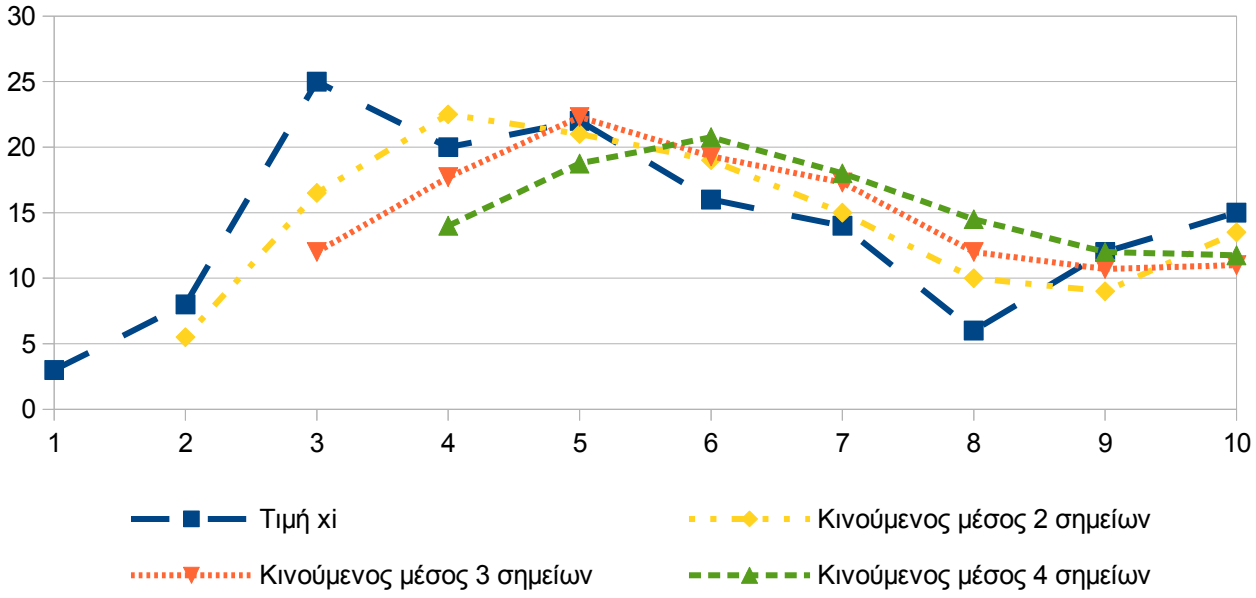
Συμπληρώνουμε τον πίνακα με τις τιμές του κινούμενου μέσου 3 σημείων.

Χρόνος	1	2	3	4	5	6	7	8	9	10
Τιμή	3	8	25	20	22	16	14	6	12	15
Κινούμενος μέσος 3 σημείων			12	17,7	22,3	19,3	17,3	12	10,7	11

Το διάγραμμα 26 παρουσιάζει τις αρχικές τιμές μαζί με τον κινούμενο μέσο όρο 3 σημείων. Στο διάγραμμα 27, σελίδα 121, παρουσιάζεται η ιδιότητα του κινούμενου μέσου όρου να “απορροφά” εύκολα τις μεγάλες διακυμάνσεις της χρονοσειράς με την κατάλληλη αύξηση των στοιχείων της σειράς που χρησιμοποιούνται για τον υπολογισμό του.



Διάγραμμα 26: Διάγραμμα των τιμών και του κινούμενου μέσου τριών σημείων



Διάγραμμα 27: Όσο μεγαλώνει το πλήθος των σημείων που χρησιμοποιείται για τον υπολογισμό του κινούμενου μέσου όρου τόσο πιο ομαλή γίνεται η γραφική παράσταση του.

Πίνακας 4.1: Υπολογισμός και γραφική αναπαράσταση κινούμενου μέσου όρου στον υπολογιστή

Δεν υπάρχει ενσωματωμένη συνάρτηση, μπορεί να υπολογιστεί όπως περιγράφεται αναλυτικά στην προηγούμενη παράγραφο**.**

Ορίζουμε **x = c(3, 8, 25, 20, 22, 16, 14, 6, 12, 15).** Φορτώνουμε τη βιβλιοθήκη zoo με την εντολή **library(zoo)** και μετά χρησιμοποιούμε τη συνάρτηση **rollmean(x,2)** για τον υπολογισμό κινούμενο μέσο 2 σημείων κλπ. Με την εντολή **ts.plot(x)** προκύπτει το διάγραμμα του διανύσματος x ως χρονοσειρά.

Ένα περισσότερο εξευγενισμένο διάγραμμα προκύπτει με τις εντολές

plot(x, ann=FALSE, type="n")

abline(h=0, col=gray(.90))

lines(x, col="green4", lty="dotted")

points(x, bg="limegreen", pch=21)

title(main="Διάγραμμα χρονοσειράς",xlab="Τιμή", col.main="blue", col.lab=gray(.8), cex.main=1.2, cex.lab=1.0, font.main=4, font.lab=3)

Για να προκύψει το κοινό διάγραμμα της χρονοσειράς και του κινούμενου μέσου όρου μπορεί να χρησιμοποιηθεί ο κώδικάς

x.ts = ts(x)

y.ts = ts(y)

require(graphics)

ts.plot(x.ts, y.ts,gpars=list(xlab="Χρόνος", ylab="Τιμή", lty=c(1:2)))

Το R – Project μπορεί να δημιουργήσει εύκολα πολλά διαγράμματα. Κάποια από αυτά επιδεικνύονται με τις εντολές

example(plot.ts)

example(ts.plot)

library(zoo)

example(plot.zoo)

library(lattice)

example(xyplot.zoo)

4.3.2 Μέθοδος της ευθείας των ελαχίστων τετραγώνων

Περισσότερο πολύπλοκη μέθοδος που απαιτεί υπολογιστή είναι η μέθοδος της ευθείας των ελαχίστων τετραγώνων με την οποία απλά υπολογίζεται η ευθεία γραμμικής παλινδρόμησης της εξαρτημένης μεταβλητής πάνω στο χρόνο (δηλαδή ο χρόνος t παίρνει το ρόλο της ανεξάρτητης μεταβλητής στο γραμμικό μοντέλο). Η εξίσωσή της ευθείας ελαχίστων τετραγώνων, έχει τη μορφή

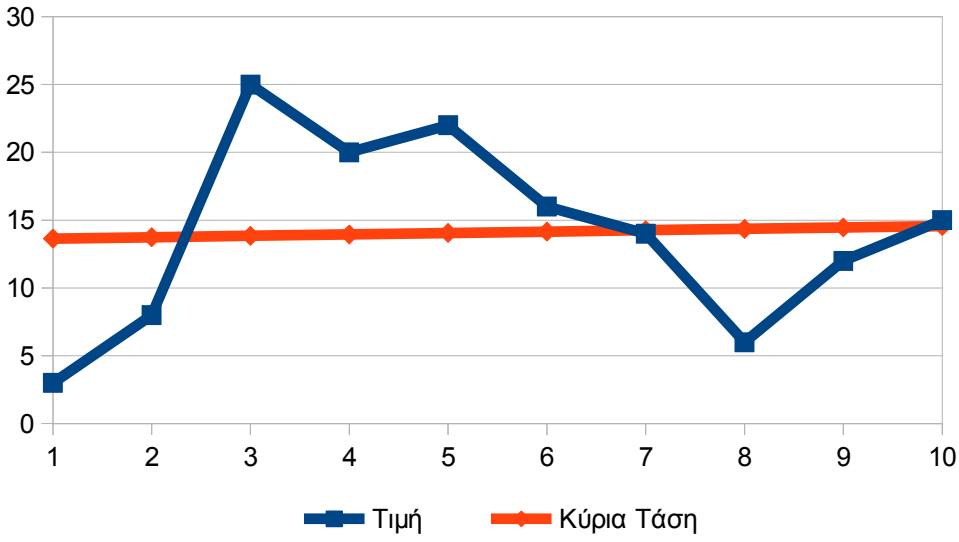
X = α * t+β.

Οι συντελεστές του γραμμικού μοντέλου υπολογίζονται από τις τιμές της μεταβλητής X ως εξής :

$$\alpha = \frac{\sum_{i=1}^{v}(t_i - \bar{t})(x_i - \bar{x})}{\sum_{i=1}^{v}(t_i - \bar{t})^2} \quad \text{και} \quad \beta = \bar{x} - \alpha \cdot \bar{t}$$

Στον πίνακα που ακολουθεί παρουσιάζονται οι απαραίτητοι υπολογισμοί, ενώ στο διάγραμμα 28, σελίδα 123, εμφανίζεται η κύρια τάση των χρονικών δεδομένων του παραδείγματος.

Χρόνος	1	2	3	4	5	6	7	8	9	10	Σύνολο
Τιμή	3	8	25	20	22	16	14	6	12	15	
(xi – x)(ti - t)	50,0	21,4	-27,3	-8,9	-4,0	1,0	-0,1	-20,3	-7,4	4,1	8,5
(ti - t)^2	20,3	12,3	6,3	2,3	0,3	0,3	2,3	6,3	12,3	20,3	82,5
Κύρια Τάση	13,6	13,7	13,8	13,9	14,0	14,2	14,3	14,4	14,5	14,6	



Διάγραμμα 28: Κύρια τάση της χρονοσειράς του παραδείγματος

Ακόμα περισσότερο το ίδιο μπορεί να γίνει προς την κατεύθυνση ενός μη γραμμικού μοντέλου εξέλιξης στο χρόνο όπου ακολουθείται η γενικότερη μέθοδος προσαρμογής καμπύλης στα

δεδομένα, καμπύλη που την εντοπίζει ο ερευνητής ανάλογα με τα δεδομένα του

προβλήματος.

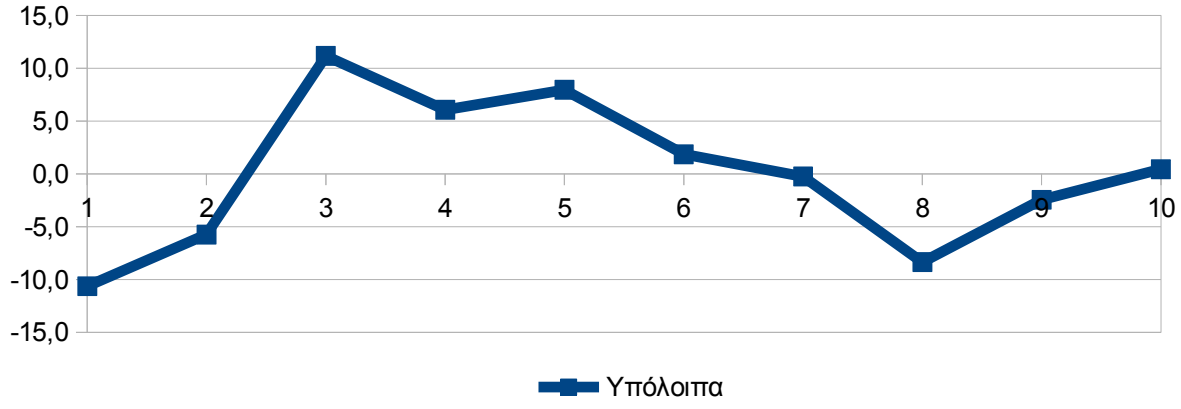
4.4 Προσδιορισμός της συνιστώσας της κυκλικής εναλλαγής

Ο τρόπος με τον οποίο βρίσκουμε την κυκλική μεταβολή – αν αυτή υπάρχει – είναι η μέθοδος των υπολοίπων (ή καταλοίπων) (residuals).

Στον πίνακα που ακολουθεί εμφανίζονται οι υπολογισμοί των υπολοίπων και των σχετικών (ως προς την πραγματική τιμή) υπολοίπων ενώ στα διαγράμματα 29 και 30, σελίδα 124 παρουσιάζεται η εξέλιξη των υπολοίπων και των σχετικών υπολοίπων με την πάροδο των χρονικών στιγμών.

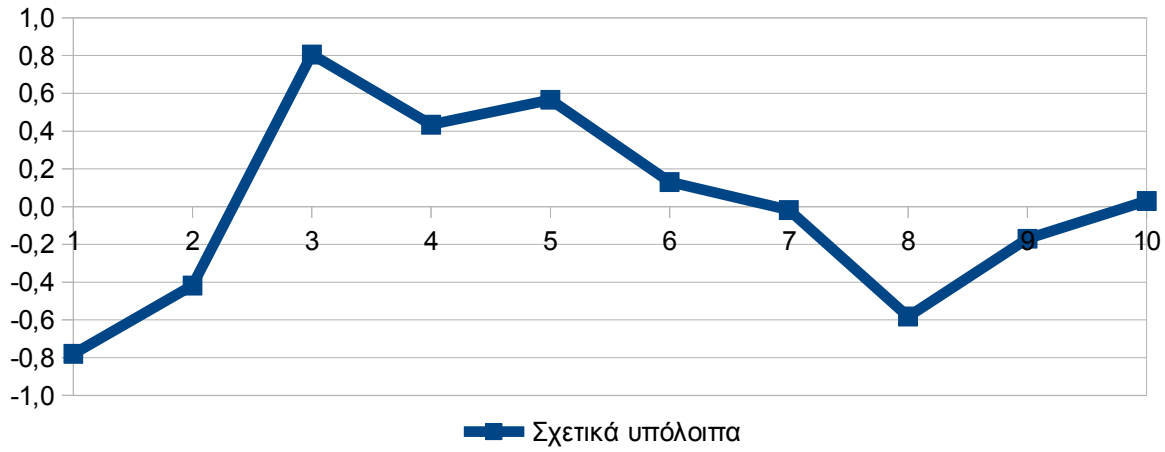
Χρόνος	1	2	3	4	5	6	7	8	9	10
Τιμή	3	8	25	20	22	16	14	6	12	15
Κύρια Τάση	13,6	13,7	13,8	13,9	14,0	14,2	14,3	14,4	14,5	14,6
Υπόλοιπα	-10,6	-5,7	11,2	6,1	8,0	1,8	-0,3	-8,4	-2,5	0,4
Σχετικά υπόλοιπα	-0,8	-0,4	0,8	0,4	0,6	0,1	0,0	-0,6	-0,2	0,0

Πίνακας 4.2: Υπόλοιπο = Τιμή - Κύρια τάση
Σχετικό υπόλοιπο = (Τιμή – Κύρια τάση) / Κύρια τάση

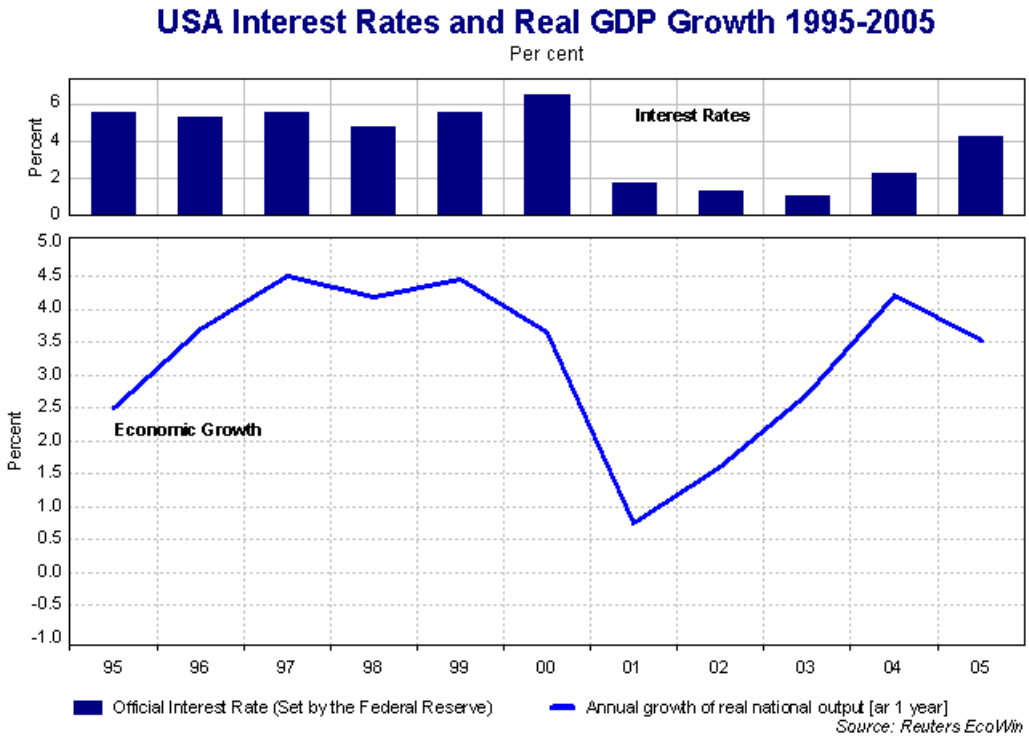


Διάγραμμα 29: Διάγραμμα υπολοίπων της χρονοσειράς

Το διάγραμμα σχετικών υπολοίπων δίνει περισσότερες πληροφορίες σε σχέση με το απλό διάγραμμα υπολοίπων καθώς δεν εξαρτάται από το σύστημα μέτρησης των τιμών της μεταβλητής.



Διάγραμμα 30: Διάγραμμα σχετικών υπολοίπων της χρονοσειράς



Διάγραμμα 31: Παράδειγμα κυκλικής μεταβολής. Πηγή : http://tutor2u.net/economics/revision-notes/a2-macro-cyclical-fluctuations.html

Η παρατήρηση της κυκλικής συνιστώσας μίας χρονοσειράς μπορεί να οδηγήσει σε καλύτερη κατανόηση των παραγόντων που την επηρεάζουν και στη λήψη των κατάλληλων

αποφάσεων για μεταβολή της μελλοντικής της πορείας. Χαρακτηριστικό παράδειγμα είναι η σχέση τραπεζικών επιτοκίων και ανάπτυξης όπως αποτυπώνεται στο διάγραμμα 31, σελίδα 125.

4.5 Προσδιορισμός της εποχιακής συνιστώσας

Η εποχιακή συνιστώσα γίνεται άμεσα εμφανής από το χρονοδιάγραμμα. Με οπτική παρατήρηση γίνεται φανερό αν υπάρχει επανάληψη των τιμών ανά εβδομάδα, μήνα, τρίμηνο, τετράμηνο, εξάμηνο ή ως προς οποιαδήποτε μονάδα μέτρησης έχει νόημα στο πλαίσιο που τίθεται η χρονοσειρά.

Ο πιο απλός τρόπος για να εντοπιστεί η συνεισφορά της εποχικότητας στην τιμή της χρονοσειράς είναι ο υπολογισμός του **κεντρικού** κινούμενου μέσου όρου των γειτονικών όρων σε πλήθος που να καλύπτει μία πλήρη εποχική μεταβολή. Έτσι για παράδειγμα αν έχουμε δεδομένα ημερησίων πωλήσεων και παρατηρούμε εβδομαδιαία επανάληψη τότε για να εκτιμήσουμε αυτή την περιοδικότητα θα υπολογίζαμε για τη Δευτέρα, τη μέση τιμή των πωλήσεων από την προηγούμενη Παρασκευή έως και την επόμενη Πέμπτη, για την ήμερα Τρίτη, θα υπολογίζαμε από το Σάββατο έως την Παρασκευή κτλ. Ανάλογα, αν συλλέγαμε μηνιαία δεδομένα και παρατηρούσαμε ετήσια περιοδικότητα τότε θα υπολογίζαμε το κινούμενο μέσο όρο 12 μηνών κλπ.

Παρατήρηση : Πρέπει να τονιστεί πως όταν το πλήθος των χρονικών περιόδων είναι άρτιο (όπως συμβαίνει στους 12 μήνες) τότε δεν μπορούμε να πάρουμε ίσο πλήθος περιόδων πριν και μετά, περίπτωση κατά την οποία χωρίζουμε ανισομερώς τα δύο διαστήματα. Το σφάλμα της εκτίμησης θα είναι μικρό.

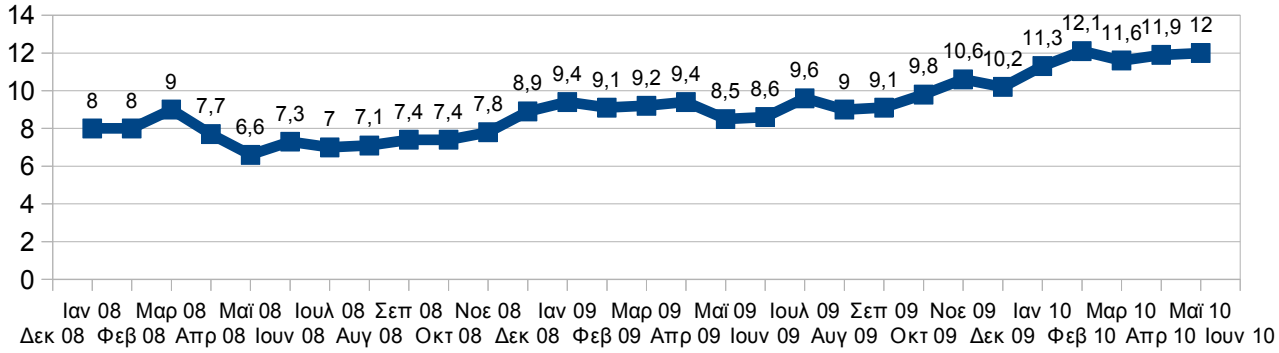
Η διαφορά που θα προκύπτει μεταξύ της πραγματικής τιμής και του κινούμενου μέσου όρου μπορεί να θεωρηθεί η περιοδική (εποχιακή) συνιστώσα της χρονοσειράς.

Παράδειγμα

Στον παρακάτω πινάκα δίνεται το ποσοστό ανεργίας στην Ελλάδα από τον Ιανουάριο του 2008 έως το Μάιο του 2010. Τα ίδια δεδομένα παρουσιάζονται στο διάγραμμα 32,σελίδα 127 που ακολουθεί.

Μήνας	Ποσοστό	Μήνας	Ποσοστό	Μήνας	Ποσοστό
Ιαν 08	8	Ιαν 09	9,4	Ιαν 10	11,3
Φεβ 08	8	Φεβ 09	9,1	Φεβ 10	12,1
Μαρ 08	9	Μαρ 09	9,2	Μαρ 10	11,6
Απρ 08	7,7	Απρ 09	9,4	Απρ 10	11,9
Μαϊ 08	6,6	Μαϊ 09	8,5	Μαϊ 10	12
Ιουν 08	7,3	Ιουν 09	8,6		
Ιουλ 08	7	Ιουλ 09	9,6		
Αυγ 08	7,1	Αυγ 09	9		
Σεπ 08	7,4	Σεπ 09	9,1		
Οκτ 08	7,4	Οκτ 09	9,8		
Νοε 08	7,8	Νοε 09	10,6		
Δεκ 08	8,9	Δεκ 09	10,2		

Να εκτιμηθεί η εποχιακή συνιστώσα της ανεργίας με τη μέθοδο του κινούμενου μέσου όρου.



Διάγραμμα 32: Ποσοστό ανεργίας στην Ελλάδα (Ιανουάριος 2008 έως Μάιος 2010)

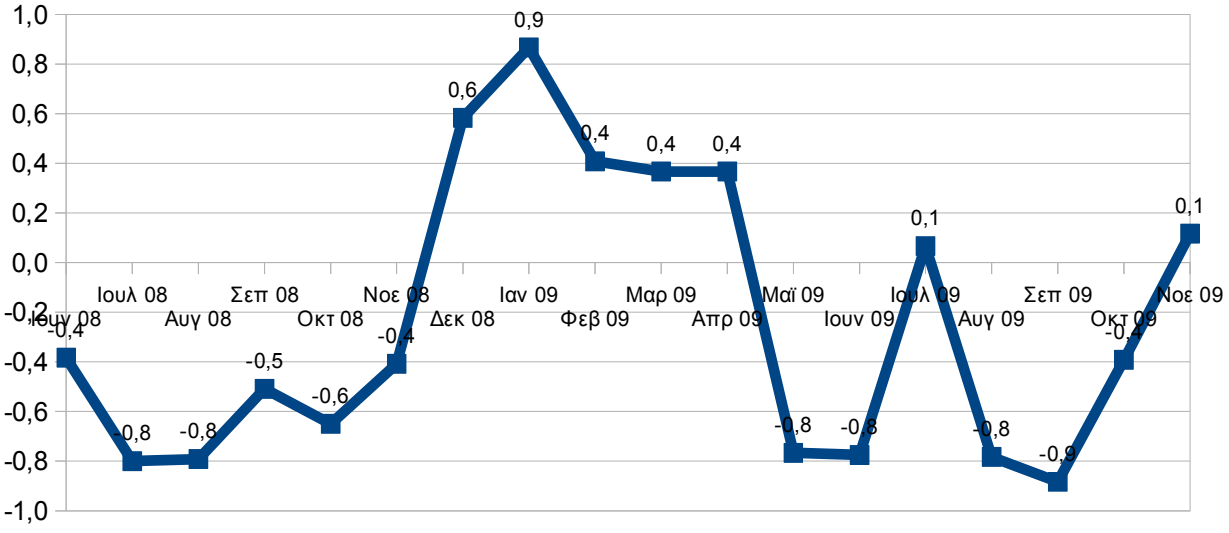
Λύση

Είναι φανερό πως η ανεργία στην Ελλάδα, λόγω της τουριστικής απασχόλησης η οποία συμβαίνει κατά κύριο λόγο το καλοκαίρι, είναι φαινόμενο που μεταβάλλεται σε ετήσια βάση. Με αυτήν την εκτίμηση καταλαβαίνουμε πως ο κινούμενος μέσος όρος πρέπει να υπολογίζεται από το σύνολο των 12 μηνών του έτους. Καθώς το πλήθος είναι ζυγό επιλέγουμε τους 5 προηγούμενους μήνες και τους 6 επόμενους. Αυτό σημαίνει πως η πρώτη τιμή για τον κεντρικό κινούμενο μέσο όρο θα υπολογιστεί για τον 6ο μήνα των δεδομένων δηλαδή για τον Ιούνιο του 2008 ενώ ο τελευταίος μήνας υπολογισμού θα είναι ο Νοέμβριος του 2009.

Οι υπολογισμοί γίνονται εύκολα με κάποιον υπολογιστή και δίνουν τον επόμενο πίνακα

	Ποσοστό	Κεντρικός κινούμενος μέσος	Διαφορά (εποχιακή συνιστώσα της ανεργίας)
Ιουν 08	7,3	7,7	-0,4
Ιουλ 08	7	7,8	-0,8
Αυγ 08	7,1	7,9	-0,8
Σεπ 08	7,4	7,9	-0,5
Οκτ 08	7,4	8,1	-0,7
Νοε 08	7,8	8,2	-0,4
Δεκ 08	8,9	8,3	0,6
Ιαν 09	9,4	8,5	0,9
Φεβ 09	9,1	8,7	0,4
Μαρ 09	9,2	8,8	0,4
Απρ 09	9,4	9,0	0,4
Μαϊ 09	8,5	9,3	-0,8
Ιουν 09	8,6	9,4	-0,8
Ιουλ 09	9,6	9,5	0,1
Αυγ 09	9	9,8	-0,8
Σεπ 09	9,1	10,0	-0,9
Οκτ 09	9,8	10,2	-0,4
Νοε 09	10,6	10,5	0,1

Στην επόμενη σελίδα παρουσιάζεται το διάγραμμα 33 της διαφοράς του παρατηρούμενου ποσοστού ανεργίας από την τιμή του κεντρικού κινούμενου μέσου το οποίο δεν είναι παρά μία καλή εκτίμηση της εποχιακής συνιστώσας της ανεργίας.



