ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ
ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

ΣΤΕΛΙΟΣ ΖΗΜΕΡΑΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΤΕΥΘΥΝΣΗ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΗΣ-ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΣΑΜΟΣ

2014

ΕΙΣΑΓΩΓΗ

Ετυμολογία **στατίζω**

(ελληνική λέξη που σημαίνει διαπιστώνω)

Η Στατιστική είναι η επιστήμη η οποία ασχολείται με τον σχεδιασμό πειραμάτων, τη συλλογή και ανάλυση στατιστικών δεδομένων με σκοπό την εξαγωγή συμπερασμάτων που αφορούν τα χαρακτηριστικά ενός πληθυσμού.

Στατιστική: Επιστήμη λήψης αποφάσεων σε καθεστώς αβεβαιότητας

ΕΙΣΑΓΩΓΗ

Το αντικείμενο της στατιστικής συνίσταται στην αντικείμενο αποτελεσματική αξιοποίηση πληροφοριών μετά από κατάλληλη συλλογή, επεξεργασία, οργάνωση, παρουσίαση και ανάλυση στατιστικών δεδομένων.

Αποτελεί μέρος της επιστημονικής έρευνας με σκοπό την ερμηνεία των δεδομένων και μοντελοποίησή τους σε μοντέλα όσο το δυνατό αξιόπιστα που να ερμηνεύουν με καταλληλότητα τα πραγματικά δεδομένα.

ΕΙΣΑΓΩΓΗ

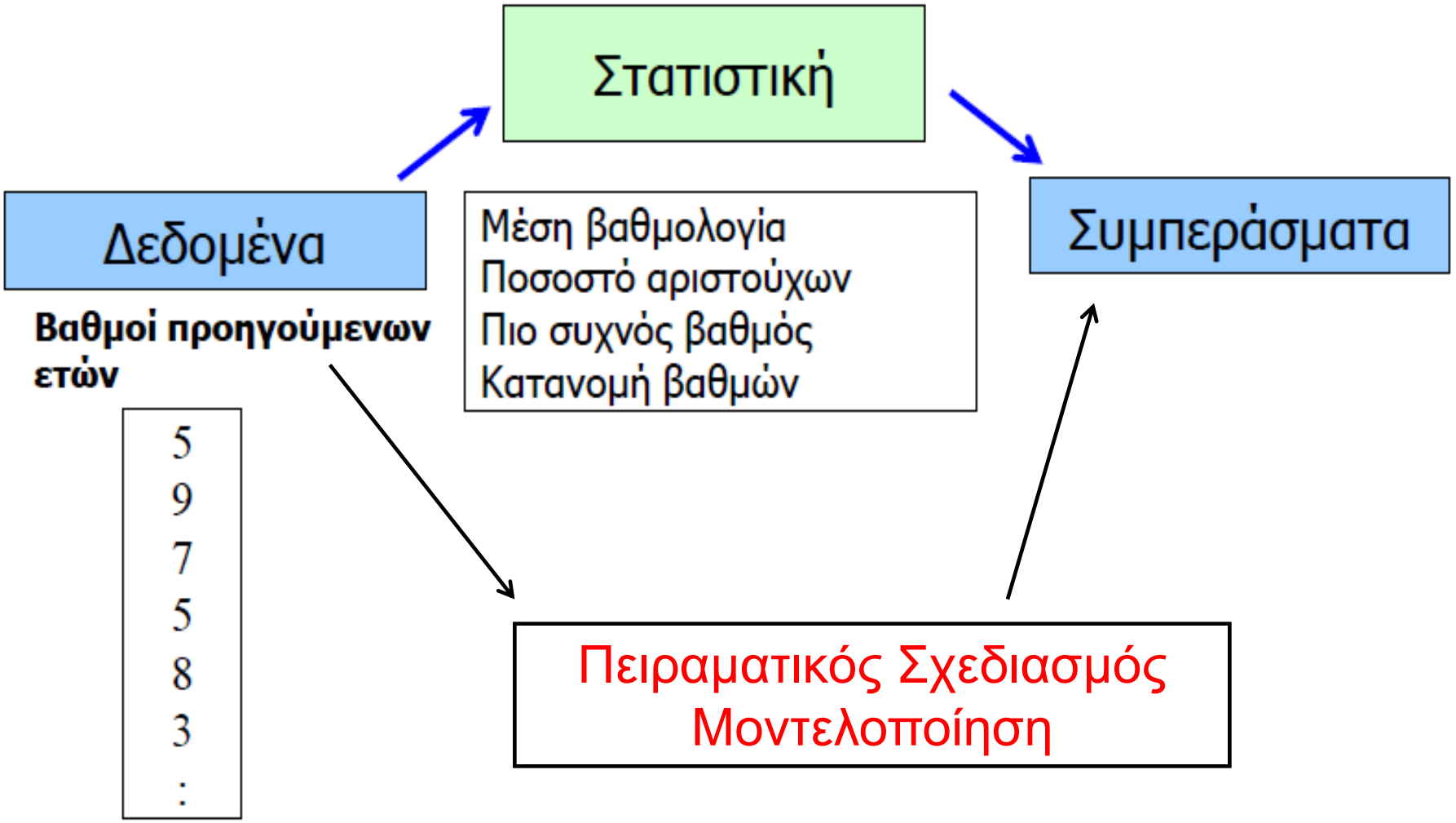
Στατιστική είναι η επιστήμη η οποία

- Περιγράφει με σαφή και ακριβή τρόπο διάφορα μετρήσιμα οικονομικά, δημογραφικά, κοινωνικά, πολιτικά και άλλα φαινόμενα καθώς και τη διαχρονική τους εξέλιξη (Περιγραφική)
- Μελετά τους νόμους που διέπουν τις συνολικές εκδηλώσεις των τυχαίων φαινομένων. (Πιθανότητες).
- Εκτιμά διαφόρους παραμέτρους ενός πληθυσμού ή προβλέπει τη διαχρονική εξέλιξη των φαινομένων στο άμεσο μέλλον μετά από αντικειμενική αξιοποίηση του παρελθόντος (Εκτιμητική).

