ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

5^{ο} ΕΞΑΜΗΝΟ

ΣΤΑΤΙΣΤΙΚΑ ΠΑΚΕΤΑ Ι

ΔΙΔΑΣΚΩΝ ΣΤΕΛΙΟΣ ΖΗΜΕΡΑΣ

Σάμος 2003

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1 ΓΕΝΙΚΟΙ ΟΡΙΣΜΟΙ

1.1 ΣΤΑΤΙΣΤΙΚΗ - ΣΤΑΤΙΣΤΙΚΟΣ ΑΝΑΛΥΤΗΣ

Στην εποχή της τεχνολογικής ανάπτυξης και πληροφορικής έκρηξης, ως **Στατιστική** ορίζεται η επιστήμη που απασχολείται με την συγκέντρωση, παρουσίαση, αξιολόγηση και επεξεργασία συμπερασμάτων. Η ιδιαιτερότητα της ως επιστήμη είναι ότι ενώ σε θεωρητικό επίπεδο η χρήση μαθηματικών μοντέλων είναι αναγκαία για την επίλυση προβλημάτων, στο επίπεδο εφαρμογών χρησιμοποιεί το πλαίσιο όλων σχεδόν των άλλων γνωστικών περιοχών (ιατρική, χρηματοοικονομικά, μάρκετινγκ, αστρονομία, αρχαιολογία, ψυχολογία κ.α.). Τα άτομα που ασχολούνται με την στατιστική ονομάζονται **στατιστικοί αναλυτές** (ΔΕΝ ΥΠΑΡΧΕΙ ΟΡΟΣ ΣΤΑΤΙΣΤΙΚΟΛΟΓΟΣ).

Δουλειά του *στατιστικού αναλυτή* είναι να αναπτύσσει μοντέλα κατάλληλα τόσο ως προς τις ερωτήσεις όσο προς τα δεδομένα. Επιπλέον πρέπει να μπορεί να αντλήσει από τα μοντέλα αυτά σχετικές πληροφορίες που πηγάζουν απο τα δεδομένα.

1.2 ΕΡΓΑΛΕΙΑ ΣΤΑΤΙΣΤΙΚΗΣ

Η ανάγκη για καταγραφή και επεξεργασία δεδομένων οδήγησε στην δημιουργία αρχικά στατιστικών υπηρεσιών κατάλληλων για την αποθήκευσης και αρχειοθέτηση δεδομένων καθώς και την αποδοτικότερη επεξεργασία και συλλογή αποτελεσμάτων. Η διαδικασία της επεξεργασίας δεδομένων και συλλογής αποτελεσμάτων περιλαμβάνει ένα προστάδιο, το στάδιο της ανάλυσης των δεδομένων.

Με τον όρο αυτό εννοούνται όχι μόνο οι τεχνικές και ο μέθοδοι επεξεργασίας πληροφοριών που προέκυψαν από πραγματικά ή εικονικά πειράματα ή παρακολούθηση φαινομένων αλλά και η θεσμοθέτηση κοινά παραδεκτών τεχνικών με σκοπό την εκτίμηση των χαρακτηριστικών του πληθυσμού.

Παράλληλα η χρήση των υπαρχόντων στατιστικών – μαθηματικών εργαλείων ανάλυσης δεδομένων οδήγησε στην αποδοχή κοινών μεθοδολογιών για την αποδοτικότερη επεξεργασία δεδομένων. Ο διαχωρισμός των υπάρχων μεθοδολογιών περιλαμβάνει:

- Περιγραφική στατιστική,
- Στατιστική συμπερασματολογία – Επαγωγική στατιστική,
- Ανάλυση παλινδρόμησης και διακύμανσης,
- Στοχαστική ανάλυση,
- Μπεϋζιανή ανάλυση,
- Πολυμεταβλητή ανάλυση – Ανάλυση κατηγορικών δεδομένων,
- Μη- παραμετρική στατιστική.

Ενδεικτικά αναφέρεται το συγκεκριμένο μάθημα ως ύλη περιλαμβάνει τις πρώτες τρεις ομάδες (Περιγραφική Στατιστική, Στατιστική συμπερασματολογία, Ανάλυση παλινδρόμησης).

Είναι παραδεκτό ότι όλες οι ερευνητικές τεχνικές που παράγουν δεδομένα επιδέχονται στατιστική επεξεργασία. Μερικές από τα κύρια εργαλεία της επεξεργασίας είναι: (1) Περιγραφική Στατιστική και (2) Στατιστική Συμπερασματολογία (Επαγωγική).

- Σκοπός της **Περιγραφικής Στατιστικής** είναι γενικά η άθροιση και σύνοψη δεδομένων. Ειδικότερα αποτελεί ένα στατιστικό εργαλείο με σκοπό την συγκέντρωση ταξινόμηση και παρουσίαση πρωτογενών δεδομένων σε κατανοητή μορφή. Γίνεται με την χρήση **πινάκων** (συχνοτήτων, διπλής εισόδου), γραφημάτων (ραβδογράμματα, θηκογράμματα, διασποράς), και στατιστικών μέτρων (μέτρα κεντρικής τάσης, μέτρα κύμανσης, και μεταβλητότητας).
- Σκοπός της **Στατιστικής Συμπερασματολογίας** είναι η διεξαγωγή από τα δεδομένα νόμων, κανόνων και συμπερασμάτων των οποίων η ισχύς ξεπερνά το επίπεδο των παρατηρήσεων. Οι προτεινόμενοι κανόνες καθορίζουν ένα **μαθηματικό μοντέλο** με σκοπό την καλύτερη και απλούστερη ερμηνεία των δεδομένων.

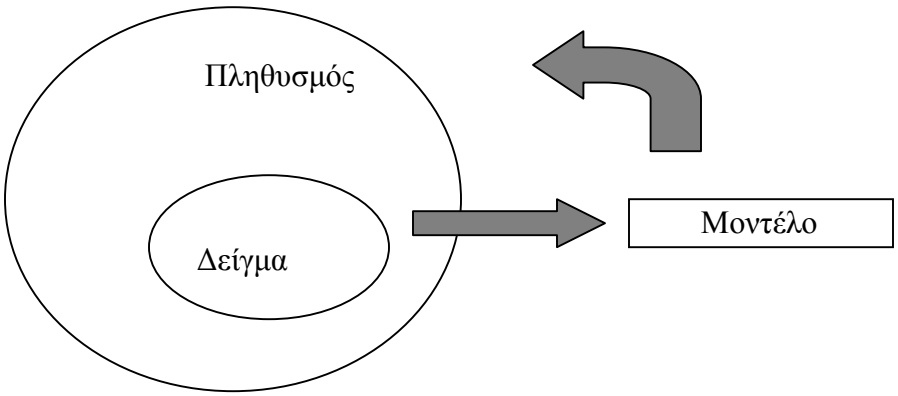
1.3 ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Τα μαθηματικά μοντέλα αποτελούν σήμερα την πιο διαδεδομένη μέθοδο μελέτης φυσικών, κοινωνικών, οικονομικών, ιατρικών φαινομένων. Σε γενικό πλαίσιο, χρησιμοποιούνται για την ανάλυσης και μελέτη τέτοιου είδους φαινομένων καθώς και την παράλληλη διεξαγωγή αποτελεσμάτων. Μαθηματικό μοντέλο μπορεί να θεωρηθεί προσομοίωση των πραγματικών φαινομένων τα οποία ακολουθούν συγκεκριμένους κανόνες. Οι κανόνες αυτοί και γενικότερα τα χαρακτηριστικά γνωρίσματα προσπαθούν να αντικατασταθούν από ανάλογους μαθηματικούς συσχετισμούς. Συχνά η πολυπλοκότητα των υπό μελέτη φαινομένων μας αναγκάζει να προβούμε σε απλοποιήσεις και παραδοχές. Η βασική απαίτηση είναι το μαθηματικό μοντέλο να εξηγεί με τον απλούστερο και καταλληλότερο τρόπο το συγκεκριμένο πρόβλημα που εξετάζεται.

Η επιτυχία στην κατασκευή ενός συγκεκριμένου μαθηματικού μοντέλου, έγκειται στην κατάλληλη χρήση της απλοποίησης και της παραδοχής, στο σωστό χρόνο, και στο σωστό στάδιο έτσι ώστε να περιέχει όσο το δυνατό περισσότερη πραγματικότητα. Η κατασκευή του μοντέλου στηρίζεται αρχικά στην παρατήρηση, την εμπειρία και την διαίσθηση που οδηγούν στην εξήγηση και στην διατύπωση θεωριών οι οποίες περιγράφουν με αντιπροσωπευτικό τρόπο το συγκεκριμένο φαινομένου (προβλήματος). Μετά τον έλεγχό τους ακολουθεί η αναπροσαρμογή τους, η ανατροφοδότησή τους με καινούργια στοιχεία και η επαναδιατύπωσή τους με σκοπό τον έλεγχο (και σύγκριση) του προτεινόμενου μοντέλου με τα ήδη υπάρχοντα.

Κάθε μοντέλο (θεωρητικά ή πρακτικά) εξηγεί ένα συγκεκριμένο (ή ομάδα συγκεκριμένων) φαινομένων. Κάθε φαινόμενο εξελίσσεται όχι αφηρημένα αλλά υποκείμενα, αυτόνομες μονάδες παρατήρησης, το σύνολο των οποίων ορίζει τον **πληθυσμό**. Το φαινόμενο αναλύεται σε επιμέρους μετρήσιμα χαρακτηριστικά, τις **μεταβλητές**, στις οποίες αντιστοιχούμε τιμές. Η αντιστοίχηση αυτή ονομάζεται **μέτρηση** και γίνεται με την χρήση εργαλείων γενικού χαρακτήρα. Πρακτικά, τις περισσότερες

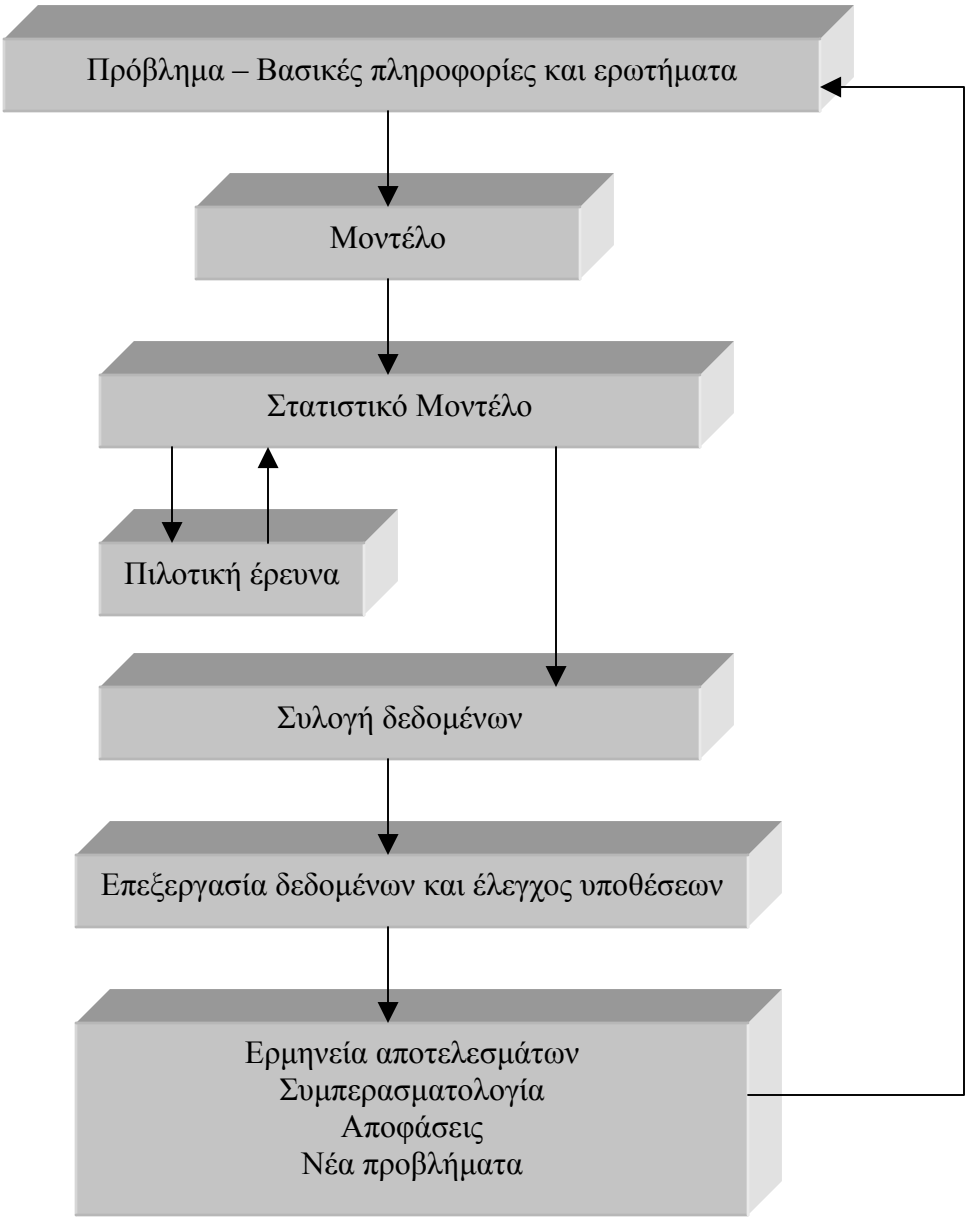
φορές, είναι αδύνατη η μελέτη του πληθυσμού; στην περίπτωση αυτή ένα υπο-σύνολο του πληθυσμού λαμβάνεται με σκοπό την ανάλυση και διεξαγωγή συμπερασμάτων. Το υπό-σύνολο αυτό ονομάζεται **δείγμα** του πληθυσμού. Σχηματικά η διαδικασία μοντελοποίησης ενός προβλήματος δίνεται από το σχήμα 1



Σχήμα 1-1: Σχηματική παρουσίαση ορισμού διαδικασίας μοντελοποίησης.

Μεταφέροντας ένα πρόβλημα σε υποθέσεις θα πρέπει να επιλέξουμε από τα χαρακτηριστικά του, εκείνα τα οποία αφενός είναι σημαντικά αφετέρου μετρήσιμα. Η χρήση της στατιστικής συμπερασματολογίας καθώς και μεθόδων ελέγχου μοντέλων αποτελεί σημαντικό βήμα για την καταλληλότερη επιλογή του συγκεκριμένου μοντέλου όπου αφενός θα εξηγεί με απλό και κατανοητό το συγκεκριμένο πρόβλημα αφετέρου θα είναι απλό στην χρήση και τον χρόνο λήψης αποφάσεων. Τα στάδια επεξεργασίας και ανάλυσης των δεδομένων δίνονται στο σχήμα 2.

Μεταφέροντας ένα πρόβλημα σε υποθέσεις, θα πρέπει να επιλέξουμε από τα χαρακτηριστικά του (μεταβλητές) εκείνα που είναι **σημαντικά** και **μετρήσιμα**. Εστιάζοντας την προσοχή σε τμήματα που είναι σημαντικά προσπαθούμε να κατασκευάσουμε μοντέλο / μοντέλα διαφορετικά από τα υπάρχοντα διακρινόμενα για την απλότητα, σαφήνεια και εφαρμογή των υπό μελέτη στοιχείων του προβλήματος. Ένα μοντέλο είναι καλύτερο όταν μας πληροφορεί πληρέστερα για το πρόβλημα προς ανάλυση ενώ μας επιτρέπει να κάνουμε ακριβέστερες προβλέψεις. Είναι ιδιαίτερα σημαντικό να αναφερθεί ότι μετά τον έλεγχο προσαρμογής των δεδομένων που συγκεντρώθηκαν με το στατιστικό μοντέλο που δημιουργήσαμε, μπορούμε να υπολογίσομε ένα διάστημα εμπιστοσύνης για τις παραμέτρους του πληθυσμού και να προχωρήσουμε στον έλεγχο υποθέσεων γι' αυτές.



Σχήμα 1-2 : Στάδια επεξεργασίας και ανάλυσης δεδομένων

1.4 ΠΛΗΘΥΣΜΟΣ – ΔΕΙΓΜΑ

Με τον όρο **πληθυσμό** ορίζουμε το σύνολο ατόμων ή αντικειμένων (ή άλλων οντοτήτων) όπου βασικός σκοπός είναι η μελέτη, ανάλυση και διεξαγωγή

αποτελεσμάτων, τα οποία θα ερμηνεύουν με τον καλύτερο τρόπο το υπό-μελέτη σύνολο. Μπορούμε να ορίσουμε διαφορετικούς πληθυσμούς ανάλογα με τα χαρακτηριστικά των μελών του (ανθρώπων, φυτών, ζώων, ποδοσφαιρικών ομάδων, εκλογικών τμημάτων κ.α.).

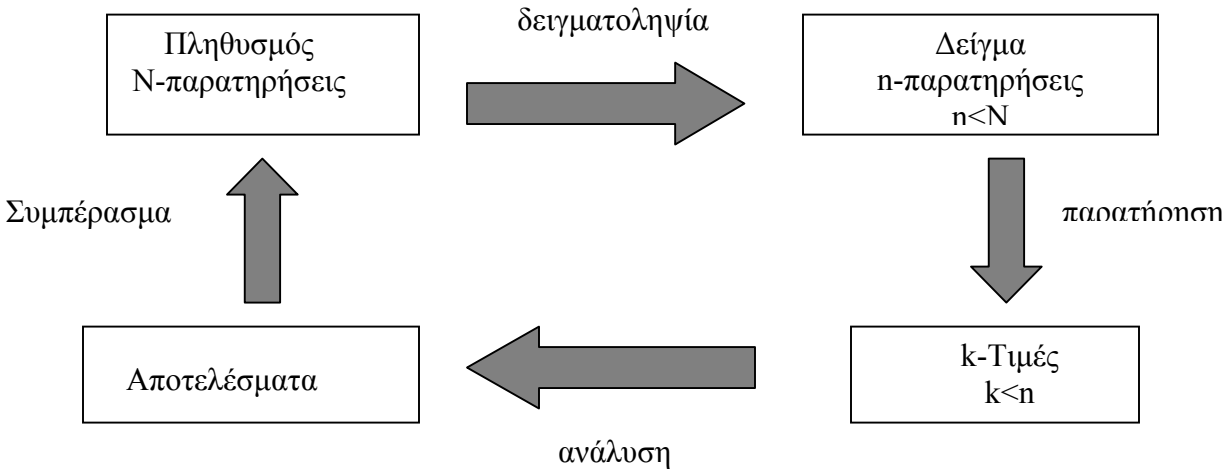
Για κάθε στατιστική μελέτη ο πληθυσμός πρέπει να είναι καλά ορισμένος, να περιγράφεται δηλαδή όσο το δυνατό καλύτερα με βάση τα κοινά χαρακτηριστικά που τον αποτελούν με σκοπό τον γενικότερο διαχωρισμό αν ένα στοιχείων αποτελεί μέλος του ή όχι. Παραδείγματα όπως το σύνολο των Ελλήνων, οι μαθητές της 3ης Λυκείου, οι φοιτητές του Πανεπιστημίου Αιγαίου αποτελούν μερικά χαρακτηριστικά παραδείγματα πληθυσμών.

Ένας πληθυσμός μπορεί να χαρακτηριστεί ως **υπαρκτός** όπου καθορίζεται από *συγκεκριμένα χαρακτηριστικά γνωρίσματα* (ύψος φοιτητών ενός Πανεπιστημίου) ή ως **ιδεατός** όπου καθορίζεται από *αφηρημένα (ασαφή) χαρακτηριστικά γνωρίσματα* (δείκτης ικανοποίησης φοιτητών ενός Πανεπιστημίου το 2001 σχετικά με ένα συγκεκριμένο μάθημα με σκοπό την πρόβλεψη του ιδίου δείκτη για τους φοιτητές του 2002). Ιδεατοί πληθυσμοί χρησιμοποιούνται τις περισσότερες φορές σε μελέτες έρευνας αγοράς ή κλινικές μελέτες .

Πληθυσμός με πεπερασμένο πλήθος στατιστικών χαρακτηριστικών ονομάζεται **πεπερασμένος**. Σε αντίθετη περίπτωση ονομάζεται **μη πεπερασμένος** ή **άπειρος**. Επειδή τις περισσότερες φορές οι πληθυσμοί που μελετώνται είναι μεγάλοι σε μέγεθος, καταλήγουμε στην διαδικασία της δειγματοληψίας. Στην περίπτωση αυτή ο ερευνητής είναι αυτός που ορίζει τον **υπό-μελέτη πληθυσμό** ανάλογα με τις ανάγκες και απαίτησης της μελέτης.

Επειδή ο πληθυσμός προς ανάλυση είναι τόσο μεγάλος σε μέγεθος, οι αναλυτές καταφεύγουν στον ορισμό ενός μικρότερου μέρους (**δείγμα**) με σκοπό την διερεύνησή του. Η διαδικασία αυτή επιλογής συγκεκριμένου δείγματος ονομάζεται ***δειγματοληψία***. Για αξιόπιστη στατιστική ανάλυση το δείγμα πρέπει να είναι **αντιπροσωπευτικό** δηλαδή ***οι τιμές του να είναι αντιπροσωπευτικές των τιμών του πληθυσμού ώστε τα αποτελέσματα της ανάλυσης να είναι αξιόπιστα***. Για να είναι ένα δείγμα αντιπροσωπευτικό ενός πληθυσμού θα πρέπει πρώτα ο πληθυσμός να έχει ορισθεί με σαφήνεια. Στην περίπτωση αυτή το δείγμα πρέπει να είναι τυχαίο δηλαδή κάθε στοιχείο του να έχει την ίδια πιθανότητα να επιλεγεί στο δείγμα με οποιοδήποτε άλλο (δεν πρέπει να υπάρχει μεροληψία στην επιλογή των χαρακτηριστικών του πληθυσμού που ανήκουν στο δείγμα).

Ειδικότερα η λέξη δείγμα αναφέρεται σε δύο διαφορετικές έννοιες. **Δείγμα** είναι το υπό-σύνολο των υποκειμένων που επιλέγονται από το πληθυσμό για να χρησιμοποιηθούν στην ανάλυση αλλά είναι επίσης το σύνολο των παρατηρήσεων (τιμές) που χρησιμοποιούνται στην ανάλυση. Το δείγμα των υποκειμένων πρέπει να επιλεγεί από το πληθυσμό με τρόπο ώστε να εξασφαλίζεται η δυνατότητα γενίκευσης των συμπερασμάτων που θα βγάλουμε. Το δείγμα των τιμών είναι πλήρως καθορισμένο μετά την επιλογή του δείγματος των υποκειμένων. Για την στατιστική ανάλυση χρησιμοποιούμε το δείγμα των τιμών αλλά τα συμπεράσματα αφορούν τον πληθυσμό των υποκειμένων. Σχηματικά η διαδικασία της στατιστικής συμπερασματολογίας με βάση το δείγμα δίνεται στο σχήμα 3.



Σχήμα 3: Διαδικασία της στατιστικής συμπερασματολογίας

1.5 ΜΕΤΑΒΛΗΤΕΣ – ΤΙΜΕΣ ΜΕΤΑΒΛΗΤΩΝ

Κάθε πληθυσμός έχει διαφορετικά **χαρακτηριστικά** (ιδιότητες) κάποια από τα οποία ενδιαφερόμαστε να μελετήσουμε. Τα χαρακτηριστικά αυτά, τα οποία μεταβάλλονται από πληθυσμό σε πληθυσμό ονομάζονται **μεταβλητές**. Για παράδειγμα οι άνθρωποι μπορούν να διαφέρουν ως προς την ηλικία, το φύλο, το βάρος, την οικογενειακή κατάσταση, την περιοχή που διαμένουν. Για να **μετρήσουμε** τα μεταβαλλόμενα χαρακτηριστικά χρειάζεται να χρησιμοποιήσουμε είτε κατάλληλα εργαλεία μέτρησης είτε κατάλληλες κωδικοποιήσεις με σκοπό την διεξαγωγή **μετρήσεων** για κάθε χαρακτηριστικό (μεταβλητή). Οι μετρήσεις αυτές ονομάζονται **τιμές** των μεταβλητών.

Είναι προφανές ότι οι τιμές διαφοροποιούνται μεταξύ των ατόμων ή μεταξύ των στατιστικών χαρακτηριστικών. Οι τιμές των μεταβλητών μπορεί να είναι πραγματικοί αριθμοί (βάρος – 50 κιλά) ή συμβουλευτικές εκφράσεις (φύλο – άνδρας).