Διάγραμμα 33: Εποχιακή συνιστώσα της ανεργίας

4.6 Συνιστώσα του τυχαίου σφάλματος της χρονοσειράς

Η συνιστώσα του τυχαίου σφάλματος της χρονοσειράς δεν είναι τίποτα άλλο παρά ότι απομένει αν από την τιμή της χρονοσειράς αφαιρέσουμε την κύρια τάση, την κυκλική

εναλλαγή και την εποχιακή συνιστώσα. Το μέγεθος του τυχαίου σφάλματος δίνει το μέγεθος της αξιοπιστίας της πρόβλεψης της εξέλιξης της χρονοσειράς στο μέλλον, χωρίς να υπάρχει ωστόσο κάποιος κανόνας αξιολόγησης της.

4.7 Αυτοδιακύμανση και αυτοσυσχέτιση

Τόσο η αυτοδιακύμανση όσο και η αυτοσυσχέτιση είναι δύο στατιστικές ποσότητες που μας επιτρέπουν να ανιχνεύσουμε τυχόν περιοδικότητα μίας χρονοσειράς με ένα μετρήσιμο τρόπο.

Η αυτοδιακύμανση είναι η συνδιακύμανση της μεταβλητής που ορίζει τη χρονοσειρά με ένα αντίγραφό της χρονικά μετατοπισμένο κατά πλήθος χρονικών στιγμών που επιλέγεται από τον ερευνητή.

Η αυτοσυσχέτιση είναι η κανονικοποιημένη ως προς το πιθανό εύρος τιμών εκδοχή της αυτοδιακύμανσης.

Στην πράξη, υπολογίζουμε την αυτοσυσχέτιση της χρονοσειράς με τον εαυτό της σε όλες τις πιθανές μετατοπίσεις και παρατηρούμε το σύνολο των τιμών που προκύπτουν. Η συνδιακύμανση στα πρώτα βήματα, θα μειώνεται από το 1 μέχρι κάποιο σημείο όπου θα αρχίσει να ανέρχεται ξανά (αν υπάρχει κάποια περιοδικότητα). Στο σημείο όπου η αυτοσυσχέτιση θα ξαναγίνει θετική και ισχυρή, θα ορίζεται η περίοδος της περιοδικής συνιστώσας της χρονοσειράς.

4.7.1 Αυτοδιακύμανση

Θυμίζουμε πως ο τύπος της συνδιακύμανσης δύο μεταβλητών δίνεται από τον τύπο (παράγραφο 3.2, σελίδα 93) :

$$s_{XY}^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(x_i - \bar{x})(y_i - \bar{y}) \quad \text{(τύπος συνδιακύμανσης)}$$

Αν x_1, x_2, ..., x_κ είναι οι παρατηρήσεις μίας μεταβλητής στις χρονικές στιγμές 1, 2, ..., κ, τότε οι μετατοπισμένες παρατηρήσεις κατά τ στιγμές $(0 \le \tau \le \kappa)$ θα είναι οι $x_{1+\tau}$, $x_{2+\tau}$, ..., x_κ . Αν τις τοποθετήσουμε κατά ζεύγη, μαζί με τις αρχικές θα πάρουμε τον εξής πίνακα

Θέση	**1**	**2**	**3**	**...**	**κ-τ-1**	**κ-τ**	**κ-τ+1**	**...**	**κ-1**	**κ**

Αρχική χρονοσειρά	x_1	x_2	x_3	…	$x_{κ-τ-1}$	$x_{κ-τ}$	$x_{κ-τ+1}$	…	$x_{κ-1}$	$x_κ$
Μετατοπισμένη χρονοσειρά	$x_{1+τ}$	$x_{2+τ}$	$x_{3+τ}$	…	$x_{κ-1}$	$x_κ$				

Παρατηρούμε πως υπάρχουν πλέον μόνο κ – τ – 1 ζευγάρια τιμών από τα οποία μπορεί να υπολογισθεί η συνδιακύμανση. Τοποθετώντας στον τύπο της συνδιακύμανσης, στη θέση των τιμών x_i τα αρχικά στοιχεία της χρονοσειράς και στη θέση των τιμών y_i τις μετατοπισμένες παρατηρήσεις παίρνουμε τον τύπο της αυτοδιακύμανσης :

$$C_{XX} = \frac{1}{\kappa - \tau}\sum_{i=1}^{\kappa-\tau}(x_i - \bar{x})(x_{i+\tau} - \bar{x}_\tau)$$ (τύπος δειγματικής συνδιακύμανσης)

όπου $\bar{x} = \frac{1}{\kappa}\sum_{i=1}^{\kappa} x_i$ και $\bar{x}_\tau = \frac{1}{\kappa - \tau}\sum_{i=1+\tau}^{\kappa} x_i = \frac{1}{\kappa - \tau}\sum_{i=1}^{\kappa-\tau} x_{i+\tau}$

Παρατήρηση : Ο παραπάνω τύπος δεν είναι αμερόληπτος εκτιμητής της αυτοδιακύμανσης καθώς ο υπολογισμός των μέσων τιμών της χρονοσειράς και της μετατόπισής της κατά τ μονάδες γίνεται από το ίδιο το δείγμα. Αν με κάποιον άλλο τρόπο είναι γνωστή η μέση τιμή όλης της χρονοσειράς (έστω μ αυτή) τότε εναλλακτικά χρησιμοποιούμε τον τύπο

$$C_{XX} = \frac{1}{\kappa - \tau}\sum_{i=1}^{\kappa-\tau}(x_i - \mu)(x_{i+\tau} - \mu)$$ (αμερόληπτος τύπος συνδιακύμανσης)

Στη γλώσσα της θεωρίας πιθανοτήτων επιπλέον γράφουμε :

$$C_{XX}(t,s) = E[(X_t - \mu_t)(X_s - \mu_s)] = E[X_t X_s] - \mu_t \mu_s$$

όπου η διαφορά των s και t είναι αυτή που υποθέτει ο ερευνητής πως καθορίζει την περιοδική συμπεριφορά της χρονοσειράς.

4.7.2 Αυτοσυσχέτιση

Αν $x_1, x_2, ..., x_κ$ είναι οι παρατηρήσεις μίας μεταβλητής στις χρονικές στιγμές 1, 2, ..., κ, τότε η δειγματική αυτοσυσχέτιση με διαφορά (τ) ορίζεται να είναι

$$R_\tau = \frac{1}{(\kappa - \tau)\sigma^2}\sum_{i=1}^{\kappa-\tau}(x_i - \mu)(x_{i+\tau} - \mu) \ ,$$

όπου μ και σ η μέση τιμή και η τυπική απόκλιση της χρονοσειράς. Καθώς στην πράξη σπάνια γνωρίζουμε τις τιμές αυτές τις αντικαθιστούμε με τις αντίστοιχες δειγματικές ποσότητες, δηλαδή χρησιμοποιούμε τον τύπο

$$R_\tau = \frac{1}{(\kappa - \tau)s^2} \sum_{i=1}^{\kappa-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x}_\tau)$$

εκτίμηση που δεν είναι η καλύτερη δυνατή καθώς δεν είναι αμερόληπτη αλλά είναι η μόνη που μπορούμε να υλοποιήσουμε σε αυτήν την περίπτωση.

Παρατηρήσεις

1. Εύκολα αποδεικνύεται πως (τ = t-s = σταθερό) :

$$C_{XX}(\tau) = E[(X(t) - \mu)(X(t+\tau) - \mu)] = E[X(t)X(t+\tau)] - \mu^2 = R_\tau - \mu^2 \ .$$

2. Αν η χρονοσειρά είναι συνεχής, τότε αναπαριστάται από μία συνεχής συνάρτηση X(t). Σε αυτήν την περίπτωση η αυτοσυσχέτιση ορίζεται ως

$$C(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T X(t)X(t+\tau)dt$$

Παράδειγμα 1

Θα χρησιμοποιήσουμε τα δεδομένα ανεργίας 29 συνεχόμενων μηνών του προηγούμενου παραδείγματος. Είναι φανερό πως η κατάλληλη διαφορά για την περιοδικότητα είναι 12 χρονικές στιγμές.

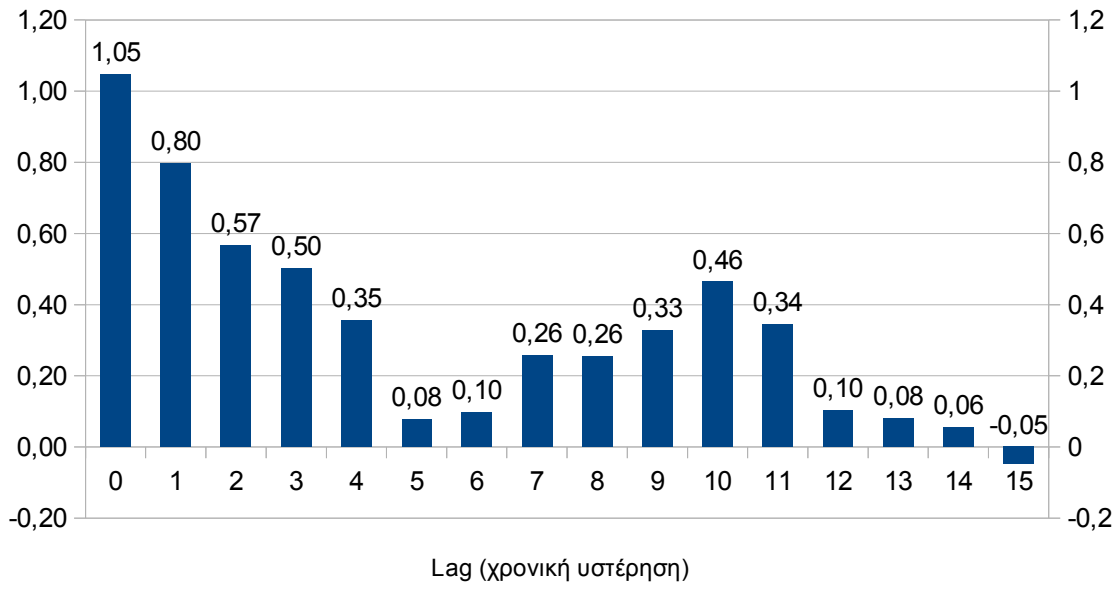
Οι υπολογισμοί γίνονται εύκολα με το LibreOffice Calc και παρουσιάζονται στον πίνακα της επόμενης σελίδας.

Οι συναρτήσεις που χρησιμοποιούνται είναι οι COVAR() και οι CORREL() σε δύο στήλες που περιέχουν τη χρονοσειρά και τη μετατόπισή της κατά 0 έως και 15 χρονικές στιγμές.

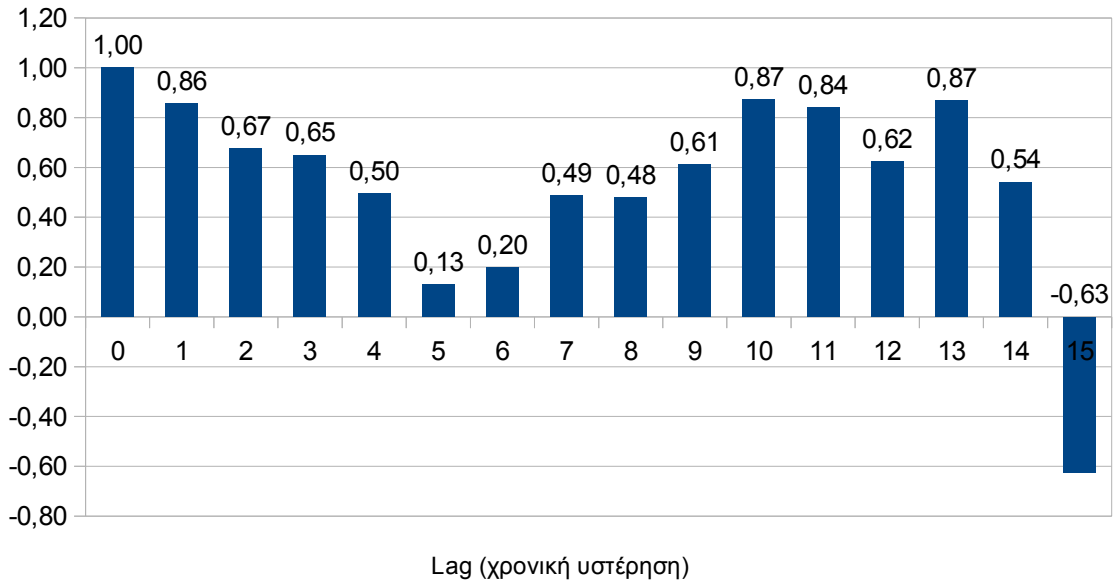
Από το διάγραμμα 35, σελίδα 133, παρατηρούμε πως ισχυρή αυτοσυσχέτιση εμφανίζεται μεταξύ των χρονικών υστερήσεων 10 έως και 13.

Από την εμπειρία μας και λόγω της φύσης του φαινομένου της ανεργίας, αντιλαμβανόμαστε πως η ορθότερη επιλογή για την περιοδική ερμηνεία του φαινομένου μεταξύ των αριθμών 10, 11, 12 και 13 είναι η υστέρηση 12, δηλαδή η ετήσια περιοδικότητα.

		Lag (χρονική υστέρηση)															
Μήνας	**Ποσοστό**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Ιουν 08	7,3	7,3															
Ιουλ 08	7	7	7,3														
Αυγ 08	7,1	7,1	7	7,3													
Σεπ 08	7,4	7,4	7,1	7	7,3												
Οκτ 08	7,4	7,4	7,4	7,1	7	7,3											
Νοε 08	7,8	7,8	7,4	7,4	7,1	7	7,3										
Δεκ 08	8,9	8,9	7,8	7,4	7,4	7,1	7	7,3									
Ιαν 09	9,4	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3								
Φεβ 09	9,1	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3							
Μαρ 09	9,2	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3						
Απρ 09	9,4	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3					
Μαϊ 09	8,5	8,5	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3				
Ιουν 09	8,6	8,6	8,5	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3			
Ιουλ 09	9,6	9,6	8,6	8,5	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3		
Αυγ 09	9	9	9,6	8,6	8,5	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3	
Σεπ 09	9,1	9,1	9	9,6	8,6	8,5	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7	7,3
Οκτ 09	9,8	9,8	9,1	9	9,6	8,6	8,5	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1	7
Νοε 09	10,6	10,6	9,8	9,1	9	9,6	8,6	8,5	9,4	9,2	9,1	9,4	8,9	7,8	7,4	7,4	7,1
Αυτοδιακύμανση		**1,05**	**0,80**	**0,57**	**0,50**	**0,35**	**0,08**	**0,10**	**0,26**	**0,26**	**0,33**	**0,46**	**0,34**	**0,10**	**0,08**	**0,06**	**-0,05**
Αυτοσυσχέτιση		**1,00**	**0,86**	**0,67**	**0,65**	**0,50**	**0,13**	**0,20**	**0,49**	**0,48**	**0,61**	**0,87**	**0,84**	**0,62**	**0,87**	**0,54**	**-0,63**



Διάγραμμα 34: Αυτοδιακύμανση του ποσοστού ανεργίας



Διάγραμμα 35: Αυτοσυσχέτιση του ποσοστού ανεργίας

Παράδειγμα 2

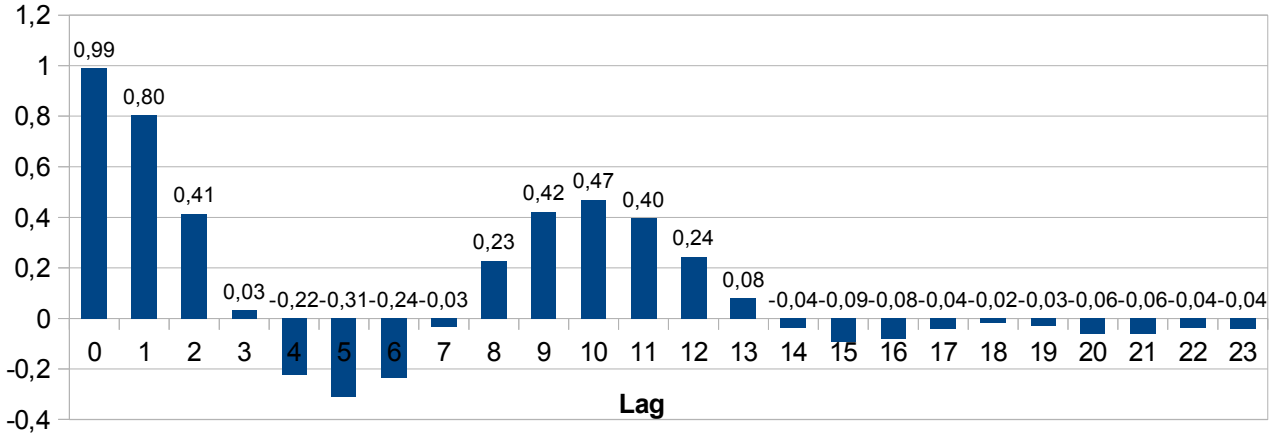
Στον επόμενο πίνακα παρουσιάζεται το πλήθος ηλιακών κηλίδων από το έτος 1771 έως το 1870.

Ηλιακές Κηλίδες									
Έτος	**Πλήθος**	**Έτος**	**Πλήθος**	**Έτος**	**Πλήθος**	**Έτος**	**Πλήθος**	**Έτος**	**Πλήθος**
1771	8	1791	15	1811	47	1831	10	1851	48
1772	17	1792	40	1812	30	1832	24	1852	42
1773	36	1793	62	1813	16	1833	83	1853	28
1774	50	1794	98	1814	7	1834	132	1854	10
1775	64	1795	125	1815	38	1835	131	1855	8
1776	67	1796	96	1816	74	1836	118	1856	2
1777	71	1797	67	1817	101	1837	90	1857	0
1778	48	1798	64	1818	82	1838	67	1858	1
1779	28	1799	54	1819	66	1839	60	1859	5
1780	8	1800	39	1820	35	1840	47	1860	12
1781	13	1801	21	1821	31	1841	41	1861	14
1782	57	1802	7	1822	7	1842	21	1862	35
1783	122	1803	4	1823	20	1843	16	1863	46
1784	138	1804	23	1824	92	1844	6	1864	41
1785	103	1805	55	1825	154	1845	4	1865	30
1786	86	1806	94	1826	126	1846	7	1866	24
1787	65	1807	96	1827	85	1847	14	1867	16
1788	37	1808	77	1828	68	1848	34	1868	7
1789	24	1809	59	1829	38	1849	45	1869	4
1790	11	1810	44	1830	23	1850	43	1870	2

Η χρονική εξέλιξη του πλήθους παρουσιάζεται στο επόμενο διάγραμμα 36. Είναι φανερό πως υπάρχει περιοδικότητα χωρίς ωστόσο να υπάρχει κάποια γενικότερη αυξητική ή φθίνουσα τάση.

Διάγραμμα 36: Πλήθος ηλιακών κηλίδων

Από το διάγραμμα 37 είναι φανερό ότι η αυτοσυσχέτιση έχει ημιτονοειδή μεταβολή κάτι που οφείλεται στην περιοδικότητα του φαινομένου. Επίσης, η περιοδικότητα είναι ίση με 10 με 11 χρόνια καθώς σε αυτές τις διαφορές εμφανίζεται η μεγαλύτερη θετική συσχέτιση.



Διάγραμμα 37: Αυτοσυσχέτιση του πλήθους ηλιακών κηλίδων με χρονική μετατόπιση από 1 έως και 23 έτη

Πίνακας 4.3: Υπολογισμός και γραφική αναπαράσταση κινούμενου μέσου όρου στον υπολογιστή

	Δεν υπάρχει ενσωματωμένη συνάρτηση, μπορεί να υπολογιστεί όπως περιγράφεται αναλυτικά στην προηγούμενη παράγραφο**.**
	**x = c(7.3, 7, 7.1, 7,4, 7.4, 7.8, 8.9, 9.4, 9.1, 9.2, 9.4, 8.5, 8.6, 9.6, 9, 9.1, 9.8, 10.6)** **και** **acf(x, main = "Διάγραμμα αυτοσυσχέτισης", xlab = "Lag (χρονική υστέρηση)", ylab = "Αυτοσυσχέτιση")**

4.8 Φασματική ανάλυση

Με τον όρο φασματική ανάλυση εννοούνται όλες οι μέθοδοι που προσπαθούν να ερμηνεύσουν τη συμπεριφορά της μέσω των συχνοτήτων που την αποτελούν, δηλαδή μέσω κανονικών περιοδικών συνιστωσών. Μαζί με τη χρονική ανάλυση που περιγράφηκε στις προηγούμενες παραγράφους αποτελούν εργαλεία ανάλυσης χρονοσειρών.

Το κύριο εργαλείο ανάλυσης μίας χρονοσειράς σε στοιχειώδες κανονικές περιοδικές συνιστώσες είναι η θεωρία Fourier σε διακριτή και συνεχής μορφή. Σύμφωνα με τη θεωρία Fourier αν X(t) είναι μία συνεχής περιοδική συνάρτηση με περίοδο T τότε

$$X(t)=\frac{\alpha_0}{2}+\sum_{i=1}^{\infty}\alpha_i\,\sigma\upsilon\nu\left(\frac{2\pi\mathrm{i}}{T}t\right)+\beta_i\,\eta\mu\left(\frac{2\pi\mathrm{i}}{T}t\right) \quad ,$$

όπου $\alpha_i = \frac{2}{T}\int_0^T X(t)\,\sigma\upsilon\nu\left(\frac{2\pi\mathrm{i}}{T}\right)dt$, $\beta_i = \frac{2}{T}\int_0^T X(t)\,\eta\mu\left(\frac{2\pi\mathrm{i}}{T}\right)dt$.

Θεωρώντας την χρονοσειρά $x_1, x_2, ..., x_\kappa$ που έχει συλλεχθεί ως διακριτές τιμές μίας συνεχής περιοδικής χρονοσειράς την οποία δεν γνωρίζουμε και αναζητούμε, τότε μπορούμε να βρούμε τη σειρά Fourier που διέρχεται από τις παρατηρήσεις. Αποδεικνύεται ότι στην περίπτωση αυτή :

$$X(t) = A_0 + \sum_{i=1}^{\kappa/2} A_i\,\sigma\upsilon\nu\left(\frac{2\pi\mathrm{i}}{T}t\right) + \sum_{i=1}^{\kappa/2-1} B_i\,\eta\mu\left(\frac{2\pi\mathrm{i}}{T}t\right) \quad ,$$

όπου $A_0=\frac{1}{\kappa}\sum_{i=1}^{\kappa}x_i=\bar{x}$, $A_i=\frac{2}{\kappa}\sum_{q=1}^{\kappa}x_q\,\sigma\upsilon\nu\left(\frac{2\pi\mathrm{iq}}{\kappa}\right)$, $i=1,2,\ldots,\frac{\kappa}{2}-1$, $A_{\frac{\kappa}{2}}=\frac{1}{\kappa}\sum_{q=1}^{\kappa}x_q\,\sigma\upsilon\nu\,q\pi$

και $B_i=\frac{2}{\kappa}\sum_{q=1}^{\kappa}x_q\,\eta\mu\left(\frac{2\pi\mathrm{iq}}{\kappa}\right)$, $i=1,2,\ldots,\frac{\kappa}{2}-1$

Η γενική κατεύθυνση είναι πως οι συνιστώσες που ορίζουν την πορεία της χρονοσειράς έχουν αντίστοιχους συντελεστές στατιστικά διάφορους από το μηδέν. Το θέμα της φασματικής ανάλυσης δεν θα αναπτυχθεί περαιτέρω καθώς ξεφεύγει από τους σκοπούς αυτών των σημειώσεων.

Ασκήσεις

Στον παρακάτω πίνακα δίνονται οι πωλήσεις μίας εταιρείας (σε χιλιάδες ευρώ) τα τελευταία τρία χρόνια.

Έτος	2009				2010				2011			
Τρίμηνο	1	2	3	4	1	2	3	4	1	2	3	4
Πωλήσεις	41	50	54	41	45	60	69	53	55	74	79	63

(i) Να υπολογίσετε κατάλληλους κινητούς μέσους όρους για να προσδιορίσετε την ύπαρξη τάσης στα δεδομένα μας. Στη συνέχεια, να σχεδιάσετε γραφικά τα δεδομένα.

(ii) Να κάνετε το διάγραμμα των αυτοδιακυμάνσεων και των αυτοσυσχετίσεων της χρονοσειράς.

(iii) Να σχολιάσετε τα συμπεράσματα σας.

Κεφάλαιο 5 Δοκιμασία X^2 (Chi Square Test)

Στο κεφάλαιο αυτό περιγράφεται η δοκιμασία X^2. Το Calc διαθέτει τη συνάρτηση

CHITEST() με την οποία υλοποιείται η δοκιμασία χι τετράγωνο του Pearson ή απλά δοκιμασία X^2. Οι περισσότερο συνηθισμένες εφαρμογές της δοκιμασίας αυτής είναι ως έλεγχος ομοιογένειας και ως έλεγχος ανεξαρτησίας.

5.1 Έλεγχος Ομοιογένειας X^2 (Homogeneity Test)

5.1.1 Θεωρητικό υπόβαθρο

Ο έλεγχος ομοιογένειας X^2 εφαρμόζεται σε μία ποιοτική μεταβλητή η οποία έχει k διαφορετικές τιμές και για την οποία επιθυμούμε να ελέγξουμε αν όλες οι τιμές της εμφανίζονται με ίσα ποσοστά στο σύνολο του πληθυσμού. Αν είναι δύσκολο να καταμετρηθούν όλες οι τιμές της μεταβλητής επί του πληθυσμού τότε αρκεί να συλλεχθεί ένα τυχαίο αντιπροσωπευτικό δείγμα του πληθυσμού και να εφαρμοστεί ο έλεγχος αυτός.

Η αρχική (ή μηδενική) υπόθεση στην περίπτωση του ελέγχου αυτού είναι

H_0 : Οι τιμές της μεταβλητής εμφανίζονται με ίσα ποσοστά στον πληθυσμό.

Συμβολικά, αν οι διαφορετικές τιμές της μεταβλητής είναι k σε πλήθος γράφουμε και

$$H_0 : f_1 = f_2 = .. .= f_k$$

Η αντίστοιχη εναλλακτική πρόταση είναι

$$H_1 : \text{όχι η } H_0$$

ή ισοδύναμα

H_1 :Οι τιμές της μεταβλητής δεν εμφανίζονται με ίσα ποσοστά στον πληθυσμό.

Ο έλεγχος ομοιογένειας X^2 βασίζεται στον υπολογισμό της στατιστικής ποσότητας

$$\chi^2 = \sum_{i=1}^{k} \frac{(\Pi_i - E_i)^2}{E_i}$$

όπου k είναι το πλήθος των διαφορετικών τιμών της μεταβλητής, Π_i είναι η παρατηρούμενη συχνότητα της i κατηγορίας επί του δείγματος και E_i είναι η αναμενόμενη συχνότητα της i κατηγορίας επί του δείγματος υπό την προϋπόθεση πως οι δύο μεταβλητές είναι ανεξάρτητες (δηλαδή υπό την προϋπόθεση να ισχύει η H_0).