ΕΙΣΑΓΩΓΗ

- **Η στατιστική ασχολείται με:**
  - Το σχεδιασμό της διαδικασίας συλλογής πληροφοριών
  - Τη συλλογή πληροφοριών από το σύνολο του πληθυσμού (απογραφή) ή από επιλεγμένο δείγμα (ομοιογενές σύνολο ατόμων) του πληθυσμού
  - Την οργάνωση των πληροφοριών
  - Τη συνοπτική και αποτελεσματική παρουσίασή τους
  - Την ανάλυση και εξαγωγή συμπερασμάτων
- Οι πληροφορίες για ένα χαρακτηριστικό ονομάζονται **παρατηρήσεις ή δεδομένα**

ΕΙΣΑΓΩΓΗ


Στατιστική
Δεδομένα
Μέση βαθμολογία
Ποσοστό αριστούχων
Πιο συχνός βαθμός
Κατανομή βαθμών
Συμπεράσματα
Βαθμοί προηγούμενων ετών
5
9
7
5
8
3
:
Πειραματικός Σχεδιασμός
Μοντελοποίηση

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.

ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Η βασική απαίτηση είναι το μαθηματικό μοντέλο να εξηγεί με τον απλούστερο και καταλληλότερο τρόπο το συγκεκριμένο πρόβλημα που εξετάζεται.

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 προβλήματος.

ΜΟΝΤΕΛΟΠΟΙΗΣΗ

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.

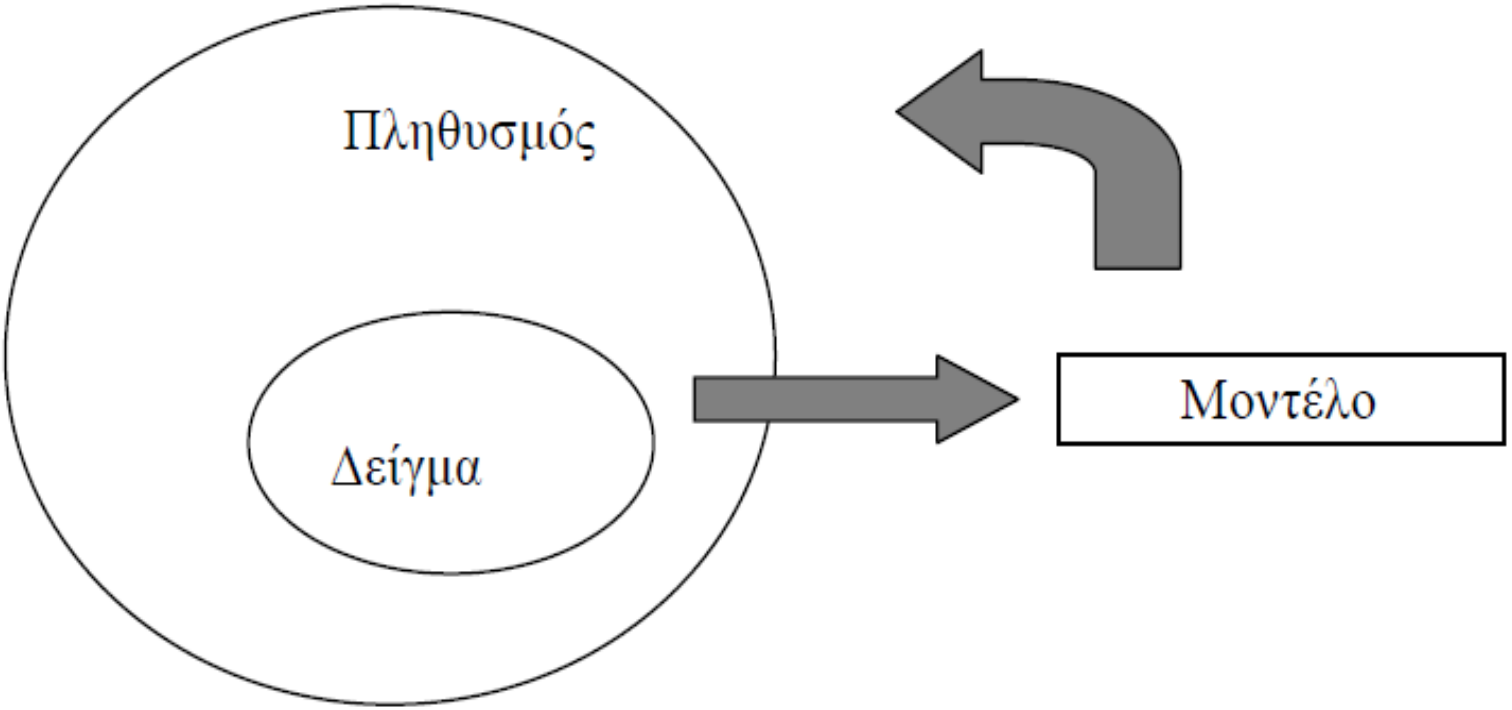
Κάθε μοντέλο (θεωρητικά ή πρακτικά) εξηγεί ένα συγκεκριμένο (ή ομάδα συγκεκριμένων) φαινομένων.