Οι μεταβλητές συμβολίζονται με κεφαλαίους λατινικούς χαρακτήρες π.χ. ΦΥΛΟ – SEX, ΗΛΙΚΙΑ – AGE, ΕΙΣΟΔΗΜΑ – INCOME, ΕΠΑΓΓΕΛΜΑ – JOB. Η πρώτη, δεύτερη, ν-στην στατιστική μονάδα συμβολίζεται με i. Ο αύξων αριθμός κινείται μεταξύ 1 και Ν όταν αναφερόμαστε σε στοιχεία πληθυσμού μεγέθους Ν (i = 1,...,N) και μεταξύ 1 και n όταν αναφερόμαστε σε στοιχεία δείγματος μεγέθους n (i = 1,...,n). Η τιμή της μεταβλητής Χ για το i-στο άτομο ή στατιστική μονάδα συμβολίζεται με x_i .

Παράδειγμα

ΑΤΟΜΑ (N) - i	ΗΛΙΚΙΑ (AGE) - x_i	ΦΥΛΟ (SEX) - y_i

1	$x_1 = 50$	y_1 = Ανδρας
2	$x_2 = 55$	y_2 = Ανδρας
3	$x_3 = 69$	y_3 = Γυναίκα
4	$x_4 = 79$	y_4 = Γυναίκα

Όταν οι τιμές μίας μεταβλητής προκύπτουν από την παρατήρηση στατιστικών μονάδων ονομάζονται ***πραγματικές τιμές*** , ενώ όταν προκύπτουν μετά από εφαρμογή ενός μαθηματικού μοντέλου ή διαδικασίας ονομάζονται **θεωρητικές τιμές** .

Όπως προαναφέρθηκε η στατιστική ανάλυση δεδομένων περιλαμβάνει ορισμό μεταβλητών καθώς και ***μετρήσεις*** των συγκεκριμένων μεταβλητών που θα εξεταστούν. Με άλλα λόγια θα πρέπει να ***μετρήσουμε*** καθένα από τα χαρακτηριστικά είτε με την χρήση συγκεκριμένου οργάνου είτε με κατάλληλη κωδικοποίηση. Και οι δύο περιπτώσεις περιλαμβάνουν εμπειρική διαδικασία μέτρησης που συνεπάγεται τον καθορισμό τιμών. Για το λόγο αυτό κατάλληλα ***εργαλεία μέτρησης*** πρέπει να ορισθούν. Τα εργαλεία μέτρησης πρέπει να είναι:

1. *Κατάλληλα*, ικανά να διακρίνουν διαφορές στα μεγέθη που χαρακτηρίζουν τις μεταβλητές.
2. *Να μην παραμορφώνουν τις τιμές.*
3. *Αντικειμενικά*, να δίνουν το ίδιο αποτέλεσμα για την μέτρηση της ίδιας τιμής οποίος και αν τα χρησιμοποιήσει.
4. *Απλά*, ώστε να μην γίνονται λάθη στην χρήση τους.

Στην διαδικασία μέτρησης εμφανίζονται **σφάλματα** που οφείλονται :

1. Ακρίβεια του εργαλείου,
2. Στην ιδιαιτερότητα του προσώπου που κάνει τις μετρήσεις,
3. Σε λανθασμένες καταγραφές.

Η ποιότητα της μελέτης εξαρτάται από την αποδοτικότητα των μεθόδων συλλογής δεδομένων που χρησιμοποιούνται , του οργάνου δηλαδή με το οποίο πραγματοποιούμε την μέτρηση της συγκεκριμένης μεταβλητής. Εύλογα αναμένονται **αξιόπιστες** και **αμερόληπτες** πληροφορίες.

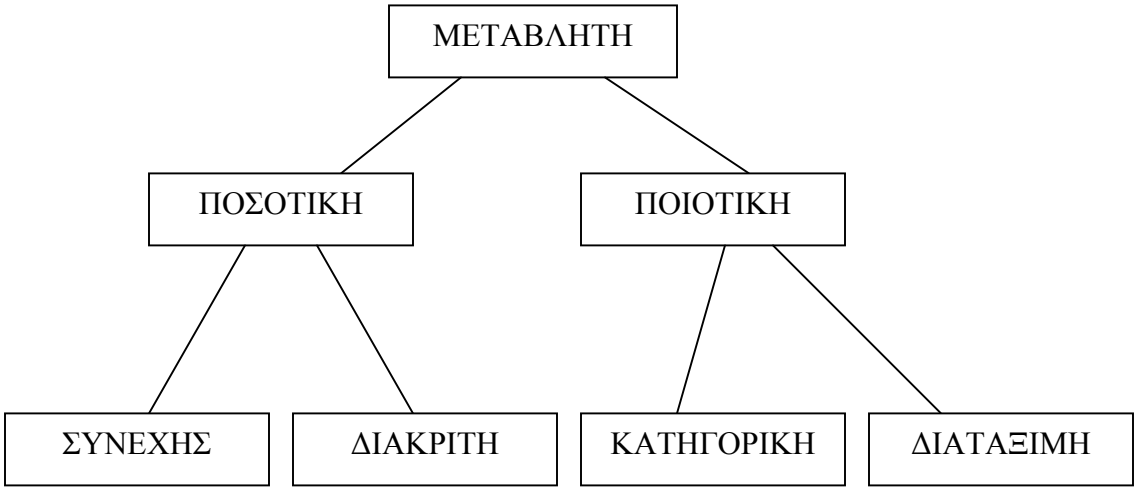
1.6 ΤΥΠΟΙ ΜΕΤΑΒΛΗΤΩΝ

Η φύση των μεταβλητών καθορίζει το είδος των δεδομένων. Οι μεταβλητές διακρίνονται σε **ποσοτικές** και **ποιοτικές** ανάλογα με το εάν οι τιμές εκφράζουν αριθμητικά ή ονομαστικά δεδομένα.

Οι ***ποσοτικές μεταβλητές*** διακρίνονται με την σειρά τους σε **συνεχείς** και **διακριτές** ανάλογα με το αν είναι συνεχείς ή διακριτές. **Συνεχείς** μπορούν να ονομαστούν επίσης οι μεταβλητές όπου για κάθε δύο τιμές της μεταβλητής μπορούμε να βρίσκουμε πάντα μια τρίτη τιμή, έστω και θεωρητική, μεταξύ τους (ύψος, βάρος). Από την άλλη μεριά, **διακριτές** μπορούν να ορισθούν οι μεταβλητές όπου μεταξύ δύο τιμών δεν υπάρχει μια τρίτη ενδιάμεση τιμή (αριθμός παιδιών σε μία οικογένεια, πλήθος πελατών σε σούπερ-μάρκετ).

Οι ποιοτικές μεταβλητές διακρίνονται σε **κατηγορικές** και **διατάξιμες**. Οι μεταβλητές, οι οποίες δίνουν την δυνατότητα στον ερευνητή να διατάξει και διαβαθμίσει

της κατηγορίες που προκύπτουν από τις τιμές ονομάζονται **διατάξιμες** (επίπεδα εκπαίδευσης, αγωνίσματα, αξιολόγηση μίας ταινίας). Οι υπόλοιπες που δεν παρέχουν την δυνατότητα διάταξης αλλά με βάση τα χαρακτηριστικά που εκφράζουν οι τιμές τους επιτρέπουν απλά και μόνο την διάκριση ορισμένων κατηγοριών ονομάζονται **κατηγορικές** (χρώμα ματιών, φύλο οικογενειακή κατάσταση). Ταξινόμηση των μεταβλητών σε κατηγορίες δίνεται στο σχήμα 4.



Σχήμα 4: Ταξινόμηση μεταβλητών

Οι μεταβλητές που λαμβάνουν την ίδια τιμή για όλες τις στατιστικές μονάδες ονομάζονται **σταθερές μεταβλητές**.

Μεταβλητές που χρησιμοποιούνται για περαιτέρω κωδικοποίηση των υπαρχόντων μεταβλητών ονομάζονται **ψευτομεταβλητές**.

Παράδειγμα

No	Φύλο	**Κωδικός**
1	Άνδρας	0
2	Γυναίκα	1
3	Άνδρας	0
4	Άνδρας	0

1.7 ΚΛΙΜΑΚΕΣ (ΕΠΙΠΕΔΑ) ΜΕΤΡΗΣΗΣ

Το επίπεδο μέτρησης των δεδομένων αποτελεί σημαντικό παράγοντα στην επιλογή της κατάλληλης στατιστικής μεθόδου και χαρακτηρίζεται από δύο ιδιότητες, τη διάταξη

των μετρήσεων και την απόσταση μεταξύ τους. Ο διαχωρισμός των επιπέδων διαμορφώνεται με βάση τον παρακάτω διαχωρισμό.

Ονομαστικές Κλίμακες, χρησιμοποιούνται για την συμβολική έκφραση ποιοτικών κατηγορικών στοιχείων και μεταβλητών. Οι τιμές αποδίδονται μόνο με ένα όνομα και καμία υπόθεση γίνεται γύρω από τις συμβολικές τιμές που λαμβάνουν.

Διάταξης Κλίμακες, μετρούν ποιοτικά διατάξιμα δεδομένα και μεταβλητές των οποίων οι τιμές μπορούν να ιεραρχηθούν με βάση κάποιο κριτήριο.

Διαστημικές Κλίμακες, αναφέρονται στην μέτρηση αριθμητικών δεδομένων τα οποία ικανοποιούν και την ιδιότητα της διάταξης και απόστασης μεταξύ τους.

Αναλογικές Κλίμακες, αναφέρονται σε αριθμητικά δεδομένα τα οποία ικανοποιούν τις παραπάνω ιδιότητες καθώς και διαθέτουν μηδενικό σημείο αναφοράς, όπου αποτελεί πραγματική μετρήσιμη κατάσταση.

2 ΕΠΙΛΟΓΗ ΣΤΑΤΙΣΤΙΚΩΝ ΤΕΧΝΙΚΩΝ

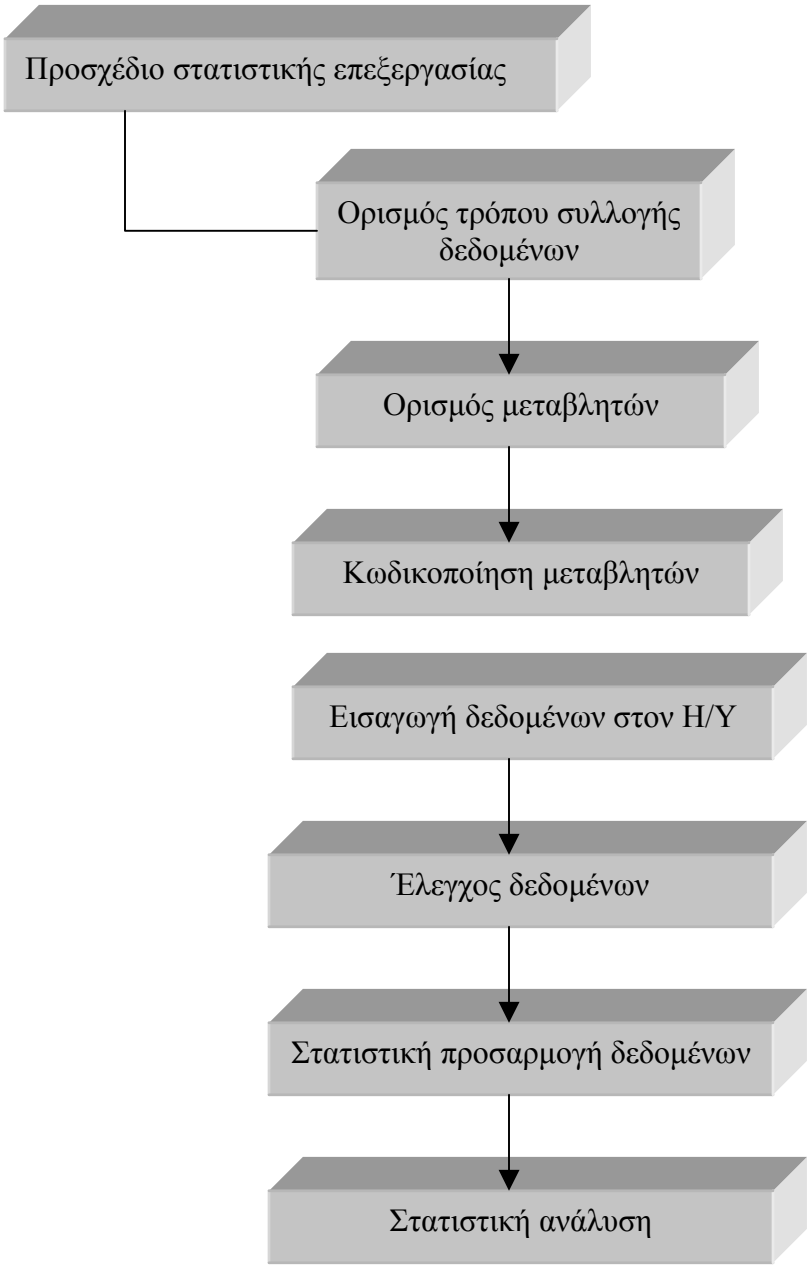
Η επιλογή της στατιστικής τεχνικής που θα ακολουθηθεί εξαρτάται από τον τύπο της μεταβλητής (κλίμακα μέτρησης) της μεταβλητής. Ανάλογα με την κλίμακα μέτρησης της μεταβλητής, οι παράμετροι του πληθυσμού προσεγγίζονται από διαφορετικό στατιστικό μέτρο, ο πίνακας συχνοτήτων δίνει διαφορετικά στοιχεία και χρησιμοποιούμε διαφορετικά σχήματα για τις γραφικές παραστάσεις. Ενδεικτικός είναι ο παρακάτω πίνακας

Κλίμακα	Στατιστική μέθοδο	
	Περιγραφική	Συμπερασματολογία
Κατηγορίας	Ποσοστά, Επικρατούσα τιμή	χ^2-τεστ, Διωνυμικός έλεγχος
Διάταξης	Εκατοστιαία σημεία, Διάμεσος	Συντελεστής συσχέτισης, ANOVA
Διαστήματος	Εύρος, Μέσος όρος, τυπική απόκλιση	Συντελεστής συσχέτισης, ANOVA, t-τεστ, παλινδρόμιση
Αναλογίας	Γεωμετρικός μέσος, Αρμονικός μέσος	Συντελεστής μεταβλητικότητας

3 ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Η διαδικασία της προετοιμασίας των δεδομένων ξεκινά με την **επιλογή** του θέματος έρευνας, συνεχίζει με τον **σχεδιασμό** συγκεκριμένης μεθοδολογίας που θα ακολουθηθεί και κλείνει με την **υλοποίηση** και **εφαρμογή** των συγκεκριμένων στατιστικών μεθόδων. Ανεπαρκής ή ελλιπής προετοιμασία των δεδομένων οδηγεί σε μεροληπτικά αποτελέσματα και λανθασμένες ερμηνείες εκθέτοντας ανεπανόρθωτα την ποιότητα της στατιστικής ανάλυσης.

Στην πρώτη φάση ελέγχεται το όργανο συλλογής τω δεδομένων. Στην δεύτερη γίνεται η αντιστοίχηση των δεδομένων με τις μεταβλητές. Ακολουθεί η απόφαση για τον τρόπο κωδικοποίησης των μεταβλητών καθώς και ο τρόπος εισαγωγής των δεδομένων στον Η/Υ (λήψη απόφασης όσο αφορά το συγκεκριμένο στατιστικό πακέτο). Στην συνέχεια ελέγχεται η λογικότητα των δεδομένων και αποφασίζεται ο τρόπος χειρισμού των παρατηρήσεων που δεν έχουν καταγραφεί (ελλείπουσες τιμές). Τελικά στάδια είναι η στατιστική προσαρμογή των δεδομένων έτσι ώστε να υπάρξει η απαιτούμενη αντιπροσωπευτικότητα του πληθυσμού και ο ορισμός της στατιστικής ανάλυσης που θα ακολουθηθεί. Τα παραπάνω στάδια παρουσιάζονται στο σχήμα 5.



Σχήμα 5: Γραφική παρουσίαση προετοιμασίας δεδομένων

Ειδικότερα τα στάδια προετοιμασίας των δεδομένων μπορούν να αναλυθούν στα παρακάτω επίπεδα:

- **Ορισμός τρόπου συλλογής δεδομένων**
  Αποτελεί το αρχικό στάδιο ανάλυσης όπου η χρήση ενός συγκεκριμένου εργαλείου συλλογής δεδομένων πρέπει να χρησιμοποιηθεί. Τα σημαντικότερα και πλέον αποδεκτά μέσα συλλογής μπορεί να είναι: ερωτηματολόγια, πρόσβαση σε βάσεις δεδομένων, προσωπικές παρατηρήσεις, WEB, στατιστικές υπηρεσίες.

- **Ορισμός μεταβλητών**
  Καθορισμός μεταβλητών από ερωτηματολόγια ή βάσεις δεδομένων. Ορισμός πρωταρχικών και δευτερεύουσων μεταβλητών. Στην απλούστερη περίπτωση κάθε πεδίο ή ερώτηση από την συλλογή δεδομένων αποτελεί μια μεταβλητή. Το όνομα της μεταβλητής είναι καλό να είναι βολικό για μελλοντική ανάλυση και να αντιπροσωπεύει τα χαρακτηριστικά που καταγράφει η μεταβλητή; π.χ. ΦΥΛΟ – SEX, ΤΑΧΥΔΡΟΜΙΚΟΣ ΚΩΔΙΚΑΣ- ZIPCODE.

- **Κωδικοποίηση μεταβλητών**
  Εννοούμε την αντιστοίχηση κωδικών σε όλες τις πιθανές τιμές μίας μεταβλητής. Οι κωδικοί είναι συνήθως αριθμοί αλλά μπορεί να είναι και χαρακτήρες π.χ. π.χ. ΦΥΛΟ – ΑΓΟΡΙ (Α) , ΚΟΡΙΤΣΙ (Κ). Ίδιες τιμές δύο διαφορετικών χαρακτηριστικών πρέπει να αντιστοιχούν ακριβώς στον ίδιο κωδικό. Για παράδειγμα δεν μπορεί το φύλλο του ερωτούμενου σε ένα ερωτηματολόγιο να το κωδικοποιούμε αλλού με Α και αλλού με α για τον άνδρα/αγόρι και Γ/γ ή Κ/κ για την γυναίκα/κορίτσι. Οι ποσοτικές μεταβλητές είναι ήδη κωδικοποιημένες. Όλοι οι χρησιμοποιούμενοι κωδικοί μιας έρευνας συνήθως καταγράφονται σε έναν πίνακα που ονομάζεται ***πίνακας κωδικοποίησης*** .

- **Εισαγωγή δεδομένων στον Η/Υ**
  Το σημαντικότερο στάδιο της ανάλυσης δεδομένων είναι η εισαγωγή τους στο Η/Υ. Η διαδικασία τις περισσότερες φορές είναι επίπονη και κουραστική, καταναλώνοντας αρκετό χρόνο εξαρτώμενος από τον αριθμό και την κωδικοποίηση των δεδομένων.

- **Έλεγχος δεδομένων**
  Πολλοί λόγοι μπορεί να οδηγήσουν στην ύπαρξη παράλογων τιμών. Όποιες τιμές εμφανίζονται **ακραίες** ή **λανθάνουσες** πρέπει να ελέγχονται σχολαστικά.

- **Χειρισμός ελλειπουσών τιμών**
  Σαν ελλείπουσες τιμές χαρακτηρίζονται εκείνες οι τιμές οι οποίες δεν έχουν καταγραφεί. Για τις ποιοτικές μεταβλητές ένας πρακτικός τρόπους αντιμετώπισης του προβλήματος είναι ο καθορισμός ακόμα μιας κατηγορίας για την συγκεκριμένη μεταβλητή ή οποία μπορεί να συμπεριληφθεί στην ανάλυση. Για τις ποσοτικές μεταβλητές σχετικοί τρόποι αντιμετώπισης του προβλήματος θα αναλυθούν σε άλλα κεφάλαια.

- **<u>Στατιστική προσαρμογή δεδομένων</u>**
  Περιλαμβάνει στην κατασκευή νέων μεταβλητών που είναι απαραίτητες για την ανάλυση. Η δημιουργία βουβών μεταβλητών μπορεί να επαπροσδιορισεί μόνο ποιοτικά δεδομένα Οι πιθανές τιμές των μεταβλητών είναι τις περισσότερες φορές 0 ή 1.

4 ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ

Όταν πρέπει να χρησιμοποιήσουμε κάποια στατιστική τεχνική ***<u>πρέπει</u>*** :

- Να γνωρίζουμε τις προϋποθέσεις της
- Να ελέγξουμε κατά πόσο είναι δυνατό να ισχύουν αυτές οι προϋποθέσεις (χρήση πιλοτικής έρευνας)
- Να μπορούμε να διατυπώσουμε τις υποθέσεις που ελέγχονται.
- Να μπορούμε να ερμηνεύσουμε τα αποτελέσματα σε σχέση με το σύστημα που μελετάμε.

Όταν πρέπει να χρησιμοποιήσουμε κάποια στατιστική τεχνική ***<u>δεν πρέπει</u>*** :

- Να χρησιμοποιούμε τεχνικές επειδή κάποιος άλλος έτσι έκανε.
- Να προσπαθούμε να καταγράψουμε αποτελέσματα που συμφωνούν με τις δικές μας απόψεις.
- Να επιλέγουμε υποσύνολα από τα δεδομένα που μας φαίνεται ότι υποστηρίζουν τις υποθέσεις μας.
- Να παρουσιάζουμε αποτελέσματα που αδυνατούμε να τα ερμηνεύσουμε.

5 ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ

Από τα ερωτηματολόγια, τα δεδομένα καταγράφονται και κωδικοποιούνται σε φύλλα δεδομένων ή φύλλα κωδικοποίησης με μορφή

	Α.Α	Μεταβλ. 1	……	Μεταβλ. n
Αντικ. 1	1		……	
Αντικ. 2	2		……	
:	:		……	
Αντικ. m	m		……	

Στην συνέχεια ακολουθεί η μεταφορά των δεδομένων σε ειδικά πακέτα στατιστικής ανάλυσης (SPSS, MINITAB,…) ή στην απλούστερη περίπτωση σε προγράμματα λογιστικών φύλλων (EXCEL). Αν ο όγκος των δεδομένων είναι μεγάλος ή η δομή τους σύνθετη τότε χρησιμοποιούμε προγράμματα δημιουργίας βάσης δεδομένων όπως ACCESS.
Τα στάδια για την εισαγωγή των δεδομένων στον Η/Υ είναι τα ακόλουθα:

- ***Κωδικοποίηση*** – την σωστή μεταφορά των δεδομένων στον Η/Υ.
- ***Πληκτρολόγηση*** – μεταφορά δεδομένων σε ηλεκτρονική μορφή.
- ***Έλεγχος σφαλμάτων*** – ανάλυση για τον εντοπισμό και την διόρθωση λαθών κατά την πληκτρολόγηση.

5.1 ΚΩΔΙΚΟΠΟΙΗΣΗ

Συνήθως σε μελέτες μεγάλου μεγέθους η κωδικοποίηση και πληκτρολόγηση γίνεται από ανεξάρτητες ομάδες ατόμων που καθοδηγούνται από ένα ειδικευόμενο αναλυτή. Για να αποφευχθούν τα σφάλματα πριν την κωδικοποίηση πρέπει να αποφασιστούν τα ακόλουθα:

✓ Ποίες μεταβλητές θα **εισαχθούν**, **ονόματα** και **σειρά**

Τα ονόματα είναι σύντομα. Επιπλέον πρέπει να θυμίζουν το περιεχόμενο του χαρακτηριστικού – μεταβλητή. Καλό θα ήταν να τεθεί ένας συγκεκριμένος κανόνας και να τον ακολουθήσουμε σε όλες τις μεταβλητές. Τέλος θα ήταν χρήσιμο η σειρά των μεταβλητών να είναι ίδια με το ερωτηματολόγιο.