Οι αναμενόμενες συχνότητες E_i είναι όλες ίσες με τον λόγο του συνόλου των παρατηρήσεων προς το πλήθος των διαφορετικών τιμών, δηλαδή

$$E_i = \frac{Σύνολο\ Παρατηρήσεων}{k}$$

Εικόνα 13: Αναμενόμενες συχνότητες

Προσέξτε πως η μικρότερη τιμή που μπορεί να πάρει το στατιστικό χ^2 είναι το μηδέν και τη λαμβάνει όταν κάθε παρατηρούμενη συχνότητα είναι ίση με κάθε αναμενόμενη. Όσο μεγαλύτερη είναι η διαφορά μεταξύ των $Π_i$ και E_i επί του δείγματος τόσο μεγαλύτερη θα είναι η τιμή του στατιστικού χ^2. Είναι φανερό πως όσο μεγαλύτερη είναι η διαφορά μεταξύ των $Π_i$ και E_i τόσο περισσότερο απίθανο είναι να ισχύει η αρχική υπόθεση, ενώ όσο πιο μικρή είναι αυτή η διαφορά τόσο μεγαλώνει η πιθανότητα να ισχύει!

Το μέγεθος της διαφορετικότητας μεταξύ $Π_i$ και E_i απεικονίζεται στο μέγεθος του στατιστικού χ^2 το οποίο είναι απλά ένας μη αρνητικός πραγματικός αριθμός.

Αποδεικνύεται πως το στατιστικό χ^2 ακολουθεί την X^2 κατανομή με k-1 βαθμούς ελευθερίας (γράφουμε και $\boldsymbol{\chi^2 \sim X^2_{k-1}}$). Για να κρίνουμε αν η τιμή του στατιστικού χ^2 που υπολογίσαμε είναι “μικρή” ή “μεγάλη” αρκεί να συμβουλευτούμε τον πίνακα τιμών της κατανομής αυτής και να βρούμε την πιθανότητα το στατιστικό χ^2 να πάρει τιμές μεγαλύτερες από αυτήν που υπολογίστηκε στο δείγμα μας.

Συνοπτικά, υπολογίζουμε την πιθανότητα

$$\boldsymbol{p = P(\chi^2 > \chi^2_{\text{δείγμα}} \mid \chi^2 \sim X^2_{k-1}).}$$

Η πιθανότητα p είναι το στοιχείο το οποίο θα αξιολογήσουμε για να αποφασίσουμε αν απορρίπτουμε ή όχι την αρχική υπόθεση της ανεξαρτησίας.

Συνήθως το όριο το οποίο λαμβάνεται ως όριο απόρριψής είναι το 0,05 ή 5%, ωστόσο αυτό εξαρτάται από την επιστημονική περιοχή και πολλές φορές λαμβάνεται μεγαλύτερο (0,1ή 10%)

Συνοπτικά, σαν έναν γενικό κανόνα έχουμε τον εξής :

Αν $p<0,05$ τότε η αρχική υπόθεση H_0 απορρίπτεται

Αν η αρχική υπόθεση απορριφθεί τότε συνάγουμε πως οι διαφορετικές τιμές της μεταβλητής δεν εμφανίζονται με τα ίδια ποσοστά στον πληθυσμό. Στην περίπτωση αυτή η παρατήρηση του πίνακα συχνοτήτων διευκρινίζει την εικόνα δείχνοντας τις τιμές που εκπροσωπούνται περισσότερο ή λιγότερο από το αναμενόμενο

Παρατήρηση : Ο στατιστικός έλεγχος που περιγράφουμε ονομάζεται έλεγχος *ομοιογένειας* διότι αντιστοιχεί στην αρχική υπόθεση πως όλες οι αναμενόμενες συχνότητες είναι ίσες, δηλαδή πως η κατανομή είναι *ομοιογενής*. Αν μεταβάλλουμε τον ορισμό των αναμενόμενων συχνοτήτων (Εικόνα 13) μπορούμε να ελέγξουμε οποιαδήποτε αρχική υπόθεση κατανομής των συχνοτήτων ανάμεσα στις διαφορετικές τιμές της μεταβλητής! Θα μπορούσαμε για παράδειγμα να ελέγξουμε την υπόθεση πως το Γαλάζιο εμφανίζεται δύο φορές περισσότερο από το Κίτρινο, τρεις φορές περισσότερο από το Πράσινο και ίσες φορές με το Κόκκινο. Αν θεωρήσουμε ως μονάδα το Γαλάζιο τότε το Κίτρινο θα είναι 1/2, το Πράσινο 1/3, το Κόκκινο 1 με το σύνολο τους να είναι 1+1/2+1/3+1 = 17/6 και τις αντίστοιχες αναμενόμενες συχνότητες στο δείγμα των 22 παρατηρήσεων, 22*6/17 = 7,76 Γαλάζιο, 22*3/17 = 3,89 Κίτρινο, 22*2/17 = 2,59 Πράσινο και 22*6/17 = 7,76 Κόκκινο. Η περαιτέρω ανάλυση προχωρά με ανάλογο τρόπο!

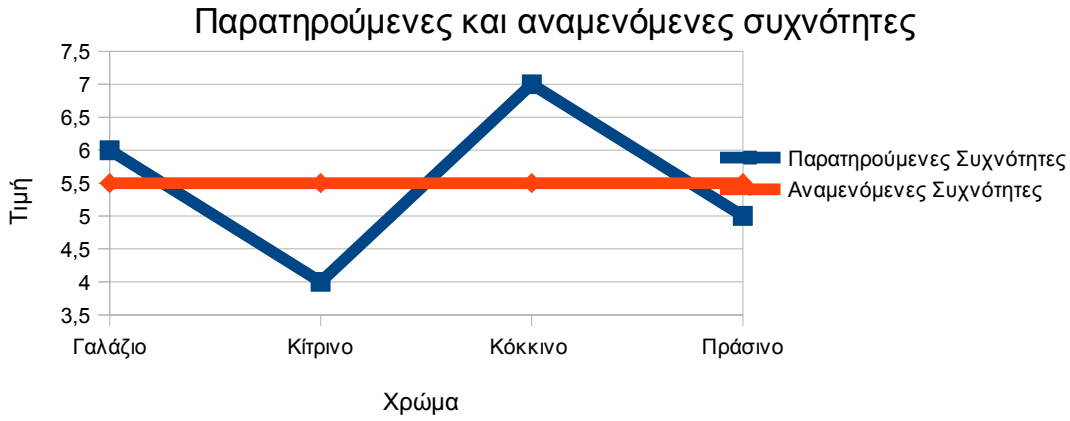
5.1.2 Υλοποίηση της δοκιμασίας στο Calc

Ο έλεγχος ομοιογένειας υλοποιείται με τη βοήθεια του πίνακα της εικόνας 13. Οι αναμενόμενες συχνότητες προκύπτουν διαιρώντας το σύνολο των παρατηρήσεων με το

Συνάρτηση	**COUNTIF(B$4:B$25;D4)**	**E$39/COUNT(E$35:E$38)**	**(E35-F35)^2/F35**
Χρώμα	**Παρατηρούμενες Συχνότητες**	**Αναμενόμενες Συχνότητες**	**(Παρ. – Αναμ.)^2 / Αναμ**
Γαλάζιο	**6**	**5,5**	**0,05**
Κίτρινο	**4**	**5,5**	**0,41**
Κόκκινο	**7**	**5,5**	**0,41**
Πράσινο	**5**	**5,5**	**0,05**
Σύνολο	**22**	**22**	**0,91**

Εικόνα 14: Έλεγχος ομοιογένειας X^2

πλήθος των διαφορετικών τιμών (δηλαδή τέσσερα στην περίπτωση αυτή). Η διαφοροποίηση των αναμενόμενων και των παρατηρούμενων συχνοτήτων είναι δυνατό να παρασταθεί γραφικά με το κατάλληλο διάγραμμα διασποράς (Εικόνα 15)



Εικόνα 15: Αναπαράσταση διαφοράς μεταξύ παρατηρούμενων και αναμενόμενων συχνοτήτων

Μετά, αν και δεν είναι απαραίτητο, είναι καλό για την ολοκληρωμένη παρουσίαση του αποτελέσματος να υπολογίσουμε την τιμή του στατιστικού χ^2 το οποίο υπολογίζεται ως άθροισμα της τελευταίας στήλης του πίνακα της εικόνας 13.

Συνάρτηση	**CHITEST(E35:E38;F35:F38)**
Τιμή p	**0,82**

Εικόνα 16: Υπολογισμός της πιθανότητας p

Η ζητούμενη πιθανότητα p υπολογίζεται από το στατιστικό χ^2, τους βαθμούς ελευθερίας και την συνάρτηση CHIDIST() του Calc. Υπολογίζουμε

$$\mathbf{p = P(\chi^2 > 0,91 \mid \chi^2 \sim X^2_3) = CHIDIST(0,91;3) = 0,82}$$

Εναλλακτικά, μπορούμε να υπολογίζεται άμεσα την πιθανότητα p με τη συνάρτηση **CHITEST()** δίνοντας ως πρώτο όρισμα τις παρατηρούμενες συχνότητες και ως δεύτερο όρισμα τις αναμενόμενες συχνότητες. Το αποτέλεσμα φυσικά προκύπτει το ίδιο! (Εικόνα 14)

Συμπέρασμα ελέγχου ομοιογένειας χι τετράγωνο
Η τιμή της πιθανότητας (0,82=82%) είναι μεγαλύτερη από τη μικρότερη τιμή αποδοχής (0,05=5%) άρα η υπόθεση H_0 δεν απορρίπτεται. Με απλά λόγια μπορούμε να δεχθούμε πως τα χρώματα Γαλάζιο, Κίτρινο, Κόκκινο και Πράσινο εμφανίζονται με ίδια ποσοστά στο γενικό πληθυσμό από τον οποίον προήλθε το δείγμα των 22 παρατηρήσεων.

5.1.3 Προϋποθέσεις εφαρμογής της δοκιμασίας X^2 ως έλεγχος ομοιογένειας

Η δοκιμασία X^2 είναι μη παραμετρική δηλαδή δεν προϋποθέτει πως το δείγμα των τιμών της τυχαίας μεταβλητής που μελετούμε προέρχεται από πληθυσμό στον οποίο η μεταβλητή ακολουθεί κάποια συγκεκριμένη κατανομή η οποία προσδιορίζεται από κάποιες παραμέτρους θέσης και διασποράς.

Είναι καλό να υπάρχει αρκετά μεγάλο πλήθος παρατηρήσεων. Μικρό πλήθος παρατηρήσεων ισοδυναμεί με μεγάλο σφάλμα Τύπου ΙΙ δηλαδή υπάρχει σημαντική πιθανότητα να μην απορριφθεί η αρχική υπόθεση H_0 ενώ θα έπρεπε να απορριφθεί!

Μία προϋπόθεση η οποία πρέπει σύμφωνα με τη βιβλιογραφία να τηρείται στη δοκιμασία αυτή είναι οι αναμενόμενες συχνότητες να είναι σε όλες τις περιπτώσεις μεγαλύτερες από τη μονάδα. Επιπλέον, όχι περισσότερες από το ένα πέμπτο (20%) από τις αναμενόμενες συχνότητες δεν πρέπει να είναι μικρότερες από 5.

Αν οι παραπάνω προϋποθέσεις δεν τηρούνται τότε το αποτέλεσμα της δοκιμασίας X^2 δεν θεωρείται ιδιαίτερα έγκυρο και είναι καλό αν είναι δυνατόν να επιβεβαιωθεί με δεύτερη ανεξάρτητη δειγματοληψία.

Ευκόλως εννοούμενο είναι βέβαια πως για να έχει νόημα η εφαρμογή της δοκιμασίας πρέπει το δείγμα των παρατηρήσεων να είναι αντιπροσωπευτικό του πληθυσμού.

Πίνακας 5.1: Δοκιμασία X^2 ως δοκιμασία ομοιογένειας

	Δεν υπάρχει δεσμευμένη συνάρτηση στο Calc. Ο έλεγχος μπορεί να υλοποιηθεί με τον τρόπο που περιγράφεται στην παράγραφο 5.1.2
	Δεν υπάρχει δεσμευμένη συνάρτηση στο R – Project. Ο έλεγχος μπορεί να υλοποιηθεί με τις παρακάτω εντολές : **x = c(6, 4, 7, 5)** **e = sum(x) / length(x)** **v = sum((x – e)^2 / e)** **pchisq(v, df=3, lower.tail=FALSE)** Το τελικό αποτέλεσμα ταυτίζεται με αυτό του Calc

5.2 Έλεγχος Ανεξαρτησίας X^2 (Independent Test)

5.2.1 Θεωρητικό υπόβαθρο

Ο έλεγχος ανεξαρτησίας X^2 χρησιμοποιείται όταν θέλουμε να επιβεβαιώσουμε ή να

αναιρέσουμε την υποψία πως δύο ποιοτικές ή διατακτικές μεταβλητές είναι εξαρτημένες, δηλαδή πως η τιμή που θα έχει η μία μεταβλητή εξαρτάται από την τιμή που θα έχει η άλλη.

Η αρχική (ή μηδενική) υπόθεση στην περίπτωση του ελέγχου αυτού είναι

H_0 : Οι δύο μεταβλητές είναι ανεξάρτητες.

Η αντίστοιχη εναλλακτική πρόταση είναι

H_1 : όχι η H_0,

ή ισοδύναμα

H_1 : Οι δύο μεταβλητές είναι εξαρτημένες.

Η δοκιμασία αυτή είναι συνήθως το πρώτο βήμα στη μελέτη ενός ζεύγους ποιοτικών μεταβλητών. Προσέξτε πως αν τελικά απορριφθεί η αρχική υπόθεση τόσο συνάγουμε μία αφηρημένη εξάρτηση μεταξύ των μεταβλητών η οποία πρέπει να διευκρινιστεί με περαιτέρω έρευνα ή/και θεωρητική ερμηνεία η οποία θα εξαρτάται από την επιστημονική περιοχή στην οποία ορίζονται οι μεταβλητές!

Ο έλεγχος ανεξαρτησίας X^2 βασίζεται στον υπολογισμό της στατιστικής ποσότητας

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(\Pi_{i,j} - E_{i,j})^2}{E_{i,j}}$$

όπου *k* είναι το πλήθος των διαφορετικών τιμών της μίας μεταβλητής, *l* είναι το πλήθος των διαφορετικών τιμών της άλλης μεταβλητής, $\Pi_{i,j}$ είναι η παρατηρούμενη συχνότητα της i κατηγορίας της μίας μεταβλητής και της j κατηγορίας της άλλης (επί του δείγματος) και $E_{i,j}$ είναι η αναμενόμενη συχνότητα της i κατηγορίας της μίας μεταβλητής και της j κατηγορίας της άλλης (επί του δείγματος) υπό την προϋπόθεση πως οι δύο μεταβλητές είναι ανεξάρτητες (δηλαδή υπό την προϋπόθεση να ισχύει η H_0).

Για τον υπολογισμό των αναμενόμενων συχνοτήτων $E_{i,j}$ ορίζουμε τα γεγονότα

A={Η μεταβλητή X_1 παίρνει την τιμή i}, B={Η μεταβλητή X_2 παίρνει την τιμή j}.

Υπολογίζουμε,

P(Το ζεύγος των μεταβλητών (X_1, X_2) έχει την τιμή (i,j)) = P(A · B) =

$$P(A)\cdot P(B) = \frac{\Sigma\acute{\upsilon}νολο\ περιπτώσεων\ i}{\Sigma\acute{\upsilon}νολο\ Τιμών} \cdot \frac{\Sigma\acute{\upsilon}νολο\ περιπτώσεων\ j}{\Sigma\acute{\upsilon}νολο\ Τιμών}$$

και

$$E_{i,j} = P(\text{Το ζεύγος των μεταβλητών } (X_1, X_2) \text{ έχει την τιμή } (i,j)\ ) \cdot \Sigma\acute{\upsilon}νολο\ Τιμών =$$

$$\frac{\Sigma\acute{\upsilon}νολο\ περιπτώσεων\ i \cdot \Sigma\acute{\upsilon}νολο\ περιπτώσεων\ j}{\Sigma\acute{\upsilon}νολο\ Τιμών} = \frac{Άθροισμα\ γραμμής\ i \cdot Άθροισμα\ στήλης\ j}{\Sigma\acute{\upsilon}νολο\ Τιμών}$$

Η μικρότερη τιμή που μπορεί να πάρει το στατιστικό χ^2 είναι το μηδέν. Όσο μεγαλύτερη είναι η διαφορά μεταξύ των $\Pi_{i,j}$ και $E_{i,j}$ επί του δείγματος τόσο μεγαλύτερη θα είναι η τιμή του στατιστικού χ^2. Είναι επίσης φανερό πως όσο μεγαλύτερη είναι η διαφορά μεταξύ των $\Pi_{i,j}$ και $E_{i,j}$ τόσο περισσότερο απίθανο είναι να ισχύει η αρχική υπόθεση, ενώ όσο πιο μικρή είναι αυτή η διαφορά τόσο μεγαλώνει η πιθανότητα να ισχύει!

Το μέγεθος της διαφορετικότητας μεταξύ $\Pi_{i,j}$ και $E_{i,j}$ απεικονίζεται στο μέγεθος του στατιστικού χ^2 το οποίο είναι ένας πραγματικός αριθμός. Το στατιστικό χ^2 αποδεικνύεται πως ακολουθεί την X^2 κατανομή με (k-1)(l-1) βαθμούς ελευθερίας (γράφουμε και $\boldsymbol{\chi^2 \sim X^2_{(k-1)(l-1)}}$). Αυτό σημαίνει πως για να κρίνουμε αν η τιμή του στατιστικού χ^2 που υπολογίσαμε είναι “μικρή” ή “μεγάλη” αρκεί να συμβουλευτούμε τον πίνακα τιμών της κατανομής αυτής και να βρούμε την πιθανότητα το στατιστικό χ^2 να πάρει τιμές μεγαλύτερες από αυτήν που υπολογίστηκε στο δείγμα μας.

Συνοπτικά, υπολογίζουμε την πιθανότητα

$$\mathbf{p = P(\chi^2 > \chi^2_{\text{δείγμα}} \mid \chi^2 \sim X^2_{(k-1)(l-1)}).}$$

Η πιθανότητα p είναι το στοιχείο το οποίο θα αξιολογήσουμε για να αποφασίσουμε αν απορρίπτουμε ή όχι την αρχική υπόθεση της ανεξαρτησίας των δύο μεταβλητών.

Συνήθως το όριο το οποίο λαμβάνεται ως όριο απόρριψης είναι το 0,05 ή 5% ωστόσο αυτός ο κανόνας δεν είναι αυστηρός ενώ το όριο αποχής εξαρτάται από την επιστημονική περιοχή και πολλές φορές λαμβάνεται μεγαλύτερο.

Συνοπτικά, σαν έναν γενικό κανόνα έχουμε τον εξής :

Αν $p<0,05$ τότε η αρχική υπόθεση H_0 απορρίπτεται

Αν η αρχική υπόθεση απορριφθεί τότε συνάγουμε πως το είδος της τιμής που θα έχει η μία μεταβλητή “σχετίζεται” με το είδος της τιμής που θα έχει η άλλη μεταβλητή, χωρίς να

ξεκαθαρίζει το είδος της σχέσης που θα υπάρχει. Στην περίπτωση αυτή η παρατήρηση του πίνακα συχνοτήτων βοηθάει στη διευκρίνηση της εικόνας.

5.2.2 Βασικά βήματα στο Calc.

Για την εφαρμογή στο Calc της δοκιμασίας X^2 ως έλεγχο ανεξαρτησίας απαιτείται η συμπλήρωση ενός διμεταβλητού πίνακα συχνοτήτων από τις συνδυαστικές παρατηρήσεις που έχουμε συλλέξει. (Εικόνα 17)

Επιπλέον, απαιτείται η συμπλήρωση ενός διμεταβλητού πίνακα ίδιας διάστασης στον οποίο θα καταχωρηθούν οι αναμενόμενες συχνότητες (Εικόνα 18). Η συμπλήρωση του πίνακα με τις αναμενόμενες συχνότητες περιγράφεται παρακάτω.

Χρησιμοποιώντας τα δεδομένα των δύο πινάκων μπορούμε να υπολογίσουμε την τιμή του στατιστικού χ^2 (Εικόνα 19) κάτι που δεν είναι απαραίτητο αλλά είναι χρήσιμο για την ολοκληρωμένη παρουσίαση των αποτελεσμάτων της δοκιμασίας.

Τέλος, χρησιμοποιώντας την κατάλληλη συνάρτηση του Calc υπολογίζουμε την πιθανότητα p (Εικόνα 21) την οποία αξιολογούμε και ερμηνεύουμε, συνήθως συγκρίνοντας της με το 0,05=5% και ανάλογα αποφασίζουμε την απόρριψη ή μη της στατιστικής υπόθεσης.

5.2.3 Παράδειγμα υλοποίησης της δοκιμασίας στο Calc.

Καταγράφηκε το χρώμα ματιών και το φύλο από 60 σπουδαστές και οι παρατηρήσεις συγκεντρώθηκαν στο διμεταβλητό πίνακα συχνοτήτων της εικόνας 17. Οι παρατηρήσεις είναι εξήντα σε πλήθος ώστε να ικανοποιείται η πρώτη προϋπόθεση εφαρμογής της δοκιμασίας (αναμενόμενες συχνότητες >=5, υποπαράγραφος 5.2.4).

Διμεταβλητός Πίνακας Συχνοτήτων						
		Χρώμα Ματιών				
		Καστανά	**Μαύρα**	**Μπλε**	**Πράσινα**	**Σύνολο**
Φύλο	**Αγόρι**	**18**	**6**	**6**	**6**	**36**
	Κορίτσι	**12**	**6**	**6**	**0**	**24**
	Σύνολο	**30**	**12**	**12**	**6**	**60**

Εικόνα 17: Τα αρχικά συνδυαστικά δεδομένα των δύο ποιοτικών μεταβλητών

Ο επόμενος πίνακας (Εικόνα 18) περιέχει τις αναμενόμενες συχνότητες κάθε κελιού. Η αναμενόμενη συχνότητα κάθε κελιού υπό την υπόθεση πως ισχύει η υπόθεση H_0 της ανεξαρτησίας προκύπτει από τον τύπο

$$Αναμενόμενη\ Συχνότητα\ =\ \frac{Άθροισμα\ Γραμμής \cdot Άθροισμα\ Στήλης}{Συνολικό\ Άθροισμα}$$

Στην εικόνα 16 παρουσιάζονται γραφικά και συνοπτικά οι διαφορές των αναμενόμενων και των παρατηρήσιμων συχνοτήτων για τα αγόρια και τα κορίτσια. Το επόμενο βήμα το οποίο είναι μη αναγκαίο αλλά επιθυμητό για την ολοκληρωμένη παρουσίαση της δοκιμασίας είναι ο υπολογισμός του στατιστικού χ^2 ο οποίος παρουσιάζεται στον πίνακα της εικόνας 19.

Η καταχώρηση κάθε κελιού προκύπτει από τον τύπο

$$\frac{(Παρατηρούμενη\ Συχνότητα\ του\ κελιού\ -\ Αναμενόμενη\ Συχνότητα\ του\ κελιού)^2}{Αναμενόμενη\ Συχνότητα\ του\ κελιού},$$

ενώ αθροίζοντας το σύνολο όλων των στοιχείων προκύπτει η ζητούμενη τιμή.

Τέλος στον πίνακα της εικόνας 21 παρουσιάζεται η πιθανότητα p εμφάνισης τιμής στο στατιστικό χ^2 μεγαλύτερη από αυτήν που υπολογίστηκε στο δείγμα μας. Η p υπολογίζεται με χρήση της συνάρτησης CHITEST() ωστόσο, έχοντας υπολογίσει την τιμή του στατιστικού χ^2 η οποία προκύπτει να είναι ίση με 5 (Εικόνα 19) το ίδιο αποτέλεσμα θα μπορούσε να προκύψει αν χρησιμοποιούσαμε τη συνάρτηση του Calc CHIDIST(5;3) (5 είναι η τιμή του στατιστικού χ^2 και 3 = (4-1)(2-1) οι βαθμοί ελευθερίας) η οποία επιστρέφει την τιμή της κατανομής X^2 στο σημείο 5 για 3 βαθμούς ελευθερίας. Μία δοκιμή μπορεί να σας πείσει!

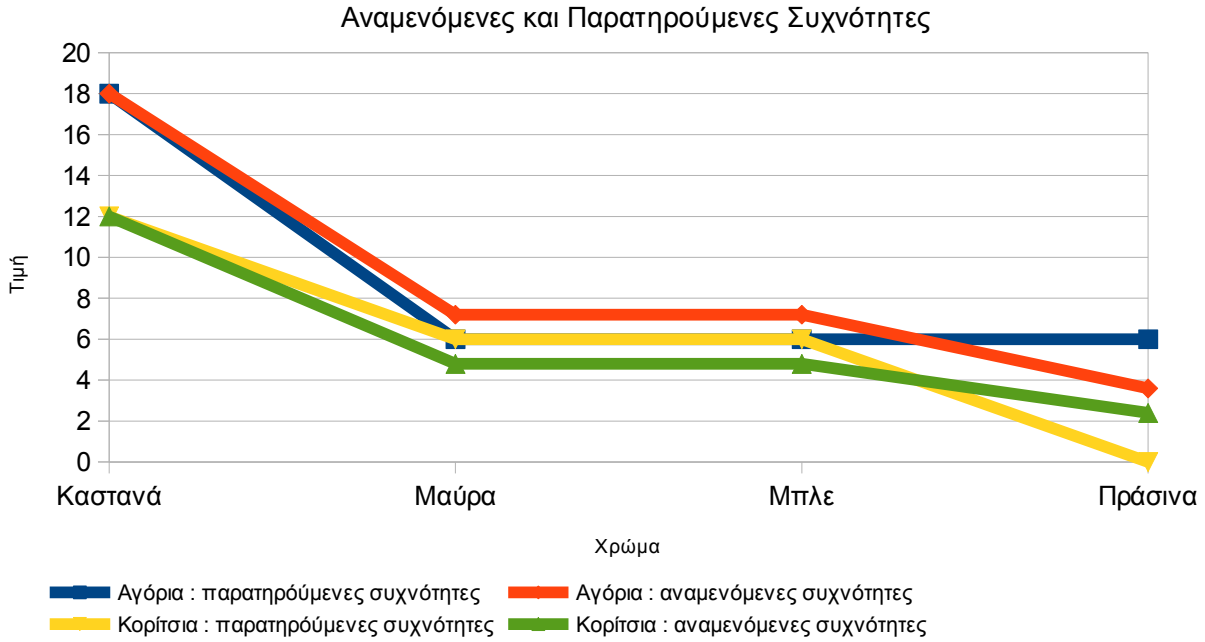
Υπολογίζουμε **p = 0,17 = 17%** κάτι που σημαίνει πως δεν ήταν ιδιαίτερα απίθανη η εμφάνιση των διαφορετικότητας μεταξύ των παρατηρήσιμων και των αναμενόμενων συχνοτήτων μεταξύ των δύο πινάκων. Συνήθως, το όριο βάσει του οποίου κρίνουμε πως ήταν ιδιαίτερα απίθανη η εμφάνιση των διαφορών αυτών και ως εκ τούτου συνάγουμε την εξάρτηση των μεταβλητών, είναι το 0,05=5%.