ΜΟΝΤΕΛΟΠΟΙΗΣΗ

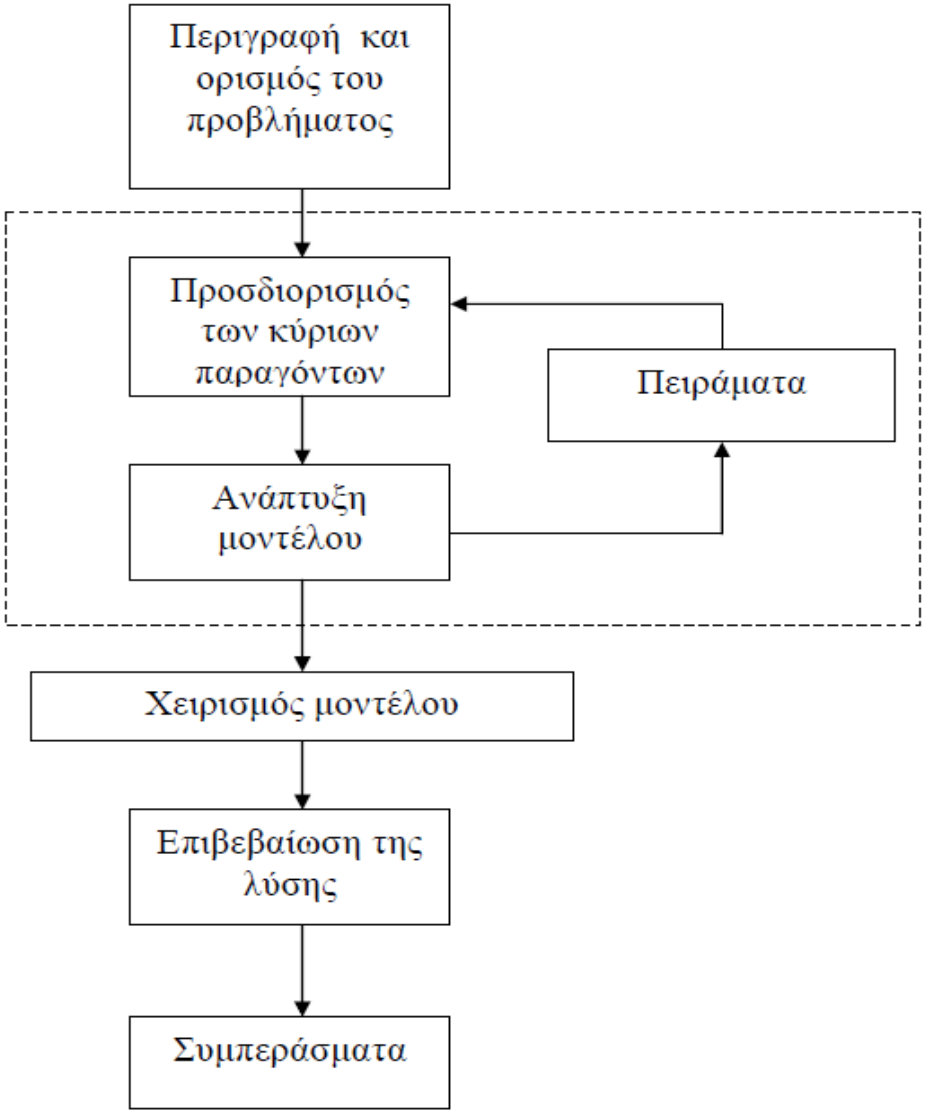
Κάθε φαινόμενο εξελίσσεται όχι αφηρημένα αλλά υποκείμενα, αυτόνομες μονάδες παρατήρησης, το σύνολο των οποίων ορίζει τον **πληθυσμό. Το φαινόμενο αναλύεται σε επιμέρους μετρήσιμα χαρακτηριστικά, τις μεταβλητές, στις οποίες αντιστοιχούμε τιμές. Η αντιστοίχηση αυτή ονομάζεται μέτρηση και γίνεται με την 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.**