✓ **Τρόποι** κωδικοποίησης **ποιοτικών** και **διατάξιμων** μεταβλητών

Πρώτα ορίζουμε τους κωδικούς – νούμερα για κάθε επίπεδο της μη ποσοτικής μεταβλητής. Όταν έχουμε δίτιμες μεταβλητές χρησιμοποιούμε την κωδικοποίηση 0-1 με κωδικό 1 για την επιτυχία (επιτυχία στη στατιστική 1 – ΝΑΙ, 0-ΟΧΙ). Όταν έχουμε πολλές κατηγορικές μεταβλητές με τα ίδια επίπεδα χρησιμοποιούμε κοινό τρόπο κωδικοποίησης. Σε διατάξιμες μεταβλητές χρησιμοποιούμε κωδικούς που ξεκινούν από το 1 για το μικρότερο επίπεδο και αυξάνουν κατά μια μονάδα κάθε ανώτερο επίπεδο. Εναλλακτικά μπορούμε να χρησιμοποιούμε κωδικούς συμμετρικούς στο μηδέν έτσι ώστε αρνητικές τιμές να αποδίδουν αρνητικές γνώμες.
Παράδειγμα: Ερώτηση: Το μάθημα της στατιστικής σας ικανοποιεί;

Απαντήσεις	**Κώδικες**	**Εναλλακτικοί Κώδικες**
Πάρα πολύ	1	3
Πολύ	2	2
Αρκετά	3	1
Μέσα	4	0
Λίγο	5	-1
Πολύ Λίγο	6	-2

Καθόλου 7 -3

✓ **Αγνοούμενες τιμές** ή **μη καταχωρημένες τιμές**

Οι μη καταχωρημένες ή αγνοούμενες τιμές αναφέρονται σε ερωτήσεις ή μεταβλητές που οι τιμές λείπουν κυρίως γιατί οι ερωτώμενοι δεν συμπληρώσαν το αντίστοιχο πεδίο του ερωτηματολογίου. Κωδικοποιούνται με τιμές 9.99 ή 999 για να μπορέσουμε να εντοπίσουμε και παραλείψεις που οφείλονται σε πληκτρολόγηση. Εναλλακτικά αφήνουμε κενό το αντίστοιχο πεδίο. Στην περίπτωση αυτή δεν μπορούμε να εντοπίσουμε παραλείψεις που οφείλονται σε πληκτρολόγηση. Επίσης οι αγνοούμενες τιμές πρέπει να διαφοροποιούνται από τις απαντήσεις του τύπου "Δεν απαντώ" ή "Δεν ισχύει"
Οι αγνοούμενες τιμές συνήθως : **Αφαιρούνται** από την ανάλυση,
Αντικαθίστανται από μέσες τιμές,
Προβλέπονται από μοντέλα.

✓ **Δημιουργία μεταβλητής αναγνώρισης ερωτηματολογίου**

Είναι σημαντικό να δημιουργηθεί μια μεταβλητή η οποία να αντιστοιχεί μοναδικά σε κάθε ερωτηματολόγιο – μεταβλητή. Συνήθως ονομάζεται Α.Α (Αύξων αριθμός) και ξεκινά από το 1. Ο ίδιος αριθμός θα πρέπει να σημειώνεται στα ερωτηματολόγια. Η χρησιμότητα του Α.Α. Εντοπίζεται στον έλεγχο σφαλμάτων, στην αρχειοθέτηση ερωτηματολογίων και στην διασταύρωση στοιχείων.

5.2 ΕΛΕΓΧΟΣ ΣΦΑΛΜΑΤΩΝ ΠΛΗΚΤΡΟΛΟΓΗΣΗΣ

Μετά την εισαγωγή των δεδομένων στον Η/Υ, ακολουθεί ο έλεγχος της πιστότητας των δεδομένων για τυχόν εντοπισμό σφαλμάτων. Σαν **ύποπτες** τιμές ορίζονται εκείνες οι οποίες μπορεί να οφείλονται σε εσφαλμένη πληκτρολόγηση και μπορούν να χαρακτηρισθούν οι ακόλουθες περιπτώσεις:

✓ **Ακραίες τιμές**

Ορίζονται ως οι τιμές οι οποίες είναι ασυνεπείς με το σύνολο των υπολοίπων τιμών. Μπορεί να οφείλονται σε λανθασμένη πληκτρολόγηση αλλά μπορεί και όχι. Στην δεύτερη περίπτωση μπορεί να χαρακτηρίζει την μη κανονικότητα της. Όταν έχουμε κατηγορικές μεταβλητές οι ακραίες τιμές οφείλονται σε λάθος πληκτρολόγηση. Όταν εντοπίσουμε ακραίες τιμές επιστρέφουμε στα ερωτηματολόγια ελέγχοντας την ορθότητά τους.

✓ **Αντιστροφή ψηφίων**

Αντί για 34 μπορεί να πληκτρολογήσουμε 43. Τέτοια σφάλματα είναι δύσκολο να εντοπισθούν στις ποσοτικές μεταβλητές. Στις ποιοτικές ο εντοπισμός τους μπορεί να είναι πιο εύκολος ειδικά αν ο αντίστροφος κωδικός δεν ανήκει στους προεπιλεγμένους

κωδικούς της μεταβλητής. Συνήθως τα σφάλματα αυτού του τύπου εντοπίζονται ως ακραίες τιμές.

- ✓ **Επαναλήψεις τιμών (Διπλοεγγραφές)**

Επαναλήψεις των ίδιων αριθμών ή κωδικών είναι σύνηθες φαινόμενο. Ο εντοπισμός τους μπορεί να γίνει μόνο οπτικά και διασταυρώνονται με τα ερωτηματολόγια.

- ✓ **Λάθος καταχωρήσεις**

Αναφέρονται σε καταχωρήσεις που γίνονται σε λάθος στήλες (λάθος μεταβλητή). Αυτά τα σφάλματα μπορούν να εντοπισθούν ως ακραίες τιμές αν το εύρος των τιμών των τιμών δύο διαδοχικών μεταβλητών είναι διαφορετικό.

Έλεγχος των λαθών μπορεί να γίνει με:

- o Εκτύπωση μέσης τιμής, τυπικής απόκλισης, ελάχιστης και μέγιστης τιμής μίας μεταβλητής.
- o Κατανομές συχνοτήτων για κάθε μεταβλητή.
- o Εκτύπωση και έλεγχο των πληκτρολογούμενων δεδομένων κάθε μεταβλητής για ύπαρξη επαναλαμβανόμενων τιμών
- o Διασταύρωση ερωτηματολογίων και πληκτρολογούμενων δεδομένων στην περίπτωση ακραίων τιμών.
- o Δειγματοληπτική διασταύρωση με ερωτηματολόγια.

Παρακάτω παρουσιάζονται δύο παραδείγματα κωδικοποίησης δεδομένων

Παράδειγμα 1

Έστω τα παρακάτω χαρακτηριστικά (μεταβλητές) που έχουν ληφθεί από διαδικασία δειγματοληψίας μέσα από ερωτηματολόγια:
Επίθετο αθλητή, Όνομα αθλητή, Φύλο, Βάρος (κιλά), Ύψος (εκατοστά), Διεύθυνση, Ταχ. Κώδικας, Σχολείο, Πόλη, Νομός, Δρόμος 30μ, Δρόμος 1000μ, Άλμα χωρίς φόρα με χέρια (εκατοστά), Άλμα χωρίς φόρα χωρίς χέρια (εκατοστά).
Κωδικοποίηση των μεταβλητών αποτελείται από δύο στάδια.
Πρώτο στάδιο χαρακτηρισμός μεταβλητών σε ποσοτικές και ποιοτικές. Άρα
Ποσοτικές: Βάρος (κιλά), Ύψος (εκατοστά), Δρόμος 30μ, Δρόμος 1000μ, Άλμα χωρίς φόρα με χέρια (εκατοστά), Άλμα χωρίς φόρα χωρίς χέρια (εκατοστά).
Ποιοτικές: Φύλο, Διεύθυνση, Ταχ. Κώδικας, Σχολείο, Πόλη, Νομός.
Άλλες Μεταβλητές: Επίθετο αθλητή, Όνομα αθλητή.

Μερικές από τις ποιοτικές μεταβλητές μπορούν να χαρακτηριστούν και ως ποιοτικές διακριτές όπου παίρνουν τιμές κωδικών (1,2,3,...) όπως:
Φύλο: Αγόρι, Κορίτσι
Σχολείο: Δημοτικό, Γυμνάσιο, Λύκειο, Τεχνικό.
Πόλη: Αθήνα, Πειραιάς, Πάτρα, Θεσσαλονίκη κ.α.

Νομός: Αττικής, Βοιωτίας, Θράκης, Πελοποννήσου κ.α.

Δεύτερο στάδιο είναι η μετατροπή των μεταβλητών σε γνώριμα ονόματα για την εισαγωγή στο Η/Υ.
Ποσοτικές: Βάρος (κιλά)-Weight, Ύψος (εκατοστά)-Hight, Δρόμος 30μ-Min30, Δρόμος 1000μ-Min1000, Άλμα χωρίς φόρα με χέρια (εκατοστά)-LenHand, Άλμα χωρίς φόρα χωρίς χέρια (εκατοστά)-Lenght.
Ποιοτικές: Φύλο-Sex, Διεύθυνση-Address, Ταχ. Κώδικας-Post Code, Σχολείο-School, Πόλη-Town, Νομός-Area.
Άλλες Μεταβλητές: Επίθετο αθλητή-Surname, Όνομα αθλητή-Name

Παράδειγμα 2
Έστω η καταγραφή διαφορετικών χαρακτηριστικών μίας χώρας. Η κωδικοποίηση αρχίζει με τον ορισμό των μεταβλητών
Όνομα χώρας, Πληθυσμός, Πυκνότητα πληθυσμού, Ποσοστό αστικού πληθυσμού, Θρησκεία, Οικονομική ομάδα.
Καθορισμός ποσοτικών και ποιοτικών μεταβλητών
Ποσοτικές: Πληθυσμός, Πυκνότητα πληθυσμού, Ποσοστό αστικού πληθυσμού.
Ποιοτικές: Θρησκεία, Οικονομική ομάδα
Άλλες Μεταβλητές: Όνομα χώρας.

Καθορισμός των ποιοτικών διακριτών
Θρησκεία: Καθολικοί, Μουσουλμάνοι, Χριστιανοί, Προτεστάντες,κ.α.
Οικονομική ομάδα: Οικονομικά Ισχυρά κράτη, Ανατολική Ευρώπη, Ασία, Αφρική, Μέση Ανατολή, Λατινική Αμερική, κ.α.

Μετατροπή των μεταβλητών
Ποσοτικές: Πληθυσμός-Population, Πυκνότητα πληθυσμού-Density, Ποσοστό αστικού πληθυσμού-Urban.
Ποιοτικές: Θρησκεία-Religion, Οικονομική ομάδα-Region

6 ΤΡΟΠΟΙ ΣΥΓΚΡΙΣΗΣ ΔΕΔΟΜΕΝΩΝ

Η εισαγωγή των δεδομένων αποτελεί σημαντικό κομμάτι της ανάλυσης και επεξεργασίας τους. Βασική διαδικασία είναι η ομαδοποίηση των σε πίνακες διπλής εισόδου με σκοπό την καλύτερη κατανόηση της δομής των. Με βάση τους πίνακες, τις περισσότερες φορές, συγκρίσεις μεταξύ ιδίων ή διαφορετικών ομάδων λαμβάνει χώρα. Στις περιπτώσεις σύγκρισης των δεδομένων εμφανίζονται δύο βασικές κατηγορίες (ομάδες) μεταβλητών: **ανεξάρτητες μεταβλητές** και **εξαρτημένες μεταβλητές**.

Ανεξάρτητες ονομάζονται οι μεταβλητές όπου επηρεάζονται άμεσα από το πείραμα και δεν επιδρούν στην σχεδιασμό του μοντέλο. Καθορίζονται πριν από τον σχεδιασμό του μοντέλου (ή της σύγκρισης) και ελέγχονται πριν από το πείραμα (πειραματικές ομάδες).

Εξαρτημένες ονομάζονται οι μεταβλητές οι οποίες επιδρούν είτε σε άλλες μεταβλητές είτε στον σχεδιασμό συγκεκριμένου πειράματος. Καθορίζονται κατά την διάρκεια του πειράματος.

Βασικό χαρακτηριστικό των πειραμάτων είναι ότι από την στιγμή που υπάρχουν ανεξάρτητες μεταβλητές, υπάρχει η δυνατότητα αλληλεπίδρασης και συσχέτισης μεταξύ διαφορετικών παραγόντων (factors).Οι ανεξάρτητες μεταβλητές μπορούν να αποτελέσουν τις μεταβλητές που θα συγκριθούν με τα αποτελέσματα των διαφορετικών πειραμάτων. Η διαδικασία αυτή ονομάζεται control διαδικασία.

Παράγοντες (factors) ονομάζεται ένα σύνολο από κατηγορίες ή περιορισμούς. Μερικοί παράγοντες χαρακτηρίζονται ως ανεξάρτητες μεταβλητές. Αυτό σημαίνει ότι τα πειράματα επιδρούν κατά τέτοιο τρόπο έτσι ώστε τα αποτελέσματα να χρησιμοποιηθούν και να συσχετισθούν σε σχέση με τις εξαρτώμενες μεταβλητές του γενικότερου πειράματος. Μερικοί παράγοντες όπως ηλικία ή βάρος θεωρούνται βασικές μεταβλητές και τις περισσότερες φορές χρησιμοποιούνται στον σχεδιασμό της πειραματικής ομάδας (control).

Παρακάτω παρατείθονται διαφορετικοί τρόποι παρουσίασης και σύγκρισης διαφορετικών ομάδων με την βοήθεια πινάκων διπλής εισόδου.

1. ***Σύγκριση μεταξύ διαφορετικών μεταβλητών (ενός παράγοντα)***

Η μεταβλητή συγκρίνεται μόνο με μία συνθήκη κάθε φορά *(the one factor between subjects experiments)*

	Παράγοντας		
Επίπεδα	Control	παραγοντας 1	παραγοντας 2
Ομάδες	Ομάδα 1	Ομάδα 2	Ομάδα 3

Παράδειγμα: Σύγκριση της χρήσης διαφορετικών φαρμάκων με την βασική ομάδα (control)

	Φαρμακο		
Επίπεδα	Control	Φαρμακο 1	Φαρμακο 2
Ομάδες	Ομάδα 1	Ομάδα 2	Ομάδα 3

2. ***<u>Σύγκριση μεταξύ ιδίων μεταβλητών (ενός παράγοντα)</u>***

Η μεταβλητή συγκρίνεται μόνο με όλες τις συνθήκες (επαναλαμβανόμενα πειράματα) *(the one factor within subjects experiments)*

	Παράγοντας		
Επίπεδα	Παραγ. 1	Παραγ. 2	Παραγ. 3
Ομάδες	Τιμές (ίδιο πείραμα για όλες τις μεταβλητές)		

Παράδειγμα: Θέλουμε να ελέγξουμε την ακρίβεια ενός όπλου με βάση 3 σχήματα (Κύκλος, Τετράγωνο, Τρίγωνο)

	Σχήμα		
Επίπεδα	Κύκλος	Τετράγωνο	Τρίγωνο
Ομάδες	Τιμές (ίδιο πείραμα για όλες τις μεταβλητές)		

3. ***<u>Σύγκριση μεταξύ διαφορετικών μεταβλητών (δύο παράγοντες)</u>***

Η μεταβλητή συγκρίνεται με περισσότερες από μία συνθήκη κάθε φορά *(the two factor between subjects experiments)*

	Παράγοντας		
Επίπεδα	Control	Παραγ. 1	Παραγ. 2
Ομάδες 1	Group 1	Group 2	Group 3
Ομάδες 2	Group 4	Group 5	Group 6

Παράδειγμα: Θέλουμε να ελέγξουμε την επίδραση 3 φαρμάκων σε άτομα που έχουν κάνει κάποια εξάσκηση και σε άτομα που δεν έχουν εξασκηθεί.

	Φάρμακο		
Επίπεδα	Φαρμακο 1	Φαρμακο 2	Φαρμακο 3
Εξασκηση	Group 1	Group 2	Group 3
Μη εξασκηση	Group 4	Group 5	Group 6

4. ***<u>Σύγκριση μεταξύ ιδίων μεταβλητών (δύο παράγοντες)</u>***

Η μεταβλητή συγκρίνεται μόνο με όλες τις συνθήκες (επαναλαμβανόμενα πειράματα) τόσες φορές όσες και οι παράγοντες. *(the two factor within subjects experiments)*

	Παράγοντας		
Επίπεδα 1	Παραγ. 1	Παραγ. 2	Παραγ. 3
Επίπεδα 2	Παραγοντας 4		
Ομάδες	Τιμές (ίδιο πείραμα για όλες τις μεταβλητές)		

Παράδειγμα: Θέλουμε να ελέγξουμε την ακρίβεια ενός όπλου με βάση 3 σχήματα (Κύκλος, Τετράγωνο, Τρίγωνο) και 2 χρώματα (Κόκκινο, Μπλε) για άνδρες και γυναίκες

	Σχήμα		
Επίπεδα 1	Κύκλος	Τετράγωνο	Τρίγωνο
Επίπεδα 2	Κόκκινο-Μπλε	Κόκκινο-Μπλε	Κόκκινο-Μπλε
Άνδρες	Τιμές (ίδιο πείραμα για όλες τις μεταβλητές)		
Γυναίκες			

7 ΓΡΑΦΗΜΑΤΑ-ΔΙΑΓΡΑΜΜΑΤΑ

Γράφημα ονομάζεται μια γραφική αναπαράσταση μίας ή περισσοτέρων μεταβλητών. Τα γραφήματα είναι χρήσιμα για να βλέπουμε και να καταλαβαίνουμε το σχήμα της κατανομής μίας μεταβλητής. Είναι χρήσιμα επίσης για να δούμε οπτικά τη σχέση ανάμεσα σε δύο ή περισσότερες μεταβλητές.

Κύριος **<u>στόχος</u>** της στατιστικής ανάλυσης είναι να αντληθούν όσο το δυνατό περισσότερες πληροφορίες από τα δεδομένα. Θα πρέπει να εξηγήσουμε και όχι να δώσουμε ερμηνείες στα στοιχεία. Βασικά χαρακτηριστικά για την δημιουργία ενός γραφήματος είναι:

- ✓ Να ξεκαθαρίσουμε τους στόχους και τις προτεραιότητες σε ότι αφορά το μήνυμα που θέλουμε να δώσουμε.
- ✓ Να επιλέξουμε το κατάλληλο είδος γραφικής παράστασης.
- ✓ Να ενημερώσουμε τον αναγνώστη σχετικά με την φύση των απεικονιζόμενων

πληροφοριών με σαφή τίτλο.

- ✓ Να κατασκευάσουμε ένα σχεδιάγραμμα το οποίο να είναι: παραστατικό, σαφές και ακριβές.

Θα πρέπει να δοθεί **προσοχή**:

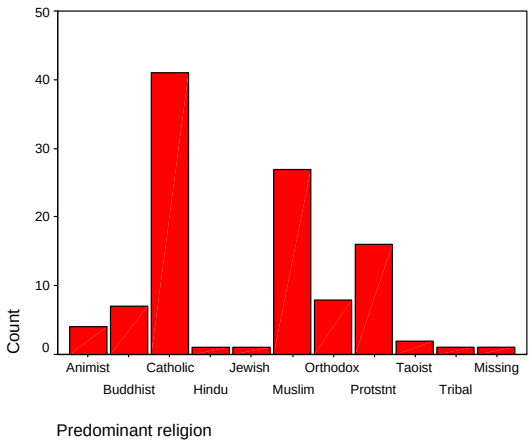
- ✓ Στις διαστάσεις του διαγράμματος.
- ✓ Στις χρησιμοποιούμενες γραμμοσκιάσεις και χρωματισμούς.
- ✓ Στην ορθολογική χρήση των εργαλείων λογισμικού.
- ✓ Στις τυχόν διευκρινήσεις που είναι απαραίτητες να δοθούν

Ένας οδηγός διαγραμμάτων δίνεται στον παρακάτω πίνακα

Τύπος γραφήματος	Τύπος μεταβλητών	Διάγραμμα
Μιάς διάστασης	Κατηγορική	Ραβδόγραμμα, Κυκλικό
	Αριθμητική	Ιστόγραμμα, Διάγραμμα μίσχου-φύλλου, Διάγραμμα πλαισίου
Δύο διαστάσεων	Δύο ποσοτικές	Διάγραμμα σημείων
	Δύο ποιοτικές	Ραβδόγραμμα
	Μια ποσοτική – μιά ποιοτική	Διάγραμμα πλαισίου, Διάγραμμα σφαλμάτων
Πολλών διαστάσεων	Ποσοτικές - ποιοτικές	Πίνακες διαγραμμάτων σημείων, Αστεροειδή γραφήματα, Πρόσωπα του Chernoff

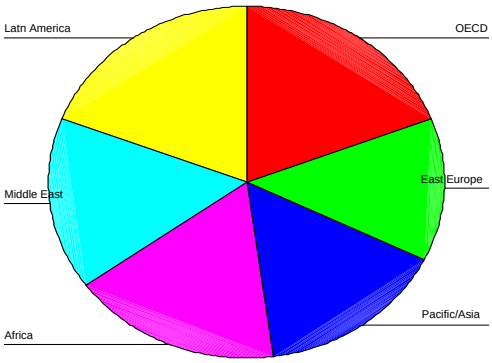
7.1 ΡΑΒΔΟΓΡΑΜΜΑ (Bar chart)

Περιγράφει τις κατηγορίες μίας ποιοτικής μεταβλητής με ράβδους. Το ύψος της κάθε ράβδου συνήθως είναι ανάλογο του πραγματικού αριθμού ή του ποσοστού που αντιστοιχεί σε κάθε κατηγορία.



7.2 ΚΥΚΛΙΚΟ ΔΙΑΓΡΑΜΜΑ (Pie chart)

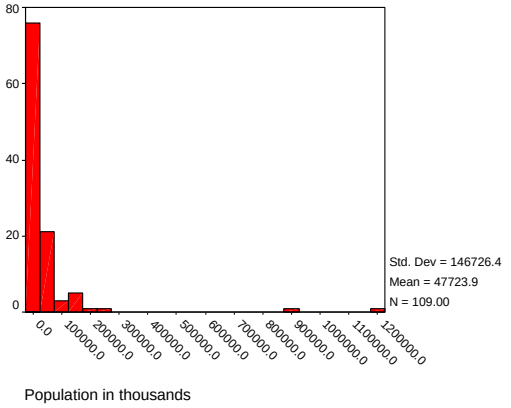
Περιγράφει τις κατηγορίες μίας ποιοτικής μεταβλητής με κομμάτια μίας πίτας (κύκλου). Το κομμάτι κάθε κατηγορίας είναι ανάλογο του αριθμού των αντικειμένων που ανήκουν σε κάθε κατηγορία. Παρουσιάζουν μία στατική εικόνα ενός δείγματος ή πληθυσμού.

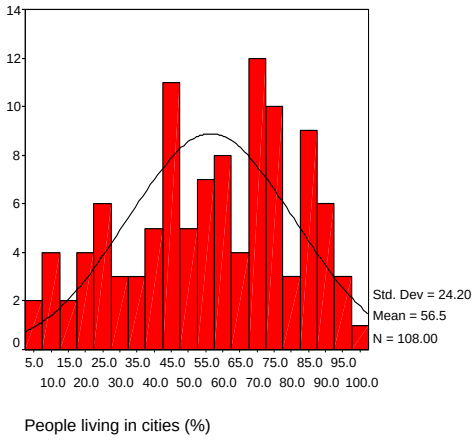


7.3 ΙΣΤΟΓΡΑΜΜΑ (Histogram)

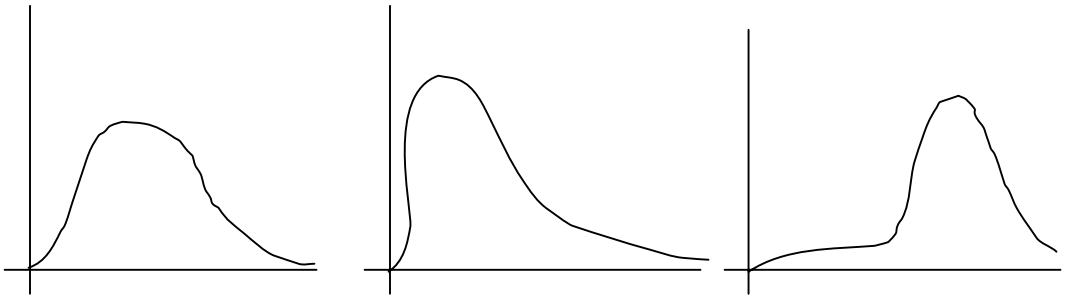
Απεικονίζει την κατανομή μίας ποσοτικής μεταβλητής με την βοήθεια ράβδων. Κάθε ράβδος αντιστοιχεί σε ένα διάστημα τιμών και το ύψος είναι ανάλογο των αντικειμένων που ανήκουν σε αυτό το διάστημα Συχνά απεικονίζουμε και την γραμμή της κανονικής κατανομής γραμμής για σύγκριση. Στον άξονα x εμφανίζονται είτε μεμονωμένες τιμές είτε σύνολο τιμών. Το ύψος κάθε τμήματος αναπαραστά τη συχνότητα με την οποία εμφανίζεται αυτή η τιμή ή το διάστημα. Αντίθετα με το διάγραμμα μπάρας στο οποίο ο άξονας x παριστάνει κατηγορίες, ο οριζόντιος άξονας ενός ιστογράμματος δείχνει μια αριθμητική κλίμακα. Επίσης χρησιμοποιείται για να δείξει το σχήμα μίας μεταβλητής. Το ιστόγραμμα κατασκευάζεται με βάση το πίνακα συχνοτήτων ομαδοποιημένων μετρήσεων. Αν η πρώτη ή η τελευταία ομάδα είναι

ανοικτή παραλείπεται στην κατασκευή του ιστογράμματος. Το μήκος των ομάδων επηρεάζει την εικόνα του ιστογράμματος. Όσο πιο μεγάλο είναι το μήκος των ομάδων τόσο πιο ασαφές και ακαθόριστο είναι το σχήμα της κατανομής των μετρήσεων. Αν οι ομάδες είναι πολλές και έχουν μικρό μήκος το ιστόγραμμα παρουσιάζει ανωμαλίες που αντανακλούν την μεταβλητικότητα της δειγματοληψίας.