Αναμενόμενες Συχνότητες υπό την υπόθεση πως οι δύο μεταβλητές είναι ανεξάρτητες						
		Χρώμα Ματιών				
		Καστανά	Μαύρα	Μπλε	Πράσινα	Σύνολο
Φύλο	Αγόρι	18	7,2	7,2	3,6	36
	Κορίτσι	12	4,8	4,8	2,4	24
	Σύνολο	30	12	12	6	60
Κάθε καταχώρηση προκύπτει ως : Άθροισμα Γραμμής * Άθροισμα Στήλης / Συνολικό Άθροισμα						
π.χ. Η αναμενόμενη συχνότητα των Αγοριών με Καστανά μάτια προκύπτει ως 30*36/60 κτλπ						

Εικόνα 18: Ο πίνακας αναμενόμενων συχνοτήτων, απαραίτητος για την εφαρμογή της δοκιμασίας X^2 στο Calc

Διμεταβλητός Πίνακας Συχνοτήτων						
		Χρώμα Ματιών				
		Καστανά	Μαύρα	Μπλε	Πράσινα	Σύνολο
Φύλο	Αγόρι	0,00	0,20	0,20	1,60	2,00
	Κορίτσι	0,00	0,30	0,30	2,40	3,00
	Σύνολο	0,00	0,50	0,50	4,00	5,00

Εικόνα 19: Υπολογισμός της τιμής του στατιστικού χ^2 (μη αναγκαίος υπολογισμός αλλά επιθυμητός για την ολοκληρωμένη παρουσίαση της δοκιμασίας)



Εικόνα 20: Γραφική αναπαράσταση των διαφορών μεταξύ παρατηρήσιμων και αναμενόμενων συχνοτήτων

p	Συνάρτηση
0,17	**CHITEST(N7:Q8;N16:Q17)**

Εικόνα 21: Η πιθανότητα p και η συνάρτηση από την οποία υπολογίστηκε

Ο έλεγχος ανεξαρτησίας Χι τετράγωνο συνοπτικά
Η εξάρτηση των μεταβλητών ανιχνεύεται μέσω της διαφορετικότητας των παρατηρούμενων από τις αναμενόμενες συχνότητες. Όσο μεγαλύτερες είναι οι διαφορές αυτές, τόσο μεγαλύτερη θα είναι η τιμή του στατιστικού χι τετράγωνο
Η πιθανότητα να ισχύει η στατιστική υπόθεση (δηλαδή να είναι ανεξάρτητες οι μεταβλητές) και να πάρει το στατιστικό χι τετράγωνο την τιμή που υπολογίσαμε είναι 0,17 ή 17%.
Τελικό Συμπέρασμα : Η στατιστική υπόθεση δεν απορρίπτεται. Πιο συγκεκριμένα, οι διαφορετικότητα μεταξύ των δύο πινάκων δεν είναι τόσο μεγάλη ώστε να συνάγουμε πως το χρώμα ματιών έχει εξάρτηση με το φύλο.

Πίνακας 5.2: Δοκιμασία X^2 ως έλεγχος ανεξαρτησίας

Δεν υπάρχει δεσμευμένη συνάρτηση στο Calc. Ο έλεγχος μπορεί να υλοποιηθεί με τον τρόπο που περιγράφεται στην παράγραφο 5.2.3.

Ο παρακάτω κώδικας υλοποιεί τη δοκιμασία ανεξαρτησίας του παραδείγματος.

```
.Table <- matrix(c(18, 6, 6, 6, 12, 6, 6, 0), 2, 4, byrow=TRUE)
rownames(.Table) <- c('Αγόρι', 'Κορίτσι')
colnames(.Table) <- c('Καστανά', 'Μαύρα', 'Μπλε', 'Πράσινα')
.Table
chisq.test(.Table, correct=FALSE)
remove(.Table)
```

5.2.4 Προϋποθέσεις εφαρμογής της δοκιμασίας X^2 ως έλεγχος ανεξαρτησίας

Απαιτούνται οι προϋποθέσεις της παραγράφου 5.1.3 οι οποίες αναφέρονται στη δοκιμασία X^2 ως έλεγχο ομοιογένειας. Επιπλέον, οι αποκλίσεις των συχνοτήτων (δηλαδή οι διαφορές $\Pi_{i,j}$-$E_{i,j}$) πρέπει να ακολουθούν την κανονική κατανομή. Αυτό βέβαια δεν ισχύει για τις ίδιες τις συχνότητες, είτε τις αναμενόμενες, είτε τις παρατηρήσιμες.

Ένας πρόσθετος κανόνας που προτείνεται στη βιβλιογραφία όταν και οι δύο ποιοτικές μεταβλητές έχουν από δύο τιμές (δηλαδή ο πίνακας είναι 2x2) είναι οι τρεις από τις τέσσερις αναμενόμενες τιμές να είναι μεγαλύτερες από 10. (Στην περίπτωση που οι μεταβλητές έχουν από δύο τιμές μπορεί να υπολογιστεί επικουρικά και ο συντελεστής Φ του Cramer.) Τέλος, δεν πρέπει να υπάρχουν κελιά με μηδενική τιμή.

Αν υπάρχει μεγάλη απόκλιση από τις παραπάνω προϋποθέσεις τότε είναι καλό το αποτέλεσμα του ελέγχου να μην λαμβάνεται υπόψη ή να επαληθεύεται με δεύτερη δειγματοληψία αν αυτό είναι δυνατόν.

5.3 Δοκιμασία Fisher

Η δοκιμασία Fisher (Fisher's exact test) εφαρμόζεται για τον έλεγχο της ανεξαρτησίας δύο ποιοτικών δίτιμων μεταβλητών (δηλαδή μεταβλητών οι οποίες έχουν μόνο δύο τιμές όπως το φύλο ή η επιτυχία στις εξετάσεις κ.α.). Μπορεί να εφαρμοστεί στη θέση της δοκιμασίας X^2 όταν το δείγμα των παρατηρήσεων είναι μικρό κάτι που στην πράξη σημαίνει πως η δοκιμασία X^2 έχει μεγάλο περιθώριο σφάλματος. Ένας απλός κανόνας που επιβάλει την εφαρμογή της δοκιμασίας Fisher είναι η αναμενόμενη συχνότητα σε οποιοδήποτε κελί να είναι μικρότερη από δέκα. Φυσικά, αυτό δεν απαγορεύει την εφαρμογή της μεθόδου όταν όλες οι αναμενόμενες τιμές είναι μεγαλύτερες από δέκα και μπορεί να χρησιμοποιηθεί

παράλληλα με την δοκιμασία X^2 για επαλήθευση των αποτελεσμάτων.

Στο παράδειγμα που ακολουθεί θέλουμε να βρούμε αν υπάρχει συσχέτιση μεταξύ του φύλου των εφήβων και της δίαιτας. Ρωτήθηκαν σχετικά 24 έφηβοι, 12 αγόρια και 12 κορίτσια για το αν τη στιγμή της ερώτησης βρισκόταν σε περίοδο δίαιτας. Τα αποτελέσματα εμφανίζονται στον Πίνακα 5.3

	Αγόρι	**Κορίτσι**	**Σύνολο**
Δίαιτα	**1**	**9**	**10**
Όχι δίαιτα	**11**	**3**	**14**
Σύνολο	**12**	**12**	**24**

Πίνακας 5.3: Δεδομένα δοκιμασίας Fisher

Εύκολα συμπεραίνουμε πως τα δεδομένα του παραδείγματος δεν είναι κατάλληλα για τη δοκιμασία X^2 γιατί οι αναμενόμενες συχνότητες είναι μικρότερες από δέκα (π.χ. για τα αγόρια που βρίσκονται σε δίαιτα θα περιμέναμε συχνότητα 10*12/24 = 5 <10). Άρα θα εφαρμόσουμε τη δοκιμασία Fisher.

Η στατιστική υπόθεση που επιθυμούμε να ελέγξουμε είναι η

H_0 : Τα αγόρια και τα κορίτσια είναι το ίδιο πιθανό να βρίσκονται σε δίαιτα,

έναντι της εναλλακτικής

H_1 : Όχι η H_0.

Το ερώτημα που απαντά η δοκιμασία Fisher είναι το εξής : Γνωρίζοντας πως 10 από τους 24 μαθητές οι 10 βρίσκονται σε δίαιτα και οι 14 δεν βρίσκονται σε δίαιτα, ποια είναι η πιθανότητα να είναι τόσο άνισα κατανεμημένοι οι 10 και οι 14 μαθητές μεταξύ αγοριών και κοριτσιών; Ισοδύναμα, αν επιλέγαμε 10 μαθητές τυχαία από το δείγμα ποια η πιθανότητα οι εννέα να είναι μέσα από τα 12 κορίτσια και μόνο ένας να είναι μέσα από τα 12 αγόρια;

Στη γενική περίπτωση όπου οι καταχωρήσεις του πίνακα είναι οποιεσδήποτε συχνότητες a, b, c και d, ο Πίνακας 4.1 συμπληρώνεται ως εξής :

	Αγόρι	**Κορίτσι**	**Σύνολο**
Δίαιτα	**a**	**b**	**a+b**
Όχι δίαιτα	**c**	**d**	**c+d**
Σύνολο	**a+c**	**b+d**	**n**

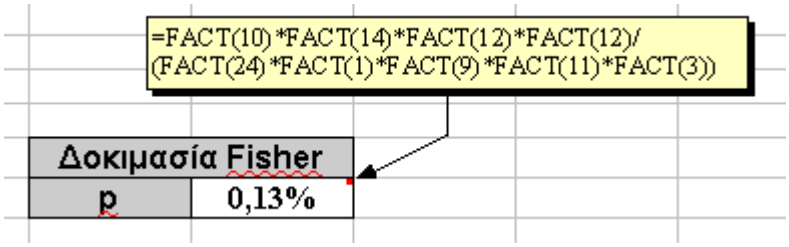
Ο Fisher έδειξε ότι η πιθανότητα να εμφανιστεί η ζητούμενη κατανομή συχνοτήτων ισούται

με

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} \quad ,$$

όπου το σύμβολο του θαυμαστικού (!) είναι ο τελεστής παραγοντικό δηλαδή $n! = 1 \cdot 2 \cdot \ldots \cdot n$ και ανάλογα για τις υπόλοιπες συχνότητες.

Ο υπολογισμός στα δεδομένα του παραδείγματος δίνει p = 0,0013 ή 0,13% (Εικόνα 15) άρα μπορούμε με μεγάλη ασφάλεια να απορρίψουμε την στατιστική υπόθεση και να συμπεράνουμε πως δεν είναι ίδιο το ποσοστό των αγοριών και των κοριτσιών που βρίσκονται σε δίαιτα.



Εικόνα 22: Δοκιμασία Fisher : Υπολογισμός της πιθανοφάνειας της παρατηρούμενης κατανομής συχνοτήτων

Πίνακας 5.4: Δοκιμασία Fisher

Δεν υπάρχει δεσμευμένη συνάρτηση στο Calc. Ο έλεγχος μπορεί να υλοποιηθεί με τον τρόπο που περιγράφεται στην παράγραφο 5.3.

```
.Table <- matrix(c(1, 9, 11, 3), 2, 2, byrow=TRUE)
rownames(.Table) <- c('Δίαιτα', 'Όχι δίαιτα')
colnames(.Table) <- c('Αγόρι', 'Κορίτσι')
fisher.test(.Table)
```

Παρατήρηση : Για τα δεδομένα του πίνακα 5.3 θα προκύψει p = 0,037, διαφορετικό από το 0,013 που υπολογίζει το Calc. Ο λόγος είναι πως το το R – Project χρησιμοποιεί εκτιμητή μέγιστης πιθανοφάνειας (maximum likehood estimator) για να υπολογιίσει το p. Με ?fisher.test παρέχονται περισσότερες πληροφορίες.

5.4 Παρουσίαση της δοκιμασίας X^2

Μία τυπική παρουσίαση της δοκιμασία X^2 ως τεστ ανεξαρτησίας είναι

“Για την ανίχνευση της εξάρτησης του φύλου και του χρώματος ματιών εφαρμόστηκε η δοκιμασία X^2. Η δοκιμασία απέτυχε να καταδείξει στατιστικά σημαντική διαφοροποίηση ($\chi^2(3, N = 60) = 5.00$, $p = .16$.”

Ανάλογα η κατάδειξη μίας εξάρτησης μπορεί να περιγραφεί ως

“Η δοκιμασία X^2 με τη διόρθωση κατά Yates, κατέδειξε πως το ποσοστό των μαθητών που κάνουν δίαιτα διαφέρει σημαντικά ανάμεσα στα δύο φύλα ($\chi^2(1, N = 24) = 8.4$, $p = 0.003$, ο λόγος πιθανοτήτων (odd ratio) είναι ίσος με 33.0).”

Το αποτέλεσμα της δοκιμασίας Fisher, παρουσιάζεται ανάλογα :

“Η δοκιμασία Fisher κατέδειξε στατιστική εξάρτηση του φύλου και της εφαρμογής δίαιτας $p = .037$”.

Επιπρόσθετα μπορεί να εμφανιστεί και το διάστημα εμπιστοσύνης, χωρίς ωστόσο να είναι απαραίτητο.

Κεφάλαιο 6 Έλεγχος ισότητας μέσης τιμής

Στο κεφάλαιο αυτό παρουσιάζονται οι στατιστικές δοκιμασίες που αφορούν τη μέση τιμή από έναν ή δύο πληθυσμούς.

Υπάρχουν δύο κατηγορίες δοκιμασιών για τον έλεγχο μέσων τιμών, οι παραμετρικές και οι μη παραμετρικές. Οι παραμετρικές δοκιμασίες είναι οι πλέον συνηθισμένες και χρησιμοποιούνται όταν έχουμε δείγμα μεγάλου πλήθους (≥ 30) είτε όταν το δείγμα των τιμών που επεξεργαζόμαστε είναι μικρό (<30) αλλά γνωρίζουμε ή μπορούμε να δείξουμε πως

1) (Ένα δείγμα – One Sample T Test) Το δείγμα τιμών μας προέρχεται από πληθυσμό ο οποίος κατανέμεται κανονικά.
2) (Δύο ανεξάρτητα δείγματα – Independent Samples T Test) Τα δύο ανεξάρτητα δείγματα μας προέρχονται από δύο ανεξάρτητους πληθυσμούς οι οποίοι κατανέμονται κανονικά. (Ωστόσο, δείτε και την παράγραφο των περιορισμών της μεθόδου!)
3) (Ζευγαρωτές παρατηρήσεις – Paired Samples T Test) Οι διαφορές των ζευγαρωτών παρατηρήσεων προέρχονται από πληθυσμό ο οποίος ακολουθεί την κανονική κατανομή.

Ο έλεγχος της κανονικότητας των τιμών γίνεται είτε παρατηρώντας το ιστόγραμμα των τιμών της μεταβλητής είτε το διάγραμμα p-p είτε το διάγραμμα q-q.

Επίσης, είναι απαραίτητο οι τιμές της τυχαίας μεταβλητής να είναι αριθμοί ενώ επιπλέον πρέπει να είναι δυνατή η ερμηνεία ενός σταθερού διαστήματος σε οποιοδήποτε σημείο της κλίμακας. Για παράδειγμα η απάντηση στην ερώτηση “Πόσες ημέρες την εβδομάδα συμβαίνει το γεγονός Χ” είναι αποδεκτή ως τυχαία μεταβλητή παραμετρικών ελέγχων καθώς η διαφορά μεταξύ των τιμών 2 και 3 είναι ακριβώς η ίδια με τη διαφορά μεταξύ των τιμών 5 και 6 (δηλαδή μία ημέρα και στις δύο περιπτώσεις) ενώ η απάντηση στην ερώτηση “Βαθμολογήστε από 1 έως 5 το άγχος που νιώθετε όταν βλέπετε την εικόνα Υ” δεν είναι αποδεκτή γιατί η διαφορά μεταξύ των τιμών 2 και 3 δεν μπορεί να συγκριθεί με τη διαφορά μεταξύ 3 και 4 καθώς δεν υπάρχει μία κοινά αποδεκτή κλίμακα άγχους!

6.1 Παραμετρικές στατιστικές δοκιμασίες

Παραμετρικές ονομάζονται οι στατιστικές δοκιμασίες οι οποίες προϋποθέτουν πως η κρίσιμη στατιστική ποσότητα προέρχεται από κάποιον πληθυσμό ο οποίος ακολουθεί την

κανονική κατανομή.

Η κανονικότητα της κατανομής της στατιστικού μπορεί να τεκμηριωθεί με τους παρακάτω τρόπους

1. Με απλή παραδοχή του γεγονότος αυτού από προηγούμενη ανάλυση της ίδιας μεταβλητής σε άλλες εργασίες
2. Με ποιοτική (μη αριθμητική) ανάλυση της μεταβλητής στην οποία θα τεκμηριώνεται η ύπαρξη πολλών ανεξάρτητων πηγών σφάλματος στη μέτρηση της τιμής της μεταβλητής είτε
3. Με γραφικό τρόπο παραθέτοντας το ιστόγραμμα των τιμών του δείγματος είτε ένα από τα διαγράμματα p-p και q-q.

Είναι φανερό πως όσο περισσότερα από τα παραπάνω στοιχεία παρουσιαστούν σε μία έρευνα τόσο περισσότερο αποδεκτή θα κριθεί η εφαρμογή της μεθόδου άρα και τόσο περισσότερο αξιόπιστο το τελικό συμπέρασμα!

Στη βιβλιογραφία μπορούν να βρεθούν στατιστικοί έλεγχοι με τους οποίους τεκμηριώνεται η προέλευση των τιμών ενός δείγματος από πληθυσμό κανονικής κατανομής με γνωστότερους τους ελέγχους Shapiro – Wilk και Kolmogorov – Smirnov αλλά η χρήση τους δεν συνιστάται καθώς οι έλεγχοι αυτοί είναι ιδιαίτερα ευαίσθητοι σε διαφορές μη κρίσιμες για το Student Test και πολλές φορές αρνούνται την αυτονόητη κανονικότητα που παρουσιάζεται σε ένα ιστόγραμμα (όταν για παράδειγμα το δείγμα είναι πολύ μεγάλο και υπάρχει πολύ μικρή απόκλιση από την κανονική κατανομή) ενώ δεν απορρίπτουν την υπόθεση της κανονικότητας ενώ η οπτική παρατήρηση τείνει στο αντίθετο γεγονός (όπως όταν υπάρχει μεγάλη απόκλιση από την κανονικότητα σε μικρό πλήθος παρατηρήσεων)

6.1.1 Έλεγχος ισότητας μέσης τιμής ενός δείγματος (One Sample T Test)

6.1.1.1 Παράδειγμα στατιστικού ελέγχου ισότητας μέσης τιμής

Η επιτροπή εμπορίου επιθυμεί να ελέγξει αν μία εταιρεία συσκευασίας γάλακτος νομιμοποιείται να γράφει στις συσκευασίες της πως το προϊόν ζυγίζει 500 γραμμάρια. Ο νόμος προβλέπει πως η εταιρεία νομιμοποιείται να αναγράφει την τιμή 500 γραμμάρια αν συμβαίνει το μέσο βάρος όλης της παραγωγής να είναι ακριβώς τόσο! (Παρατηρήστε πως λόγω των μικρών σφαλμάτων που υπάρχουν στην διαδικασία τυποποίησης σε συσκευασίες του γάλακτος, το βάρος μίας συσκευασίας δεν μπορεί να είναι ακριβώς 500

γραμμάρια αλλά θα είναι λίγο μεγαλύτερο ή λίγο μικρότερο).

Καθώς είναι αδύνατον να ζυγίσουμε το σύνολο όλων των συσκευασιών γάλακτος που βγήκαν ή θα βγούνε στο μέλλον από τη γραμμή παραγωγής αναγκαζόμαστε να πάρουμε ένα δείγμα από την παραγωγή και με βάση αυτό, χρησιμοποιώντας τον κατάλληλο στατιστικό έλεγχο να ελέγξουμε αν δεχόμαστε ή απορρίπτουμε τον ισχυρισμό της εταιρείας πως το μέσο βάρος όλης της παραγωγής είναι ίσο με 500 γραμμάρια..

Στην εικόνα 23 δίνονται τα δεδομένα τα οποία συλλέχθηκαν από δέκα συσκευασίες γάλακτος μιας γραμμής συσκευασίας της γαλακτοβιομηχανίας.

Σε πραγματικές συνθήκες το δείγμα που θα λαμβάναμε θα είχε μέγεθος μεγαλύτερο από δέκα (πενήντα θα ήταν ικανοποιητικό μέγεθος) αλλά για λόγους οικονομίας χώρου, στο παράδειγμα μας υποθέτουμε ότι αυτό είναι δέκα.

Θα ελέγξουμε αν η στατιστική υπόθεση

H_0 : Η μέση τιμή όλης της παραγωγής είναι ίση με 500 γραμμάρια

απορρίπτεται έναντι της εναλλακτικής (ερευνητικής) υπόθεσης

H_1 : Η μέση τιμή όλης της παραγωγής είναι μικρότερη από 500 γραμμάρια,

ή συνοπτικά την στατιστική υπόθεση **$H_0 : \mu = 500$** έναντι της **$H_1 : \mu < 500$.**

Παρατηρήστε πως η εναλλακτική (ερευνητική) υπόθεση είναι μονόπλευρη ανισότητα. Ο λόγος που γίνεται αυτό είναι γιατί στη συγκεκριμένη περίπτωση το σημαντικό είναι να μην είναι ελλιποβαρείς οι συσκευασίες καθώς αυτό είναι το παράνομο γεγονός το οποίο αναζητά η επιτροπή εμπορίου. Οι έλεγχοι αυτού του τύπου λέγονται μονόπλευροι. Στην περίπτωση στην οποία θα ήταν κατακριτέα η διαφορά και στις δύο περιπτώσεις ανισότητας τότε θα γράφαμε **$H_1 : \mu \neq 500$** και η δοκιμασία θα ονομαζόταν δίπλευρη.

6.1.1.2 Υλοποίηση της δοκιμασίας στο Calc.

Βάρος Συσκευασίας
490
503
499
492
500
501
489
478
498
508

Εικόνα 23: Δεδομένα .

Υπολογίζουμε τη μέση τιμή και την τυπική απόκλιση του δείγματος (Εικόνα 24). Η μέση τιμή (495,8 γραμμάρια) προκύπτει μικρότερη από την ισχυριζόμενη τιμή των 500 γρ. Η διαφορά μεταξύ της υποτιθέμενης μέσης τιμής και της παρατηρούμενης δειγματικής μέσης τιμής είναι **δ=4,2 γραμμάρια**. Όμως, είναι λογικό πως από τη στιγμή που οι συσκευασίες δεν έχουν βάρος ακριβώς ίσο με 500 γραμμάρια, είναι αναμενόμενο πως ούτε το δειγματικό μέσο βάρος θα προκύψει ακριβώς ίσο με 500 γραμμάρια.

Μέση Τιμή (AVERAGE(B4:B13))	**495,8**
Τυπική Απόκλιση (STDEV(B4:B13))	**8,64**

Εικόνα 24: Περιγραφικά Στατιστικά του δείγματος

Στο σημείο αυτό πρέπει να κρίνουμε αν η δειγματική διαφορά των 4,2 γραμμαρίων είναι "*μεγάλη*" ή "*μικρή*". Αν αποφασίσουμε ότι είναι "*μεγάλη*" τότε θα πρέπει να απορρίψουμε την υπόθεση πως το μέσο βάρος όλων των συσκευασιών είναι ίσο με 500 γραμμάρια ενώ αντίθετα αν καταλήξουμε πως είναι "*μικρή*" τότε δεν πρέπει να την απορρίψουμε (κάτι που πολλές φορές καταγράφεται μη ορθά ως "η αρχική υπόθεση γίνεται δεκτή" αντί του ορθότερου "η αρχική υπόθεση δεν απορρίπτεται".)

Την κρίση για το μέγεθος της διαφοράς μπορούμε να την κάνουμε αν γνωρίζουμε το εύρος των τιμών που περιμέναμε να πάρει αυτή όταν πάρουμε δείγματα από μία παραγωγή με μέσο βάρος 500 γραμμάρια. Ισοδύναμα, με στατιστική ορολογία αρκεί να γνωρίζουμε την κατανομή της διαφοράς δ υπό την προϋπόθεση πως ισχύει η αρχική υπόθεση H_0.

Η πληροφορία αυτή μας δίνεται από το κεντρικό οριακό θεώρημα το οποίο μας ενημερώνει πως η διαφορά δ θα ακολουθεί την κατανομή Student με 9 βαθμούς

ελευθερίας (9 = 10 – 1 = n – 1). Συμβολικά,

$$\bar{X} - 500 = \delta \sim T_9\left(0, \frac{8{,}64}{\sqrt{10}}\right) \Leftrightarrow \delta \sim T_9(0, 2.73) \Leftrightarrow \delta \sim T_9(0, 1.65^2)$$

Με απλά λόγια η τελευταία ισοδυναμία μας δίνει την πληροφορία πως αν πάρουμε πάρα πολλά δείγματα μεγέθους n=10 από μία αλυσίδα συσκευασίας γάλακτος με μέση τιμή 500 γραμμάρια τότε οι διαφορές των μέσων τιμών θα είναι συγκεντρωμένες γύρω από το μηδέν ενώ σχεδόν το σύνολο των διαφορών θα βρίσκονται σε εύρος τριών τυπικών αποκλίσεων από το μηδέν δηλαδή σε απόσταση 3*(9/7)^{1/2} = 3 * 1,13 = 3,39 από το μηδέν [η διακύμανση της κατανομής Student T_ν είναι $\sigma^2 = \nu/(\nu-2)$].

Η πιθανότητα να παρατηρήσουμε μία τόσο μεγάλη διαφορά είναι το κριτήριο το οποίο θα χρησιμοποιήσουμε για να αποφασίσουμε αν η διαφορά δ=4,2 είναι "*μεγάλη*" ή "*μικρή*" και κατ' επέκταση αν θα απορρίψουμε ή όχι την αρχική υπόθεση πως το μέσο βάρος της παραγωγής είναι 500 γραμμάρια. Η πιθανότητα αυτή συμβολίζεται p και ορίζεται ως

$$p = P(\delta > 4{,}2 \,||\, \mu = 500)$$

Πρώτα πρέπει να υπολογιστεί το στατιστικό μέγεθος

$$t = \frac{|\bar{x} - 500|\sqrt{n}}{s}$$

το οποίο υπολογίζεται σε ξεχωριστό κελί όπως φαίνεται στην εικόνα 25.

Τιμή στατιστικού	**1,54**
(ABS((F6-C17)*SQRT(COUNT(B4:B13))/F7))	

Εικόνα 25: Στατιστικό t

Η ζητούμενη πιθανότητα υπολογίζεται χρησιμοποιώντας τη συνάρτηση **TDIST()** η οποία απαιτεί τρία ορίσματα. Το πρώτο είναι η τιμή του στατιστικού t, η δεύτερη είναι το πλήθος των βαθμών ελευθερίας το οποίο ισούται με το μέγεθος του δείγματος μείον ένα (δηλαδή 9 = 10 – 1 στην περίπτωση μας) ενώ το τρίτο παίρνει τιμή ένα (1) ή δύο (2) ανάλογα αν ο στατιστικός έλεγχος είναι μονόπλευρος ή δίπλευρος. Αυτό καθορίζεται από το είδος της εναλλακτικής υπόθεσης και στην περίπτωση μας όπως αναφέραμε στην εισαγωγή είναι μονόπλευρος.

p (δίπλευρος έλεγχος)	0,16

Εικόνα 26: Η πιθανότητα p βάσει της οποίας θα γίνει δεκτή ή θα απορριφθεί η στατιστική υπόθεση

Το αποτέλεσμα εφαρμογής της συνάρτησης το οποίο το συμβολίζουμε με p είναι πιθανότητα, δηλαδή ένας αριθμός μεταξύ 0 και 1. Το p είναι η πιθανότητα να προκύψει δειγματική διαφορά μεγαλύτερη από την παρατηρούμενη από μία παραγωγή με μέσο βάρος 500 γραμμάρια. Η ερμηνεία του p είναι άμεση και είναι η εξής :

Αν η πιθανότητα αυτή είναι μεγάλη τότε δεν έχουμε ισχυρό έρεισμα να απορρίψουμε τον ισχυρισμό. Αν όμως είναι μικρή (συνήθως το όριο λαμβάνεται να είναι το 5%=0,05) τότε απορρίπτουμε τον ισχυρισμό διότι προκύπτει πως ήταν ιδιαίτερα απίθανη η δειγματοληψία με τέτοιο μέσο βάρος.