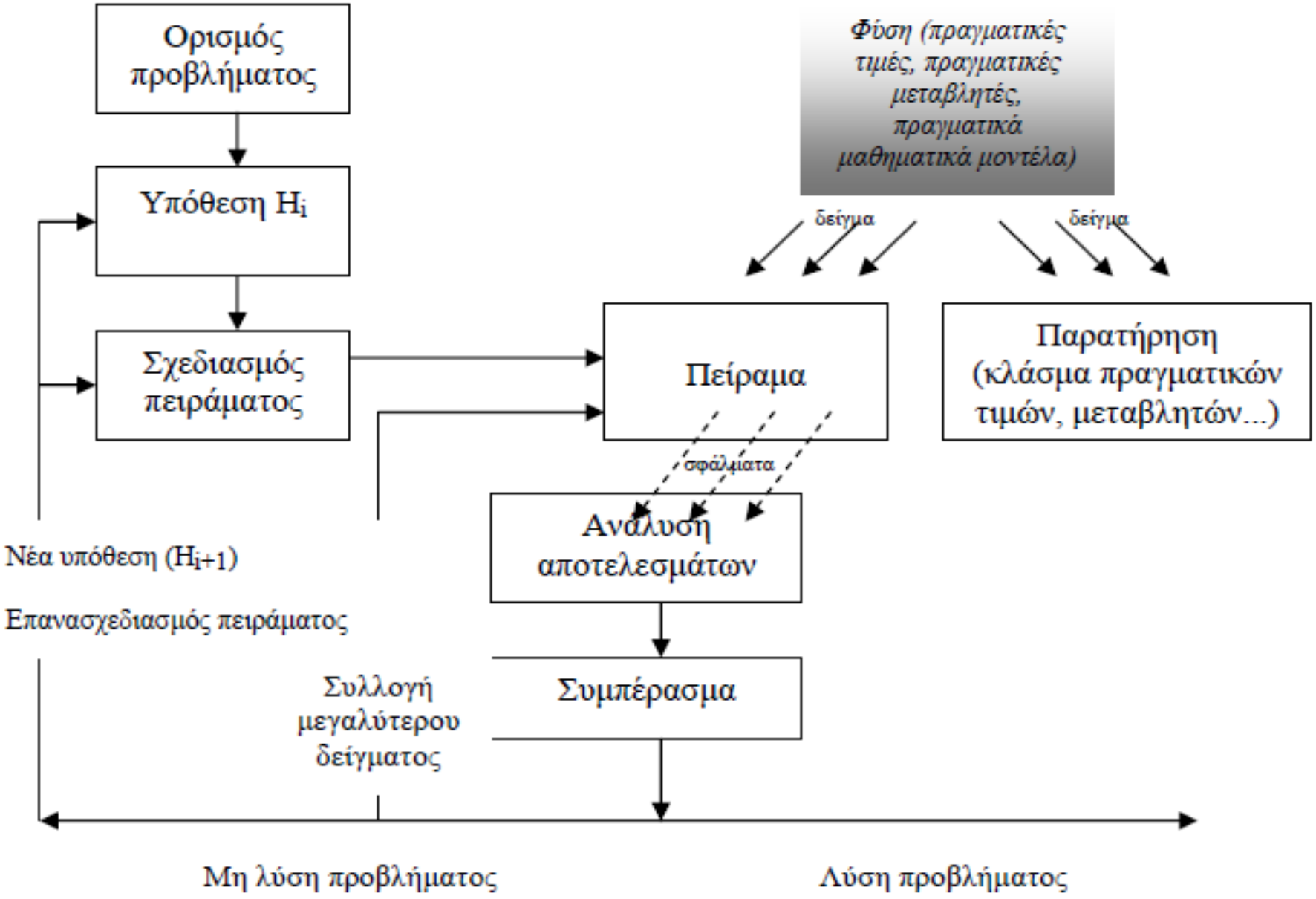
ΜΟΝΤΕΛΟΠΟΙΗΣΗ



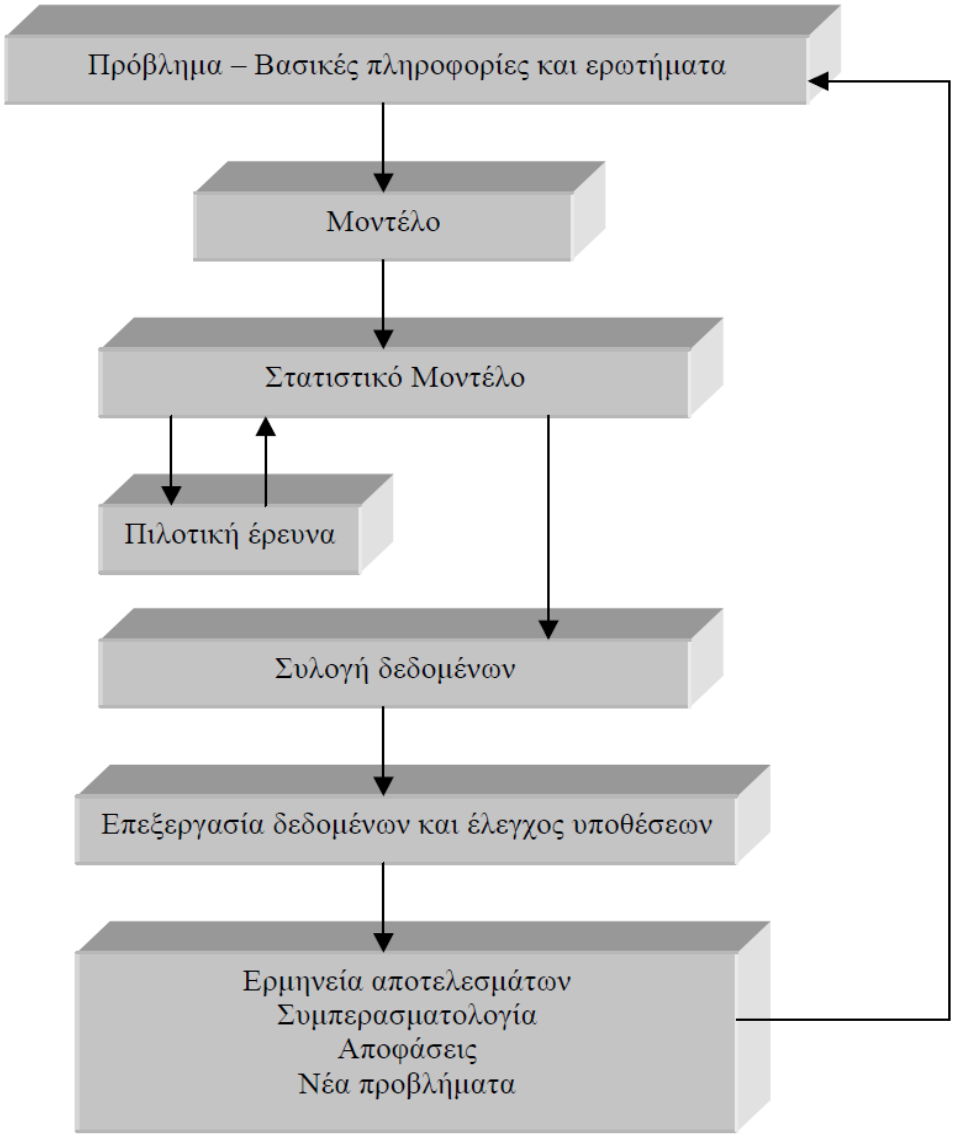
ΕΠΙΣΤΗΜΟΝΙΚΗ ΣΚΕΨΗ



ΕΠΙΣΤΗΜΟΝΙΚΗ ΣΚΕΨΗ



ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ



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.

ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

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.

ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

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 δευτερευουσών 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.

ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

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 .

ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

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**

Στατιστική προσαρμογή δεδομένων

Περιλαμβάνει στην κατασκευή νέων μεταβλητών που είναι απαραίτητες για την ανάλυση. Η δημιουργία βουβών μεταβλητών μπορεί να επαναπροσδιοριστεί μόνο ποιοτικά δεδομένα

ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Όταν πρέπει να χρησιμοποιήσουμε κάποια στατιστική τεχνική ***πρέπει :***

- 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.

ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Όταν πρέπει να χρησιμοποιήσουμε κάποια στατιστική τεχνική ***δεν πρέπει :***

- 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.

ΕΙΣΑΓΩΓΗ

- **τυχαία μεταβλητή** (**τ.μ.**): οποιοδήποτε χαρακτηριστικό του οποίου η τιμή αλλάζει στα διάφορα στοιχεία του πληθυσμού