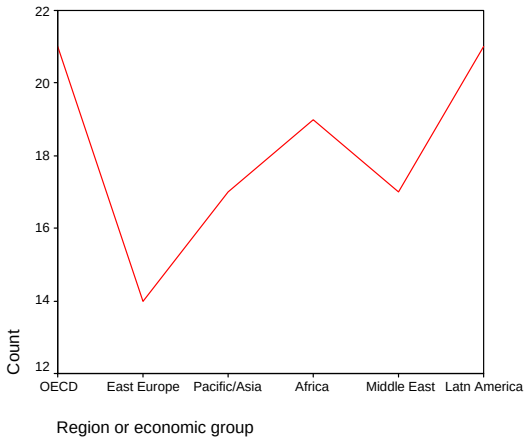
Επίσης το ιστόγραμμα δείχνει την ασυμμετρία (skewness) μίας κατανομής που είναι μέτρο της συμμετρίας της. Αν μία κατανομή είναι συμμετρική, η αριστερή πλευρά της είναι ένα κατοπτρικό είδωλο της δεξιάς. Αν οι τιμές μίας κατανομής συσσωρεύονται στις μικρές τιμές η κατανομή είναι θετικά ασύμμετρη. Αν οι τιμές της κατανομής βρίσκονται στο άλλο άκρο της κλίμακας η κατανομή είναι αρνητικά συμμετρική.



συμμετρική Θετικά συμμετρική Αρνητικά συμμετρική

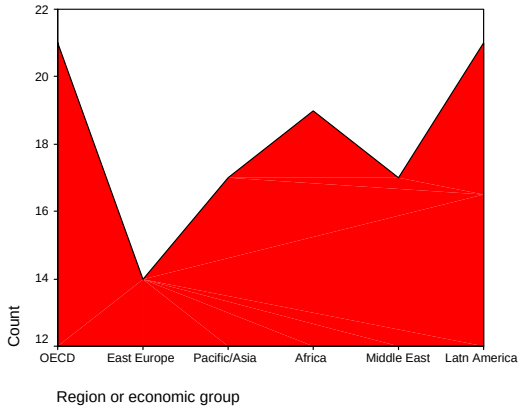
7.4 ΠΟΛΥΓΩΝΟ ΣΥΧΝΟΤΗΤΩΝ

Το πολύγωνο συχνοτήτων αποτελεί το δεύτερο τρόπο γραφικής σύμπτυξης και παρουσίασης των μετρήσεων. Κατασκευάζεται ενώνοντας με ευθείες γραμμές τα μέσα των πάνω πλευρών του ορθογωνίων του ιστογράμματος. Παρουσιάζει με πιο ευδιάκριτο τρόπο απ'ότι το ιστόγραμμα τη μορφή των στοιχείων. Χρησιμοποιείται μόνο για την διαγραμματική απεικόνιση συνεχών δεδομένων και ενδείκνυται κυρίως για την ταυτόχρονη σύγκριση δύο ή περισσοτέρων κατανομών. Η κατασκευή του σχεδιάζεται είτα με βάση το ιστόγραμμα της εξεταζόμενης κατανομής είτε ανεξάρτητα από αυτό.



7.5 ΕΜΒΑΔΟΓΡΑΜΜΑ

Σύγκριση με βάση το εμβαδόν κάποιας περιοχής. Τα γραφήματα μπορεί να είναι απλά ή συσσωρευμένα τα οποία συγκρίνουν την ίδια μεταβλητή, ή μεταβλητές μεταξύ τους ή περιπτώσεις μεταξύ τους.



7.6 ΔΙΑΓΡΑΜΜΑ ΜΙΣΧΟΥ-ΦΥΛΛΟΥ (Steam-and-leaf)

Η τεχνική αυτή οδηγεί σε σχηματική αναπαράσταση παρόμοια με εκείνη του ιστογράμματος χωρίς να οδηγεί σε απώλεια πληροφοριών. Το διάγραμμα αυτό δίνει την δυνατότητα ανασύστασης και ανάκλησης των μετρήσεων των αρχικών δεδομένων του δείγματος με ακρίβεια πράγμα το οποίο δεν επιτυγχάνεται με το ιστόγραμμα ή τους πίνακες συχνοτήτων. Χρησιμοποιείται για την επεξεργασία μέτριου αριθμού παρατηρήσεων (περίπου 150). Η παρουσίαση του σχήματος μοιάζει με εκείνου του ιστογράμματος αλλά η τεχνική κατάρτισης δεν είναι η ίδια.

Το φυλλόγραμμα:

- ✓ Εμφανίζει τα δεδομένα σε όλο το εύρος των παρατηρηθέντων μετρήσεων.

- ✓ Παρουσιάζει την συγκέντρωση των παρατηρήσεων (συχνότητες).
- ✓ Δείχνει την μορφή της κατανομής.
- ✓ Εμφανίζει τυχόν ακραίες και εκτροπές παρατηρήσεις.
- ✓ Επιτρέπει την επισήμανση της απουσίας συγκεκριμένων τιμών ή μετρήσεων.

Παράδειγμα 1:
Μας δίνονται ηλικίες 10 ατόμων 27 34 34 43 21 38 46 38 22 35 και ζητάμε να βρούμε το διάγραμμα μίσχου-φύλλου.
Τα βήματα που ακολουθούμε είναι τα έξης

- ❖ Διατάσουμε τα δεδομένα σε αύξουσα σειρά
  21 22 27
  34 34 35 38 38
  43 46
- ❖ Θεωρούμε ότι κάθε παρατήρηση αποτελείται από δύο τμήματα, το αρχικό ψηφίο και το επόμενο ψηφίο π.χ. 21 – 2 αρχικό, 1 επόμενο.
- ❖ Κατασκευάζουμε τον παρακάτω πίνακα για όλα τα νούμερα

κορμός	φύλλο
2	1 2 7
3	4 4 5 8 8
4	3 6

Ως **φύλλο** κάθε στοιχείου λαμβάνεται το τελευταίο ή τα δύο τελευταία ψηφία της τιμής της παρατήρησης και ως **κορμός** το πρώτο ή τα εναπομείναντα πρώτα ψηφία. Π.χ 789 – 7|89 ή 78|9

Παράδειγμα 2:
Μας δίνονται οι επιδόσεις 20 φοιτητών σε ένα τεστ 841 855 764 700 1000 1117 946 920 899 872 928 873 873 1070 855 755 855 709 1105 928 και ζητάμε να βρούμε το διάγραμμα μίσχου-φύλλου.

- ❖ Κατασκευάζουμε τον παρακάτω πίνακα για όλα τα νούμερα

κορμός	φύλλο
7	00 09 55 64
8	44 55 55 55 72 73 73 99
9	20 28 28 46
10	00 70
11	05 17

Διαπιστώνουμε ότι

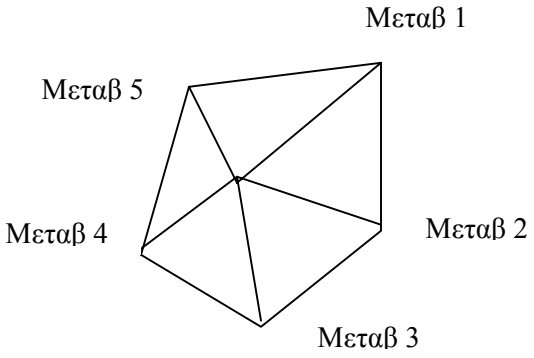
- Η μικρότερη τιμή είναι το 700 και μεγαλύτερη το 1117.
- Η πολυπληθέστερη ομάδα είναι εκείνη που περιλαμβάνει μετρήσεις από 800 – 900.
- Τα τέσσερα φύλλα του μίσχου 7 δηλώνουν ότι ισάριθμα άτομα πέτυχαν επίδοση από 700 – 800 μονάδες.
- Τα 2 φύλλα του τελευταίου μίσχου δηλώνουν ότι βρέθηκαν 2 άτομα που πέτυχαν επίδοση 1100 ή καλύτερη.

Αν περιστρέψουμε νοητά το ανωτέρω διάγραμμα σε οριζόντια θέση θα ήταν σα να σκιαγραφούσαμε 5 παραλληλόγραμμα το πρώτο με συχνότητα 4 μονάδες, το δεύτερο με συχνότητα 8 μονάδες κ.λ.π.

7.7 ΑΣΤΕΡΟΕΙΔΗ ΔΙΑΓΡΑΜΜΑΤΑ

Κάθε παρατήρηση απεικονίζεται σαν ένα αστέρι το οποίο αποτελείται από τόσες ακτίνες όσες και οι μεταβλητές που χρησιμοποιούμε. Το μήκος του είναι ανάλογο των τιμών των μεταβλητών. Τα αστεροειδή διαγράμματα μας βοηθούν να κατατάξουμε τις παρατηρήσεις σε ομοειδής ομάδες με το μέγεθος και το σχήμα των αστέρων. Τα μήκη των αστέρων καθορίζονται με δύο τρόπους

- Από τις πραγματικές τιμές
- Από τις αναδιαβαθμισμένες τιμές. Στην περίπτωση αυτή οι τιμές αλλάζουν έτσι ώστε η ελάχιστη τιμή να γίνει 0 (όχι ακτίνα) και η μέγιστη 1 (μεγάλη ακτίνα).

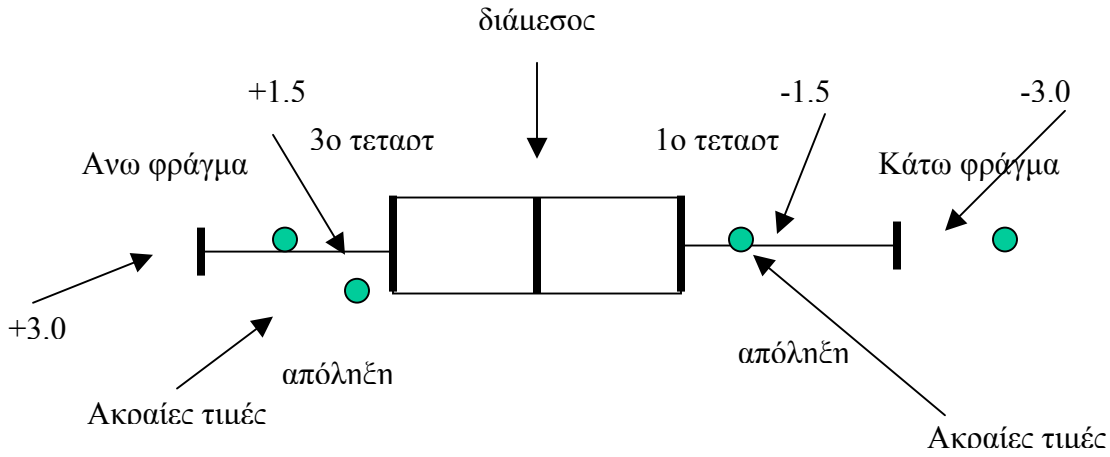


7.8 ΠΡΟΣΩΠΑ ΤΟΥ CHERNOFF

Είναι γραφήματα ανάλογα των αστεροειδών μόνο που αντί η κάθε παρατήρηση να απεικονίζεται σαν ακτίνα απεικονίζεται σαν πρόσωπο. Κάθε μεταβλητή καθορίζει ένα συγκεκριμένο χαρακτηριστικό του προσώπου. Η απεικόνιση αυτή είναι διαδεδομένη στις κοινωνικές επιστήμες και την ψυχολογία.

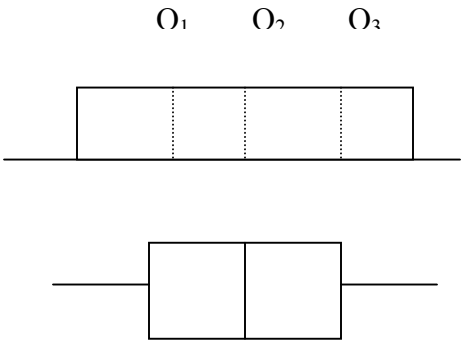
7.9 ΔΙΑΓΡΑΜΜΑΤΑ ΠΛΑΙΣΙΟΥ-ΑΠΟΛΗΞΕΩΝ (Box plot)

Περιλαμβάνουν περιληπτικά την κατανομή των ποσοτικών μεταβλητών. Κάθε πλαίσιο-κουτί απεικονίζει το 1ο τεταρτημόριο, την διάμεσο και το 3ο τεταρτημόριο. Οι απολήξεις υποδεικνύουν τα όρια των ακραίων τιμών. Οι τιμές εκτός των φραγμάτων των απολήξεων θεωρούνται ακραίες και υποδεικνύονται στο γράφημα με ξεχωριστά σημεία. Συμμετρικά διαγράμματα πλησιάζουν την κανονική κατανομή.

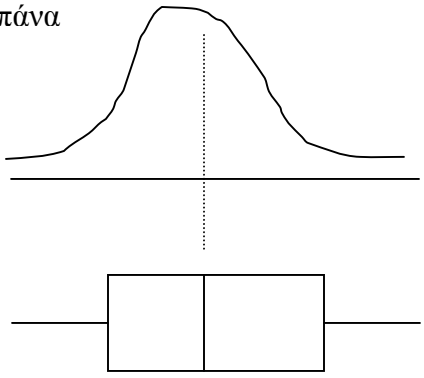


Αναφερόμενοι στο σχήμα της κατανομής των δεδομένων με βάση το διάγραμμα box-plot, έχουμε τις παρακάτω σχεδιάσεις (Q_1 – 3ο τεταρτημόριο, Q_2 – διάμεσος, Q_3 – 1ο τεταρτημόριο

Ομοιόμορφη

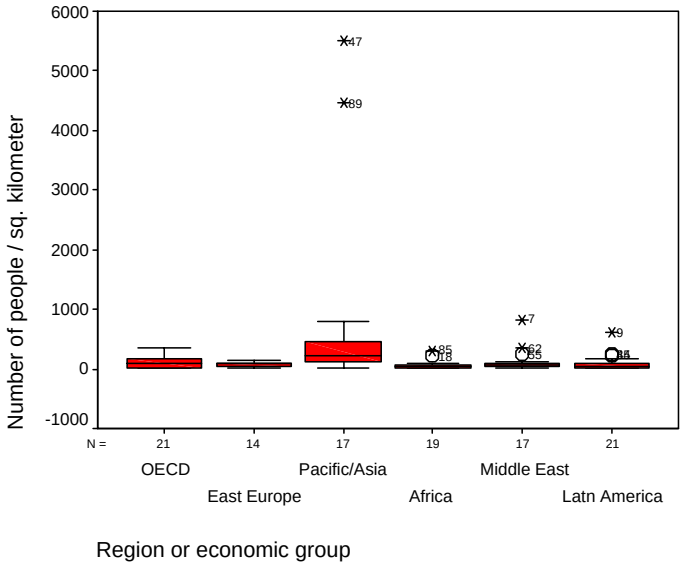


Καμπάνα



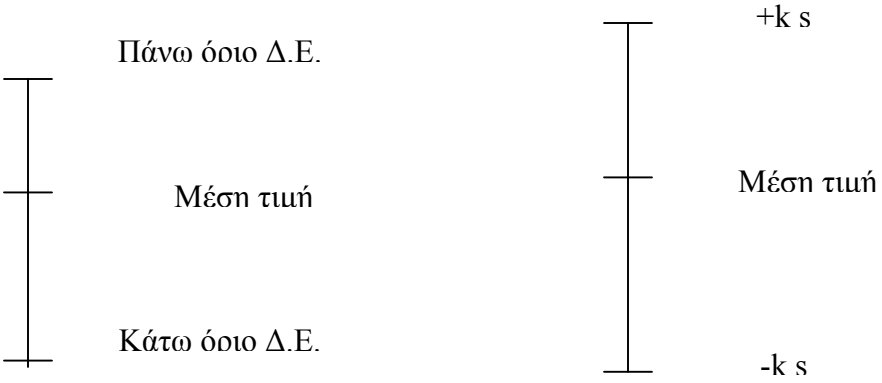
Δεξιά συμμετρική

Αριστερα συμμετρική

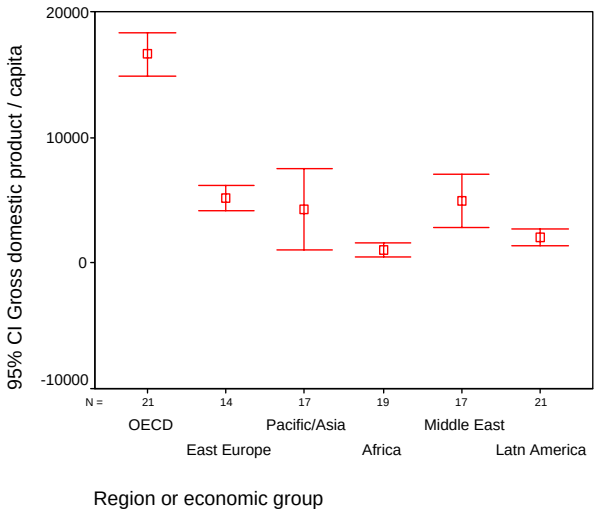


7.10 ΔΙΑΓΡΑΜΜΑΤΑ ΣΦΑΛΜΑΤΩΝ (Error-bars)

Είναι διαγράμματα που περιγράφουν περιληπτικά την κατανομή μίας ποσοτικής μεταβλητής σε διάφορα επίπεδα. Συνήθως αναπαραστά διαστήματα εμπιστοσύνης για το μέσο, αλλά εναλλακτικά μπορεί να χρησιμοποιηθεί και για την τυπική απόκλιση. Μοιάζουν με τα θηκογράμματα, αλλά συγκρίνουν διαστήματα εμπιστοσύνης και όχι κατανομές. Το σχήμα τους είναι μια ράβδος κατανεμημένη ισομερώς γύρο από το τη μέση τιμή των τιμών. Το μήκος κάθε ράβδου ισούται με το 100(1-α)% διάστημα εμπιστοσύνης γύρο από το μέσο.



όπου k μπορεί να πάρει τιμές 1 (για 70%), 2 (για 95%) και 3 (για 99%).



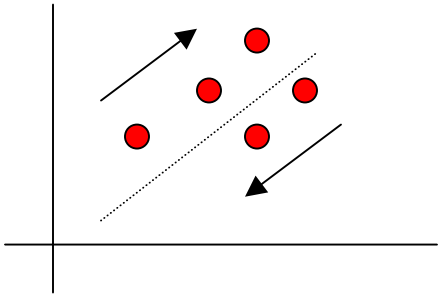
7.11 ΔΙΑΓΡΑΜΜΑΤΑ ΣΗΜΕΙΩΝ (Scatter plots)

Ονομάζονται και **διαγράμματα** διασποράς. Αποτελούν το γραφικό τρόπο αναζήτησης σχέσεων μεταξύ μεταβλητών. Περιγράφουν τη δυσδιάστατοι κατανομή δύο ποσοτικών μεταβλητών. Κάθε σημείο απεικονίζει ένα ζευγάρι τιμών των υπό εξέταση μεταβλητών (συσχέτιση μεταβλητών). Εντοπίζονται εύκολα συσχετίσεις και ακραίες τιμές.

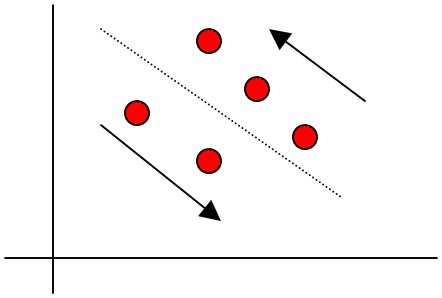
Συσχέτιση αναφέρεται ο βαθμός με τον οποίο σχετίζονται (συμμεταβάλλονται) δύο μεταβλητές. Ο βαθμός σχέσης με την οποία σχετίζονται δύο μεταβλητές ονομάζεται απλή συσχέτιση και ο βαθμός σχέσης με την οποία σχετίζονται περισσότερες μεταβλητές ονομάζεται πολλαπλή συσχέτιση. Η απλή συσχέτιση ασχολείται με το βαθμό το οποίο τα σημεία συγκεντρώνονται γύρω από μια ευθεία χωρίς να προσδιορίζεται ποιά είναι ακριβώς αυτή η γραμμή που διέρχεται μέσα από το νέφος των σημείων.

Γραμμική συσχέτιση εμφανίζεται όταν στο διάγραμμα διασκορπισμού τα σημεία όλων των παρατηρήσεων τείνουν να συγκεντρώνονται γύρο από μια ευθεία και **μη γραμμική** όταν τα σημεία τείνουν να συγκεντρωθούν γύρο από μια καμπύλη.

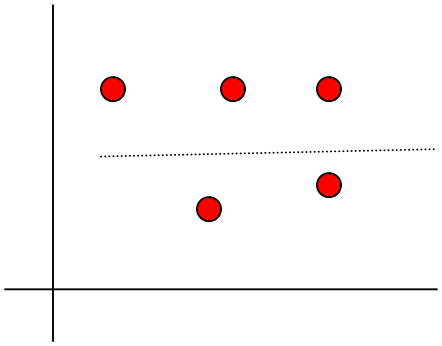
- **Θετική συσχέτιση** όταν δύο μεταβλητές τείνουν να μεταβάλλονται προς την ίδια κατεύθυνση. Στην περίπτωση αυτή οι τιμές τείνουν να αυξάνονται ή να μειώνονται.