Στην συγκεκριμένη περίπτωση καθώς είναι p=0,16=16% > 5% = 0,05 δεν απορρίπτουμε τον ισχυρισμό της εταιρείας, δηλαδή δεν έχουμε στατιστική ένδειξη πως η παραγωγή είναι ελλιποβαρής άρα η εταιρεία δικαιούται να συνεχίσει να γράφει στις συσκευασίες της πως το βάρος κάθε συσκευασίας είναι 500 γραμμάρια!

6.1.1.3 Συνοπτικά βήματα για τον έλεγχο μέσης τιμής για ένα δείγμα (One Sample T Test).

I. Τοποθετούμε τα δεδομένα του δείγματος σε μία στήλη του Calc.

II. Καταγράφουμε στο χαρτί τη στατιστική υπόθεση η οποία είναι της μορφής H_0 : Η μέση τιμή του πληθυσμού από τον οποίο προήλθε το δείγμα είναι ίση με μ_0.

III. Ορίζουμε το όριο αποδοχής a της στατιστικής υπόθεσης. Συνήθως το a παίρνει την τιμή 0,05.

IV. Χρησιμοποιώντας τα δεδομένα μας υπολογίζουμε τη μέση τιμή και την τυπική απόκλιση.

V. Υπολογίζουμε το στατιστικό t.

VI. Χρησιμοποιώντας τη συνάρτηση TDIST() υπολογίζουμε την πιθανοφάνεια p της δειγματικής μέσης τιμής.

VII. Απορρίπτουμε ή όχι την στατιστική υπόθεση ανάλογα με το αν η p είναι μικρότερη ή όχι από το όριο σφάλματος a που θέσαμε στην έρευνα μας. (Συνήθως a=0,05=5%).

Πίνακας 6.1: Δοκιμασία t-test ενός δείγματος

Όπως περιγράφεται στην παράγραφο 6.1.1.2

x = c(490, 503, 499, 492, 500, 501, 489, 478, 498, 508)

t.test(x, mu = 500)

Με ?t.test εμφανίζονται και οι υπόλοιπες επιλογές

6.1.2 Έλεγχος ισότητας μέσης τιμής δύο ανεξάρτητων δειγμάτων (Independent Samples T-Test)

6.1.2.1 Εισαγωγή

Ο έλεγχος ισότητας δύο ανεξάρτητων δειγμάτων ή δοκιμασία Student ή απλά T Test είναι ο πλέον δημοφιλής τρόπος για να ελέγξουμε αν η μέση τιμή μίας μεταβαλλόμενης ποσότητας είναι ίδια σε δύο ανεξάρτητους πληθυσμούς παίρνοντας αντιπροσωπευτικά δείγματα από τους πληθυσμούς αυτούς. Για την υλοποίηση του ελέγχου λαμβάνονται υπόψη οι μέσες τιμές των δύο δειγμάτων, το μέγεθος των διασπορών των δύο δειγμάτων και το μέγεθος των δύο δειγμάτων. Το τελικό αποτέλεσμα προσδιορίζεται από την τιμή που θα έχει η πιθανότητα p να ισχύει η ισότητα των μέσων τιμών έχοντας παρατηρήσει τις δειγματικές τιμές.

Θεωρώντας σταθερούς όλους τους παράγοντες που επηρεάζουν την πιθανότητα p μπορούμε να πούμε πως η πιθανότητα p αυξάνεται όταν

1. αυξάνεται η διαφορά των δειγματικών μέσων,
2. αυξάνεται το μέγεθος των δύο δειγμάτων των πληθυσμών,
3. μειώνεται η δειγματική διακύμανση σε ένα ή και στα δύο δείγματα.

Για το T Test δύο ανεξάρτητων δειγμάτων η αρχική υπόθεση είναι η

H_0 : οι μέσες τιμές των πληθυσμών είναι ίσες ($H_0 : \mu_1=\mu_2$),

ενώ η εναλλακτική υπόθεση είναι η

H_1 : οι μέσες τιμές των πληθυσμών δεν είναι ίσες, είναι σημαντικά διαφορετικές ($H_1 : \mu_1\neq\mu_2$).

Η αποδοχή ή απόρριψη της H_0 καθορίζεται από την τιμή που λαμβάνει η πιθανότητα p η οποία με τη σειρά της προσδιορίζεται από την τιμή που λαμβάνει το στατιστικό t, το οποίο θα οριστεί στην επόμενη παράγραφο. Αν η πιθανότητα p είναι ιδιαίτερα μικρή (στις περισσότερες περιπτώσεις το όριο είναι το 0,05) τότε απορρίπτουμε την πρόταση H_0 πως οι μέσες τιμές των πληθυσμών είναι ίσες. Σε μία εργασία θα γράφαμε πως οι μέσες τιμές είναι σημαντικά διαφορετικές (ή στατιστικά διαφορετικές).

6.1.2.2 Προϋποθέσεις εφαρμογής του ελέγχου T-Test δύο ανεξάρτητων δειγμάτων.

Ισχύει ότι και για τον αντίστοιχο έλεγχο ενός δείγματος. Αν τα δύο δείγματα είναι μεγαλύτερα από 30 τότε η εφαρμογή της δοκιμασίας νομιμοποιείται άμεσα ενώ αν τα δείγματα είναι μικρά τότε πρέπει να τεκμηριωθεί ο λόγος για τον οποίον οι μεταβλητές ακολουθούν την κανονική κατανομή είτε με ποιοτική ανάλυση των παραγόντων που συμβάλουν στην μεταβλητότητα των τιμών τους είτε με αναφορά σε ανάλογες μελέτες της ίδιας μεταβλητής από προηγούμενες έρευνες είτε με γραφικό τρόπο χρησιμοποιώντας το ιστόγραμμα, το διάγραμμα p-p ή το διάγραμμα q-q.

Επιπλέον, η δοκιμασία αυτή απαιτεί μία σχετική ομοιότητα των διακυμάνσεων των δύο πληθυσμών. Περισσότερα για αυτό σε επόμενη παράγραφο!

6.1.2.3 Παράδειγμα στατιστικού ελέγχου ισότητας μέσης τιμής

Συγκρίνουμε την ικανότητα των μαθητών δύο σχολείων Α και Β στο μάθημα των μαθηματικών. Θέλουμε να αποδείξουμε πως τα δύο σχολεία έχουν μαθητές ίδιας ικανότητας στο μάθημα αυτό. Η ικανότητα στα μαθηματικά θα μετρηθεί με την επίδοση τους στα ίδια θέματα στα οποία οι μαθητές θα κληθούν να εξεταστούν. Καθώς δεν ήταν δυνατή η εξέταση του συνόλου των μαθητών των δύο σχολείων επιλέχθηκε ένα αντιπροσωπευτικό δείγμα από 10 μαθητές από το σχολείο Α και ένα αντιπροσωπευτικό δείγμα από 12 μαθητές από το σχολείο Β. Αυτοί εξετάστηκαν και οι επιδόσεις τους παρουσιάζονται στον πίνακα της εικόνας 27.

Χρησιμοποιώντας τα δεδομένα των δύο αντιπροσωπευτικών δειγμάτων πρέπει να αποφασίσουμε αν οι μαθητές των δύο σχολείων τα καταφέρνουν το ίδιο καλά στο μάθημα των μαθηματικών. Με στατιστική ορολογία πρέπει να ελέγξουμε την αρχική υπόθεση

H_0: Οι μέσες τιμές επιδόσεις των μαθητών των δύο σχολείων είναι ίσες,

και να βρούμε αν η H_0 απορρίπτεται ή όχι βάσει των στοιχείων του δείγματος. Καθώς δεν

υποψιαζόμαστε ανωτερότητα των μαθητών του ενός σχολείου έναντι του άλλου ορίζουμε ως εναλλακτική υπόθεση την

H_1 : Οι μέσες τιμές επιδόσεις των μαθητών των δύο σχολείων δεν είναι ίσες,

δηλαδή υλοποιούμε μία δίπλευρη δοκιμασία.

Σχολείο Α	Σχολείο Β
12	20
13	19
10	17
14	10
15	14
13	13
20	17
19	19
17	16
16	15
	18
	19

Εικόνα 27: Δεδομένα στατιστικού ελέγχου

Υπολογίζοντας τις μέσες επιδόσεις των μαθητών που αντιπροσωπεύουν κάθε σχολείο (Εικόνα 28) βρίσκουμε πως οι μαθητές από το σχολείο Β υπερτερούν από τους μαθητές του σχολείου Α κατά 1,52 μονάδες (Εικόνα 29). Η διαφορά των 1,52 μονάδων δείχνει πως οι μαθητές του δείγματος του σχολείου Β είναι καλύτεροι όσον αφορά τη μέση επίδοση τους από τους μαθητές του σχολείου Α.

	A	B
Μέση Τιμή	14,9	16,42
Τυπική Απόκλιση	3,14	2,97

Εικόνα 28: Περιγραφικά στατιστικά ομάδων

Διαφορά μέσων τιμών	1,52

Εικόνα 29: Δειγματική διαφορά

Το ερώτημα που αβίαστα τίθεται είναι αν η διαφορά **δ = 1,52** μονάδες μεταξύ των δειγματικών μέσων τιμών είναι τόσο μεγάλη ώστε να μην είναι δυνατό να αποδοθεί στο

στατιστικό σφάλμα της δειγματοληψίας δηλαδή στη φυσιολογική απόκλιση που θα περιμέναμε να έχουν οι μέσες τιμές των δύο δειγμάτων μεταξύ τους.

Αν αποφασίσουμε πως το μέγεθος της διαφοράς είναι πολύ μεγάλο για να οφείλεται στο τυχαίο σφάλμα της δειγματοληψίας τότε αυτό θα σημαίνει πως υπάρχει μία “συστημική” διαφοροποίηση στις επιδόσεις η οποία θα οφείλεται στη διαφορετική ικανότητα των μαθητών των δύο σχολείων στα μαθηματικά και πιο συγκεκριμένα στην περίπτωση μας πως οι μαθητές του σχολείου Β είναι περισσότερο ικανοί στα μαθηματικά από τους μαθητές του σχολείου Α.

Αν αντίθετα, καταλήξουμε πως μία τέτοια διαφορά στις δειγματικές μέσες επιδόσεις ήταν αρκετά πιθανό να εμφανιστεί λόγω του τυχαίου σφάλματος της δειγματοληψίας τότε μπορούμε να ολοκληρώσουμε την έρευνα μας χωρίς να απορρίψουμε την ισοδυναμία των μαθητών των δύο σχολείων στα μαθηματικά. (Προσέξτε πως αναφέρουμε “δεν απορρίπτουμε” και όχι “αποδεχόμαστε” καθώς η αποδοχή προϋποθέτει ισχυρότερα επιχειρήματα από έναν απλό στατιστικό έλεγχο. Επιπλέον, είναι περισσότερο σεμνό και δεν δημιουργεί εσφαλμένες εντυπώσεις ενός απόλυτα αντικειμενικού ελέγχου στο σύνολο των ανεξάρτητων πληθυσμών των δύο σχολείων!)

Για να χαρακτηρίσουμε το μέγεθος της διαφοράς δ ως “μεγάλο” ή “μικρό” πρέπει να γνωρίζουμε το εύρος των τιμών που περιμέναμε να πάρει αυτή σε μία αντίστοιχη τυχαία δειγματοληψία 10 και 12 μαθητών αντίστοιχα από δύο σχολεία Α και Β με ίση μέση επίδοση στα μαθηματικά.

Ισοδύναμα, με στατιστική ορολογία πρέπει να γνωρίζουμε την κατανομή της διαφοράς δ υπό την προϋπόθεση πως ισχύει η αρχική υπόθεση H_0.

6.1.2.4 Θεωρητική ανάλυση και λύση

Ο καλύτερος τρόπος για την μέτρηση της πραγματικής κατανομής της διαφοράς δ είναι η συνεχής επανάληψη της δειγματοληψίας 10 και 12 μαθητών αντίστοιχα από τα δύο σχολεία Α και Β, η μέτρηση της δειγματικής διαφοράς σε κάθε δειγματοληψία και η παρατήρηση της κατανομής του συνόλου των δειγματικών διαφορών που προκύπτει.

Καθώς αυτό δεν είναι καθόλου εύκολο να γίνει καταφεύγουμε στη θεωρία κατανομών η οποία μας ενημερώνει πως η διαφορά δ διαιρούμενη με την “κοινή” τυπική απόκλιση $s_{\text{κοινή}}$ ακολουθεί την κατανομή Student.

Η “κοινή” διακύμανση $s_{\text{κοινή}}$ στο Calc υπολογίζεται από τον τύπο

$$s^2_{\text{κοινή}} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{3,14^2}{10} + \frac{2,97^2}{12} = 1,72 = 1,31^2$$

Τύπος 9: Υπολογισμός "κοινής" διακύμανσης

ενώ οι βαθμοί ελευθερίας df της κατανομής είναι ίσοι με $n_1+n_2 - 2 = 10+12 - 2 = 20$. Συνοπτικά, γράφουμε

$$t = \frac{\delta}{1,31} \sim T_{20}$$

Γνωρίζουμε από τη θεωρία κατανομών πως η διακύμανση της κατανομής Student T_ν είναι $\sigma^2 = \nu/(\nu-2)$. Για ν=20 υπολογίζουμε πως η κατανομή Student T_{20} έχει διακύμανση 20/18 = 1,11 και αντίστοιχη τυπική απόκλιση σ = 1,05.

Με απλά λόγια η τελευταία ισοδυναμία μας δίνει την πληροφορία πως αν επαναλάβουμε πολλές φορές τη δειγματοληψία 10 και 12 μαθητών αντίστοιχα από τα σχολεία Α και Β, μετρήσουμε τη διαφορά δ των δειγματικών μέσων επιδόσεων και διαιρέσουμε ότι βρούμε με το 1,31 θα πάρουμε τιμές οι οποίες θα κατανέμονται συμμετρικά γύρω από το μηδέν, οι περισσότερες από αυτές θα βρίσκονται στην περιοχή του μηδέν, το πλήθος τους θα μειώνεται όσο απομακρυνόμαστε από το μηδέν ενώ σχεδόν το σύνολο από αυτές θα βρίσκονται σε απόσταση 1,05 x 3 = 3,15 πριν και μετά το μηδέν.

Ολοκληρώνουμε την έρευνα μας υπολογίζοντας βάσει της παραπάνω θεωρητικής περιγραφής και της κατανομής T_{20}, την πιθανότητα p να βρούμε μία διαφορά η οποία θα βρίσκεται σε απόσταση μεγαλύτερη από 1,52.

Υπολογίζουμε

$$p = P(|\delta| > 1,52) = P(|t| > 1,16) = TDIST(1,16;20;2) = 0,26$$

όπου χρησιμοποιήσαμε τη συνάρτηση TDIST() του Calc η οποία δίνει την ζητούμενη πιθανότητα. Αν ο έλεγχος ήταν μονόπλευρος (στην περίπτωση που εξ'αρχής υποθέταμε πως το ένα σχολείο π.χ. το Α είναι καλύτερο από το Β) θα τοποθετούσαμε τη μονάδα (1) ως τελευταίο όρισμα στη συνάρτηση κάτι που θα είχε ως αποτέλεσμα ακριβώς το ήμισυ του αποτελέσματος που πήραμε, δηλαδή 0,13.

6.1.2.5 Σύντομη λύση δίχως ανάλυση

Αν και δεν προτείνεται υπάρχει η δυνατότητα άμεσου υπολογισμού του τελικού αποτελέσματος δίχως περιττές πράξεις χρησιμοποιώντας τη συνάρτηση **TTEST()** η οποία παίρνει τέσσερα ορίσματα από τα οποία τα δύο πρώτα είναι τα κελιά όπου περιέχονται τα δεδομένα του πρώτου και του δεύτερου δείγματος, στο τρίτο καταχωρείται το είδος του στατιστικού ελέγχου (1: μονόπλευρο ή 2:δίπλευρο) ενώ στο τελευταίο τοποθετείται ο τύπος του T-Test που θα εφαρμοστεί. Στην περίπτωση της παραγράφου αυτής καθώς τα δύο δείγματα προέρχονται από ανεξάρτητους πληθυσμούς η τιμή που πρέπει να καταχωρηθεί ως τέταρτο και τελευταίο όρισμα είναι το 2 (για ίσες διακυμάνσεις πληθυσμών) ή το 3 (για άνισες διακυμάνσεις πληθυσμών). Καθώς από την ανάλυση στις δειγματικές διακυμάνσεις αποφασίσαμε πως οι διακυμάνσεις των πληθυσμών είναι ίσες τοποθετούμε ως τελευταίο όρισμα το 2 και υπολογίζουμε την πιθανότητα p ίση με 0,26 ολοκληρώνοντας τη δοκιμασία (Εικόνα 30).

Πιθανότητα p (Δεχόμαστε ίση διασπορά στους δύο πληθυσμούς)	0,26

Εικόνα 30: Πιθανοφάνεια διαφοράς μεταξύ των ομάδων

Πρέπει ωστόσο να σημειωθεί πως η ολοκληρωμένη παρουσίαση της δοκιμασίας επιβάλει στην παρουσίαση των αποτελεσμάτων να περιλαμβάνονται όλες οι στατιστικές ποσότητες που προηγήθηκαν του υπολογισμού του p ώστε κάθε τρίτος αναγνώστης να είναι σε θέση να επιβεβαιώσει το τελικό αποτέλεσμα και να κρίνει την ορθότητα της διαδικασίας..

6.1.2.6 Έλεγχος της ισότητας των διακυμάνσεων

Η παραμετρική δοκιμασία Independent Samples T Test είναι απολύτως αξιόπιστη όταν οι διακυμάνσεις των πληθυσμών είναι ίσες ενώ δεν συμβαίνει το ίδιο όταν οι διακυμάνσεις δεν είναι ίσες. Επιπλέον, το σφάλμα αξιοπιστίας της μεθόδου είναι ανάλογο της διαφοράς στα μεγέθη των δύο δειγμάτων και αντιστρόφως ανάλογο της διαφοράς των μεγεθών των δύο δειγμάτων. Από την άλλη, όταν τα δύο δείγματα είναι ίδιου μεγέθους τότε η εφαρμογή της δοκιμασίας είναι αποδεκτή όταν η μία τυπική απόκλιση είναι το πολύ διπλάσια της άλλης.

Όταν το ένα δείγμα έχει αρκετά μεγαλύτερη διακύμανση και αρκετά μικρότερο μέγεθος από το άλλο η χρήση της δοκιμασίας κρίνεται απαγορευτική και μπορεί να οδηγήσει σε

εσφαλμένα τελικά συμπεράσματα**.** Ενδεικτικά αναφέρουμε πως η εξομοιωτική της εφαρμογής της δοκιμασίας σε δύο δείγματα τα οποία προέρχονται από την κανονική κατανομή αλλά το ένα δείγμα έχει αρκετά μεγαλύτερη διακύμανση (λόγος μεταξύ τυπικών αποκλίσεων ίσος με 5) από το άλλο και αρκετά μικρότερο μέγεθος από το άλλο (λόγος δειγματικών μέγεθός ίσος με 1/5) οδηγεί σε υπολογισμό της τιμής του p ίση με 0,05 ενώ η πραγματική τιμή του υπολογίζεται ίση με 0,22!

Αν δεν μπορούμε να αποφασίσουμε πως οι διακυμάνσεις των πληθυσμών είναι ίσες τότε πρέπει να γίνει ξεχωριστός στατιστικός έλεγχος υπόθεσης για την ισότητα των διακυμάνσεων των πληθυσμών. Οι πιο γνωστές δοκιμασίες ελέγχου ισότητας διακυμάνσεων είναι η δοκιμασία λόγου (F Test), η πιο ασφαλής δοκιμασία Levene, η δοκιμασία του Bartlett και η δοκιμασία Brown-Forsythe.

Η δοκιμασία του λόγου (F Test) παρουσιάζεται στη συνέχεια. Η αρχική υπόθεση της δοκιμασίας αυτής είναι

H_0 : Οι διακυμάνσεις των δύο πληθυσμών είναι ίσες ($\sigma_1 = \sigma_2$).

Το στατιστικό που πρέπει να υπολογιστεί από τα στοιχεία των δειγμάτων είναι το

$$F = \frac{s_L^2}{s_S^2}$$

όπου s_L^2 η μεγαλύτερη από τις δύο δειγματικές διακυμάνσεις και s_S^2 η μικρότερη από τις δύο δειγματικές διακυμάνσεις. Το στατιστικό F αποδεικνύεται πως ακολουθεί την κατανομή F(9,11) ($9 = n_L - 1 = 10 - 1$ και $11 = n_S - 1 = 12 - 1$).

Στην περίπτωση των δεδομένων της εικόνας 28 $F = 3{,}14^2/2{,}97^2 = 1{,}18$ και η πιθανότητα εμφάνισης αυτής της τιμής υπολογίζεται στο Calc με χρήση της συνάρτησης **FDIST(1,18;9;11)** η οποία δίνει ως αποτέλεσμα 0,39 = 39% > 5% = 0,05 επιβεβαιώνοντας πως η αρχική υπόθεση H_0 δεν απορρίπτεται και μπορούμε να προχωρήσουμε στην υλοποίηση της δοκιμασίας

Τέλος, δεν πρέπει να παραγνωρίσουμε πως βοηθά και η ποιοτική ανάλυση της μεταβλητής. Για παράδειγμα, στην περίπτωση της επίδοσης των μαθητών στα μαθηματικά (Εικόνα 28) φαίνεται ρεαλιστική η υπόθεση πως η διακύμανση της επίδοσης στο ένα σχολείο θα είναι ίδια με την διακύμανση στο άλλο σχολείο καθώς οι ίδιοι παράγοντες που επιδρούν στους μαθητές του σχολείου Α εισάγοντας μεταβλητότητα στην επίδοση τους στα

μαθηματικά, επιδρούν εξίσου και στους μαθητές του σχολείου Β και εισάγουν ανάλογη μεταβλητότητα στις επιδόσεις τους!

6.1.2.7 Θεωρητική παρατήρηση *

Το Calc όπως και το Excel χρησιμοποιούν τον ίδιο τύπο (Τύπος 9, σελίδα 162) για τον υπολογισμό της "κοινής" διακύμανσης τόσο όταν δεχόμαστε ίσες διακυμάνσεις πληθυσμών, όσο και όταν δεχόμαστε άνισες διακυμάνσεις. Ωστόσο, η θεωρία προτείνει διαφορετικούς τρόπους υπολογισμού για την "κοινή" διακύμανση $s_{κοινή}$ και το πλήθος των βαθμών ελευθερίας df που απαιτούνται για τον υπολογισμό της πιθανότητας p ανάλογα με το αν δεχθούμε ή όχι πως οι διακυμάνσεις των πληθυσμών είναι ίσες.

Αν δεχθούμε πως οι διακυμάνσεις των δύο ανεξάρτητων πληθυσμών είναι ίσες τότε η "κοινή" διακύμανση (ονομάζεται και συνδυασμένο τυπικό σφάλμα της διαφοράς των δύο μέσων τιμών – pooled standard error of the difference) υπολογίζεται ως

$$s^2_{κοινή} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

και οι βαθμοί ελευθερίας df είναι ίσοι με n_1+n_2-2, ενώ αν δεχθούμε πως οι διακυμάνσεις των δύο ανεξάρτητων πληθυσμών δεν είναι ίσες τότε η "κοινή" διακύμανση (ονομάζεται και μη συνδυασμένο τυπικό σφάλμα της διαφοράς των δύο μέσων τιμών – unpooled standard error of the difference) υπολογίζεται ως

$$s^2_{κοινή} = \frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}$$

και οι βαθμοί ελευθερίας είναι ίσοι με

$$df = \frac{(n_1-1)(n_2-1)}{(n_1-1)(1-c)^2+(n_2-1)c^2} \quad \text{όπου} \quad c = \frac{s_1^2/n_1}{s_1^2/n_1+s_2^2/n_2}$$

Αντιλαμβάνεται ο αναγνώστης πως το Calc (και το Excel) χρησιμοποιεί ένα συνδυασμό των παραπάνω τρόπων καθώς υπολογίζει τους βαθμούς ελευθερίας df από την πρώτη περίπτωση και την "κοινή" διακύμανση από τη δεύτερη περίπτωση. Η διαφορά αυτή δεν είναι κρίσιμη όταν οι διακυμάνσεις δεν είναι πολύ διαφορετικές αλλά μπορεί να αλλάζει το τελικό αποτέλεσμα όταν η διαφορά των διακυμάνσεων είναι μεγάλη, περίπτωση κατά την οποία είναι καλό ο ερευνητής να αναζητήσει και μη παραμετρικούς τρόπους ελέγχου της υπόθεσης του!

6.1.2.8 Τελικά σχόλια

Τελευταίο βήμα του στατιστικού ελέγχου είναι η ερμηνεία της πιθανότητας p η οποία με τα συγκεκριμένα δεδομένα υπολογίστηκε να είναι 0,26 = 26% (Εικόνα 30). Η πιθανότητα αυτή σημαίνει πως αν δεχθούμε την υπόθεση ότι οι δύο πληθυσμοί (δηλαδή τα δύο σχολεία) έχουν μαθητές με ίδια μέση επίδοση, και προχωρήσουμε σε 100 τυχαίες δειγματοληψίες δέκα και δώδεκα μαθητών αντίστοιχα από τα δύο σχολεία, περιμένουμε στις 26 από αυτές τις δειγματοληψίες να έχουμε διαφορά στις μέσες επιδόσεις των μαθητών μεταξύ των δύο σχολείων μεγαλύτερη από 1,52 μονάδες. Καθώς, αυτή η πιθανότητα είναι αρκετά μεγάλη (πιο συγκεκριμένα, μεγαλύτερη από 5% που τίθεται συνήθως ως κατώτερο όριο) δεν μπορούμε να απορρίψουμε την υπόθεση της ισότητας των δύο μέσων τιμών, ή ισοδύναμα η διαφορά των 1,52 μονάδων είναι αρκετά πιθανό να οφείλεται στο τυχαίο σφάλμα της δειγματοληψίας.

6.1.2.9 Βασικά βήματα του ελέγχου

I. Τοποθετούμε τα δεδομένα του δύο δειγμάτων σε δύο, όχι κατά ανάγκη συνεχόμενες στήλες του Calc.

II. Καταγράφουμε στο χαρτί τη στατιστική υπόθεση η οποία είναι της μορφής H_0: Η μέση τιμή του πληθυσμού Α είναι ίση με τη μέση τιμή του πληθυσμού Β.

III. Χρησιμοποιώντας τα δεδομένα μας υπολογίζουμε τις τυπικές αποκλίσεις των δύο πληθυσμών τις οποίες συγκρίνουμε και αποφασίζουμε αν μπορούμε να δεχθούμε πως τα δύο δείγματα προέρχονται από δύο πληθυσμούς με ίση ή άνιση διακύμανση.

IV. Χρησιμοποιώντας τη συνάρτηση TTEST(), υπολογίζουμε την πιθανοφάνεια p της διαφοράς που παρατηρήθηκε μεταξύ των δύο μέσων τιμών.