- **δεδομένα**: ένα σύνολο τιμών μιας τ.μ. που έχουμε στη διάθεση μας
- **πληθυσμός**: μια ομάδα ή μια κατηγορία στην οποία αναφέρεται η τ.μ.
- **δείγμα**: ένα υποσύνολο του πληθυσμού που μελετάμε

- **παράμετρος**: ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στον πληθυσμό

- **στατιστικό**: ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στο δείγμα

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

- Αφορά μεθοδολογία οργάνωσης, σύνοψης και παρουσίασης των δεδομένων, τα οποία συλλέξαμε, με εύκολο και πληροφοριακό τρόπο
- Η περιγραφική στατιστική περιλαμβάνει αριθμητικές και γραφικές τεχνικές
- Η μέθοδος που εφαρμόζεται εξαρτάται από το είδος της πληροφορίας που θέλουμε να μελετήσουμε
  - ✓ μέτρα κεντρικής τάσης και/ή
  - ✓ μέτρα μεταβλητότητας
  - ✓ μέτρα μορφής

ΤΥΠΟΙ ΔΕΔΟΜΕΝΩΝ

(α) αριθμητικά και ονομαστικά δεδομένα

αριθμητικά ή ποσοτικά δεδομένα (numeric, quantitative data) είναι πραγματικοί αριθμοί που αναφέρονται στην τιμή κάποιου χαρακτηριστικού (π.χ. ύψος, βάρος, μισθός κ.λ.π).

ονομαστικά ή ποιοτικά δεδομένα (nominal, qualitative data) είναι η πληροφορία που παρέχεται από ένα άτομο και περιγράφεται από κάποια λέξη που εκφράζει κατάσταση ή ιδιότητα

(β) συνεχή και διακριτά δεδομένα

Απαντήσεις σχετικά με το ύψος κάποιων ατόμων (συνεχή, continuous data) ή σχετικά με τον αριθμό των παιδιών που διαθέτει κάποιος (διακριτά, discrete data)
Τα *συνεχή* είναι μετρίσιμα (καλύπτουν ολόκληρο διάστημα τιμών χωρίς κενά, ενώ τα *διακριτά* είναι απαριθμήσιμα (οι τιμές ξεχωρίζουν η μια από την άλλη)