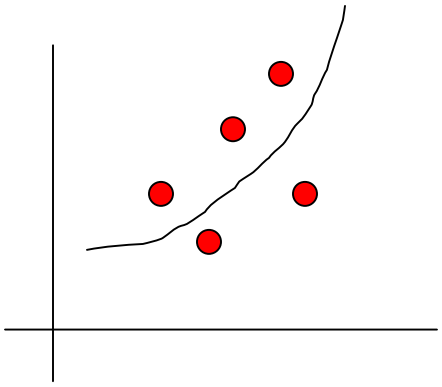
- **Αρνητική συσχέτιση** όταν δύο μεταβλητές τείνουν να μεταβάλλονται προς αντίθετη κατεύθυνση. Στην περίπτωση αυτή οι τιμές της μίας μεταβλητής τείνουν να αυξάνονται και της άλλης να μειώνονται

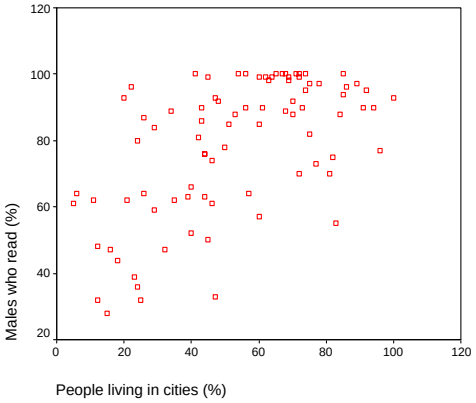


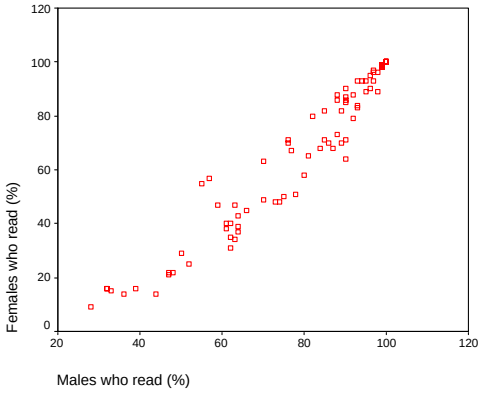
- **Μηδενική συσχέτιση** όταν οι μεταβολές των τιμών της μίας μεταβλητής δεν συνδέονται με τις μεταβολές της άλλης. Τα σημεία του νέφους είναι διασκορπισμένα σε όλο το μήκος του διαγράμματος



- **<u>Μη γραμμική συσχέτιση</u>** όταν οι μεταβολές των τιμών της μίας μεταβλητής συνδέονται με τις μεταβολές της άλλης με μη γραμμική μορφή. Τα σημεία του νέφους τείνουν να συγκεντρωθούν γύρο από μια καμπύλη





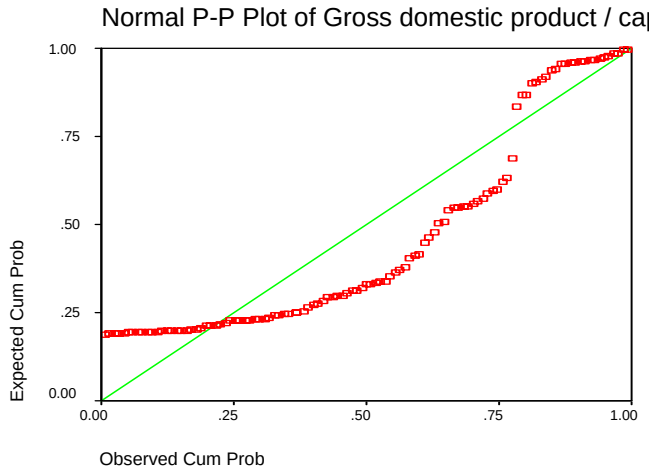


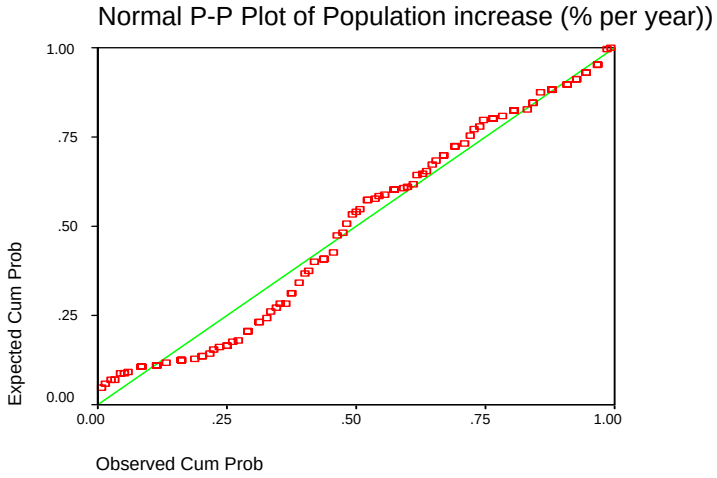
7.12 ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΑ ΔΙΑΓΡΑΜΜΑΤΑ (P-P, Q-Q)

Προσπαθούμε να διαπιστώσουμε (με γραφικούς τρόπους) πόσο κοντά σε κάποια κατανομή που προσδιορίζουμε είναι τα δεδομένα. Συνήθως ενδιαφερόμαστε για την κανονική κατανομή. Όσο πιο κοντά στην διχοτόμο της γωνίας των αξόνων είναι συγκεντρωμένα τα σημεία τόσο περισσότερο ενισχύεται η υπόθεση ότι τα δεδομένα ακολουθούν την κανονική κατανομή. <u>Πιθανοθεωρητικά διαγράμματα</u> είναι διαγράμματα ενός δείγματος σε σχέση με τα δεδομένα που περιμέναμε να πάρουμε αν θεωρούσαμε από μια κανονική κατανομή.

- **<u>Normal P-P</u>**

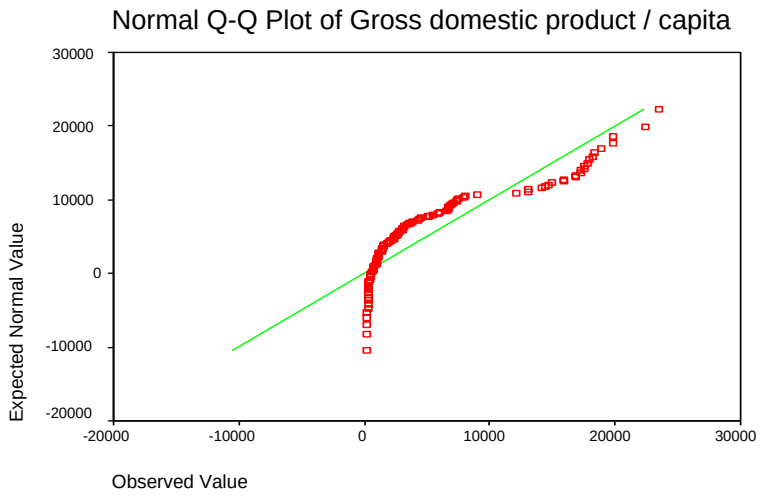
Τα γραφήματα αυτά παριστάνουν σε σύστημα οριζοντίων και καθέτων αξόνων τις τιμές της παρατηρούμενης αθροιστικής συχνότητας (άξονας x) και της υποτιθέμενης κανονικής κατανομής (άξονας y) που ακολουθεί η μεταβλητή που εξετάζεται. Όσο πιο κοντά στην διχοτόμο της γωνίας των αξόνων είναι συγκεντρωμένα τα σημεία τόσο περισσότερο ενισχύεται η υπόθεση ότι τα δεδομένα ακολουθούν την κανονική κατανομή.

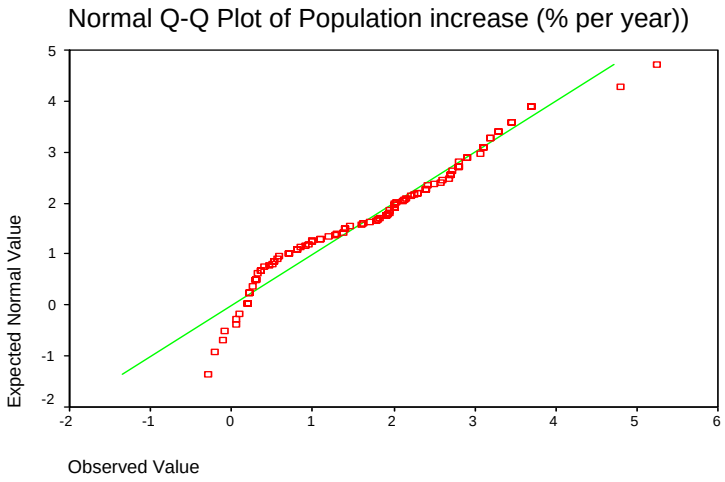




- **Normal Q-Q**

Τα γραφήματα αυτά παριστάνουν τα εκατοστημόρια της παρατηρούμενης ως προς την αναμενόμενη κανονική κατανομή. Τα εκατοστημόρια υπολογίζονται με βάση διαφορετικούς αλγόριθμους και τοποθετούνται στον κάθετο άξονα y, ενώ α αντίστοιχα παρατηρούμενα στον οριζόντια άξονα x

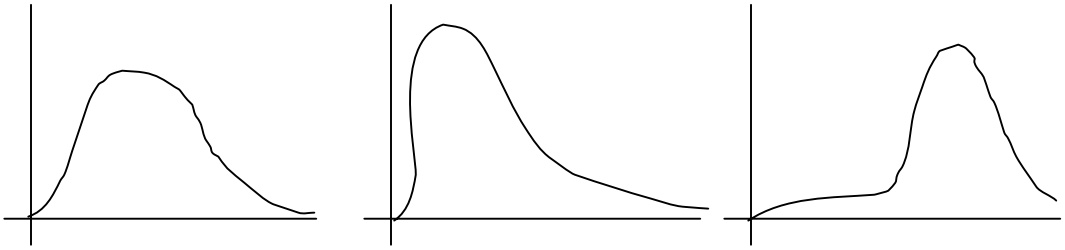




8 ΚΑΜΠΥΛΕΣ ΣΥΧΝΟΤΗΤΩΝ

Χρησιμοποιείται στην περίπτωση μεγάλου όγκου δεδομένων. Καταρτίζουμε ένα πίνακα συχνοτήτων και στην συνέχεια σχεδιάζουμε το αντίστοιχο ιστόγραμμα και πολύγονο συχνοτήτων. Αυξανόμενων του πλήθους των τάξεων η πολυγωνική γραμμή προσεγγίζει μια ομαλή καμπύλη η οποία ονομάζεται ***καμπύλη συχνοτήτων.*** Στην ουσία η καμπύλη συχνοτήτων αποτελεί την γραφική απεικόνιση επί του επιπέδου των ορθογωνίων αξόνων άπειρων σημείων των οπίων οι τετμημένες αντιστοιχούν στις άπειρες τιμές της συνεχούς μεταβλητής. Μπορούμε να διακρίνουμε 4 οικογένειες θεωρητικών κατανομών οι οποίες λαμβάνουν το όνομά τους από το σχήμα που παρουσιάζουν.

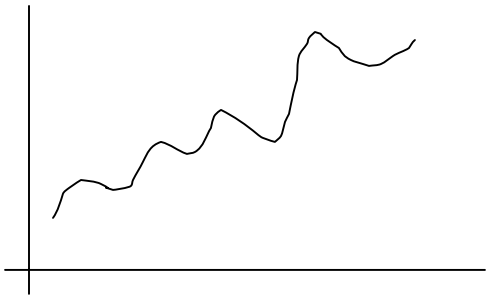
- Μονοκόρυφες κατανομές (unimodal distributions)



- Κατανομές σχήματος U (U-shape distributions)

- Κατανομές σχήματος J (J-shape distributions)

- Διάφορες σύνθετες κατανομές ιδιάζουσας μορφής.



8.1 ΜΟΝΟΚΟΡΥΦΕΣ

Το εμβαδόν της επιφάνειας που περικλείεται από τις καμπύλες και το οριζόντιο άξονα αντιστοιχεί στο σύνολο των συχνοτήτων. Οποιοδήποτε τμήμα του μπορεί αν εκτιμηθεί και να δώσει την αναλογία των δεδομένων που περιλαμβάνονται μεταξύ δύο τιμών της μεταβλητής.

8.2 ΣΥΜΜΕΤΡΙΚΗ-ΚΩΝΟΕΙΔΗΣ

Ο άξονας συμμετρίας χωρίζει την επιφάνεια σε δύο ίσα μέρη καθένα περιλαμβάνοντας το 50% των παρατηρήσεων. Τα δύο άκρα της καμπύλης ονομάζονται ουρές και τείνουν ασυμπωτικά προς τον άξονα x. (**κανονική** ή **τυποποιημένη κανονική κατανομή**).

9 ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Η δυσκολία χειρισμού πολυπληθών δεδομένων επιβάλει την επινόηση μεθόδων με την χρήση των οποίων είναι δυνατή η συμπύκνωση των πληροφοριών. Για το σκοπό αυτό χρησιμοποιούνται μέθοδοι διάταξης συχνοτήτων (σχετικών, απόλυτων και αθροιστικών) καθώς και διαφορετικά μέτρα (τάσης, θέσεως, διασποράς, λοξότητας και κύρτωσης) μιας ομάδας παρατηρήσεων.

9.1 ΣΥΧΝΟΤΗΤΕΣ

Συχνότητα (f_i) μιας μεταβλητής x_i της μεταβλητής X ονομάζεται ο φυσικός αριθμός n_i που θανερώνει πόσες φορές παρουσιάζεται στο δείγμα η συγκεκριμένη τιμή.
Σχετική συχνότητα ονομάζεται το πηλίκο της συχνότητας n_i προς το πλήθος των παρατηρήσεων $f_i = \frac{n_i}{\sum_{l=1}^{k} n_l}$
Αθροισμα συχνοτήτων λέγεται το άθροισμα των συχνοτήτων που είναι μικρότερες ή ίσες της τιμής αυτής. Ισούται με την τιμή 1.
Άθροισμα ποσοστιαίων συχνοτήτων λέγεται το άθροισμα των συχνοτήτων % που είναι μικρότερες ή ίσες της τιμής αυτής. Ισούται με την τιμή 100.

- *Αριθμητικές τεχνικές ομαδοποίησης ποσοτικών δεδομένων*

1. Διάταξη **συνεχών δεδομένων** κατά αύξουσα σειρά
2. Υπολογισμός του εύρους των τιμών ($R=E_{max}-E_{min}$)

3. Υπολογισμός του πλάτους διαστημάτων $\delta = \frac{R}{k}$ όπου κ ο επιθυμητός αριθμός ή $\delta = \frac{R}{1+3.322\log n}$ όπου n συνολικός αριθμός παρατηρήσεων.
4. Τα ανώτερα και κατώτερα όρια των διαστημάτων συμβολίζονται με U_i και L_i . Η διαφορά μεταξύ 2 ορίων προσδιορίζει το **πλάτος** δ_i με $\delta_i = U_i - L_i$, το δε ημιάθροισμα καλείται **κεντρική τιμή διαστήματος** w_i με $w_i = \frac{U_i + L_I}{2}$.

Παράδειγμα: Το βάρος 10 ατόμων σε κιλά δίνεται από τις παρακάτω τιμές:
65.1, 66.8, 67.3, 67.9, 70.0, 72.3, 72.3, 73.0, 73.0, **89.9**

R=89.9-65.1=24.8
$\delta = \frac{R}{k} = \frac{24.8}{3} = 8.26 \approx 8$ όπου κ=3 ο επιθυμητός αριθμός των κλάσεων που θέλουμε.

Αριθμός τάξεων (i)	Τάξεις	Απόλυτες συσχότητες
1	[65-73)	7
2	[73-81)	2
3	[81-90)	1

❖ *Αριθμητικές τεχνικές ομαδοποίησης ποιοτικών δεδομένων*

1. Διάταξη **ποιοτικών δεδομένων** κατά αύξουσα σειρά
2. Καταμέτρηση των τιμών
3. Κατάρτιση πίνακα απολύτων και σχετικών συχνοτήτων.

9.2 ΜΕΤΡΑ ΠΕΡΙΓΡΑΦΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Κάθε πληθυσμός χαρακτηρίζεται από ορισμένες ιδιότητες οι οποίες προσδιορίζουν την φυσιογνωμία και την δομή του. Οι ιδιότητες αυτές εκφράζονται μέσω κάποιων αριθμητικών μεγεθών-μέτρων. Για ένα σύνολο ποσοτικών δεδομένων, το στατιστικό περιγραφικό μέτρο είναι ένας και μόνο αριθμός, ο οποίος υπολογίζεται από τα διαθέσιμα στοιχεία και παρέχει πληροφορίες σχετικά με την μορφή και δομή των δεδομένων. Το κύριο πλεονέκτημα των μέτρων είναι ότι μπορούν να χρησιμοποιηθούν για την διενέργεια εκτιμήσεων και τον έλεγχο στατιστικών υποθέσεων.

Τα στατιστικά μέτρα πρέπει να τηρούν ορισμένες ιδιότητες οι οποίες είναι γνωστές ως **συνθήκες του Yule**.

1. Η τιμή του μέτρου πρέπει να υπολογίζεται με αντικειμενικό τρόπο και με την συμμετοχή όλων των παρατηρήσεων.

2. Η τιμή του να υπολογίζεται με τον κατά δυνατότερο απλούστερο τρόπο και να έχει συγκεκριμένη σημασία ώστε να είναι κατανοητή ακόμα και σε άτομα που δεν είναι εξοικειωμένα με στατιστικούς υπολογισμούς.
3. Το στατιστικό μέτρο να είναι όσο το δυνατό λιγότερο ευαίσθητο σε κυμάνσεις της δειγματοληψίας και να μπορεί να υπολογισθεί σε επόμενους στατιστικούς υπολογισμούς.

Μπορούμε να διακρίνουμε **5 είδη στατιστικών μέτρων** που αντιστοιχούν σε ισάριθμους τρόπους περιγραφής της κατανομής των δεδομένων.

1. Τα δεδομένα εμφανίζουν μια τάση να περιστρέφονται γύρω από μια κεντρική τιμή η εκφράζει την τυπική ή μέση τιμή τους. (**μέτρα κεντρικής τάσης**- μέσος όρος, αρμονικός, γεωμετρικός)
2. Μέτρα με σκοπό τον εντοπισμό της θέσης της κατανομής κατά μήκους του άξονα των τιμών της μεταβλητής (**μέτρα θέσης**- επικρατούσα τιμή, διάμεσος, ποσοστημόρια)
3. Ποσοτική αξιολόγηση της συγκέντρωσης των τιμών μιας μεταβλητής γύρω από την κεντρική τιμή τους (**μέτρα διασποράς**- διακύμανση, τυπική απόκλιση, συντελεστής μεταβλητότητας)
4. Το είδος της θετικής ή αρνητικής ασυμμετρίας (**μέτρα λοξότητας**- συντελεστής ασυμμετρίας)
5. Προσδιορισμός της κατανομής των τιμών γύρω από την κεντρική τιμή σε σχέση με τις ακραίες τιμές προσδιορίζοντας μονοκόρυφη καμπύλη με αιχμηρότητα (**μέτρα κύρτωσης**- συντελεστής κύρτωσης)

9.2.1 ΜΕΤΡΑ ΚΕΝΤΡΙΚΗΣ ΤΑΣΗΣ

• Αριθμητικός μέσος

Έστω η τυχαία μεταβλητή X και n παρατηρήσεις $\{X_1,\ldots,X_n\}$ τότε ορίζουμε την μέση τιμή των n παρατηρήσεων σαν

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

πιο συγκεκριμένα, εάν οι στατιστικές μονάδες X_i αποτελούν το σύνολο ενός πληθυσμού μεγέθους $n = N$ το μ είναι ο πληθυσμιακός μέσος ενώ εάν $n < N$ δηλαδή οι παρατηρήσεις X_i αποτελούν απλώς ένα δείγμα μεγέθους n από τον προηγούμενο πληθυσμό, τότε το $2^{ο}$ μέλος στην σχέση συνήθως συμβολίζεται με $\overline{X}_n$ και ονομάζεται δειγματικός μέσος με τύπο

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Στην περίπτωση ομαδοποιημένων παρατηρήσεων (όπου συχνότητες περιπλέκονται) τότε το άθροισμα του αριθμητή και των δύο μέσων αντικαθίστανται από το τύπο $\sum_{i=1}^{N} f_i x_i$.

Ιδιότητες

- Αν όλες οι τιμές των παρατηρήσεων είναι ίσες με μια ποσότητα τότε ο μέσος είναι ίσος με την ποσότητα αυτή: $x_1 = x_2 = \cdots x_n = a \Rightarrow \bar{x} = a$

- Η τιμή του μέσου βρίσκεται μεταξύ ελάχιστης και μέγιστης τιμής $x_{\min} \le \bar{x} \le x_{\max}$

- Το άθροισμα των αποκλίσεων των τιμών από τον μέσο είναι πάντα ίσο με 0 ανεξάρτητα από N, τις τιμές των δεδομένων και την τιμή του μέσου. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ ανεξάρτητα $\mathrm{x_i}, n, \bar{x}$

- Το άθροισμα των τετραγώνων των αποκλίσεων των τιμών της μεταβλητής από μία ποσότητα α είναι ελάχιστο όταν οι αποκλίσεις αυτές υπολογίζονται από τον αριθμητικό μέσο των δεδομένων $\sum_{i=1}^{n} (x_i - a)^2 \to \min \Leftrightarrow \mathrm{a} = \bar{\mathrm{x}}$

- Αν σε όλες τις τιμές ποσοτικών δεδομένων προσθέσουμε μια σταθερά α τότε ο μέσος θα αυξηθεί (ή μειωθεί) κατά την ποσότητα αυτή.

- Αν πολλαπλασιάσουμε όλες τις τιμές ποσοτικών δεδομένων με μια σταθερά α τότε ο αριθμητικός μέσος πολλαπλασιάζεται με την ποσότητα αυτή.

• Σταθμικός αριθμητικός μέσος

Εάν ομαδοποιήσουμε τα δεδομένα μας $\{X_1, \ldots, X_n\}$ σε k κλάσεις της μορφής $(L_j, U_j]$ με πλάτος δ_j και με αντίστοιχες απόλυτες συχνότητες f_j για $1 \le j \le k$, τότε ο σταθμικός αριθμητικός μέσος δίνεται από τον τύπο

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{k} f_i w_i, \quad n = \sum_{i=1}^{k} f_i$$

όπου για <u>συνεχή δεδομένα</u> το $w_i = (L_i + U_i)/2$ είναι το κέντρο της i κλάσης ενώ για <u>ομαδοποιημένα διακριτά δεδομένα</u> $w_i = X_i$

• Γεωμετρικός μέσος

<u>Ο γεωμετρικός μέσος</u> n παρατηρήσεων $\{X_1, \ldots, X_n\}$ είναι η $n - οστη$ ρίζα του γινομένου τους δηλαδή:

$$G = \sqrt[N]{\prod_{I=1}^{N} x_i} \qquad \text{για τον πληθυσμό}$$

$$g = \sqrt[n]{\prod_{I=1}^{n} x_i} \qquad \text{για τον δείγμα}$$

Η τιμή του γεωμετρικού μέσου εκφράζεται σε μονάδες μέτρησης της μεταβλητής αλλά δεν μπορεί να υπολογιστεί αν υπάρχει έστω και μια μέτρηση με μηδενική ή αρνητική τιμή. Ο γεωμετρικός μέσος μπορεί να γραφεί ως

$$G = \sqrt[N]{\prod_{I=1}^{N} x_i} = \sqrt[N]{x_1 x_2 \cdots x_N} = (x_1 x_2 \cdots x_N)^{1/N} \Rightarrow \log G = \frac{1}{N}(\log x_1 + \log x_2 + \cdots + \log x_N) = \frac{\sum_{I=1}^{N} \log x_i}{N}$$

όπου η τελευταία έκφραση δηλώνει ότι ο λογάριθμος του γεωμετρικού μέσου είναι ίσος με τον αριθμητικό μέσο των λογαρίθμων των τιμών των παρατηρήσεων.

Ιδιότητες

- Το γινόμενο των πηλίκων είναι ίσο με 1 ανεξάρτητα από το N και τις τιμές των παρατηρήσεων δηλαδή $\prod_{i=1}^{N} u_i = \prod_{i=1}^{N}\left(\frac{u_i}{G}\right) = 1 \quad \text{ανεξάρτητα } x_i, N$

- Το άθροισμα των τετραγώνων των λογαρίθμων των λόγων u_i γίνεται ελάχιστο μόνο όταν οι λόγοι αυτοί υπολογίζονται με βάση το γεωμετρικό τους μέσο $\sum_{i=1}^{N} \log\left(\frac{x_i}{a}\right)^2 \rightarrow \min \Leftrightarrow a = G$

• Σταθμικός γεωμετρικός μέσος

Επί συνόλου n-παρατηρήσεων σε μια κατανομή συχνοτήτων με κ-τάξεις δίνεται από τον τύπο

$$g = \left(w_1^{f_1} w_2^{f_2} w_3^{f_3} \cdots w_k^{fk}\right)^{1/n}$$

όπου οι κεντρικές τιμές των τάξεων εμφανίζονται αντίστοιχα με συχνότητες με άθροισμα 1.

Ο λογάριθμος του σταθμικού γεωμετρικού μέσου είναι ίσος με το σταθμικό αριθμητικό μέσο που υπολογίζεται χρησιμοποιώντας ώς βάρη τους λογαρίθμους των κεντρικών τιμών των τάξεων της κατανομής.