V. Απορρίπτουμε ή όχι την στατιστική υπόθεση ανάλογα με το αν η p είναι μικρότερη ή όχι από το όριο σφάλματος a που θέσαμε στην έρευνα μας. (Συνήθως a=0,05=5%)

Πίνακας 6.2: Δοκιμασία t-test δύο ανεξάρτητων δειγμάτων
Όπως περιγράφεται στην παράγραφο 6.1.2.5.
x = c(12, 13, 10, 14, 15, 13, 20, 19, 17, 16) **y = c(20, 19, 17, 10, 14, 13, 17, 19, 16, 15, 18, 19)** **t.test(x, y)**

6.1.3 Έλεγχος ισότητας μέσης τιμής περισσότερων από δύο ανεξάρτητων δειγμάτων (ANOVA : Analysis Of Variance)

Κάποιες φορές ο ερευνητής θέλει να ελέγξει τη διαφοροποίηση μίας συνεχής μεταβλητής ανάμεσα σε περισσότερες από δύο ομάδες. Για παράδειγμα ένας καθηγητής που διδάσκει σε τέσσερα τμήματα Α' Λυκείου από ισάριθμα σχολεία Α, Β, Γ και Δ δίνει σε όλα τα τμήματα το ίδιο διαγώνισμα μαθηματικά και θεωρώντας το τμήμα του κάθε ενός σχολείου ως αντιπροσωπευτικό του συνόλου των μαθητών του αντίστοιχου σχολείου επιθυμεί να ελέγξει αν τα τέσσερα σχολεία έχουν στατιστικά διαφορετική επίδοση στα μαθηματικά.

Μία λύση που μπορεί να σκεφτεί είναι να εφαρμόσει τη δοκιμασία Student για δύο ανεξάρτητα δείγματα (Independent Samples T – Test) μεταξύ όλων των συνδυασμών των σχολείων (δηλαδή $C(4,2)=\frac{4!}{2!2!}=6$ φορές) ωστόσο αυτή η λύση δεν είναι η ενδεδειγμένη καθώς αν ορίσουμε επίπεδο απόρριψης της στατιστικής υπόθεσης το 5%, υπάρχει πιθανότητα 23,2% να προκύψει τυχαία μία στατιστική διαφοροποίηση και 3,1% να προκύψουν δύο λόγω ου τυχαίου σφάλματος της δειγματοληψίας[1] .

Η διαδικασία της ανάλυσης διασποράς (Analysis of Variation – ANOVA) ελέγχει την ισότητα όλων των μέσων τιμών δίνοντας ως απάντηση ένα στατιστικό βασισμένο στην κατανομή F (προς τιμήν του Fisher που την επινόησε) και μία στατιστική σημαντικότητα p από την οποία εύκολα προκύπτει το ερευνητικό αποτέλεσμα.

Για την Ανάλυση Διακύμανσης η αρχική υπόθεση είναι η

1 Αν X το πλήθος των στατιστικά σημαντικών διαφοροποιήσεων στις 6 δοκιμές τότε $P(X=1)=\binom{6}{1}0.05\cdot 0.95^5=0,232$ και $P(X=2)=\binom{6}{2}0.05^2\cdot 0.95^4=0,031$

H_0 : η μέση επίδοση των σχολείων στα μαθηματικά είναι ίσες ($H_0 : \mu_1=\mu_2$),

ενώ η εναλλακτική υπόθεση είναι η

H_1 : η μέση επίδοση των σχολείων είναι σημαντικά διαφορετικές ($H_1 : \mu_1\neq\mu_2$).

Η αποδοχή ή απόρριψη της H_0 καθορίζεται από την τιμή που λαμβάνει η πιθανότητα p η οποία με τη σειρά της προσδιορίζεται από την τιμή που λαμβάνει το στατιστικό F. Αν η πιθανότητα p είναι ιδιαίτερα μικρή (στις περισσότερες περιπτώσεις το όριο είναι το 0,05) τότε απορρίπτουμε την πρόταση H_0 πως οι μέσες τιμές των πληθυσμών είναι ίσες.

Πιο αναλυτικά, ας υποθέσουμε πως ο παρακάτω πίνακας περιέχει τις βαθμολογίες από 10 μαθητές από τα 4 σχολεία Α, Β, Γ και Δ.

Πίνακας 6.3: Βαθμολογίες μαθητών στα μαθηματικά

	Σχολείο			
αα	**Α**	**Β**	**Γ**	**Δ**
1	19	11	13	13
2	15	13	13	13
3	08	20	15	18
4	11	13	09	07
5	13	15	20	14
6	15	11	17	14
7	14	09	16	18
8	18	09	10	11
9	20	18	20	10
10	13	14	12	10
ΜΟ	$\bar{x}_A=14,6$	$\bar{x}_B=13,3$	$\bar{x}_\Gamma=14,5$	$\bar{x}_\Delta=12,8$
ΤυπΑπ	$s_A=3,7$	$s_B=3,6$	$s_\Gamma=3,8$	$s_\Delta=3,5$
ΜΟ βαθμολογίας όλων των μαθητών :	$\bar{x}=13,8$ *(*	$s=3,6$ *)*		

Υποθέτουμε πως

- Οι βαθμολογίες είναι κανονικά κατανεμημένες.
- Οι τυπικές αποκλίσεις των δειγμάτων είναι παρόμοιες.
- Οι βαθμολογίες για κάθε δείγμα είναι ανεξάρτητες μεταξύ τους.

Πίνακας 6.4: Απόκλιση ομάδων από τη μέση τιμή				
	Σχολείο			
	Α	**Β**	**Γ**	**Δ**
Απόκλιση	14,6 – 13,8 = 0,8	13,3 – 13,8 = -0,5	14,5 – 13,8 = 0,7	12,8 – 13,8 = -1
Τετράγωνο	0,64	0,25	0,49	1

Θεωρώντας τη μέση βαθμολογία κάθε ενός δείγματος ως αντιπροσωπευτική όλων των μαθητών του δείγματος, υπολογίζουμε το άθροισμα των τετραγωνικών αποκλίσεων από τη μέση τιμή όλων των μαθητών (between-group" sum of squares):

$$SS_B = 10\left((\bar{x}_A-\bar{x})^2+(\bar{x}_B-\bar{x})^2+(\bar{x}_\Gamma-\bar{x})^2+(\bar{x}_\Delta-\bar{x})^2\right) = 10\left(0{,}8^2+(-0{,}5)^2+0{,}7^2+(-1)^2\right) = 23{,}8$$

Καθώς χρησιμοποιήσαμε τη δειγματική μέση τιμή για την εκτίμηση της μέσης τιμής του πληθυσμού, υπολογίζουμε τη μέση τετραγωνική απόκλιση μεταξύ των ομάδων διαιρώντας το παραπάνω άθροισμα με το πλήθος των ομάδων μείον ένα.

Μέση τετραγωνική απόκλιση μεταξύ των ομάδων $MSS_B = \frac{23{,}8}{4-1} = 7{,}93$

Υπολογίζουμε την απόκλιση κάθε παρατήρησης από τη μέση τιμή της ομάδας που ανήκει.

Πίνακας 6.5: Απόκλιση βαθμολογίας από το κέντρο του δείγματος				
	Σχολείο			
αα	**Α**	**Β**	**Γ**	**Δ**
1	4,4	-2,3	-1,5	0,2
2	0,4	-0,3	-1,5	0,2
3	-6,6	6,7	0,5	5,2
4	-3,6	-0,3	-5,5	-5,8
5	-1,6	1,7	5,5	1,2
6	0,4	-2,3	2,5	1,2
7	-0,6	-4,3	1,5	5,2
8	3,4	-4,3	-4,5	-1,8
9	5,4	4,7	5,5	-2,8
10	-1,6	0,7	-2,5	-2,8

Υπολογίζουμε το άθροισμα των τετραγωνικών αποκλίσεων των παρατηρήσεων από τη μέση τιμή του δείγματος όπου ανήκουν :

$$SS_W = 4{,}4^2 + 0{,}4^2 + (-6{,}6)^2 + \ldots + (-2{,}8)^2 = 480{,}6$$

Καθώς χρησιμοποιήσαμε τις τέσσερις δειγματικές μέσες τιμές για την εκτίμηση της μέσης τιμής κάθε μίας ομάδας, υπολογίζουμε τη μέση τετραγωνική απόκλιση στο εσωτερικό των ομάδων (within-group sum of squares) διαιρώντας το παραπάνω άθροισμα με το 40 – 4 = 36 και βρίσκουμε

Μέση τετραγωνική απόκλιση στο εσωτερικό των ομάδων

$$MSS_W = \frac{480,6}{40-4} = 13,35$$

Θεωρούμε το λόγο $F = \frac{MSS_B}{MSS_W}$. Αν δεν υπάρχει επίδραση του σχολείου στη βαθμολογία των μαθητών τότε θα περιμέναμε ο λόγος αυτός να είναι κοντά στο 0. Υπολογίζουμε

$$F = \frac{MSS_B}{MSS_W} = \frac{7,93}{13,35} = 0,59 \ .$$

Καθώς το στατιστικό F ακολουθεί την κατανομή F με 3 και 36 βαθμούς ελευθερίας υπολογίζουμε την πιθανότητα η τιμή αυτή να οφείλεται στο τυχαίο σφάλμα της δειγματοληψίας από κάποιον σχετικό πίνακα να είναι p = 0,658 ή 65,8%. (Συνάρτηση FDIST(0,54;3;36) του Calc)

Καθώς η πιθανότητα αυτή είναι ιδιαίτερα μεγάλη (και οπωσδήποτε μεγαλύτερη από το 0,05 = 5%) δεν απορρίπτουμε τη στατιστική υπόθεση, δηλαδή συμπεραίνουμε πως δεν υπάρχουν επαρκείς ενδείξεις για να συνάγουμε διαφοροποίηση στην επίδοση στα μαθηματικά μεταξύ των σχολείων Α, Β, Γ και Δ.

Πίνακας 6.6: Δοκιμασία ANOVA

Όπως περιγράφεται στην παράγραφο 6.1.3.

```
vathmoi = c(19, 15, 8, 11, 13, 15, 14, 18, 20, 13, 11, 13, 20, 13, 15, 11, 9, 9, 18, 14, 13,
13, 15, 9, 20, 17, 16, 10, 20, 12, 13, 13, 18, 7, 14, 14, 18, 11, 10, 10)
taksi = c('A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B',
'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'D', 'D', 'D', 'D', 'D', 'D', 'D', 'D', 'D', 'D')
taksif = as.factor(taksi)
data = data.frame(v = vathmoi, t = taksif)
plot(data$v ~ data$t) (Συγκριτικό διάγραμμα θηκογραμμάτων των κατανομών)
mymodel = aov(data$v ~ data$t)
summary(mymodel)
pairwise.t.test(data$v, data$t, p.adj = "none") Σύγκριση μεταξύ των τάξεων ανά 2.
```

Πίνακας 6.7: Επέκταση : Δοκιμασία ANOVA με δύο παράγοντες (two – way ANOVA)

```
vathmoi = c(19, 15, 8, 11, 13, 15, 14, 18, 20, 13, 11, 13, 20, 13, 15, 11, 9, 9, 18, 14, 13,
13, 15, 9, 20, 17, 16, 10, 20, 12, 13, 13, 18, 7, 14, 14, 18, 11, 10, 10)
taksi = c('A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'B',
'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'D', 'D', 'D', 'D', 'D', 'D', 'D', 'D', 'D', 'D')
gender = c('m', 'm', 'f', 'f', 'f', 'f', 'f', 'f', 'f', 'm', 'm', 'm', 'f', 'm', 'f', 'f', 'f', 'f', 'm', 'm', 'm',
'm', 'f', 'f', 'm', 'm', 'm', 'f', 'm', 'm', 'm', 'm', 'f', 'f', 'f', 'm', 'f', 'm', 'm', 'f')
taksif = as.factor(taksi)
genderf = as.factor(gender)
data = data.frame(v = vathmoi, t = taksif, g = genderf)
plot(data$v ~ data$t)
Έλεγχος για διαφοροποίηση στο φύλο και στο σχολείο
mymodel = aov(data$v ~ data$t + data$g)
summary(mymodel)
Έλεγχος για διαφοροποίηση και στους δύο παράγοντες
mymodel = aov(data$v ~ data$t * data$g)
summary(mymodel)
A = subset(data, t == "A") Επιλέγει στο Α μόνο τους μαθητές της Α τάξης.
```

Παρατήρηση

Η ανάλυση διακύμανσης είναι μία γενική μέθοδος για την ανίχνευση διαφοροποιήσεων στις τιμές μίας συνεχούς μεταβλητής κάτω από την επιρροή παραγόντων. Η υλοποίηση της ανάλυσης διακύμανσης αρχίζει από την καταγραφή του μοντέλου το οποίο υποτίθεται πως εξηγεί τις διαφοροποιήσεις στις τιμές της μεταβλητής. Το μοντέλο καταγράφεται ως μία μαθηματική εξίσωση και ανήκει σε έναν από τους παρακάτω τύπους

1. Μοντέλο σταθερών παραγόντων (fixed effects model) που εφαρμόζεται στις περιπτώσεις όπου αναζητείται η διαφοροποίηση των τιμών ανάμεσα σε περισσότερες από δύο ομάδες που ορίζονται από κάποιο παράγοντα. Ως παράδειγμα μπορεί να αναφερθεί ο καθηγητής που θέλει να ανιχνεύσει τις διαφοροποιήσεις στις επιδόσεις των μαθητών σε δέκα διαφορετικά σχολεία και επιλέγει να εξετάσει ένα τυχαίο δείγμα μαθητών από κάθε ένα σχολείο. Η μεταβλητή είναι η επίδοση στη γραπτή δοκιμασία ενώ ο σταθερός παράγοντας είναι το σχολείο φοίτησης.
2. Μοντέλο τυχαίων παραγόντων (random effects model) όπου δεν είναι γνωστό το πλήθος των περιπτώσεων που περιέχει ένας παράγοντας οπότε αναγκαστικά εκτιμάται από το ίδιο το δείγμα. Αν ο καθηγητής της προηγούμενης παραγράφου είχε ως στόχο την ανίχνευση των διαφοροποιήσεων στην επίδοση στα μαθηματικά στο σύνολο των σχολείων της επικράτειας και επέλεγε δέκα σχολεία με τυχαίο τρόπο από το σύνολο των σχολείων της επικράτειας τότε υπάρχει ενδεχόμενο να μην έχουν εμφανιστεί στο δείγμα των σχολείων διαφοροποιημένα σχολεία ως προς το σύνολο και αυτή η ασάφεια εισάγει επιπλέον απροσδιοριστία.
3. Μικτό μοντέλο σταθερών και τυχαίων παραγόντων (mixed effects model) στο οποίο υπάρχουν σταθεροί και τυχαίοι παράγοντες. Για παράδειγμα αν ο καθηγητής είχε μοιράσει τρία διαφορετικά βιβλία μαθηματικών στα 10 σχολεία (που επιλέχθηκαν με τυχαία δειγματοληψία από το σύνολο των σχολείων της επικράτειας) και ενδιαφερόταν να ανιχνεύσει τόσο τη διαφοροποίηση μεταξύ των σχολείων όσο και τη διαφοροποίηση μεταξύ των 3 βιβλίων τότε θα είχε έναν σταθερό παράγοντα που επηρεάζει την επίδοση των μαθητών (τα 3 διαφορετικά βιβλία) και ένα τυχαίο παράγοντα που είναι το σχολείο φοίτησης.

Η περαιτέρω ανάλυση της μεθόδου ξεφεύγει από τα πλαίσια αυτών των σημειώσεων.

6.1.4 Έλεγχος ισότητας μέσης τιμής ζευγαρωτών παρατηρήσεων (Paired

Samples T-Test)

Ζευγαρωτές ονομάζονται δύο παρατηρήσεις οι οποίες προέρχονται από το ίδιο υποκείμενο. Στο παράδειγμα που αναλύουμε παρακάτω (Εικόνα 31) τα υποκείμενα είναι τριάντα χώρες του κόσμου και οι παρατηρήσεις είναι το προσδόκιμο ζωής των ανδρών και των γυναικών από κάθε χώρα. Ένα άλλο παράδειγμα θα μπορούσε να ήταν ο έλεγχος μίας ιατρικής εξέτασης σε τριάντα ασθενείς πριν και μετά τη λήψη ενός φαρμάκου κλπ.

Με τον έλεγχο ισότητας μέσης τιμής ζευγαρωτών παρατηρήσεων ελέγχουμε αν η μέση τιμή της μίας μέτρησης είναι σημαντικά διαφορετική από τη μέση τιμή της δεύτερης μέτρησης.

Για το T – Test δύο εξαρτημένων δειγμάτων η αρχική υπόθεση είναι η

H_0 : οι μέσες τιμές των πληθυσμών είναι ίσες ($H_0 : \mu_1=\mu_2$),

ενώ η εναλλακτική υπόθεση είναι η

H_1 : οι μέσες τιμές των πληθυσμών δεν είναι ίσες, είναι σημαντικά διαφορετικές ($H_1 : \mu_1\neq\mu_2$)

Η απόφαση λαμβάνεται βάσει της τιμής της πιθανότητας p η οποία με τη σειρά της υπολογίζεται βάσει της τιμής που λαμβάνει το στατιστικό

$$t = \sqrt{N}\frac{\bar{X}_\delta}{s_\delta}$$

όπου

1. N το πλήθος των παρατηρήσεων
2. $\bar{X}_\delta$ η μέση τιμή της διαφοράς των παρατηρήσεων
3. s_δ η τυπική απόκλιση της διαφοράς των παρατηρήσεων

Αποδεικνύεται πως το στατιστικό t ακολουθεί την κατανομή Student με N-1 βαθμούς ελευθερίας. Χρησιμοποιώντας την πληροφορία αυτή μπορούμε και υπολογίζουμε την πιθανότητα

$p = P(t > t_{\delta\epsilon\acute{\iota}\gamma\mu\alpha} \,||\, t \sim T_{N-1})$

η οποία αποτελεί και το κριτήριο αποδοχής ή απόρριψης της H_0.

Δεδομένα		
Χώρα	**Προσδόκιμο ζωής γυναικών**	**Προσδόκιμο ζωής ανδρών**
Αίγυπτος	63	60
Αυστρία	79	73
Αφγανιστάν	44	45
Βέλγιο	79	73
Βολιβία	64	59
Δομινικανή Δημοκρατία	70	66
Ελ Σαλβαδόρ	69	64
Ελλάδα	80	75
Ζάμπια	45	44
Ινδία	59	58
Ισημερινός	73	67
Καμερούν	58	55
Καναδάς	81	74
Κίνα	69	67
Κουβέιτ	78	73
Λευκορωσία	76	66
Λιθουανία	77	68
Μαλαισία	72	66
Μποτσουάνα	66	60
Νησιά Μπαρμπάντος	78	73
Νιγηρία	57	54
Νικαράγουα	67	61
Ολλανδία	81	75
Περού	67	63
Ρωσία	74	64
Σενεγάλη	58	55
Σομαλία	55	54
Τανζανία	45	41
Τσεχία	77	69
Χιλή	78	71

Εικόνα 31: Δεδομένα στατιστικού ελέγχου

Στην εικόνα 31 εμφανίζεται το προσδόκιμο ζωής των γυναικών και των ανδρών όπως μετρήθηκε το 1995 σε τριάντα τυχαία επιλεγμένες χώρες του κόσμου.

	Γυναίκες	Άνδρες
Μέση Τιμή	67,97	63,1
Τυπική Απόκλιση	11,12	9,32

Εικόνα 32: Προσδόκιμο ζωής

Υπολογίζοντας τη μέση τιμή (Εικόνα 32) βρίσκουμε πως οι γυναίκες των χωρών του

δείγματος ζούνε κατά μέσο όρο 4,87 έτη παραπάνω από τους άνδρες των χωρών του δείγματος (Εικόνα 33). Άμεσα, τίθεται το ερώτημα αν η διαφορά αυτή είναι αρκετά μεγάλη ώστε να συμπεράνουμε πως οι γυναίκες του συνόλου των χωρών ζούνε παραπάνω από τους άνδρες του συνόλου των χωρών.

Διαφορά μέσων τιμών	4,87

Εικόνα 33: Δειγματική διαφορά

Η πιθανότητα p υπολογίζεται χρησιμοποιώντας τη συνάρτηση **TTEST()** η οποία παίρνει τέσσερα ορίσματα από τα οποία τα δύο πρώτα είναι τα κελιά όπου περιέχονται τα δεδομένα του πρώτου και του δεύτερου δείγματος, στο τρίτο καταχωρείται το είδος του στατιστικού ελέγχου (1: μονόπλευρο ή 2:δίπλευρο) ενώ στο τελευταίο τοποθετείται ο τύπος του T-Test που θα εφαρμοστεί. Στην περίπτωση του ελέγχου ζευγαρωτών παρατηρήσεων η επιλογή που πρέπει να τοποθετηθεί στην τέταρτη παράμετρο είναι ο αριθμός 1.

Πιθανότητα p	0,00000000005

Εικόνα 34: Πιθανοφάνεια της διαφοράς

Το αποτέλεσμα της εφαρμογής της συνάρτησης TTEST() είναι ο μικροσκοπικός αριθμός που φαίνεται στην εικόνα 34 (έγινε ορατός επιλέγοντας την εμφάνιση έντεκα δεκαδικών ψηφίων στο κελί αυτό!), ο οποίος σημαίνει πως η πιθανότητα εμφάνισης της διαφοράς των 4,87 ετών μεταξύ του προσδόκιμου ζωής ανδρών και γυναικών λόγω του σφάλματος της τυχαίας επιλογής των τριάντα χωρών είναι μόλις 0,00000000005 ή 0,000000005%. Καθώς η πιθανότητα αυτή είναι εξαιρετικά μικρή έχουμε κάθε δικαίωμα να απορρίψουμε την υπόθεση πως οι μέσες τιμές των δύο εξαρτημένων μεταβλητών είναι ίσες, ισοδύναμα μπορούμε με μεγάλη ασφάλεια να καταλήξουμε στο συμπέρασμα πως οι γυναίκες σε όλες τις χώρες του κόσμου ζούνε περισσότερο από τους άνδρες σε όλες τις χώρες.

Πίνακας 6.8: Δοκιμασία t-test ζευγαρωτών παρατηρήσεων

Όπως περιγράφεται στην παράγραφο 6.1.4.

```
x = c(63, 79, 44, 79, 64, 70, 69, 80, 45, 59, 73, 58, 81, 69, 78, 76, 77, 72, 66, 78, 57, 67, 81, 67, 74, 58, 55, 45, 77, 78)
y = c(60, 73, 45, 73, 59, 66, 64, 75, 44, 58, 67, 55, 74, 67, 73, 66, 68, 66, 60, 73, 54, 61, 75, 63, 64, 55, 54, 41, 69, 71)
t.test(x, y, paired = TRUE)
```

Πίνακας 6.9: Γενίκευση : Περισσότερες από δύο επαναλαμβανόμενες μετρήσεις (Repeated measures)

Παράδειγμα : 10 μαθητές εξετάστηκαν στα μαθηματικά στην αρχή, στο μέσο και στο τέλος του σχολικού έτους. Αναζητούμε διαφοροποίηση μεταξύ των τριών μετρήσεων.

```
math1 = c(10, 8, 11, 15, 18, 20, 13, 14, 14, 11)
math2 = c(9, 11, 13, 14, 20, 20, 15, 15, 13, 12)
math3 = c(13, 12, 13, 15, 20, 19, 16, 16, 12, 10)
data = data.frame(m1 = math1, m2 = math2, m3 = math3)
exetaseis = c("1η Περιόδος", "2η Περίοδος", "3η Περίοδος")
exetaseisf = as.factor(exetaseis)
vathmoiexetasewn = data.frame(exetaseisf)
vathmoibind = cbind(data$m1, data$m2, data$m3)
vathmoiModel <- lm(vathmoibind ~ 1)
summary(vathmoiModel)
library(car)
analysis <- Anova(vathmoiModel, idata = vathmoiexetasewn, idesign = ~exetaseisf)
summary(analysis)
```

6.1.5 Πιθανά σφάλματα στους ελέγχους υποθέσεων

Καθώς η αποδοχή ή η απόρριψη της H_0 βασίζεται σε μία πιθανότητα υπάρχει πάντα η περίπτωση να πάρουμε λάθος απόφαση. Υπάρχουν δύο περιπτώσεις σφάλματος. Η πρώτη (Σφάλμα τύπου I) συμβαίνει όταν η μηδενική υπόθεση H_0 απορρίπτεται από το στατιστικό έλεγχο ενώ στην πραγματικότητα είναι σωστή (εσφαλμένη απόρριψη, false

positive). Με απλά λόγια, οι μέσες τιμές των πληθυσμών είναι ίσες αλλά ο στατιστικός έλεγχος καταδεικνύει πως αυτό δεν ισχύει, δηλαδή πως αυτές δεν είναι ίσες! Το δεύτερο (Σφάλμα τύπου ΙΙ) συμβαίνει όταν η μηδενική υπόθεση H_0 γίνεται αποδεκτή ενώ δεν ισχύει (εσφαλμένη αποδοχή, false negative), δηλαδή οι μέσες τιμές των πληθυσμών είναι στην πραγματικότητα σημαντικά διαφορετικές και ο στατιστικός έλεγχος αποδέχεται την ισότητά τους.

Οι δύο τύποι σφάλματος έχουν αντιστρόφως ανάλογη πιθανότητα εμφάνισης. Η ελαχιστοποίηση του σφάλματος ενός τύπου μεγαλώνει την πιθανότητα εμφάνισης του σφάλματος του άλλου τύπου.

Καθώς, το σφάλμα τύπου Ι είναι ελεγχόμενο από τον ερευνητή ενώ αυτό του τύπου ΙΙ δεν είναι, έχει επικρατήσει στις περισσότερες επιστημονικές περιοχές το σφάλμα αυτό να ορίζεται στο 5% = 0,05. Το σφάλμα τύπου Ι συμβολίζεται συνήθως με **α**, ενώ το σφάλμα τύπου ΙΙ συμβολίζεται με **β.**

6.2 Παρουσιάζοντας τα αποτελέσματα ενός t -test ή μίας ANOVA

Ενδεικτική αναφορά One – Sample T – test

“Η δοκιμασία Student για ένα δείγμα κατέδειξε πως το μέσο βάρος της παραγωγής ήταν μικρότερο από 500 γραμμάρια (M = 495,5, SD = 3.70), t(30) = 8.01, *p* < .001, d = 1.44).”

Ενδεικτική αναφορά Independent – Samples T – test

“Ο έλεγχος t – test δύο ανεξαρτήτων δειγμάτων κατέδειξε πως οι γυναίκες (M = 27.0, SD = 7.21) είχαν σημαντικά υψηλότερο σκορ στο άγχος από ότι οι άνδρες (M = 24.2, SD = 7.69), t(734) = 4.30, *p* < .001, d = 0.35. ”

“Το συνολικό σκορ που συγκέντρωσαν οι γυναίκες (M = 27.0, SD = 7.21) ήταν σημαντικά υψηλότερο από αυτό των ανδρών (M = 24.2, SD = 7.69), t(340) = 4.30, *p* < .001, d = 0.35. Η δοκιμασία Levene κατέδειξε άνισες διακυμάνσεις (F = 3.56, *p* = .043), κάτι που είχε ως συνέπεια τη διόρθωση των βαθμών ελευθερίας από 734 σε 340.”