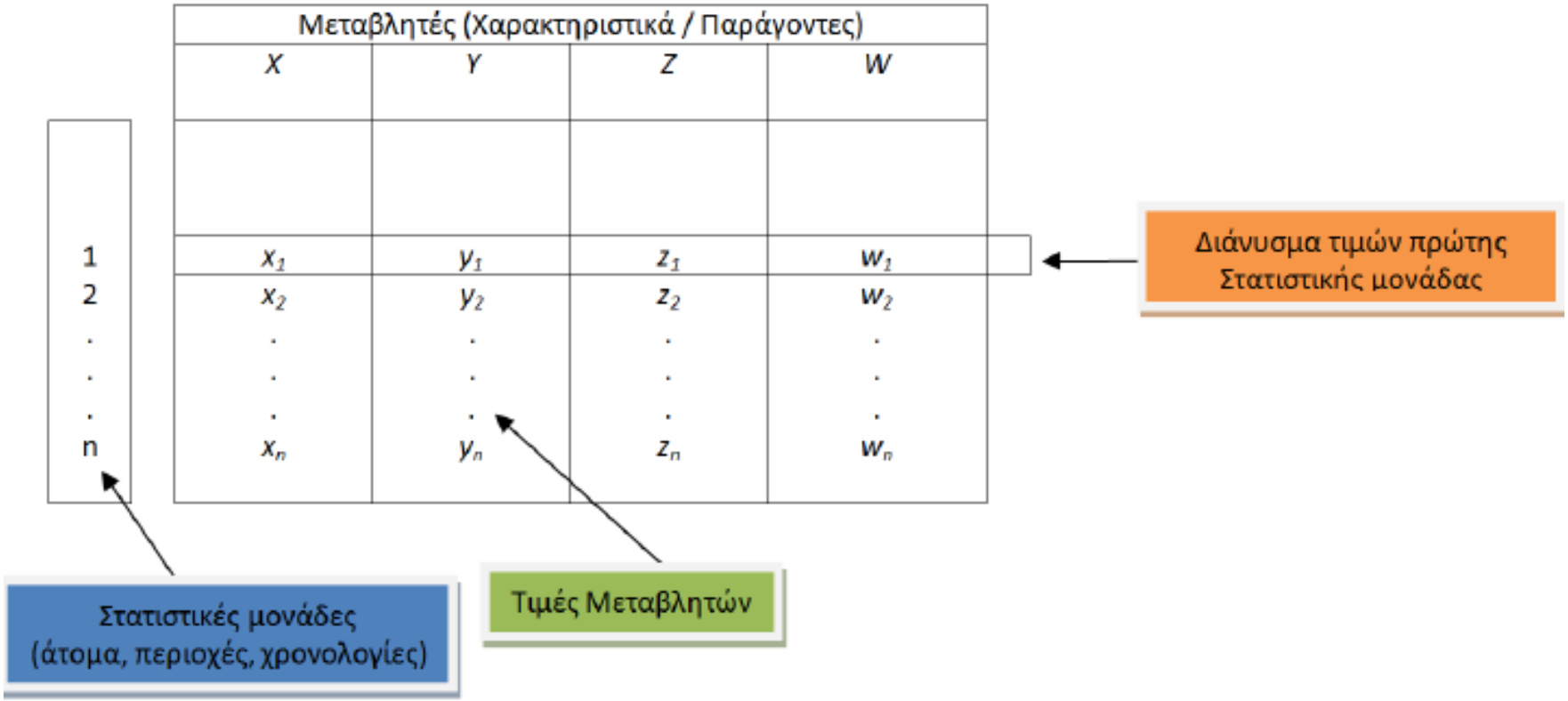
ΤΥΠΟΙ ΔΕΔΟΜΕΝΩΝ

- ✓ Οι ακριβείς περιγραφές και μετρήσεις των χαρακτηριστικών και ιδιοτήτων των στατιστικών μονάδων καλούνται τιμές των δεδομένων **(data values)** και αποτελούν αντικείμενο στατιστικής επεξεργασίας και ανάλυσης.

- ✓ Κάθε πληθυσμιακό χαρακτηριστικό καλείται **μεταβλητή (variable).**
  Οι μεταβλητές συμβολίζονται με κεφαλαίους λατινικούς χαρακτήρες.

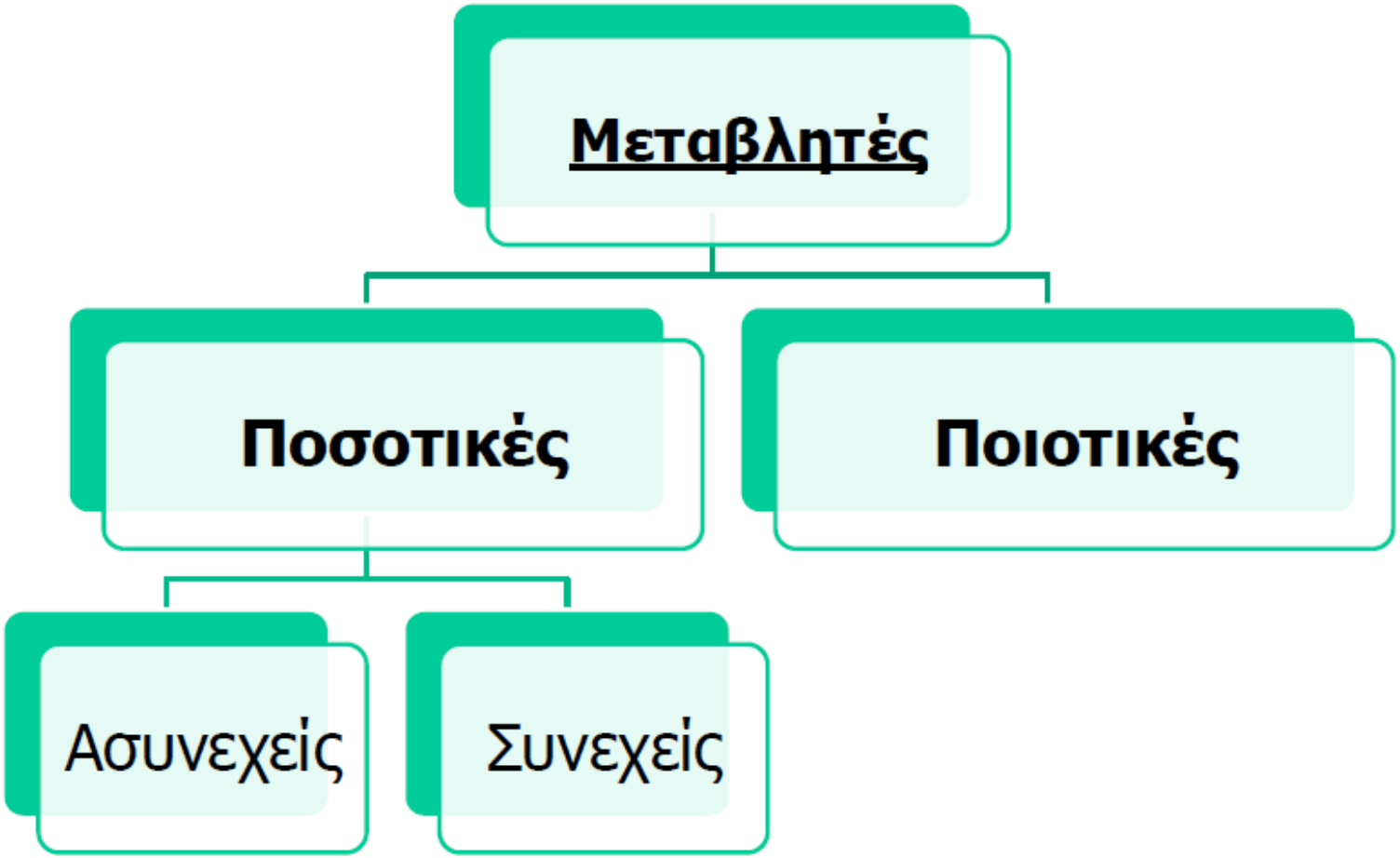
Άτομο (i)	Βάρος (κιλά) (x_i)	Φύλο (y_i)
1	x_1=72	y_1=άνδρας
2	x_2=78	y_2=άνδρας
3	x_3=65	y_3=γυναίκα
4	x_4=70	y_4=άνδρας
5	x_5=64	y_5= γυναίκα