- **Αρμονικός μέσος**

Ο αρμονικός μέσος *n* παρατηρήσεων $\{X_1,\ldots,X_n\}$ είναι το αντίστροφο του αριθμητικού μέσου των αντίστροφων τιμών των παρατηρήσεων δηλαδή

$$\frac{1}{H}=\frac{1/x_1+1/x_2+\cdots 1/x_N}{N}=\frac{\sum_{i=1}^{N}1/x_i}{N}\Rightarrow H=\frac{N}{\sum_{i=1}^{N}1/x_i}$$ για πληθυσμό

$$\frac{1}{h}=\frac{1/x_1+1/x_2+\cdots 1/x_n}{n}=\frac{\sum_{i=1}^{n}1/x_i}{n}\Rightarrow h=\frac{n}{\sum_{i=1}^{n}1/x_i}$$ για δείγμα

Η τιμή του γεωμετρικού μέσου εκφράζεται σε μονάδες μέτρησης της μεταβλητής αλλά δεν μπορεί να υπολογιστεί αν υπάρχει έστω και μια μέτρηση με μηδενική ή αρνητική τιμή

Ιδιότητες

- Το άθροισμα των αποκλίσεων των αντίστροφων τιμών από το αντίστροφο του αρμονικού μέσου είναι πάντα ίσο με 0 ανεξαρτήτως N και $x_i\prod_{i=1}^{N}u_i=\prod_{i=1}^{N}\left(\frac{1}{x_i}-\frac{1}{H}\right)=1$ ανεξάρτητα x_i, N

- Το άθροισμα των τετραγώνων των αρμονικών αποκλίσεων γίνεται ελάχιστο μόνο όταν οι αποκλίσεις υπολογίζονται από τον αρμονικό μέσο $\sum_{i=1}^{N}\left(\frac{1}{x_i}-\frac{1}{a}\right)^2\rightarrow\min\Leftrightarrow a=H$

- **Σταθμικός αρμονικός μέσος**

Ορίζεται ως το πηλίκο του πλήθους των διαθέσιμων παρατηρήσεων προς το άθροισμα των γινομένων των συχνοτήτων επί τα αντίστροφα των κεντρικών τιμών των διαστημάτων τάξεως $1/w_i$

$$h=\frac{\sum_{i=1}^{k}f_i}{\sum_{i=1}^{k}f_i\frac{1}{w_i}}$$

Παράδειγμα 1: Έστω τα δεδομένα 8,7,10,14,11. Ο υπολογισμό των μέσων δίνεται από τις ακόλουθες τιμές.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{5} = 10$$

$$g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[5]{86240} = 9.708$$

$$h = \frac{n}{\sum_{i=1}^{n} 1/x_i} = \frac{5}{0.5302} = 9.430$$

Παράδειγμα 2: Έστω τα παρακάτω ομαδοποιημένα δεδομένα εκφραζόμενα σε πίνακα διπλής εισόδου

Τάξεις	Συχνότητες (f_i)	Κέντρα (w_i)	Αριθμητικός (f_i w_i)	Γεωμετρικός (f_i log w_i)	Αρμονικός (f_i /w_i)
20-<30	2	25	50	2.796	0.080
30-40	4	35	14	6.176	0.114
40-50	4	45	200	6.796	0.080
	10		390	15.768	0.274

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{390}{10} = 39$$

$$g = \sqrt[n]{\prod_{i=1}^{n} x_i} = 37.74$$

$$h = \frac{n}{\sum_{i=1}^{n} 1/x_i} = \frac{10}{0.274} = 36.50$$

- **Σύγκριση των μέτρων κεντρικής τάσης**

Μεταξύ των 3 μέτρων κεντρικής τάσης, ο αριθμητικός μέσος επηρεάζεται περισσότερο από τους άλλους δύο από ακραίες τιμές των δεδομένων για το ίδιο σύνολο μετρήσεων, η προκύπτουσα τιμή του αριθμητικού μέσου είναι μεγαλύτερη μεταξύ των τριών ακολουθούμενη από εκείνη του γεωμετρικού και του αρμονικού δηλαδή $\bar{x} \geq g \geq h$

Παρατηρήσεις

- ✓ Οι τιμές των σταθμικών μέσων δεν επηρεάζονται αν χρησιμοποιηθούν οι σχετικές αντί τω απόλυτων συχνοτήτων.
- ✓ Οι τιμές των σταθμικών μέσων επηρεάζονται αν αλλάξει η ομαδοποίηση των δεδομένων (αν μεταβληθεί το πλήθος των τάξεων και συνεπώς το πλάτος και τα κέντρα των τάξεων
- ✓ Οι σταθμικοί μέσοι ικανοποιούν τις ιδιότητες και διατηρούν τα χαρακτηριστικά των αντίστοιχων αστάθμητων μέσων όρων.

9.2.2 ΜΕΤΡΑ ΘΕΣΗΣ

• Επικρατούσα τιμή

Λέγεται η τιμή της μεταβλητής που έχει την μεγαλύτερη συχνότητα εμφάνισης (η τιμή που επαναλαμβάνεται τις περισσότερες φορές) Συμβολίζεται με T όταν $n = N$ (επικρατούσα τιμή πληθυσμού) και με τ όταν $n < N$ (επικρατούσα τιμή δείγματος). Η επικρατούσα τιμή δεν ορίζεται μονοσήμαντα εφ' όσων υπάρχουν πολύ-κόρυφες καμπύλες συχνοτήτων[1] (δηλαδή είναι δυνατόν να υπάρξουν περισσότερα του ενός modes).

Στην περίπτωση ομαδοποιημένων δεδομένων και για μονοκόρυφη κατανομή η επικρατούσα τιμή πληθυσμού[2], δηλαδή το σημείο μέγιστης συχνότητας T της κατανομής (το μέγιστο της αντίστοιχης καμπύλης συχνοτήτων εάν πρόκειται για συνεχή δεδομένα) και κάτω από την παραδοχή της ομοιόμορφης κατανομής των παρατηρήσεων μέσα στις κλάσεις μπορεί να βρεθεί προσεγγιστικά σε μονάδες μέτρησης της μεταβλητής

$$T = U_{i-1} + \frac{(f_i - f_{i-1})\delta}{2f_i - f_{i-1} - f_{i+1}}$$

όπου U_{i-1} το κατώτερο όριο, f_i η συχνότητα της i-τάξης, f_{i-1} προηγούμενη , f_{i+1} επόμενη , δ το σταθερό πλάτος

• Διάμεσος

Η διάμεσος (median) είναι η κεντρική τιμή όταν διατάξουμε τις n μετρήσεις $\{X_1,\ldots,X_n\}$ σε αύξουσα διάταξη. Συμβολίζεται με M όταν $n = N$ (διάμεσος πληθυσμού) και με m όταν $n < N$ (διάμεσος δείγματος).
Με όρους πιθανότητας το σημείο M θα ικανοποιεί την σχέση $P(X \le M) = P(X \ge M) = 0.5$.

Μη ομαδοποιημένα δεδομένα: Έστω ότι οι παρατηρήσεις $\{X_1,\ldots,X_N\}$ ($n = N$) σε αύξουσα διάταξη είναι $X_1' \le \ldots \le X_N'$ τότε

[1] Οι μονοκόρυφες καμπύλες συχνοτήτων ονομάζοντα unimodal ενώ οι δικόρυφες bimodal
[2] Το ίδιο ισχύει και για την περίπτωση δείγματος.

$$M = \begin{Bmatrix} X'_{(N+1)/2} & N=περιττος \\ \frac{1}{2}(X'_{N/2}+X'_{N/2+1}) & N=αρτιος \end{Bmatrix}$$

Ομαδοποιημένα δεδομένα: Σε αυτή την περίπτωση δε μπορούμε να διατάξουμε τις μετρήσεις μας σε αύξουσα διάταξη. Τότε η διάμεσος M θα βρίσκεται μέσα στην i κλάση με τύπο $M = U_{i-1} + \left(\frac{n}{2} - N_{i-1}\right)\frac{\delta}{n_i}$ όπου U_{i-1} κατώτερο όριο κλάσης, n_i συχνότητα της i-κλάσης που περιέχει την διάμεσο, δ πλάτος της κλάση, N_{i-1} αθροιστική συχνότητα της προηγούμενης κλάσης, n αριθμός παρατηρήσεων.

- **Τεταρτημόρια**

Λέμε τις τιμές της μεταβλητής που χωρίζουν το σύνολο των τιμών της σε 4 ισοπληθείς ομάδες όταν οι τιμές της μεταβλητής τοποθετούνται σε αύξουσα σειρά.

- Το **πρώτο τεταρτημόριο** Q_1 είναι η τιμή της μεταβλητής που θα έχει κάτω από αυτή το 25% των παρατηρήσεων όπου δίνεται από τους τύπους

$$Q_1 = \frac{n+1}{4} \qquad \text{μη-ομαδοποιημένα δεδομένα}$$

$$Q_1 = U_{i-1} + \left(\frac{n}{4} - N_{i-1}\right)\frac{\delta}{n_i} \qquad \text{ομαδοποιημένα δεδομένα}$$

- Το **δεύτερο τεταρτημόριο** Q_2 είναι η τιμή της μεταβλητής που θα έχει κάτω από αυτή το 50% των παρατηρήσεων, διάμεσος

- Το **τρίτο τεταρτημόριο** Q_3 είναι η τιμή της μεταβλητής που θα έχει κάτω από αυτή το 75% των παρατηρήσεων όπου δίνεται από τους τύπους

$$Q_1 = \frac{3(n+1)}{4} \qquad \text{μη-ομαδοποιημένα δεδομένα}$$

$$Q_3 = U_{i-1} + \left(\frac{3n}{4} - N_{i-1}\right)\frac{\delta}{n_i} \qquad \text{ομαδοποιημένα δεδομένα}$$

Γενικά

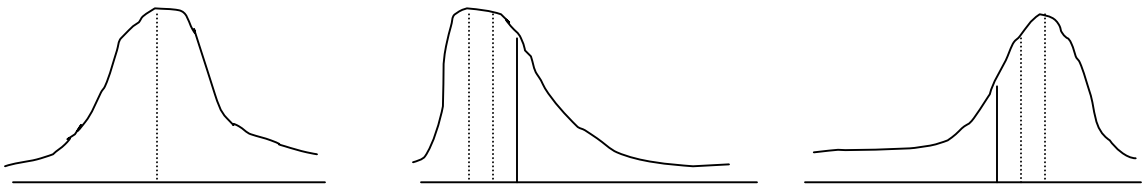
- για τον υπολογισμό **τεταρτημορίων** ο τύπος δίνεται από την σχέση $Q_k = U_{i-1} + \left(\frac{kn}{4} - N_{i-1}\right)\frac{\delta}{n_i}$
- για τον υπολογισμό **δεκατημορίων** ο τύπος δίνεται από την σχέση $D_k = U_{i-1} + \left(\frac{kn}{10} - N_{i-1}\right)\frac{\delta}{n_i}$
- για τον υπολογισμό **εκατοστημορίων** ο τύπος δίνεται από την σχέση $P_k = U_{i-1} + \left(\frac{kn}{100} - N_{i-1}\right)\frac{\delta}{n_i}$

- **Ενδοτεταρτημοριακό εύρος** είναι η διαφορά μεταξύ τρίτου και πρώτου τεταρτημορίου με τύπο $IQR = Q_3 - Q_1$. Είναι δηλαδή το κεντρικό διάστημα που περιέχει το 50% των διατεταγμένων παρατηρήσεων. Το ενδοτεταρτημοριακό εύρος αποτελεί στην ουσία μέτρο διασποράς (κύμανσης) των τιμών (στην ουσία της διακύμανσης των κεντρικών τιμών και όχι όλων των τιμών του δείγματος) και χρησιμεύει στην σύγκριση της μεταβλητότητας των τιμών μεταξύ δύο κατανομών.

- **Ημι- Ενδοτεταρτημοριακό εύρος** είναι το ήμισυ του ενδοτεταρτημοριακού εύρους.

• Σύγκριση των παραμέτρων θέσης

Το ***μέσο εύρος*** είναι εύκολο να υπολογιστεί αλλά για τον υπολογισμό του χρησιμοποιούνται μόνο οι δύο ακραίες τιμές της κατανομής και καμιά άλλη πληροφορία. Έτσι δεν μπορούμε να εμπιστευτούμε το μέσο εύρος για αντικειμενική εκπροσώπηση της κατανομής. Η ***επικρατούσα τιμή*** (ΕΤ) είναι ικανοποιητικό μέτρο θέσης αν η κατανομή είναι συμμετρική. Αν όχι δεν μπορεί να θεωρηθεί ως αντιπροσωπευτικός αριθμός της κατανομής. Η ***διάμεσος*** (Δ) δεν επηρεάζεται από ακραίες τιμές και για τον υπολογισμό της λαμβάνονται υπόψη όλες οι τιμές. Η **μέση τιμή** (ΜΤ) επηρεάζεται από ακραίες τιμές και για τον υπολογισμό της λαμβάνονται υπόψη όλες οι τιμές.
Έτσι θα μπορούσαμε να πούμε ότι η διάμεσος είναι το καλύτερο μέτρο θέσης μιας κατανομής αλλά η δυσκολία υπολογισμού της μας επιβάλει τον υπολογισμό της μέσης τιμής, η οποία έχει και μεγαλύτερη εφαρμογή στην στατιστική ανάλυση. Παρακάτω δίνεται σχηματικά η σχέση μεταξύ των μέτρων θέσης.



ΜΤ=ΕΤ=Δ ΕΤ<Δ<ΜΤ ΜΤ<Δ<ΕΤ

9.2.3 ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ

Ένα από τα βασικότερα γνωρίσματα των στατιστικών δεδομένων είναι η **μεταβλητικότητα**, δηλαδή η διαφοροποίηση των τιμών μιας μεταβλητής μεταξύ των στατιστικών μονάδων. Μικρός αριθμός διαφοροποίησης χαρακτηρίζει *ομοιογενή δεδομένα* (αναφέρεται στην ομοιότητα των δεδομένων) Στην περίπτωση πλήρης ομοιότητας, η μεταβλητότητα είναι 0 και ο μέσος αποτελεί το πληρέστερο μέτρο περιγραφής.

Οι βασικοί **λόγοι χρήσης** των μέτρων διασποράς είναι οι ακόλουθοι

- Η μεταβλητότητα αποτελεί σημαντικό χαρακτηριστικό ενός συνόλου δεδομένων και επομένως ως μέτρο του βαθμού διασποράς των παρατηρήσεων αποτελεί βασικό περιγραφικό στατιστικό.
- Η μεταβλητότητα είναι βασικό αντικείμενο πολλών στατιστικών μεθόδων. Ακόμη και στην περίπτωση ενός συνόλου η μεταβλητότητα δείχνει την δυνατότητα που έχει ένα μέτρο θέσης ως περιληπτικό στατιστικό να περιγράψει τα δεδομένα.
- Τα μέτρα μεταβλητότητας είναι χρήσιμα για τον εντοπισμό εκτρόπων παρατηρήσεων. Η ύπαρξη ακραίων τιμών επηρεάζει την τιμή του μέσου αλλά και την διασπορά των παρατηρήσεων γύρω από αυτόν.

Τα μέτρα που χρησιμοποιούνται για την μέτρηση της μεταβλητότητας μιας κατανομής διακρίνονται σε 3 κυρίως κατηγορίες:

1. Της απόστασης μεταξύ αντιπροσωπευτικών τιμών (εύρος, ενδοτεταρτημοριακό εύρος).
2. Των αποκλίσεων του πληθυσμού από μια κεντρική τιμή (μέση απόκλιση από το μέσο και τη διάμεσο, διακύμανση, τυπική απόκλιση).
3. Των αποκλίσεων όλων των μελών του πληθυσμού μεταξύ τους (μέση διαφορά).

• Εύρος

Ορίζεται από την διαφορά μεταξύ μεγαλύτερης και μικρότερης τιμής ($R=E_{max}-E_{min}$).

• Μέση απόκλιση

Ορίζεται ως ο αριθμητικός μέσος των απόλυτων τιμών των αποκλίσεων των τιμών των δεδομένων από τον αριθμητικό τους μέσο με τύπο για τον πληθυσμό και το δείγμα

$$MAA = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \qquad \text{για πληθυσμό}$$

$$mma = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad \text{για δείγμα}$$

Η μέση απόκλιση λαμβάνει **μεγάλες τιμές** στην περίπτωση που οι παρατηρήσεις (x_1, x_2,… ,x_n) βρίσκονται **μακριά** από το μέσο τους και **μηδενική τιμή** όταν οι παρατηρήσεις **ισούται μεταξύ** τους (x_1= x_2=… =x_n).

Αν στον ορισμό της μέσης απόκλισης αντικαταστήσουμε τον αριθμητικό μέσο με τη διάμεσο προκύπτουν οι ακόλουθες εκφράσεις που ισχύουν για τον πληθυσμό και το δείγμα

$$MAA^* = \frac{\sum_{i=1}^{N} |x_i - M|}{N} \qquad \text{για πληθυσμό}$$

$$mma^* = \frac{\sum_{i=1}^{n} |x_i - m|}{n} \qquad \text{για δείγμα}$$

Το άθροισμα των απολύτων αποστάσεων από την διάμεσο είναι μικρότερο από ότι από οποιοδήποτε άλλο σημείο και επομένως και από το μέσο. Άρα η **διάμεσος** είναι το σημείο ως προς το οποίο οι απόλυτες αποκλίσεις είναι οι ελάχιστες δυνατές. Άρα MMA*≤MMA και mma*≤mma όπου η ισότητα ισχύει στην περίπτωση σύμπτωσης μεταξύ μέσου και διαμέσου.

Σε περίπτωση ασυμμετρίας η **διάμεσος** περιγράφει καλύτερα τη κατανομή των δεδομένων από ότι ο μέσος με συνέπεια η mma* να εκφράζει καλύτερα την κύμανση της κατανομής.

Παράδειγμα 1: Δίνονται οι παρακάτω μετρήσεις 9, 10,12, 18, 26

Ο αριθμητικός μέσος είναι $\bar{x} = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{75}{5} = 15$. Η διάμεσος είναι η τρίτη τιμή του διατεταγμένου δείγματος $m = 12$.

Τότε η μέση δειγματική απόκλιση από το **μέσο** δίνεται από την σχέση $mma = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^{5} |x_i - 15|}{n} = \frac{28}{5} = 5.6$ και η μέση δειγματική απόκλιση από τη **διάμεσο** δίνεται από την σχέση $mma^* = \frac{\sum_{i=1}^{N} |x_i - m|}{n} = \frac{\sum_{i=1}^{5} |x_i - 12|}{n} = \frac{25}{5} = 5$. Άρα mma*<mma.

Παράδειγμα 2: Δίνονται οι τιμές 77 παρατηρήσεων μιας συνεχούς κατανομής

Κλάσεις	w_i	f_i	f_i|w-mean(x)|	f_i|w-m|
0.5-1.5	1	10	10|1-3.299|	10|13.175|
1.5-2.5	2	15	…….	…….
2.5-3.5	3	20	…….	…….

3.5-4.5	4	15	…….	…….
4.5-5.5	5	10	…….	…….
5.5-6.5	6	5	…….	…….
6.5-7.5	7	2	…….	…….
		77	99.887	95.275

Ο μέσος της κατανομής είναι $\bar{x} = \frac{\sum_{i=1}^{7} x_i}{n} = 3.299$. Η διάμεσος της κατανομής δίνεται από την σχέση $m = 2.5 + \frac{1}{20}(38.5 - 25) = 3.175$

Τότε η μέση δειγματική απόκλιση από το **μέσο** δίνεται από την σχέση $mma = \frac{\sum_{i=1}^{7} f_i |w_i - 3.299|}{\sum_{i=1}^{7} f_i} = \frac{96.887}{77} = 1.258$ και η μέση δειγματική απόκλιση από τη **διάμεσο** δίνεται από την σχέση $mma = \frac{\sum_{i=1}^{7} f_i |w_i - 3.175|}{\sum_{i=1}^{7} f_i} = \frac{95.275}{77} = 1.237$. Άρα $\text{mma}^* < \text{mma}$. Η κατανομή εκφράζει θετική ασυμμετρία ($\bar{x} > m$) με συνέπεια η mma^* να είναι προτιμότερη από την mma για την περιγραφή της μέσης τιμής.

- **Σταθμική Μέση απόκλιση**

Ονομάζεται ως ο σταθμικός αριθμητικός μέσος των απολύτων τιμών των αποκλίσεων των κέντρων των διαστημάτων κλάσεων από τον μέσο για τον πληθυσμό και το δείγμα

$$A_1 = \frac{\sum_{i=1}^{k} f_i |w_i - \mu|}{\sum_{i=1}^{k} f_i} \qquad \text{για πληθυσμό}$$

$$a_1 = \frac{\sum_{i=1}^{k} f_i |w_i - \bar{x}|}{\sum_{i=1}^{k} f_i} \qquad \text{για δείγμα}$$

Αν στον ορισμό της μέσης απόκλισης αντικαταστήσουμε τον αριθμητικό μέσο με τη διάμεσο προκύπτουν οι ακόλουθες εκφράσεις που ισχύουν για τον πληθυσμό και το δείγμα

$$A_1^* = \frac{\sum_{i=1}^{k} f_i |w_i - M|}{\sum_{i=1}^{k} f_i}$$ για πληθυσμό

$$a_1^* = \frac{\sum_{i=1}^{k} f_i |w_i - m|}{\sum_{i=1}^{k} f_i}$$ για δείγμα

- **Διακύμανση**

Ορίζεται ως ο αριθμητικός μέσος των τετραγώνων των αποκλίσεων των τιμών της μεταβλητής από τον αριθμητικό μέσο

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N^2}$$ για πληθυσμό

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - n\bar{x}^2\right)$$ για δείγμα

Η τυπική απόκλιση ενός δείγματος αποτελεί δείκτη του βαθμού διασποράς των τιμών γύρω από τον αριθμητικό μέσο.

<u>Ιδιότητες</u>

- Η διακύμανση μιας σταθερής μεταβλητής είναι ίση με το 0, $x_1 = x_2 = \cdots x_n = a \Rightarrow \sigma^2 = 0$
- Αν Y=a+Bx τότε Var(Y)=b²Var(X)
- Αν β=±1 η διακύμανση της Y=α±X είναι Var(Y)=Var(X)
- Αν α=0 η διακύμανση της Y=βX είναι Var(Y)= b²Var(X)
- Αν α=0 και β=1/γ η διακύμανση Y=(1/γ) X είναι Var(Y)=(1/γ²)Var(X)
- Το άθροισμα των τετραγώνων των αποκλίσεων των τιμών μίας μεταβλητής X από δοθέντα πραγματικό αριθμό α γίνεται ελάχιστο αν ο αριθμός α συμπίπτει με το μέσο μ των τιμών της μεταβλητής X
- Έστω πληθυσμός Π μεγέθους N που αποτελείται από κ διακριτούς μεταξύ τους υποπληθυσμούς Π_i μεγεθών N_i ώστε $N=\Sigma n_i$.Αν μ, σ^2 του πληθυσμού Π και μ_i, σ_i^2 του i-οστου πληθυσμού τότε $\sigma^2 = \frac{1}{N}\sum_{i=1}^{K} N_i {\sigma_i}^2 + \frac{1}{N}\sum_{i=1}^{k} N_i(\mu_i - \mu)^2$

- **Σταθμική διακύμανση**

Ορίζεται ως ο σταθμικός μέσος των τετράγωνων των αποκλίσεων των κεντρικών τιμών των κλάσεων από το μέσο της κατανομής.

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (w_i - \mu)^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i {w_i}^2}{\sum_{i=1}^{k} f_i} - \mu^2 \qquad \text{για πληθυσμό}$$

$$s^2 = \frac{\sum_{i=1}^{k} f_i (w_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1} = \frac{\sum_{i=1}^{k} f_i \sum_{i=1}^{k} f_i {w_i}^2 - (\sum_{i=1}^{k} f_i w_i)^2}{\sum_{i=1}^{k} f_i (\sum_{i=1}^{k} f_i - 1)} \qquad \text{για δείγμα}$$

- **Τυπική απόκλιση**

Ονομάζεται η τετραγωνική ρίζα της διακύμανσης

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2} = \sqrt{\frac{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N^2}} \qquad \text{για πληθυσμό}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i^2 - n\bar{x}^2\right)}{n-1}} \qquad \text{για δείγμα}$$

Μεγάλες τιμές των μέτρων δηλώνουν μεγάλη μεταβλητικότητα των τιμών του πληθυσμού και συνεπώς μεγάλη απόσταση των παρατηρήσεων από το μέσο. Αν όλες οι τιμές της μεταβλητής είναι ίσες μεταξύ τους θα είναι για όλες τις παρατηρήσεις $(x_i-\mu)^2=0$ και στην περίπτωση αυτή η διακύμανση και η τυπική απόκλιση θα είναι μηδενικές.
Διακύμανση και **τυπική απόκλιση** μπορούν να χρησιμοποιηθούν μόνο σε περιπτώσεις όπου οι παρατηρήσεις αναφέρονται σε κοινές κλίμακες μέτρησης.