Ενδεικτική αναφορά Paired Sample T – test

“Η δοκιμασία paired samples t – test κατέδειξε πως το σκορ της κλίμακας άγχους μετά τη στρεσογόνο εμπειρία (M = 26.4, SD = 7.41) ήταν σημαντικά μεγαλύτερο από το σκορ πριν την εμπειρία αυτή (M = 18.0, SD = 9.49), t(721) = 23.3, $p < .001$, d = 0.87. ”

Ενδεικτική αναφορά ANOVA

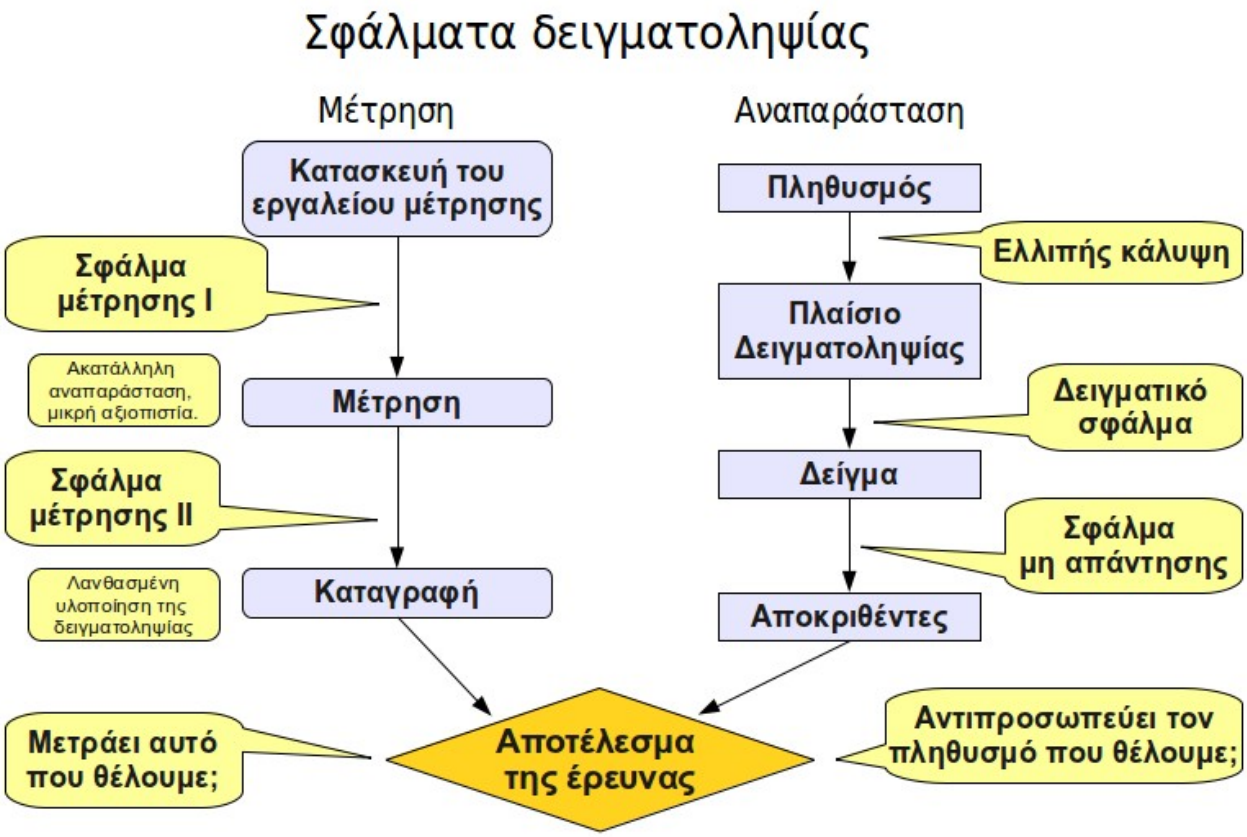
“Η ανάλυση διακύμανσης (ANOVA) έδειξε επίδραση του σχολείου φοίτησης στην επίδοση της γραπτής δοκιμασίας στα μαθηματικά F(2, 1279) = 6.15, $p = .002$, $\eta_p^2 = .010$. Η Post-hoc ανάλυση με τη δοκιμασία Tukey’s HSD κατέδειξε πως οι μαθητές των σχολείων της πόλης συγκέντρωσαν μεγαλύτερο σκορ από τους μαθητές της επαρχίας.”

Κεφάλαιο 7 Συνηθισμένα σφάλματα

Ανάλογα με το είδος τους, τα σφάλματα διακρίνονται σε (α) Σφάλματα δειγματοληψίας, (β) Στατιστικά σφάλματα, (γ) Σφάλματα ερμηνείας των αποτελεσμάτων και (δ) Σφάλματα παρουσίασης.

7.1 Σφάλματα δειγματοληψίας

Στο διάγραμμα 1 παρουσιάζονται τα σφάλματα που επηρεάζουν την αντιπροσωπευτικότητα ενός δείγματος.



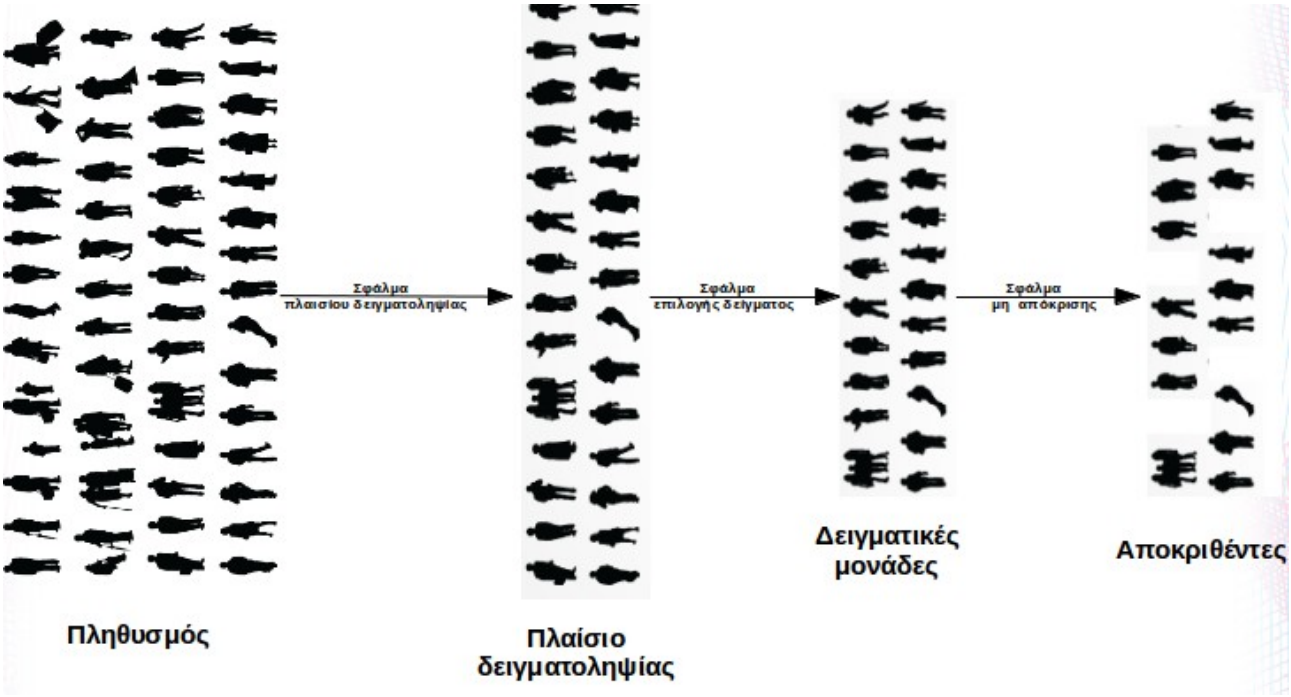
Διάγραμμα 38: Σφάλματα δειγματοληψίας

7.1.1 Σφάλματα μέτρησης

Σφάλμα μέτρησης I : Κάθε μέτρηση γίνεται με κάποιο εργαλείο όπως για παράδειγμα η μεζούρα για το μήκος ή η ζυγαριά για το βάρος. Το σφάλμα του εργαλείου μέτρησης επηρεάζει άμεσα τα συμπεράσματα της έρευνάς μας. Ιδιαίτερα, αυτό ισχύει για τις κοινωνικές επιστήμες όπου έννοιες όπως το άγχος, η επιθετικότητα κ.α. “μετρούνται” με ερωτηματολόγια (ή κλίμακες).

Σφάλμα μέτρησης II : Το εργαλείο το χειρίζεται ένας άνθρωπος ο οποίος είναι πιθανό να μην είναι πολύ προσεκτικός όταν κάνει τη μέτρηση και να εισάγει σφάλμα στις τιμές από τις οποίες θα προκύψουν τα συμπεράσματα της έρευνας. Ιδιαίτερα, όταν μετρούνται ψυχολογικά χαρακτηριστικά, η μεταβολή του περιβάλλοντος στο οποίο γίνεται η καταγραφή επηρεάζει τις αποκρίσεις των ερωτώμενων.

7.1.2 Σφάλματα αναπαράστασης του πληθυσμού



7.2 Στατιστικά σφάλματα

Δεν υπάρχει ένας τρόπος για να βρούμε τη “μέση τιμή” η οποία μπορεί να οριστεί ισοδύναμα ως η ποσότητα που πρέπει να υποκαταστήσει κάθε μία παρατήρηση ώστε το σύνολο να εξακολουθεί να έχει το ίδιο τελικό “αποτέλεσμα”. Ανάλογα με το είδος των μονάδων κάθε μεταβλητής υπολογίζουμε

- Αριθμητικό μέσο για καθαρά ποσά (π.χ. το μέσο βάρος δύο ανθρώπων που ζυγίζουν 50 και 70 κιλά αντίστοιχα, είναι τα 60 κιλά)

- Γεωμετρικό μέσο για ποσοστιαίες μεταβολές (π.χ. Αν ένα ομόλογο το 2008 είχε απόδοση 15% το 2009 είχε απόδοση -10% και το 2010 είχε απόδοση -5% η μέση του απόδοση στην τριετία είναι (1,15*0,90*0,95)1/3 = 1,675%

- Αρμονικό μέσο για ρυθμούς συχνοτήτων (π.χ. αν ταξιδέψω από Ξάνθη στη Θεσσαλονίκη και 100χμ/ώρα και επιστρέψω με 120χμ/ώρα τότε η μέση ταχύτητα του ταξιδιού είναι 2*(1/100 + 1/120)-1= 109,1χμ/ώρα)

7.3 Σφάλματα ερμηνείας των αποτελεσμάτων

- Σύγχυση μεταξύ στατιστικής σημαντικότητας και αιτιότητας. (παράδειγμα τέτοιας σύγχυσης : το να συμπεράνεις πως η κατασκευή σχολείων θα μειώσει τη βρεφική θνησιμότητα γιατί βρήκες πως το ποσοστό των ανθρώπων που γνωρίζουν γραφή και ανάγνωση είναι αρνητικά συσχετισμένο με αυτή την ποσότητα.)

- Στο ίδιο πνεύμα : σύγχυση μεταξύ συντελεστή συσχέτισης και αιτιότητας. (το να συμπεράνεις πως το βάρος του αυτοκινήτου είναι ανάλογο με την τελική του ταχύτητα – σύμπτωση καθώς οι βιομηχανίες βάζουν μεγάλες μηχανές στα μεγάλα αυτοκίνητα)

- Σύγχυση μεταξύ μη στατιστικής σημαντικότητας και μη αιτιότητας. (το να υποθέτεις πως ένα μη στατιστικά σημαντικό αποτέλεσμα σημαίνει μη αιτιότητα είναι ισοδύναμο με το να πηδάς από ένα καράβι στη θάλασσα και να συμπεραίνεις πως δεν έχεις βάρος καθώς δεν παρατήρησες κάποια αλλαγή στη στάθμη του πλοίου!)

7.4 Η Επίδραση της παλινδρόμησης και οι παρερμηνείες στις οποίες οδηγεί (Regression Effect και Regression Fallacy)

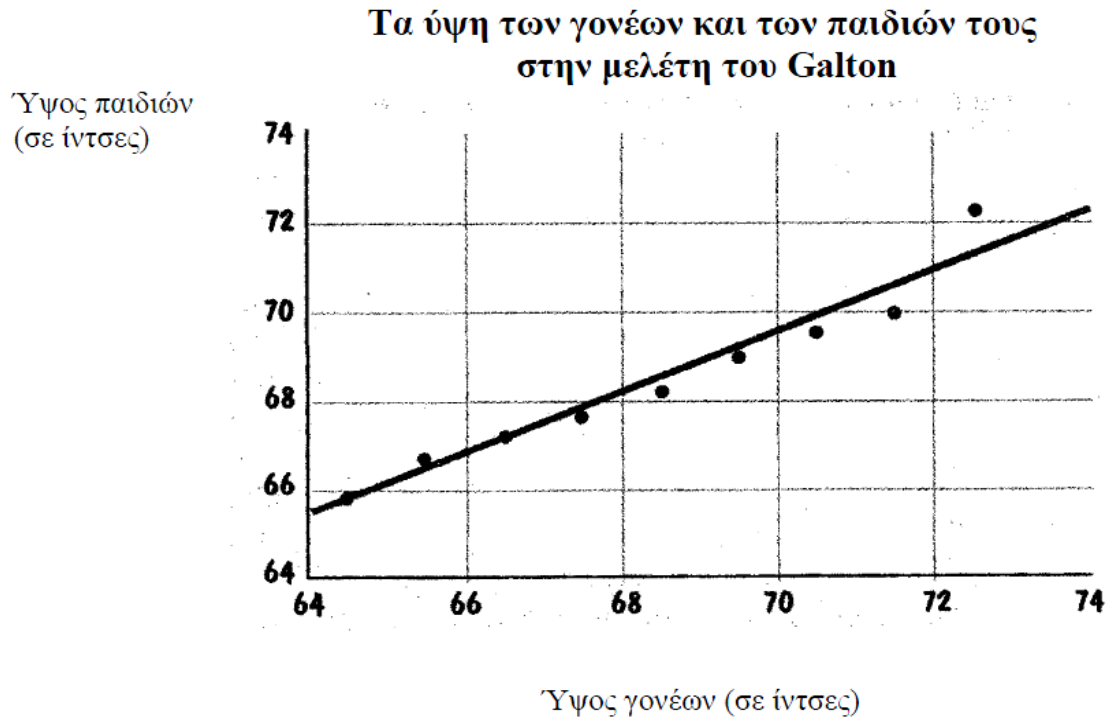
Είναι γνωστό πως η πρώτη προσπάθεια για τη μελέτη της σχέσης μεταξύ δύο μεταβλητών έγινε από τον Francis Galton (Αγγλία, 1822 - 1911) για την μελέτη της σχέσης του ύψους των παιδιών με τους γονείς τους. Από την μελέτη αυτή προήλθε και ο όρος παλινδρόμηση (regression) που ουσιαστικά αναφέρεται στην παλινδρόμηση προς την κατεύθυνση του μέσου (regression towards the mean).

Ο όρος “παλινδρόμηση προς την κατεύθυνση του μέσου” (regression towards the mean) προήλθε από την παρατήρηση του Galton ότι υπάρχει μια τάση όπου ακραίες (ως προς το μέσο τους) παρατηρήσεις της ανεξάρτητης τ.μ. αντιστοιχούν σε παρατηρήσεις της εξαρτημένης τ.μ. που δεν είναι το ίδιο ακραίες αλλά είναι πλησιέστερα προς τον μέσο τους.

Με απλούστερο τρόπο μπορεί να πει κανείς ότι ακραίες παρατηρήσεις ακολουθούνται από

λιγότερο ακραίες παρατηρήσεις (παρατηρήσεις που είναι πλησιέστερα προς το "κέντρο"). Αυτό κάνει το διάγραμμα διασποράς να έχει την μορφή μπάλας του αμερικανικού ποδοσφαίρου.

Από την μελέτη των δεδομένων ο Galton παρατήρησε ότι, ασυνήθιστα υψηλοί γονείς τείνουν να έχουν παιδιά χαμηλότερα από τους ίδιους ενώ, ασυνήθιστα χαμηλοί γονείς έχουν συνήθως υψηλότερα παιδιά (διάγραμμα 39).



Διάγραμμα 39: Τα ύψη των γονέων και των παιδιών τους

Μία συνήθης παρερμηνεία των παραπάνω παρατηρήσεων είναι η συνεπαγωγή πως όλοι οι άνθρωποι θα έχουν σε κάποια μελλοντική στιγμή το ίδιο ύψος. Αυτό φανερά όμως δεν ισχύει.

Επιπλέον, αν συνέβαινε κάτι τέτοιο θα μπορούσε κανείς να αντιστρέψει την επιχειρηματολογία παρατηρώντας ότι πάρα πολύ υψηλοί άνθρωποι έχουν γονείς κάπως χαμηλότερους από αυτούς ενώ πάρα πολύ χαμηλοί άνθρωποι έχουν κάπως υψηλότερους γονείς και να καταλήξει έτσι στο συμπέρασμα πως τα ύψη των ανθρώπων αποκλίνουν.

Που είναι το σφάλμα;

Η παρερμηνεία αυτή είναι μια λανθασμένη συλλογιστική, και οφείλεται στο φαινόμενο της παλινδρόμησης προς την κατεύθυνση του μέσου, (regression towards the mean) είναι δε ακριβώς η παρερμηνεία της προσωρινής φύσης μιας ακραίας παρατήρησης και ο χαρακτηρισμός της ως τάσης. Η κατάσταση που προκύπτει αποδίδεται στην επίδραση της παλινδρόμησης (regression effect).

Αυτό που θα πρέπει να αντιληφθούμε είναι ότι τα ύψη των ανθρώπων επηρεάζονται από τυχαίους παράγοντες και ότι, για ανθρώπους που είναι εξαιρετικά υψηλοί οι τυχαίοι παράγοντες επηρέασαν θετικά το ύψος τους και το έκαναν μεγαλύτερο από ότι αναμενόταν με βάση τα γονίδια τους. Ενδιαφέρον είναι το γεγονός πως πολλοί έχουν πέσει στο παραπάνω σφάλμα...

■ Σύμφωνα με μία μελέτη[2] που έγινε στην Αμερική παιδιά ηλικία τεσσάρων ετών με IQ 120 συνήθως, όταν ενηλικιωθούν, επιτυγχάνουν σκορ στο IQ τεστ περίπου 110. Παρομοίως, παιδιά τεσσάρων ετών με IQ σκορ 70 έχουν ένα μέσο σκορ στο IQ τεστ όταν ενηλικιωθούν 85. Αυτό δεν συνεπάγεται ότι θα υπάρχουν λιγότεροι ενήλικες απ' ότι παιδιά με πολύ υψηλά ή πολύ χαμηλά αποτελέσματα στο IQ τεστ. Παρότι όσοι άνθρωποι ξεκινούν στην παιδική ηλικία με υψηλό ή χαμηλό IQ σκορ, συνήθως, θα παλινδρομήσουν προς την κατεύθυνση του μέσου, οι θέσεις τους θα παρθούν (θα αντικατασταθούν) από άλλους οι οποίοι στην παιδική τους ηλικία θα έχουν IQ σκορ πλησιέστερα προς τον μέσο.

■ Ένα άλλο παράδειγμα λανθασμένης ερμηνείας φαινομένων που οφείλονται στην παλινδρόμηση προς την κατεύθυνση του μέσου εμφανίζεται στην αξιολόγηση των φοιτητών.

Έχει παρατηρηθεί ότι οι φοιτητές εκείνοι οι οποίοι έχουν τους υψηλότερους βαθμούς στις εξετάσεις προόδου συνήθως, δεν αποδίδουν εξίσου καλά στην τελική εξέταση ενώ, εκείνοι

2. *Christopher Jencks, Marshall Smith, Henry Acland, Nelly Jo Bane, David Cohen, Herbert Gintis,Barbara Heyns and Stephen Michelson (1972) in: A Quality Reassessment of the Effect of Family and Schooling in America, New York: Basic Books, p.59*

οι οποίοι έχουν χαμηλή βαθμολογία στην εξέταση προόδου, πολλές φορές βελτιώνουν την απόδοσή τους στην τελική εξέταση.

Θα μπορούσε αυτό να εκληφθεί ως ένδειξη ότι η απόδοση των φοιτητών συγκλίνει προς μια ανησυχητική μετριότητα με τους ασθενείς φοιτητές να βελτιώνονται και τους καλούς φοιτητές να χειροτερεύουν;

Ή, αντιστρέφοντας το προηγούμενο επιχείρημα, το γεγονός ότι αυτοί που πέτυχαν την υψηλότερη βαθμολογία στην τελική εξέταση δεν απέδωσαν εξίσου καλά στην εξέταση προόδου σημαίνει ότι η απόδοση αποκλίνει από τον μέσο; Και στις δύο περιπτώσεις η απάντηση είναι αρνητική.

■ Ένας εκπαιδευτής πιλότων παρατήρησε ότι πολύ καλές προσγειώσεις συνήθως, ακολουθούνται από προσγειώσεις που δεν είναι εξίσου καλές, ενώ μέτριες προσγειώσεις ακολουθούνται, συνήθως από καλύτερες.

Υποπίπτοντας στην λανθασμένη προσέγγιση που οφείλεται στην παρερμηνεία της παλινδρόμησης στην κατεύθυνση του μέσου ο εκπαιδευτής ισχυρίστηκε ότι η ακολουθία αυτή συμβαίνει γιατί συνήθιζε να επαινεί τις καλές προσγειώσεις και να κριτικάρει έντονα τις μέτριες.

Για το λόγο αυτό έβγαλε το συμπέρασμα, σε αντίθεση από την κοινά αποδεκτή άποψη με βάση την έρευνα για την μαθησιακή διδασκαλία, ότι ο έπαινος έχει αρνητικά αποτελέσματα στην προσπάθεια ενώ η έντονη κριτική έχει θετικά αποτελέσματα[3].

■ Ένα χαρακτηριστικό παράδειγμα του προβλήματος στον τομέα των οικονομικών δίνεται στο βιβλίο με τον προκλητικό τίτλο “Ο Θρίαμβος της Μετριότητας στις Επιχειρήσεις" (The Triumph of Mediocrity in Business”)[4].

Ο συγγραφέας «ανακάλυψε» ότι επιχειρήσεις με εξαιρετικά υψηλά κέρδη σε κάθε δεδομένη

3 *Amos Tversky and Daniel Kahmeman (1973) “On the Psychology of Prediction” Psychological Review 1973 vol. 80, 237-251*

4 *Horace Secrist : “The Triumph of Mediocrity in Business”, 1933.*

χρονιά έχουν χαμηλότερα κέρδη την επόμενη χρονιά ενώ επιχειρήσεις με πολύ χαμηλά κέρδη, εν γένει επιτυγχάνουν καλύτερα αποτελέσματα το επόμενο έτος.

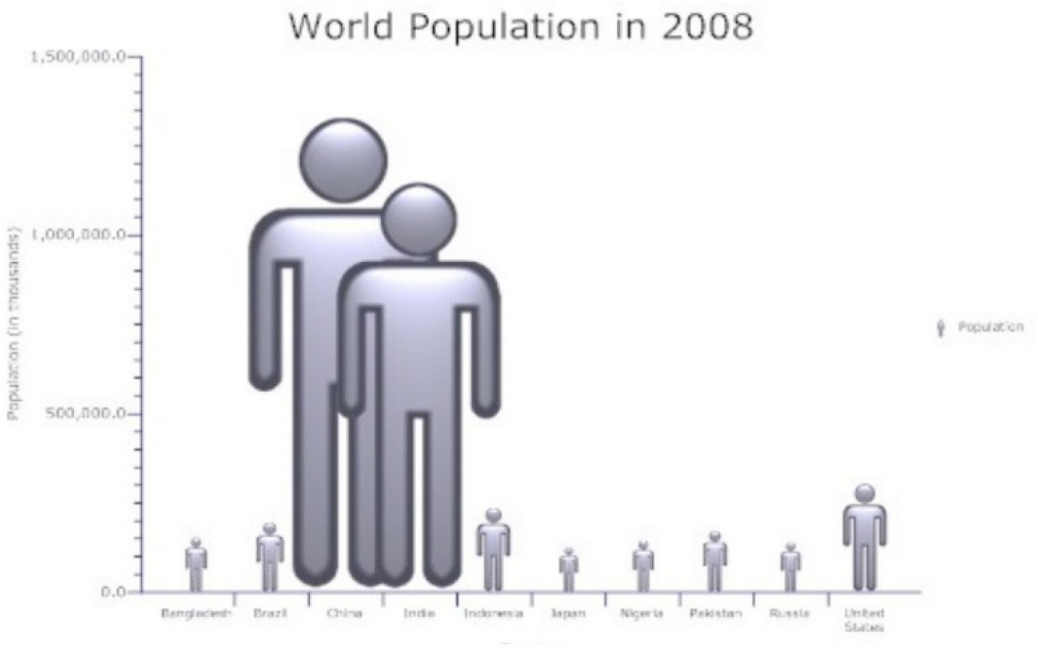
Με αυτές τις ενδείξεις κατέληξε στο συμπέρασμα ότι οι ισχυρές επιχειρήσεις γίνονται ασθενέστερες ενώ οι ασθενείς γίνονται ισχυρότερες με αποτέλεσμα σύντομα να γίνουν όλες οι επιχειρήσεις μεσαίου μεγέθους! Η λανθασμένη προσέγγιση του συγγραφέα είναι προφανής.

■ Στο Ηνωμένο Βασίλειο τοποθετήθηκαν κάμερες ταχύτητας σε σημεία όπου είχαν παρατηρηθεί πολλά ατυχήματα, έχοντας την πεποίθηση πως η ύπαρξη τους θα μείωνε σημαντικά τα ατυχήματα. Κάποιο διάστημα αργότερα πρόσεξαν πως αν και μειώθηκαν οι απώλειες ανθρώπων δεν υπήρξε παντού σημαντική μείωση στα ατυχήματα.

Κάποιοι Στατιστικοί υποστήριξαν πως το σφάλμα κεντρική τάσης είχε υπερεκτιμηθεί και πως ενδεχομένως ένα μέρος των χρημάτων που διατέθηκαν θα έδινε καλύτερο αποτέλεσμα αν διαθετόταν για άλλο σκοπό[5] .

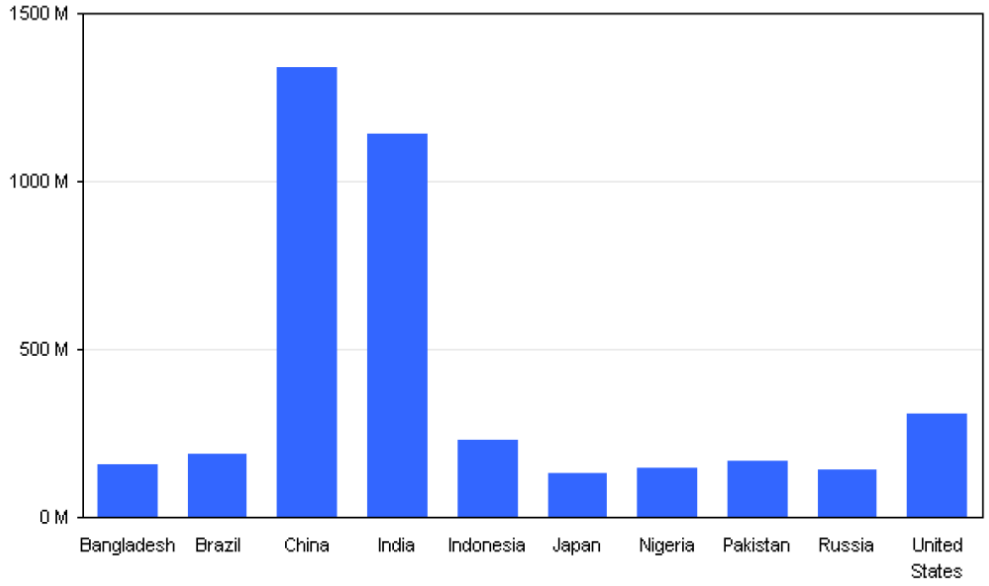
7.5 Σφάλματα παρουσίασης (εκούσια ή ακούσια!)