- **Σταθμική τυπική απόκλιση**

Ορίζεται ως η τετραγωνική ρίζα της διακύμανσης

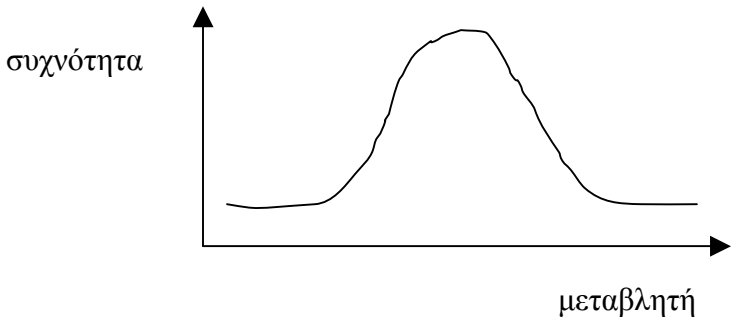
$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (w_i - \mu)^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{k} f_i w_i{}^2}{\sum_{i=1}^{k} f_i} - \mu^2} \qquad \text{για πληθυσμό}$$

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i (w_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1}} = \sqrt{\frac{\sum_{i=1}^{k} f_i w_i^2 - \sum_{i=1}^{k} f_i \bar{x}^2}{\sum_{i=1}^{k} f_i - 1}} \qquad \text{για δείγμα}$$

- **Σύγκριση των παραμέτρων διασποράς**

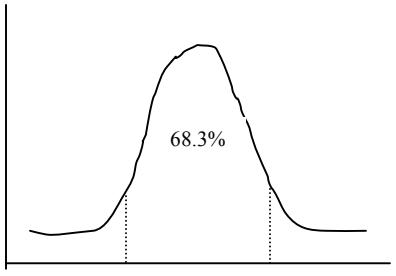
Το **εύρος** είναι πολύ απλό στον υπολογισμό, χρησιμοποιείται στον έλεγχο ποιότητας και μπορεί να χρησιμοποιηθεί για την εκτίμηση της τυπικής απόκλισης, αλλά δεν θεωρείται αξιόπιστο μέτρο διασποράς γιατί βασίζεται σε δύο ακραίες τιμές. Επίσης δεν χρησιμοποιείται για περαιτέρω στατιστική ανάλυση. Η **μέση απόλυτη απόκλιση** είναι απλή στον υπολογισμό, πλεονεκτεί έναντι του εύρους γιατί χρησιμοποιεί όλες τις τιμές για τον υπολογισμό της αλλά είναι δύσκολη η εφαρμογή της λόγο απόλυτων τιμών. Η **διακύμανση** και η **τυπική απόκλιση** είναι τα σπουδαιότερα μέτρα διασποράς γιατί για το υπολογισμό τους λαμβάνονται υπόψη όλες οι παρατηρήσεις και έχουν μεγάλη εφαρμογή στην στατιστική συμπερασματολογία. Το μειονέκτημα της διακύμανσης ότι δεν εκφράζεται στις ίδιες μονάδες με τις οποίες εκφράζονται οι τιμές της μεταβλητής εξαλείφεται με την τυπική απόκλιση. Η μέση τιμή και η τυπική απόκλιση χρησιμοποιούν όλες τις τιμές. Η μέση τιμή δίνει το κέντρο της κατανομής και η τυπική απόκλιση δίνει ένα μέτρο της διασποράς της κατανομής.

Η χρησιμότητα της μέσης τιμής και της τυπικής απόκλισης για την **κανονική κατανομή** δίνεται παρακάτω. Μια κατανομή ονομάζεται κανονική με βάση το παρακάτω σχήμα.

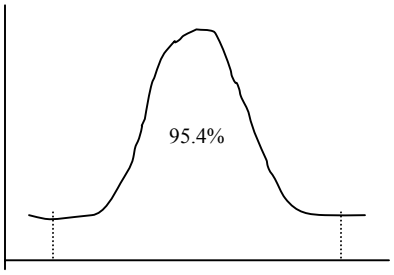


Αν η κατανομή ενός δείγματος ακολουθεί κανονική κατανομή με $N(\bar{x}, s^2)$ τότε ισχύουν τα ακόλουθα:

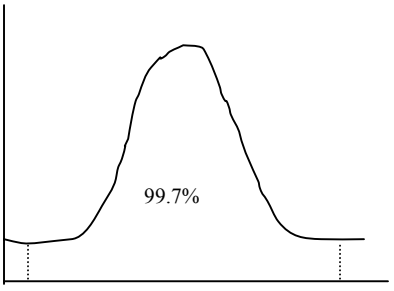
- Το 68.3% των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}+s$ και $\bar{x}-s$
- Το 95.4% των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}+2s$ και $\bar{x}-2s$
- Το 99.7% των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}+3s$ και $\bar{x}-3s$
- Το 50% των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}+\frac{2}{3}s$ και $\bar{x}-\frac{2}{3}s$



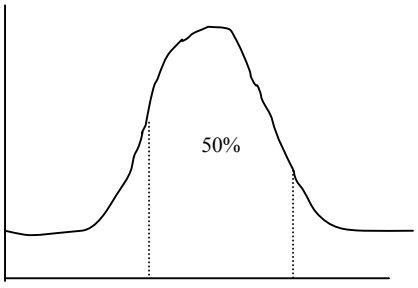
$\bar{x}-s$ $\bar{x}+s$



$\bar{x}-2s$ $\bar{x}+2s$



$\bar{x}-3s$ $\bar{x}+3s$



$\bar{x}-2/3\,s$ $\bar{x}+2/3\,s$

Τι μπορούμε να πούμε για ένα οποιοδήποτε σύνολο δεδομένων? Όταν η κατανομή δεν είναι κανονική τότε με βάση το **θεώρημα του Chebychev** έχουμε :
Για οποιοδήποτε σύνολο δεδομένων (δείγμα ή πληθυσμός) και για οποιαδήποτε σταθερά λ (με λ>1) ποσοστό τουλάχιστον $1-\frac{1}{\lambda^2}$ των δεδομένων βρίσκεται εντός λ τυπικών αποκλίσεων εκατέρωθεν του μέσου όρου δηλαδή $\bar{x}-\lambda s$ και $\bar{x}+\lambda s$. Σύμφωνα με το θεώρημα έχουμε τα ακόλουθα :

- Το 75% των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}+2s$ και $\bar{x}-2s$
- Το 89% των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}+3s$ και $\bar{x}-3s$
- Γενικά ποσοστό τουλάχιστον $(1-\frac{1}{\lambda^2})100\%$ των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}-\lambda s$ και $\bar{x}+\lambda s$

Παράδειγμα 1: για λ=2 έχουμε $1-\frac{1}{\lambda^2}=\frac{1}{4}=0.75$ δηλαδή ποσοστό 75% των τιμών της μεταβλητής ανεξάρτητα από την μορφή της κατανομής βρίσκεται στο διάστημα $\bar{x}+2s$ και $\bar{x}-2s$.

Παράδειγμα 2: Η τυπική απόκλιση ενός δείγματος αποτελεί δείκτη του βαθμού διασποράς των τιμών γύρω από τον αριθμητικό μέσο τους. Επειδή και η διακύμανση και η τυπική απόκλιση εκφράζονται σε μονάδες μέτρησης της μεταβλητής, τα δύο μέτρα μπορούν να χρησιμοποιηθούν για συγκρίσεις μεταξύ δειγμάτων ή πληθυσμών στην περίπτωση που οι παρατηρήσεις αναφέρονται σε κοινές κλίμακες μέτρησης.
Στον παρακάτω πίνακα δίνονται οι αναμενόμενοι χρόνοι 20 ασθενών σε λεπτά
Ιατρός Α: 7 10 13 16 18 18 8 19 31 46
Ιατρός Β: 16 17 19 20 20 20 20 21 21 22

Μέσος Α - $\bar{x}_A=19.6$
Μέσος Β - $\bar{x}_B=19.6$

Από το παραπάνω αποτέλεσμα είναι εμφανές ότι ο μέσος δεν είναι αντιπροσωπευτικό μέτρο σύγκρισης των τιμών των μεταβλητών Α και Β. Από τις τιμές είναι εμφανές ότι οι χρόνοι αναμονής του Ιατρού Β παρουσιάζουν μικρότερη διασπορά με διάστημα τιμών [16-22]. Αντίθετα οι τιμές του Ιατρού Α απέχουν περισσότερο από τον μέσο με μεγάλες αποκλίσεις και διάστημα τιμών [7-46]. Παρόμοια αποτελέσματα μπορούν να δοθούν από τον υπολογισμό των τυπικών αποκλίσεων για τις δύο μεταβλητές Α και Β με τιμές

Τυπική απόκλιση Α - $s_A=11.266$
Τυπική απόκλιση Β - $s_B=1.837$

Τα παραπάνω αποτελέσματα υποδηλώνουν ότι ο ιατρός Β είναι περισσότερο συνεπείς από τον Ιατρό Α, αφού η τιμή της τυπικής απόκλισης του είναι κατά πολύ μικρότερη από εκείνη του Α.

• Συντελεστής συσχέτισης

Ορίζεται ως ο λόγος της τυπικής απόκλισης προς τον αριθμητικό μέσο των μετρήσεων

$$CV=\frac{\sigma}{\mu}(100) \qquad \text{για πληθυσμό}$$

$$CV = \frac{s}{\bar{x}}(100) \qquad \text{για δείγμα}$$

Υψώνοντας τον συντελεστή συσχέτισης στο τετράγωνο έχουμε: $CV^2 = \frac{\sigma^2}{\mu^2} = \frac{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}{\mu^2} = \frac{\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\mu}\right)^2}{N} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{\mu}-1\right)$. Οι διαφορές $(x_i-\mu)$ είναι οι απόλυτες αποκλίσεις των τιμών των δεδομένων από το μέσο και επομένως τα πηλίκα $\frac{x_i-\mu}{\mu}$ αποτελούν τις αντίστοιχες *σχετικές αποκλίσεις*. Γι'αυτό το τετράγωνο του CV αποτελεί και **σχετική διακύμανση**. Στην πραγματικότητα πρόκειται για την διακύμανση των τιμών $\frac{x_i}{\mu}$ ως προς το μέσο τους που είναι μονάδα. Συνεπώς ο λόγος μεταξύ απόλυτης και τυπικής απόκλισης είναι ίσος με το λόγο μεταξύ σχετικής απόκλισης και συντελεστή μεταβλητότητας,

$$\frac{\text{απόλυτη απόκλιση}}{\sigma} = \frac{x_i-\mu}{\sigma} = z$$

$$\frac{\text{σχετική απόκλιση}}{CV} = \frac{\frac{x_i-\mu}{\mu}}{\frac{\sigma}{\mu}} = z$$

Παράδειγμα 1: Έστω ότι ο μέσος μηνιαίος μισθός 100 υπαλλήλων μιας εταιρίας Α στην Ελλάδα είναι $\bar{x}_A = 250.000$ δρχ με τυπική απόκλιση $s_A = 30.000$ δρχ, ενώ για 100 υπαλλήλους εταιρίας Β στην Αμερική είναι $\bar{x}_B = 1500$ δολάρια με τυπική απόκλιση $s_B = 400$ δολάρια. Έχοντας διαφορετικές μονάδες μέτρησης το μέτρο σύγκρισης των τιμών είναι συντελεστής μεταβλητότητας όπου για την εταιρία Α δίνεται από την σχέση

$$CV_A = \frac{s_A}{\bar{x}_A} = \frac{30000}{250000} = 0.12 = 12\%$$

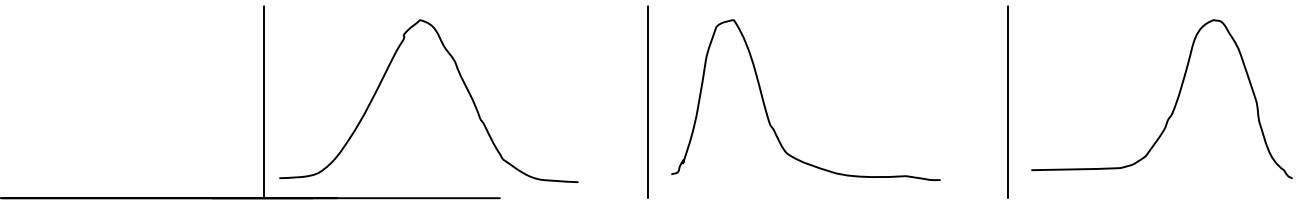
$$CV_B = \frac{s_B}{\bar{x}_B} = \frac{400}{1500} = 0.2667 = 26.67\%$$

Δηλαδή ο βαθμός διασποράς των μισθών των υπαλλήλων της εταιρείας Β είναι μεγαλύτερος από το βαθμό διασποράς των μισθών των υπαλλήλων της εταιρείας Α.

Συμπερασματικά αναφέρουμε ότι όσο μικρότερος είναι ο συντελεστής μεταβλητότητας τόσο μεγαλύτερη ομοιογένεια υπάρχει στις τιμές της μεταβλητής. Είναι φανερό ότι ο συντελεστής μεταβλητότητας δεν ενδείκνυται να χρησιμοποιείται όταν η μέση τιμή είναι κοντά στο 0.

9.2.4 ΜΕΤΡΑ ΑΣΥΜΜΕΤΡΙΑΣ ΚΑΙ ΚΥΡΤΩΣΗΣ

Πέρα από την θέση και την διασπορά μιας κατανομής σημαντικά είναι και τα μέτρα που αφορούν την **μορφή** και το **σχήμα** της κατανομής. Επομένως πολλές φορές χρειάζεται να γνωρίζουμε την μορφολογία μιας κατανομής δηλαδή την ασυμμετρία και την κύρτωσή της. Η **ασυμμετρία** αναφέρεται στην απόκλιση του διαγράμματος της κατανομής συχνοτήτων από το διάγραμμα της κανονικής κατανομής ενώ η **κυρτότητα** αναφέρεται στο πόσο πεπλατυσμένο είναι το διάγραμμα την κατανομής συχνοτήτων.
Κατανομές που έχουν μία μόνο κορυφή ονομάζονται **μονοκόρυφες**. Η ύπαρξη δύο επικρατούσων τιμών οδηγεί στον ορισμό των **δίκορυφων** κατανομών. Η έλλειψη συμμετρίας χαρακτηρίζει τις κατανομές σε **θετικά ασυμμετρικές** και **αρνητικά ασυμμετρικές**.



Θετικά ασυμμε αρνητικά ασυμμετρική

Στην περίπτωση **πλήρους συμμετρίας**, ο μέσος, η διάμεσος και η επικρατούσα τιμή συμπίπτουν και βρίσκονται στον άξονα συμμετρίας της κατανομής (Μ=Δ=ΕΤ). Στην περίπτωση **θετικής ασυμμετρίας**, ο μέσος είναι μεγαλύτερος της διαμέσου (Μ>Δ), και στην περίπτωση **αρνητικής ασυμμετρίας**, ο μέσος είναι μικρότερος της διαμέσου (Μ<Δ).

- **Συντελεστής ασυμμετρίας του Pearson**

Αν μ και σ είναι η μέση τιμή και η τυπική απόκλιση των τιμών μιας μεταβλητής (πληθυσμιακής) τότε ορίζουμε ως μέτρο ασυμμετρίας τον αριθμό

$$\beta_1 = \frac{\mu_3}{\sigma^3}$$

όπου $\mu_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3$ για απλά δεδομένα και $\mu_3 = \frac{1}{\sum_{i=1}^{N} f_i}\sum_{i=1}^{N} f_i (w_i - \mu)^3$ για ομαδοποιημένα δεδομένα, όπου w_i και f_i συμβολίζουν αντίστοιχα το κέντρο και την απόλυτη συχνότητα της i-στης κλάσης.

Είναι φανερό ότι στην περίπτωση **πλήρους συμμετρίας** δηλαδή αν οι ισαπέχουσες από την μέση τιμή μ τιμές x_i παρουσιάζουν την ίδια συχνότητα τότε οι θετικές και αρνητικές διαφορές $(x_i-\mu)^3$ έχουν άθροισμα 0, οπότε $\mu_3=0$ και $\beta_1=0$.

Αν η καμπύλη παρουσιάζει ουρά προς τα δεξιά (**θετικής ασυμμετρίας**) ή προς τα αριστερά (**αρνητικής ασυμμετρίας**) τότε οι διαφορές $(x_i-\mu)^3$ είναι θετικές ή αρνητικές αντίστοιχα, οπότε $\mu_3>0$ και $\mu_3<0$. Επίσης θα έχουμε $\beta_1>0$ (**θετικής ασυμμετρίας**) και $\beta_1<0$ (**αρνητικής ασυμμετρίας**).

Άλλα μέτρα ασυμμετρίας μπορούν να θεωρηθούν οι αριθμοί

$$S_1 = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} \quad \text{και}$$

$$S_2 = \frac{\mu - M}{\sigma}$$

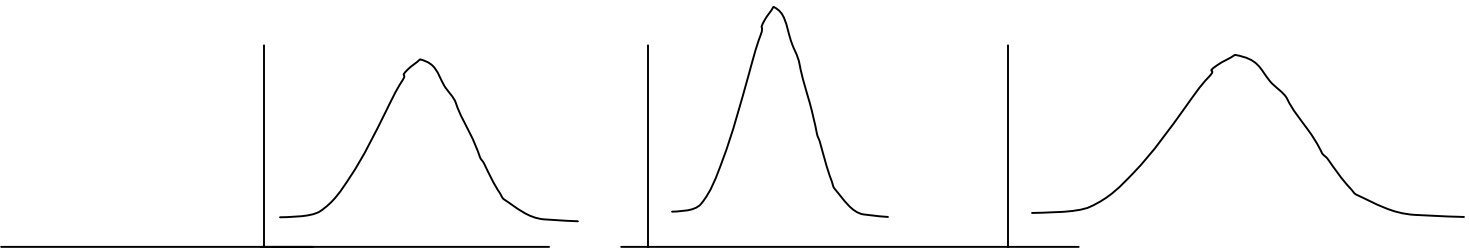
όπου M, Q_1,Q_3 είναι αντίστοιχα η διάμεσος, το 1^o και 3^o τεταρτημόριο και μ,σ είναι ο μέσος και η τυπική απόκλιση.

Για **συμμετρικές κατανομές** οι τιμές $M- Q_1$ και Q_3-M είναι ίσες, και άρα ο αριθμητής της S_1 είναι μηδενικός. Το ίδιο ισχύει και για το μέτρο S_2 αφού μ=M. Για κατανομές με ουρά προς τα δεξιά (**θετικής ασυμμετρίας**) ισχύει $S_1>0$ και $S_2>0$, ενώ για κατανομές με ουρά προς τα αριστερά (**αρνητικής ασυμμετρίας**) ισχύει $S_1<0$ και $S_2<0$. Ανακεφαλαιώνοντας έχουμε ότι

- Για **<u>συμμετρική κατανομή</u>** ισχύει $M- Q_1= Q_3-M$, $\mu=M$, $\mu_3=0$, $\beta_1=0$, $S_1=0$, $S_2=0$.
- Για **<u>θετική ασυμμετρία</u>** ισχύει $Q_3-M> M- Q_1$, $\mu>M$, $\mu_3>0$, $\beta_1>0$, $S_1>0$, $S_2>0$.
- Για **<u>αρνητική ασυμμετρία</u>** ισχύει $Q_3-M< M- Q_1$, $\mu<M$, $\mu_3<0$, $\beta_1<0$, $S_1<0$, $S_2<0$.

- **Συντελεστής κυρτότητας του Pearson**

Χαρακτηρίζει το ύψος της κορυφής της κατανομής και δίνει πληροφορίες για την αιχμηρότητα της καμπύλης. Οι κατανομές διακρίνονται σε **<u>λεπτόκυρτες</u>**, **<u>μεσόκυρτες</u>** και **<u>πλατύκυρτες</u>**.



μεσόκυρτες λεπτόκυρτες πλατύκυρτες

Αν μ και σ είναι η μέση τιμή και η τυπική απόκλιση των τιμών μιας μεταβλητής (πληθυσμιακής) τότε ορίζουμε ως μέτρο κυρτότητας τον αριθμό

$$\beta_2 = \frac{\mu_4}{\sigma^4}$$

όπου $\mu_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4$ για απλά δεδομένα και $\mu_4 = \frac{1}{\sum_{i=1}^{N} f_i}\sum_{i=1}^{N} f_i (w_i - \mu)^4$ για ομαδοποιημένα δεδομένα, όπου w_i και f_i συμβολίζουν αντίστοιχα το κέντρο και την απόλυτη συχνότητα της i-στης κλάσης.

Ο συντελεστής β_2 για **κανονική κατανομή** είναι πάντοτε 3 δηλαδή $\beta_2=3$, για **λεπτόκυρτες** είναι $\beta_2>3$ και για ***πλατύκυρτες*** είναι $\beta_2<3$. Είναι φανερό ότι όσο μεγαλύτερη είναι η διαφορά β_2-3, τόσο αιχμηρότερη είναι η καμπύλη της κατανομής των τιμών.

Ειδικότερα για την **κανονική κατανομή,** αν η κατανομή ενός δείγματος ακολουθεί κανονική κατανομή με $N(\bar{x}, s^2)$ τότε ισχύουν ότι το 68.3% των μεσαίων τιμών βρίσκεται μεταξύ $\bar{x}+s$ και $\bar{x}-s$. Αν το ποσοστό που περιέχεται στο διάστημα αυτό είναι μικρότερο του 68.3% τότε η κατανομή είναι ***πλατύκυρτη***, ενώ αν το ποσοστό που περιέχεται στο διάστημα αυτό είναι μεγαλύτερο του 68.3% τότε η κατανομή είναι **λεπτόκυρτη**,

• Ροπές

Ορίζονται είτε ως προς την αρχή των δεδομένων (το 0) είτε ως προς το μέσο (κεντρική ροπή). Επίσης διαφοροποιούνται αν τα δεδομένα είναι ταξινομημένα σε συχνότητες ή αταξινόμητα.

❖ Ορίζουμε ως ροπή i-στης τάξης ως προς το 0 τον αριθμό

$$v_t = \frac{1}{N}\sum_{i=1}^{N} x_i^t \qquad \text{για αταξινόμητα}$$

$$v_t = \frac{1}{N}\sum_{i=1}^{N} f_i x_i^t \qquad \text{για ταξινομημένα}$$

❖ Ορίζουμε ως ροπή i -στης τάξης ως προς το μέσο τον αριθμό

$$\mu_t = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^t \qquad \text{για αταξινόμητα}$$

$$\mu_t = \frac{1}{\sum_{i=1}^{k} f_i}\sum_{i=1}^{k} f_i (x_i - \mu)^t \qquad \text{για ταξινομημένα}$$

9.3 ΣΥΣΧΕΤΙΣΗ

Γενικά ο όρος συσχέτιση αναφέρεται στο βαθμό με τον οποίο σχετίζονται (συμμεταβάλλονται) δύο ή περισσότερες μεταβλητές. Ο βαθμός της σχέσης μεταξύ δύο μεταβλητών ονομάζεται **απλή συσχέτιση** και μεταξύ περισσοτέρων από δύο ονομάζεται **πολλαπλή συσχέτιση**.
Η συσχέτιση μεταξύ δύο μεταβλητών ονομάζεται **γραμμική** όταν στο διάγραμμα διασκορπισμού τα σημεία τείνουν να συγκεντρώνονται γύρω από μια ευθεία, και **μη γραμμική** όταν τα σημεία τείνουν να συγκεντρώνονται γύρω από μια καμπύλη. Η **απλή συσχέτιση** ασχολείται με το βαθμό με τον οποίο τα σημεία συγκεντρώνονται γύρω από την ευθεία χωρίς να προσδιορίζεται αυτή η ευθεία που διέρχεται από το νέφος των σημείων. Η κατεύθυνση του νέφους των σημείων σηματοδοτεί το είδος της σχέσης, ενώ η συγκέντρωση των σημείων είναι ενδεικτική του βαθμού συμμεταβολής των δεδομένων. Η γραμμική συσχέτιση μπορεί να είναι:

- **Θετική συσχέτιση** όταν δύο μεταβλητές τείνουν να μεταβάλλονται προς την ίδια κατεύθυνση. Στην περίπτωση αυτή οι τιμές τείνουν να αυξάνονται ή να μειώνονται.
- **Αρνητική συσχέτιση** όταν δύο μεταβλητές τείνουν να μεταβάλλονται προς αντίθετη κατεύθυνση. Στην περίπτωση αυτή οι τιμές της μίας μεταβλητής τείνουν να αυξάνονται και της άλλης να μειώνονται
- **Μηδενική συσχέτιση** όταν οι μεταβολές των τιμών της μίας μεταβλητής δεν συνδέονται με τις μεταβολές της άλλης. Τα σημεία του νέφους είναι διασκορπισμένα σε όλο το μήκος του διαγράμματος

Γενικά όσο πιο συγκεντρωμένα είναι τα σημεία γύρω από την ευθεία γραμμή τόσο πιο ισχυρή είναι η γραμμική σχέση που συνδέει τις δύο μεταβλητές. Αντίθετα όσο περισσότερο διασκορπισμένα εμφανίζονται τα σημεία τόσο πιο ασθενής είναι η συσχέτιση των μεταβλητών. Στην ακραία περίπτωση που όλα τα σημεία (x,y) βρίσκονται πάνω σε μια ευθεία γραμμή η συσχέτιση καλείται πλήρης.

9.3.1 ΣΥΝΔΙΑΚΥΜΑΝΣΗ

Έστω X και Y δύο μεταβλητές που ακολουθούν από κοινού κατανομή με μέσες τιμές $E[X]=\mu_X$ και $E[Y]=\mu_Y$. Η συνδιακύμανση (Cov(X,Y)) μεταξύ δύο μεταβλητών ορίζεται από την σχέση

$$\mathrm{cov}(X,Y) = E(X_i - \mu_x)(Y_i - \mu_y)$$

Από την παραπάνω σχέση προκύπτει ότι

$$\text{cov}(X,Y)=\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_x)(Y_i-\mu_y)$$ για πληθυσμό

$$\text{cov}(X,Y)=\frac{1}{n-1}\sum_{i=1}^{N}(X_i-\bar{x})(Y_i-\bar{y})$$ για δείγμα

Η συνδιακύμανση μπορεί να λάβει οποιαδήποτε θετική ή αρνητική τιμή και εκβράζεται στις μονάδες μέτρησης των μεταβλητών. Μας δίνει τις αποκλίσεις των μεταβλητών X και Y από τους μέσους τους με συνέπεια την μετατόπιση των αρχικών αξόνων του διαγράμματος διασκορπισμού σε **ομόρροπη κλίση** και **αντίρροπη κλίση**.

- **Ομμόροπη κλίση** όπου οι ψηλές τιμές της X αντιστοιχούν σε ψηλές τιμές της Y.
- **Αντίρροπη κλίση** όπου οι ψηλές τιμές της X αντιστοιχούν σε χαμηλές τιμές της Y.

Από τα ανωτέρω γίνεται εμφανές ότι η συνδιακύμανση είναι ένα στατιστικό μέτρο που περιγράφει συνοπτικά τον τρόπο με τον οποίο συνδέονται γραμμικά δύο μεταβλητές όπου

- Cov(X,Y)>0 – θετική σχέση μεταξύ X,Y
- Cov(X,Y)<0 – αρνητική σχέση μεταξύ X,Y
- Cov(X,Y)=0 – ανυπαρξία γραμμικής σχέση μεταξύ X,Y αλλά όχι ανεξαρτησία (μπορεί να υπάρχει μη γραμμική σχέση)

<u>Ιδιότητες</u>

- Ο τύπος της συνδιακύμανσης είναι συμμετρικός ως προς τις μεταβλητές X,Y δηλαδή $\text{cov}(X,Y)=\text{cov}(Y,X)$.
- Η συνδιακύμανση της μεταβλητής με τον εαυτό της είναι ίση με την διακύμανση της μεταβλητής $\text{cov}(X,X)=\text{var}(X)$.
- Η συνδιακύμανση μεταξύ μεταβλητών Z=α+βX και W=γ+δY είναι ίση με $\text{cov}(Z,W)=\beta\delta\,\text{cov}(X,Y)$.
- Η συνδιακύμανση μεταξύ μεταβλητής Z και του αθροίσματος δύο μεταβλητών X,Y είναι ίση με το άθροισμα των επιμέρους συνδιακύμανσεων της μεταβλητής Z με κάθε μια από τις μεταβλητές $\text{cov}(Z,X+Y)=\text{cov}(Z,X)+\text{cov}(Z,Y)$.

- Η διακύμανση του αθροίσματος των δύο μεταβλητών είναι ίση με το άθροισμα των δύο επι μέρους διακυμάνσεων αυξημένο κατά το διπλάσιο της συνδιακύμανσης $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X,Y)$.
- Η διακύμανση της διαφοράς των δύο μεταβλητών είναι ίση με το άθροισμα των δύο επί μέρους διακυμάνσεων μειωμένο κατά το διπλάσιο της συνδιακύμανσης $\text{var}(X-Y) = \text{var}(X) + \text{var}(Y) - 2\,\text{cov}(X,Y)$.
- Αν οι μεταβλητές είναι ανεξάρτητες Cov(X,Y)=0 τότε $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$, $\text{var}(X-Y) = \text{var}(X) + \text{var}(Y)$.

9.3.2 ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ

Η κατεύθυνση του νέφους των σημείων δηλώνει το είδος της σχέσης μεταξύ των δύο μεταβλητών ενώ η παραστατική συγκέντρωση των σημείων είναι ενδεικτική του βαθμού συμμεταβολής των δεδομένων. Η ομόρροπη ή αντίρροπη συμμεταβολή των δεδομένων υποδηλώνεται από το πρόσημο της συνδιακύμανσης.

Συντελεστής συσχέτισης (ρ) είναι ένα στατιστικό μέτρο που περιέχει όχι μόνο ένδειξη για τον τρόπο συμμεταβολής αλλά και ακριβή ποσοτική εκτίμηση του βαθμού συσχέτισης δυο μεταβλητών; είναι ανεξάρτητο από μονάδες μέτρησης και εκφράζεται σε σχετικούς όρους για συγκρίσεις. Δίνεται από τον τύπο

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{E(X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{E(X_i - \mu_x)^2}\sqrt{E(Y_i - \mu_y)^2}}$$

όπου σ_X και σ_Y σ_X είναι αντίστοιχα οι τυπικές αποκλίσεις των μεταβλητών X και Y.

Αποδεικνύεται ότι οι αριθμητικές τιμές που μπορεί να λάβει ο συντελεστής συσχέτισης είναι μεταξύ −1 και 1. Εκφράζει τον βαθμό γραμμικής συμμεταβολής δύο μεταβλητών και είναι ομόσημος με την συνδιακύμανση τους.

Ο **δειγματικός συντελεστής συσχέτισης (r)** ορίζεται κατά αναλογία με τον αντίστοιχο πληθυσμιακό ως ο λόγος της δειγματικής συνδιακύμανσης προς το γινόμενο των τυπικών αποκλίσεων των μεταβλητών X και Y στο δείγμα δηλαδή

$$r = \frac{\text{cov}(X,Y)}{s_x s_y} = \frac{\sum (X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum (X_i - \bar{x})^2}\sqrt{\sum (Y_i - \bar{y})^2}} \text{ ή}$$

$$r = \frac{\sum XY - n\bar{x}\bar{y}}{\sqrt{\sum X^2 - n\bar{x}^2}\sqrt{\sum Y^2 - n\bar{y}^2}}$$

Αποδεικνύεται ότι οι αριθμητικές τιμές που μπορεί να λάβει ο συντελεστής συσχέτισης είναι μεταξύ −1 και 1.

Σημειώνεται ότι

- **Θετικές τιμές** – ομόρροπη συμμεταβολή – θετική συσχέτιση. Όσο πλησιέστερα στο 1 τόσο ισχυρότερη η θετική συσχέτιση - ρ=1 πλήρη θετική γραμμική συσχέτιση.

- **Αρνητικές τιμές** – αντίρροπη συμμεταβολή - αρνητική συσχέτιση. Όσο πλησιέστερα στο 1 τόσο ισχυρότερη η αρνητική συσχέτιση - ρ=-1 πλήρη αρνητική γραμμική συσχέτιση.
- **Πλησιέστερα στο μηδέν –** ασθενής θετική ή αρνητική συσχέτιση, γραμμική σχέση των μεταβλητών.
- **Μηδενική** - ρ=0, μηδενική συνδιακύμανση – ανυπαρξία γραμμικής συσχέτισης.

Με βάση το βαθμό συσχέτισης το πόσο ισχυρή είναι η συσχέτιση μεταξύ δύο μεταβλητών έχουμε τις παρακάτω **διαβαθμίσεις**

- Όταν $0.8<r<1$ ή $-1< r<-0.8$ – πολλή σημαντική ή πολλή ισχυρή συσχέτιση
- Όταν $0.7< r<0.8$ ή $-0.8< r<-0.7$ – σημαντική ή ισχυρή συσχέτιση.
- Όταν $0.5< r<0.7$ ή $-0.7< r<-0.5$ – μέση συσχέτιση.
- Όταν $0.3< r<0.5$ ή $-0.5< r<-0.3$ – ασθενής συσχέτιση.
- Όταν $-0.3< r<0.3$– ανύπαρκτη συσχέτιση.
- Όταν $r=\pm 1$ – τέλεια συσχέτιση.

Ιδιότητες

- Ο συντελεστής συσχέτισης είναι συμμετρικός ως προς τις μεταβλητές X και Y $\rho_{xy} = \rho_{yx}$
- Ο ρ μιας μεταβλητής X με τον εαυτό της είναι ίσος με 1.
- Έστω μεταβλητές X,Y με μέσες τιμές E(X) και E(Y), διακυμάνσεις V(X) και V(Y) και συνδιακύμανση Cov(X,Y). Έστω επίσης οι μεταβλητές $Z=\alpha+\beta X$ και $W=\gamma+\delta Y$. Τότε $\rho_{(\alpha+\beta X)(\gamma+\delta Y)} = \rho_{(x,y)}$.

9.3.3 ΕΛΕΓΧΟΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ

Εφόσον ο συντελεστής συσχέτισης r αποτελεί εκτίμηση του πληθυσμιακού ρ θα πρέπει να αξιολογηθεί για να διαπιστωθεί αν η εκτίμηση είναι καλή. Θα πρέπει να ελεγχθεί με κάποια πιθανότητα αν ο εκτιμηθείς συντελεστής είναι στατιστικά σημαντικός και Άρα αξιόπιστος.

Για να προχωρήσουμε σε ελέγχους υποθέσεων θα πρέπει να γνωρίζουμε την κατανομή δειγματοληψίας του r.Το γεγονός ότι έχουμε θεωρήσει κανονική κατανομή για την από κοινού κατανομή των (x,y) δίνει την δυνατότητα να γνωρίζουμε την κατανομή δειγματοληψίας του συντελεστή r. Η μορφή κατανομής του r διαφοροποιείται με το αν ο πληθυσμιακός ρ είναι ίσος με το 0 ή διάφορος του 0.

Για ρ=0

Αποδεικνύεται ότι η κατανομή είναι συμμετρική γύρο από το 0 με διακύμανση εξαρτώμενη από το μέγεθος του δείγματος n και τύπο

$$\mathrm{var}(r) = \frac{1-\rho^2}{n-2}$$

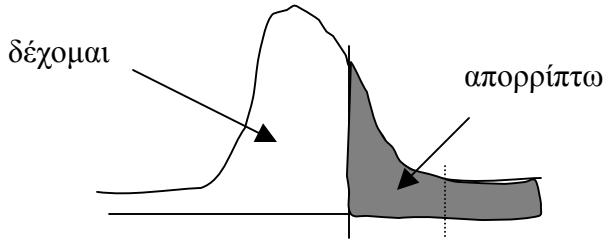
Οι θεωρητικές τιμές του r έχουν υπολογιστεί και παρουσιάζονται σε ειδικούς πίνακες με επίπεδο σημαντικότητας α και βαθμούς ελευθερίας n-2.

Ο στατιστικός έλεγχος αναφέρεται στο αν η τιμή του r που εκτιμήθηκε από τα στοιχεία του δείγματος διαφέρει σημαντικά από το ρ=0 δηλαδή H_0: ρ=0 vs H_1: ρ≠0. Τότε

αν $|r| > r_{\alpha,n-2}$ τότε ο εκτιμηθείς συντελεστής είναι στατιστικά σημαντικός

αν $|r| < r_{\alpha,n-2}$ τότε ο εκτιμηθείς συντελεστής δεν είναι στατιστικά σημαντικός

Παράδειγμα Έστω δείγμα από 20 παρατηρήσεις με δειγματικό συντελεστή συσχέτισης r=0.60. Για επίπεδο σημαντικότητας α=0.05 και βαθμούς ελευθερίας n=20-2=18 έχουμε από τους πίνακες $r_{0.05,18}$=0.4438. Αρα



$r_{0.05,18}$=0.4438 r=0.60

Άρα απορρίπτουμε την υπόθεση H_0 με αποτέλεσμα ο συντελεστής συσχέτισης r=0.60 να είναι στατιστικά σημαντικός.

Επίσης μπορεί να χρησιμοποιηθεί το στατιστικό τεστ με βάση τον τύπο

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \mapsto t_{n-2,\alpha/2}$$

όπου ακολουθεί t-κατανομή με n-2 βαθμούς ελευθερίας. Κατά συνέπεια μπορούν να χρησιμοποιηθούν οι πίνακες της t-κατανομής, η οποία ακολουθεί ασυμπτωτικά την κανονική κατανομή.

Ο στατιστικός έλεγχος αναφέρεται στο αν η τιμή του r που εκτιμήθηκε από τα στοιχεία του δείγματος διαφέρει σημαντικά από το ρ=0 δηλαδή H_0: ρ=0 vs H_1: ρ≠0. Τότε

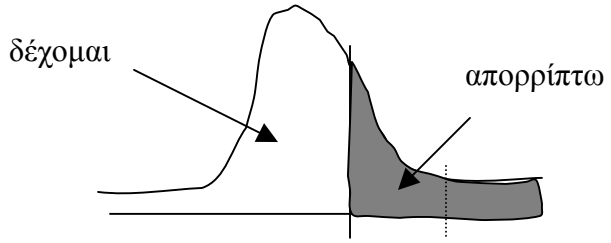
αν $|t| > t_{\alpha/2,n-2}$ τότε ο εκτιμηθείς συντελεστής είναι στατιστικά σημαντικός

αν $|t| < r_{\alpha/2,n-2}$ τότε ο εκτιμηθείς συντελεστής δεν είναι στατιστικά σημαντικός

Παράδειγμα Έστω δείγμα από 10 παρατηρήσεις με δειγματικό συντελεστή συσχέτισης r=0.836. Το στατιστικό τεστ δίνεται από τον παρακάτω τύπο

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.836}{\sqrt{\frac{1-0.836^2}{10-2}}} = 4.309$$

Για επίπεδο σημαντικότητας α=0.025 και βαθμούς ελευθερίας n=10-2=8 έχουμε από τους πίνακες $t_{0.025,8}$=2.306. Αρα



$t_{0.025,8}$=2.306 t=4.309

Αρα απορρίπτουμε την υπόθεση H_0 με αποτέλεσμα ο συντελεστής συσχέτισης r=0.836 να είναι στατιστικά σημαντικός.

Για ρ≠0

Η κατανομή δεν είναι συμμετρική και άρα δεν μπορεί να χρησιμοποιηθεί η t-κατανομή. Αποδεικνύεται ότι

$$z_r = \frac{1}{2}\ln\frac{1+r}{1-r} \mapsto N(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \frac{1}{n-3})$$

Άρα μπορούν να χρησιμοποιηθούν οι πίνακες της κανονικής κατανομής για τυποποιημένες τιμές z.

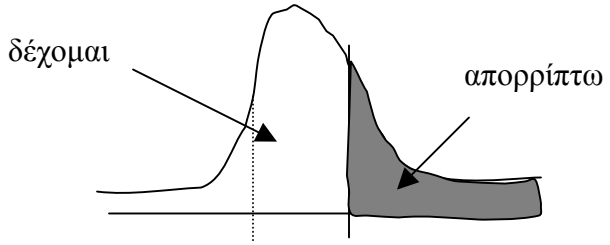
Παράδειγμα Έστω δείγμα από 19 παρατηρήσεις με δειγματικό συντελεστή συσχέτισης r=0.762. Επιθυμούμε να ελέγξουμε αν η τιμή αυτή είναι στατιστικά σημαντική από την υποθετική του πληθυσμιακού ρ=0.5. Άρα θα ελέγξουμε H_0: ρ=0.5 vs H_1: ρ≠0.5.
Από τα δεδομένα έχουμε

$$z_r = \frac{1}{2}\ln\frac{1+r}{1-r} = \ldots = 1$$

$$E[z_r] = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} = \ldots = 0.549$$

$$V[z_r] = \sqrt{\frac{1}{n-3}} = \ldots = 0.25$$

Αρα έχουμε $z = \frac{z_r - E[z_r]}{V[z_r]} = \frac{1-0.549}{0.25} = 1.804$. Για επίπεδο σημαντικότητας α=0.05 και βαθμούς ελευθερίας n=19 έχουμε από τους πίνακες $z_{0.05,19}$=1.96. Αρα



Άρα δεχόμαστε την υπόθεση H_0 με αποτέλεσμα ο πληθυσμιακός συντελεστής συσχέτισης ρ=0.5 να μην διαφέρει σημαντικά από τον δειγματικό συντελεστής συσχέτισης r=0.762.

10 ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

Στατιστική συμπερασματολογία είναι η μεθοδολογία με την οποία συγκεντρώνονται συμπεράσματα για ένα στατιστικό πληθυσμό βάση αποτελεσμάτων που λαμβάνονται από ένα δείγμα που εκλεκτικέ από το πληθυσμό.

Ο έλεγχος υποθέσεων έχει ως αποτέλεσμα μια απόφαση για την τιμή μιας παραμέτρου ενός στατιστικού πληθυσμού (μέσος, διακύμανση, αναλογία). Άλλοι έλεγχοι έχουν ως σκοπό την σύγκριση πληθυσμών ή τον βαθμό τυχαιότητας των δειγμάτων ή το είδος συσχέτισης δυο μεταβλητών.

Υπόθεση είναι μια δήλωση για έναν ή περισσότερους πληθυσμό (ους) ή τις παραμέτρους τους. Σε ένα έλεγχο υποθέσεων έχουμε την αντικαταβολή δυο υποθέσεων, της μηδενικής και της εναλλακτικής. Η **μηδενική** είναι η υπόθεση που ελέγχεται για την ορθότητα της. Η **εναλλακτική** είναι η υπόθεση η οποία είναι στην διάθεση μας όταν απορρίπτεται η μηδενική υπόθεση.

Τα στάδια του ελέγχου υποθέσεως είναι

- Ορίζεται η μηδενική υπόθεση
- Ορίζεται η εναλλακτική υπόθεση
- Ορισμός του επιπέδου σημαντικότητας α
- Ορίζεται το στατιστικό τεστ από το δείγμα
- Συλλογή δεδομένων και υπολογισμοί
- Εξαγωγή αποτελεσμάτων και λήψη αποφάσεων
- Εξαγωγή συμπερασμάτων

Ορίζεται **σφάλμα τύπου Ι** η απόρριψη της μηδενικής υπόθεσης ενώ είναι σωστή. Η πιθανότητα του σφάλματος συμβολίζεται με α και

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

Ορίζεται **σφάλμα τύπου ΙΙ** η αποδοχή της μηδενικής υπόθεσης ενώ είναι λάθος. Η πιθανότητα του σφάλματος συμβολίζεται με β και

β=P[αποδοχή της H_0 | H_0 είναι λάθος]

- **<u>Έλεγχος για την μέση τιμή πληθυσμού (μ) με γνωστή ή άγνωστη διακύμανση (σ^2) (δείγμα n>30).</u>**

$H_0 : \mu = \mu_0$ - $H_1 : \mu > \mu_0$ - $R = \{z > z_a\}$

$H_0 : \mu = \mu_0$ - $H_1 : \mu < \mu_0$ - $R = \{z < -z_a\}$ $\qquad z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ ή $z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$H_0 : \mu = \mu_0$ - $H_1 : \mu \neq \mu_0$ - $R = \{|z| > z_{a/2}\}$

- **<u>Έλεγχος για την μέση τιμή πληθυσμού (μ) με άγνωστη διακύμανση (σ^2) (δείγμα n<30)</u>**

$H_0 : \mu = \mu_0$ - $H_1 : \mu > \mu_0$ - $R = \{t > t_{n-1,a}\}$

$H_0 : \mu = \mu_0$ - $H_1 : \mu < \mu_0$ - $R = \{t < -t_{n-1,a}\}$ $\qquad t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \rightarrow t_{n-1}$

$H_0 : \mu = \mu_0$ - $H_1 : \mu \neq \mu_0$ - $R = \{|t| > t_{n-1,a/2}\}$

Προϋπόθεση: δείγμα από κανονικό πληθυσμό

- **<u>Έλεγχος για την διαφορά των μέσων 2 πληθυσμών (μ_1-μ_2) που ακολουθούν κανονική κατανομή, με γνωστές ή άγνωστες διακυμάνσεις (δείγμα n>30)</u>**

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 > 0$ - $R = \{z > z_a\}$

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 < 0$ - $R = \{z < -z_a\}$ $\qquad z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$ ή $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \rightarrow N(0,1)$

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 \neq 0$ - $R = \{|z| > z_{a/2}\}$

- **<u>Έλεγχος για την διαφορά των μέσων 2 πληθυσμών (μ_1-μ_2) που ακολουθούν κανονική κατανομή με άγνωστες ίσες διακυμάνσεις (δείγμα n<30)</u>**

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 > 0$ - $R = \{t > t_{n+m-2,a}\}$

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 < 0$ - $R = \{t < -t_{n+m-2,a}\}$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \rightarrow t_{n+m-2}$$

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 \neq 0$ - $R = \{|t| > t_{n+m-2,a/2}\}$

Προϋπόθεση: δείγμα από κανονικό πληθυσμό

- **<u>Έλεγχος για την διαφορά των μέσων 2 πληθυσμών (μ_1-μ_2) που ακολουθούν κανονική κατανομή με άγνωστες άνισες διακυμάνσεις (δείγμα n<30)</u>**

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 > 0$ - $R = \{t > t_{k,a}\}$

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 < 0$ - $R = \{t < -t_{k,a}\}$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \rightarrow t_k$$

$H_0 : \mu_1 - \mu_2 = 0$ - $H_1 : \mu_1 - \mu_2 \neq 0$ - $R = \{|t| > t_{k,a/2}\}$

όπου $n = m \Rightarrow k = 2(n-1)$

$$n \neq m \Rightarrow k = \frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{\left(\frac{s_1^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_2^2}{m}\right)^2}{m-1}}$$

- **<u>Έλεγχος για την διασπορά ενός πληθυσμού που ακολουθεί κανονική κατανομή</u>**

$H_0 : \sigma = \sigma_0$ - $H_1 : \sigma > \sigma_0$ - $R = \{x^{2*} > x^2_{n-1,a}\}$

$H_0 : \sigma = \sigma_0$ - $H_1 : \sigma < \sigma_0$ - $R = \{x^{2*} < x^2_{n-1,1-a}\}$

$$\chi^2 = \frac{(v-1)s^2}{\sigma^2} \rightarrow x^2_{n-1}$$

$H_0 : \sigma = \sigma_0$ - $H_1 : \sigma \neq \sigma_0$ - $R = \{x^{2*} > x^2_{n-1,a/2}\} - R = \{x^{2*} < x^2_{n-1,1-a/2}\}$

Προϋπόθεση: δείγμα από κανονικό πληθυσμό

- **<u>Έλεγχος για τον λόγο διασπορών δύο πληθυσμών που ακολουθούν κανονική κατανομή</u>**

$H_0 : \sigma = \sigma_0 \;\; \frac{\sigma}{\sigma_0} = 1$ - $H_1 : \sigma > \sigma_0$ - $R = \{F > F_{n_1,n_2,a}\}$

$H_0 : \sigma = \sigma_0$ - $H_1 : \sigma < \sigma_0$ - $R = \{F < F_{n_1,n_2,a}\}$ $\qquad F = \frac{s_1^2}{s_2^2} \to F_{n-1,m-1}$, $\qquad s_1^2 > s_2^2$

$H_0 : \sigma = \sigma_0$ - $H_1 : \sigma \neq \sigma_0$ - $R = \{F > F_{n_1,n_2,a/2}\}$ $\qquad F = \frac{s_2^2}{s_1^2} \to F_{n-1,m-1}$, $\qquad s_1^2 < s_2^2$

Προϋπόθεση: δείγμα από κανονικό πληθυσμό