Λανθασμένη χρήση διαγραμμάτων όπως για παράδειγμα χρήση δισδιάστατων ή τρισδιάστατων εικόνων για να παρουσιάσουμε τη διαφορά μεταξύ δύο μονοδιάστατων μεγεθών. Που είναι το σφάλμα στο διάγραμμα 40, σελίδα 185, και γιατί το διάγραμμα 41, σελίδα 186, είναι περισσότερο κατάλληλο για την περιγραφή των ίδιων δεδομένων;



Διάγραμμα 40: Παγκόσμιος πληθυσμός το 2008 (α)

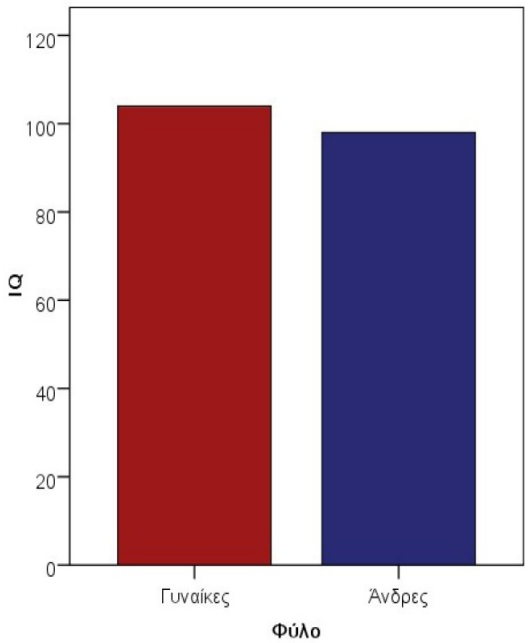
5 *http://w*

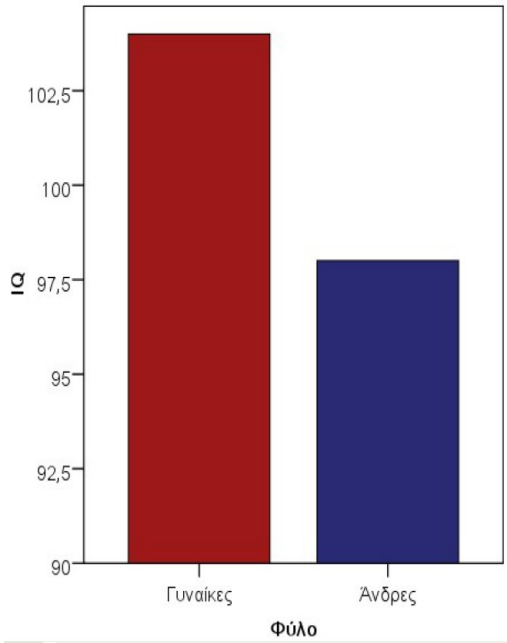


Διάγραμμα 41: Παγκόσμιος πληθυσμός το 2008 (β)

7.5.1 Λανθασμένη αρχή στον άξονα Υ σε ραβδόγραμμα.

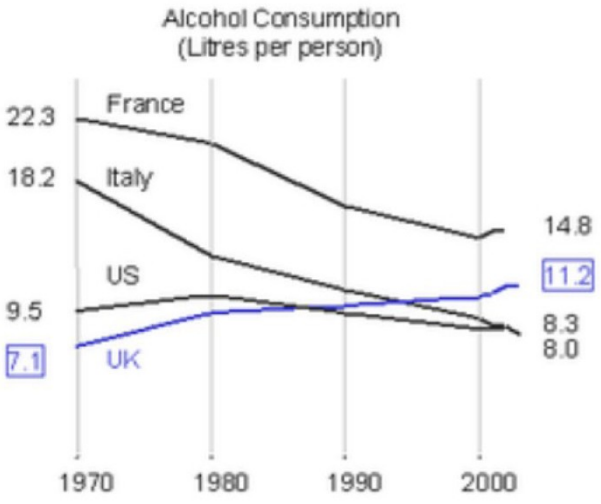
Τα επόμενα δύο ραβδογράμματα παρουσιάζουν τα ίδια στοιχεία αλλά δημιουργούν πολύ διαφορετικές εντυπώσεις. Μπορείτε να βρείτε γιατί;

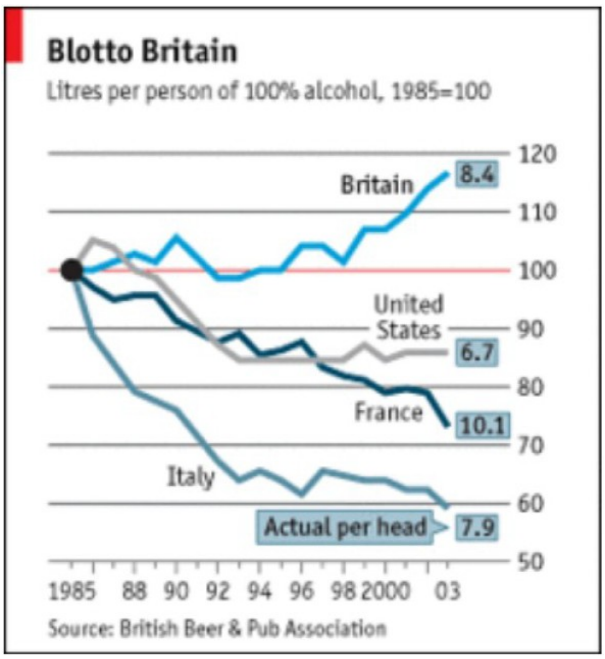




7.5.2 Λάθος επιλογή διαγράμματος.

Τα παρακάτω δύο διαγράμματα προσπαθούν να περιγράψουν τα ίδια δεδομένα. Ποιο τα καταφέρνει καλύτερα;





Ευρετήριο πινάκων

Κατάλογος εικόνων

Διαδικασίες στο Calc	
Θέμα	**Μέθοδος**
Κανονική κατανομή	Συνάρτηση **NORMDIST(αριθμός; μ; σ).** Το αποτέλεσμα είναι το εμβαδόν της γραφικής παράστασης της κανονικής κατανομής με μέση τιμή μ, τυπική απόκλιση σ από την ευθεία x = αριθμός και *αριστερά* αυτής. Για παράδειγμα αν γνωρίζεις πως το βάρος του ενήλικου πληθυσμού ακολουθεί κανονική κατανομή με μέση τιμή 65 κιλά και τυπική απόκλιση 15 κιλά και θέλεις να βρεις την πιθανότητα ένας άνθρωπος να έχει βάρος *μικρότερο* από 72 κιλά τότε αρκεί να υπολογίσεις NORMDIST(72; 65; 15) = 0,6796 ή περίπου 68%. Αντίστοιχα αν θέλεις να βρεις την πιθανότητα ένας άνθρωπος να έχει βάρος *μεγαλύτερο* από 72 κιλά θα υπολογίσεις 1 - NORMDIST(72; 65; 15) = 0,3203 ή περίπου 32%. Αντίστροφα, αν πρέπει να υπολογιστεί το σημείο του πραγματικού άξονα που αφήνει αριστερά του συγκεκριμένο μέρος του εμβαδού τότε μπορεί να χρησιμοποιηθεί η συνάρτηση **NORMINV(αριθμός; μ; σ)** π.χ. Η NORMINV(0,005;0;1) επιστρέφει -2,5758... Εύκολα καταλαβαίνουμε πως λόγω συμμετρίας μεταξύ -2,5758 και 2,5758 βρίσκεται το 99% του συνολικού εμβαδού της καμπύλης.
Διάστημα εμπιστοσύνης	Δεν υπάρχει δεσμευμένη συνάρτηση για το διάστημα εμπιστοσύνης. Μπορεί, να γίνει με συνδυασμό των συναρτήσεων AVERAGE(), STDEV() και SQRT().
Πίνακας συχνοτήτων	Υποθέτουμε πως τα αριθμητικά δεδομένα είναι στα κελιά A1:A50 Βήμα 1ο : Ορισμός “με το χέρι” των επιθυμητών ορίων των διαστημάτων σε ένα μέρος του φύλλου εργασίας. Υποθέτουμε πως τοποθετούμε στα κελιά B1:B4 τους αριθμούς 10, 20, 30, 40. Βήμα 2ο : Σε μία άλλη στήλη επιλέγουμε 5 κελιά (1 περισσότερα από το πλήθος των διαστημάτων) και με τον οδηγό συνάρτησης εισάγουμε τη συνάρτηση – πίνακα FREQUENCY με ορίσματα A1:A50 και B1:B4. Τα διαστήματα είναι της μορφής (-∞, 10], (10, 20], (20, 30], (30, 40] και (40, ∞) Η συμπλήρωση πίνακα συχνοτήτων ποιοτικής μεταβλητής περιγράφεται στην παράγραφο 2.3, σελίδα 45.
Επικρατούσα τιμή	Συνάρτηση **MODE()**. Αν υπάρχουν περισσότερες από μία τότε επιστρέφει τη μικρότερη (σε αντίθεση με το Excel που επιστρέφει τη μεγαλύτερη)
Διάμεση τιμή	Συνάρτηση **MEDIAN()**.
Αριθμητικός μέσος	Συνάρτηση **AVERAGE().**
Αρμονικός	Συνάρτηση **HARMEAN()**

μέσος	
Γεωμετρικού μέσος	Συνάρτηση **GEOMEAN()**
Ιστόγραμμα	Για το ιστόγραμμα δεν υπάρχει δεσμευμένη διαδικασία στο Calc. Μπορεί να χρησιμοποιηθεί ένα ραβδόγραμμα που θα αναπαριστά τις συχνότητες των αριθμητικών κλάσεων όπως αυτές έχουν οριστεί στο σχετικό πίνακα συχνοτήτων (Πίνακας 2.3, σελίδα 38) και αποτέλεσμα όπως αυτό του διαγράμματος 4,σελίδα 52.
Εύρος	Δεν υπάρχει ενσωματωμένη συνάρτηση, ωστόσο μπορεί εύκολα να υπολογιστεί ως **Max() - Min()**
Ενδοτεταρτημοριακό εύρος	Συνάρτηση **QUARTILE(δεδομένα; α)**. Αν α = 0 τότε επιστρέφει την ελάχιστη τιμή ενώ αν α = 4 επιστρέφει τη μέγιστη. Για α = 1 επιστρέφει το Q_1, για α = 2 επιστρέφει τη διάμεσο και για α = 3 επιστρέφει το Q_3.
Μέτρα διασποράς	Οι συναρτήσεις **VAR** και **VARP** υπολογίζουν για την διασπορά των παρατηρήσεων ενώ οι συναρτήσεις **STDEV** και **STDEVP** υπολογίζουν τη διακύμανση. Η διαφορά της VAR από τη VARP είναι ο παρονομαστής με τον οποίο διαιρείται το άθροισμα τετραγωνικών αποκλίσεων : στη VAR είναι ο (n – 1) ενώ στη VARP είναι ο n (n = πλήθος παρατηρήσεων). Δηλαδή VAR() = DEVSQ() / (COUNT() - 1) ενώ VARP() = DEVSQ() / (COUNT())
Συντελεστής ομοιογένειας	Δεν υπάρχει αντίστοιχη συνάρτηση αλλά μπορεί εύκολα να υπολογιστεί ως **STDEVP() / AVERAGE()**
Ασυμμετρία και κυρτότητα	**SKEW()** και **KURT()** Για τον υπολογισμό της KURT απαιτούνται τουλάχιστον 4 παρατηρήσεις.
Τυποποιημένες τιμές	Συνάρτηση **STANDARDIZE()** η οποία παίρνει 3 ορίσματα, την τιμή προς τυποποίηση, την μέση τιμή του δείγματος και την τυπική απόκλιση αυτού.
Συνδιακύμανση	Συνάρτηση **COVAR(δεδομένα1;δεδομένα2)**
Συντελεστής συσχέτισης	Ο συντελεστής συσχέτισης του Pearson μπορεί να υπολογιστεί από τη συνάρτηση **PEARSON(δεδομένα1;δεδομένα2).** Ο συντελεστής του Spearman ή του Kendall δεν μπορεί να υπολογιστεί άμεσα από το Calc.

Διαδικασίες στο R - Project	
Θέμα	**Μέθοδος**
Κανονική κατανομή	Η συνάρτηση **pnorm(c(72), mean=65, sd=15, lower.tail=TRUE)** δίνει το εμβαδόν της γραφικής παράστασης της κανονικής κατανομής με μέση τιμή 65, τυπική απόκλιση 15 από την ευθεία x = 72 και αριστερά αυτής. Αντίστοιχα, η **pnorm(c(72), mean=65, sd=15, lower.tail=FALSE)** δίνει το συμπληρωματικό εμβαδόν. Η συνάρτηση **qnorm(c(0.005), mean=0, sd=1, lower.tail=TRUE)** δίνει το σημείο του άξονα που περιορίζει το 0.005 = 0.5% του συνολικού εμβαδού της καμπύλης. Τα αριθμητικά αποτελέσματα είναι ίδια με αυτά του Calc.
Διάστημα εμπιστοσύνης	Το διάστημα εμπιστοσύνης προκύπτει συμπληρωματικά σε κάθε έλεγχο υποθέσεων. Γι παράδειγμα αν x = c(10, 13, 13, 18, 12, 14) και ελέγχουμε την υπόθεση πως το δείγμα προέρχεται από πληθυσμό με μέση τιμή 13 τότε το t – test για ένα δείγμα **t.test(x, alternative='two.sided', mu=13.0, conf.level=.95)** έχει ως αποτέλεσμα One Sample t-test data: x t = 1.4662, df = 9, p-value = 0.1766 alternative hypothesis: true mean is not equal to 13 95 percent confidence interval: 3.662498 56.737502 sample estimates: mean of x 30.2 Από όπου καταλαβαίνουμε πως το 95% διάστημα εμπιστοσύνης είναι από 3,7 έως 56,7 (παρεμπιπτόντως, η υπόθεση πως το δείγμα προέρχεται από έναν πληθυσμό με μέση τιμή 13 δεν απορρίπτεται! t = 1.466, df = 9, p-value = 0.177)
Πίνακας συχνοτήτων	Η συνάρτηση **table(x)** είναι η λύση για την περίπτωση που δεν απαιτείται ομαδοποίηση. Η συνάρτηση **hist(x, plot=FALSE)** αποδίδει πίνακα συχνοτήτων με αυτόματο προσδιορισμό των διαστημάτων ενώ η **hist(x, breaks = c(0, 10,20, 30), plot=FALSE)** ορίζει τα διαστήματα [0,10], (10, 20], (20, 30] . Προφανώς, αν plot = TRUE ή αν ακόμα δεν εμφανίζεται η διευκρίνηση αυτή τότε δημιουργείται το ιστόγραμμα!
Επικρατούσα τιμή	Η συνάρτηση which.max(table(x)) επιστρέφει την επικρατούσα τιμή ενός διανύσματος x (ή τη μικρότερη αυτών αν είναι περισσότερες από δύο).

	Μία πληρέστερη λύση είναι η χρήση της συνάρτησης
	smode<-function(x){
	xtab<-table(x)
	modes<-xtab[max(xtab)==xtab]
	mag<-as.numeric(modes[1]) #in case mult. modes, this is safer
	themodes<-names(modes)
	mout<-list(themodes=themodes,modeval=mag)
	return(mout)
	}
	
	η οποία επιστρέφει όλες τις επικρατούσες και την αντίστοιχη μέγιστη συχνότητα εμφάνισης. (πηγή : https://stat.ethz.ch/pipermail/r-help/2011-March/273569.html)

Διάμεση τιμή	Συνάρτηση **median(x)**. Αν **x = c(10, 20, 30, 40)** τότε **median(x) = 25**. Φυσικά, μπορεί να χρησιμοποιηθεί και η συνάρτηση **summary(x)**
Αριθμητικός μέσος	Συνάρτηση **mean(x)**. Μπορεί να χρησιμοποιηθεί και η συνάρτηση **summary(x)**
Αρμονικός μέσος	Δεν υπάρχει δεσμευμένη συνάρτηση για τον αρμονικό μέσο, ωστόσο μπορεί εύκολα να υπολογιστεί από τη συνάρτηση **1/mean(1/x)**. Αν **x = c(10, 20, 30, 40)** τότε **1/mean(1/x) = 19,2.**
Γεωμετρικού μέσος	Δεν υπάρχει δεσμευμένη συνάρτηση για το γεωμετρικό μέσο, ωστόσο μπορεί εύκολα να υπολογιστεί από τη συνάρτηση **prod(x)^(1/length(x))**. Αν **x = c(1.1, 1.15, 1.2)** τότε **prod(x)^(1/length(x))= 1,149**
Ιστόγραμμα	Ένα ιστόγραμμα προκύπτει εύκολα από τη συνάρτηση **hist(x).** Η συνάρτηση αυτή επιδέχεται μεγάλη παραμετροποίηση. Όλες οι επιλογές μπορούν να εμφανιστούν με **?hist** στο παράθυρο του R – Project. Ενδεικτικά, αναφέρουμε την εντολή **hist(x, labels = TRUE, col = 3, density = 2, breaks = c(5, 10, 15, 20, 25, 30, 35), xlab = "Τιμές", ylab = "Συχνότητα", main = "Ιστόγραμμα συχνοτήτων", ylim = c(0,5), xlim = c(0,30))** η οποία θα δημιουργήσει ένα ιστόγραμμα συχνοτήτων με όρια κλάσεων τα 5, 10, 15, 20, 25, 30, 35, εμφάνιση της συχνότητας πάνω από κάθε ράβδο, γραμμοσκιασμένες ράβδους με πράσινο χρώμα, άξονα x από 0 έως 30, άξονα y από 0 έως 5 και τίτλους όπως έχει οριστεί.
Εύρος	Η συνάρτηση **range(x)** επιστρέφει διάνυσμα με την ελάχιστη και τη μέγιστη τιμή. π.χ. Αν x = c(10, 20, 30, 40) τότε range(x) = [10, 40] άρα Έυρος = range(x)[2] - range(x)[1]
Ενδοτεταρτημ	Η συνάρτηση **summary(x)** επιστρέφει διάνυσμα με τα βασικά στατιστικά μεταξύ των

οριακό εύρος	οποίων είναι και τα τεταρτημόρια. Εναλλακτικά μπορεί να χρησιμοποιηθεί η συνάρτηση **fivenum(x)**. Οι δύο διαδικασίες δεν δίνουν πάντα τα ίδια αποτελέσματα καθώς υπολογίζουν τα τεταρτημόρια με διαφορετικές μεθόδους.
Μέτρα διασποράς	**var(x)** και **sd(x)** για τον υπολογισμό της δειγματικής διασποράς και τυπικής απόκλισης αντίστοιχα. Αντιστοιχούν στις συναρτήσεις STDEV και VAR του Calc.
Συντελεστής ομοιογένειας	Δεν υπάρχει συνάρτηση, υπολογίζεται ως **sd(x)/mean(x)**
Ασυμμετρία και κυρτότητα	Υπολογίζουμε το συντελεστή ασυμμετρίας με **skewness(x)** και το συντελεστή κυρτότητας με **kurtosis(x)**. Οι συναρτήσεις βρίσκονται (και) στο πακέτο **moments**. Αν δεν είναι ήδη εγκατεστημένο τότε το εγκαθιστούμε με την εντολή **install.packages("moments", dependencies = TRUE)** και ακολουθεί η εντολή **library(moments)** για να είμαστε σε θέση να χρησιμοποιήσουμε τις συναρτήσεις. Προσοχή : Από τον συντελεστή κυρτότητας όπως υπολογίζεται με τη συνάρτηση kurtosis δεν έχει αφαιρεθεί το 3 που αντιστοιχεί στην κανονική κατανομή.
Έλεγχος κανονικότητας κατανομής	Όπως αναφέρθηκε η εντολή **agostino.test(x, alternative = “two.sided”)** ελέγχει την απόκλιση της ασυμμετρίας της δειγματικής κατανομής από την κανονική ενώ η **anscombe.test(x, alternative = "two.sided")** κάνει το ίδιο για την κυρτότητα. Για τη δοκιμασία K-S αρκεί η συνάρτηση **ks.test(x, pnorm, mean(x), sd(x)).** π.χ. Η δοκιμή στο διάνυσμα **x = rnorm(1000)** κάνει δεκτή την μηδενική υπόθεση της κανονικότητας ενώ αν εφαρμοστεί στο διάνυσμα **x = runif(1000)** απορρίπτεται.
Τυποποιημένες τιμές	Στο βασικό πακέτο δεν υπάρχει ανάλογη συνάρτηση ωστόσο μπορεί εύκολα να υπολογιστούν οι τυποποιημένες τιμές με τη συνάρτηση **(x – mean(x))/sd(x).** Αν **x = c(10, 20, 30)** και **y = (x – mean(x))/sd(x)** τότε y = [-1, 0, 1].
Συνδιακύμανση	Συνάρτηση **cov(x, y)** Σημείωση : Οι δύο συναρτήσεις δίνουν διαφορετικά αποτελέσματα γιατί η διαίρεση γίνεται με το ν στο Calc ενώ με το ν – 1 στο R – Project.
Συντελεστής συσχέτισης	Συνάρτηση **cor(x, y, method = "pearson")** ή πιο απλά **cor(x, y)** για το συντελεστή συσχέτισης του Pearson, **cor(x, y, method = "spearman")** και **cor(x, y, method = "kendall")** για τους συντελεστές Spearman και Kendall αντίστοιχα.
Υπολογισμός και γραφική αναπαράσταση κινούμενου	Ορίζουμε **x = c(3, 8, 25, 20, 22, 16, 14, 6, 12, 15).** Φορτώνουμε τη βιβλιοθήκη zoo με την εντολή **library(zoo)** και μετά χρησιμοποιούμε τη συνάρτηση **rollmean(x,2)** για τον υπολογισμό κινούμενο μέσο 2 σημείων κλπ. Με την εντολή **ts.plot(x)** προκύπτει το διάγραμμα του διανύσματος x ως χρονοσειρά.

μέσου όρου	Ένα περισσότερο εξευγενισμένο διάγραμμα προκύπτει με τις εντολές **plot(x, ann=FALSE, type="n")** **abline(h=0, col=gray(.90))** **lines(x, col="green4", lty="dotted")** **points(x, bg="limegreen", pch=21)** **title(main="Διάγραμμα χρονοσειράς",xlab="Τιμή", col.main="blue", col.lab=gray(.8), cex.main=1.2, cex.lab=1.0, font.main=4, font.lab=3)** Για να προκύψει το κοινό διάγραμμα της χρονοσειράς και του κινούμενου μέσου όρου μπορεί να χρησιμοποιηθεί ο κώδικάς **x.ts = ts(x)** **y.ts = ts(y)** **require(graphics)** **ts.plot(x.ts, y.ts,gpars=list(xlab="Χρόνος", ylab="Τιμή", lty=c(1:2)))** Το R – Project μπορεί να δημιουργήσει εύκολα πολλά διαγράμματα. Κάποια από αυτά επιδεικνύονται με τις εντολές **example(plot.ts)** **example(ts.plot)** **library(zoo)** **example(plot.zoo)** **library(lattice)** **example(xyplot.zoo)**
Αυτοσυσχέτιση	**x = c(7.3, 7, 7.1, 7,4, 7.4, 7.8, 8.9, 9.4, 9.1, 9.2, 9.4, 8.5, 8.6, 9.6, 9, 9.1, 9.8, 10.6)** **και** **acf(x, main = "Διάγραμμα αυτοσυσχέτισης", xlab = "Lag (χρονική υστέρηση)", ylab = "Αυτοσυσχέτιση")**
Δοκιμασία X^2 ως δοκιμασία ομοιογένειας	Δεν υπάρχει δεσμευμένη συνάρτηση στο R – Project. Ο έλεγχος μπορεί να υλοποιηθεί με τις παρακάτω εντολές : **x = c(6, 4, 7, 5)** **e = sum(x) / length(x)** **v = sum((x – e)^2 / e)** **pchisq(v, df=3, lower.tail=FALSE)** Το τελικό αποτέλεσμα ταυτίζεται με αυτό του Calc
Δοκιμασία X^2 ως έλεγχος ανεξαρτησίας	Ο παρακάτω κώδικας υλοποιεί τη δοκιμασία ανεξαρτησίας του παραδείγματος. **.Table <- matrix(c(18, 6, 6, 6, 12, 6, 6, 0), 2, 4, byrow=TRUE)** **rownames(.Table) <- c('Αγόρι', 'Κορίτσι')**

	colnames(.Table) <- c('Καστανά', 'Μαύρα', 'Μπλε', 'Πράσινα') **.Table** **chisq.test(.Table, correct=FALSE)** **remove(.Table)**
Δοκιμασία Fisher	**.Table <- matrix(c(1, 9, 11, 3), 2, 2, byrow=TRUE)** **rownames(.Table) <- c('Δίαιτα', 'Όχι δίαιτα')** **colnames(.Table) <- c('Αγόρι', 'Κορίτσι')** **fisher.test(.Table)** Παρατήρηση : Για τα δεδομένα του πίνακα 5.3 θα προκύψει p = 0,037, διαφορετικό από το 0,013 που υπολογίζει το Calc. Ο λόγος είναι πως το το R – Project χρησιμοποιεί εκτιμητή μέγιστης πιθανοφάνειας (maximum likehood estimator) για να υπολογιίσει το p. Με ?fisher.test παρέχονται περισσότερες πληροφορίες.
t-test ενός δείγματος	**x = c(490, 503, 499, 492, 500, 501, 489, 478, 498, 508)** **t.test(x, mu = 500)** Με ?t.test εμφανίζονται και οι υπόλοιπες επιλογές
t-test δύο ανεξάρτητων δειγμάτων	**x = c(12, 13, 10, 14, 15, 13, 20, 19, 17, 16)** **y = c(20, 19, 17, 10, 14, 13, 17, 19, 16, 15, 18, 19)** **t.test(x, y)**
t-test ζευγαρωτών παρατηρήσεων	**x = c(63, 79, 44, 79, 64, 70, 69, 80, 45, 59, 73, 58, 81, 69, 78, 76, 77, 72, 66, 78, 57, 67, 81, 67, 74, 58, 55, 45, 77, 78)** **y = c(60, 73, 45, 73, 59, 66, 64, 75, 44, 58, 67, 55, 74, 67, 73, 66, 68, 66, 60, 73, 54, 61, 75, 63, 64, 55, 54, 41, 69, 71)** **t.test(x, y, paired = TRUE)**

Βιβλιογραφία

1. Lawrence Brown, Tony Cai and Anirban DasGupta, Interval Estimation for a Binomial Proportion Statistical Science 16, 101-133, (2001)
2. Wikipedia, η ελεύθερη εγκυκλοπαίδεια
3. http://mathworld.wolfram.com/
4. Αναγνωστοπούλου, Τ., & Κιοσέογλου, Γ. (2002). Ερωτηματολόγιο άγχους του Spielberger (State-Trait Anxiety Inventory). Στο: Α. Σταλίκας, Σ. Τριλίβα, & Π. Ρούσση (Επ.). Τα ψυχομετρικά εργαλεία στην Ελλάδα. Αθήνα: Ελληνικά Γράμματα.
5. Στατιστική – Εφαρμογές στη Διοίκηση – Οικονομία – Επιχειρήσεις, ΥΠΕΠΘ.
6. Στατιστική με το OpenOffice, Ε. Διαμαντόπουλος, 2008.
7. Στατιστική, Ο.Ε.Δ.Β., 1999.