ΤΥΠΟΙ ΔΕΔΟΜΕΝΩΝ

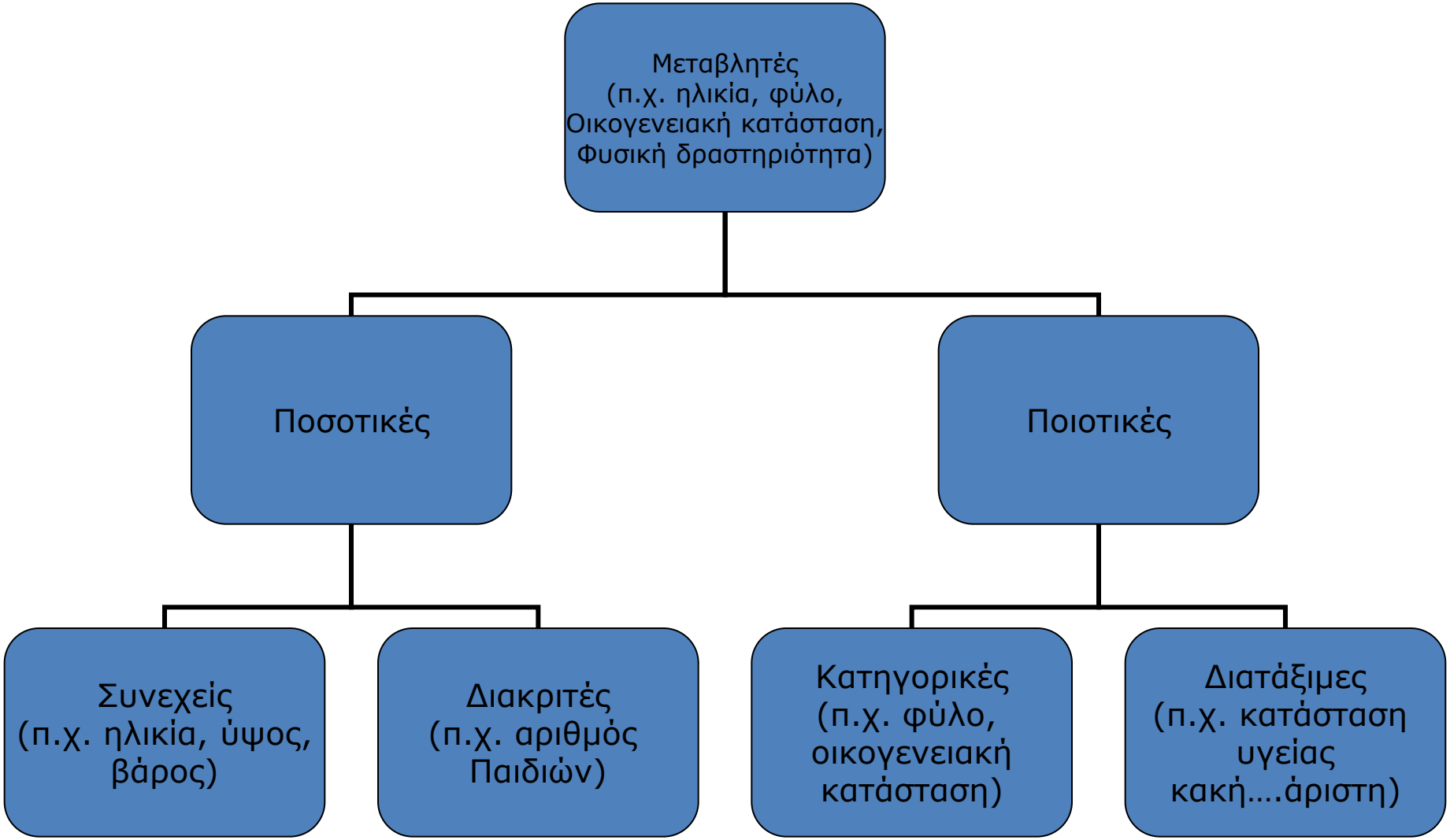


Οι μεταβλητές δίνονται σε στήλες, ενώ οι παρατηρήσεις σε γραμμές

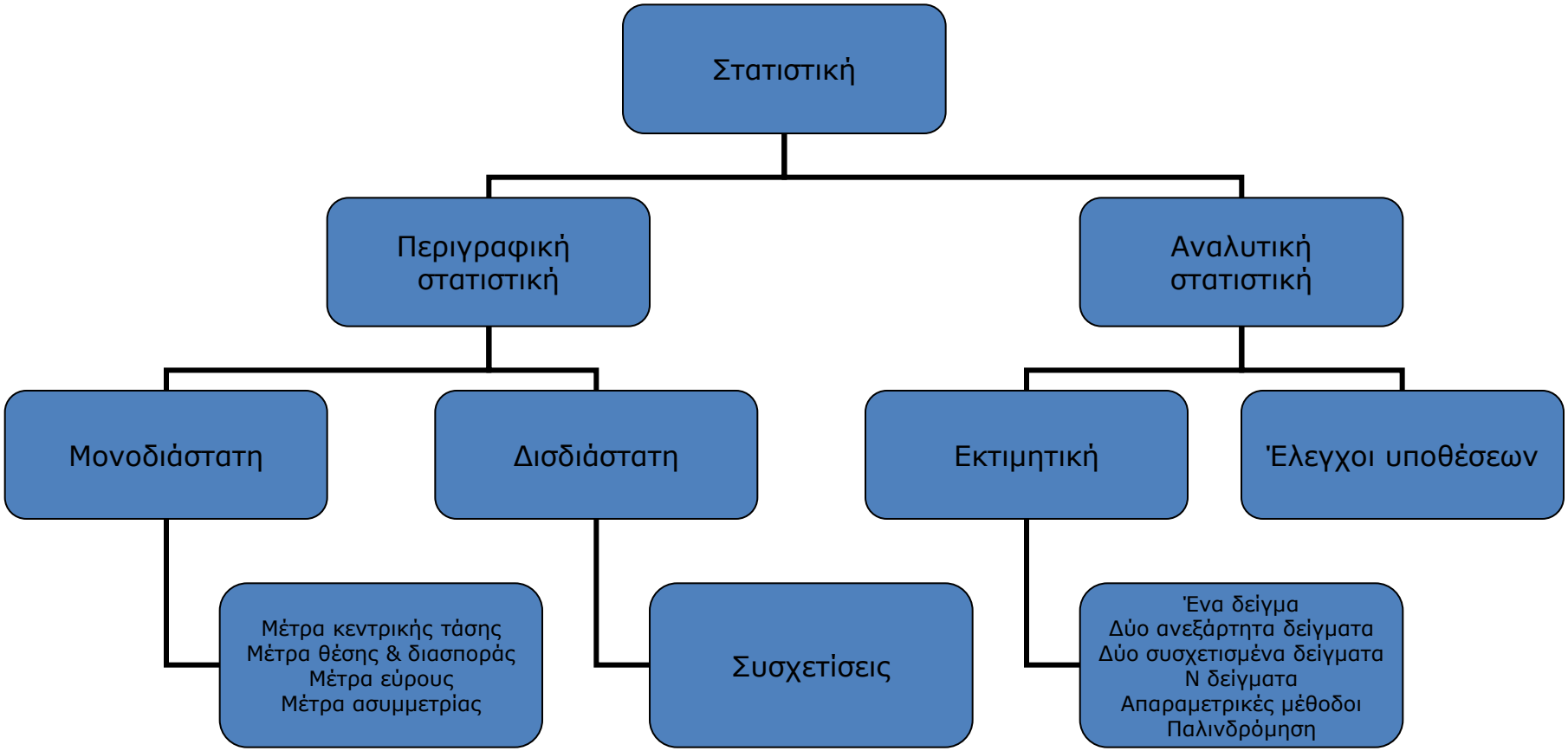
ΤΥΠΟΙ ΔΕΔΟΜΕΝΩΝ



ΤΥΠΟΙ ΔΕΔΟΜΕΝΩΝ



ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ



ΤΥΠΟΙ ΜΕΤΑΒΛΗΤΩΝ

Ποιοτικές

- Δεν επιδέχονται αριθμητική μέτρηση
- Περιγράφονται σε κατηγορίες στις οποίες ταξινομούνται οι παρατηρήσεις

  Παράδειγμα.

  Το φύλο (αγόρι, κορίτσι), οι ομάδες αίματος (A, B, O, AB)
- Κάθε παρατήρηση κατατάσσεται σε μία μόνο κατηγορία

ΤΥΠΟΙ ΜΕΤΑΒΛΗΤΩΝ

Διατάξιμες/Διαβαθμιζόμενες

- Ποιοτικές μεταβλητές με ιεράρχηση των κατηγοριών
- Αποδίδεται θέση ή σειρά στις κατηγορίες
- Δεν δίνει πληροφορίες για τη διαφορά που υπάρχει μεταξύ των κατηγοριών

Παράδειγμα.

Το αποτέλεσμα μίας νόσου (ίαση, βελτίωση, στασιμότητα, επιδείνωση, θάνατος)

ΤΥΠΟΙ ΜΕΤΑΒΛΗΤΩΝ

Ποσοτικές

επιδέχονται αριθμητική μέτρηση

- **Ασυνεχείς/διακριτές:** Λαμβάνουν ορισμένες αριθμητικές τιμές (υποσύνολο φυσικών αριθμών)

  Παράδειγμα.

  Ο αριθμός των μαθητών στις τάξεις ενός σχολείου

- **Συνεχείς:** Μπορούν να πάρουν όλες τις τιμές των πραγματικών αριθμών σε ένα διάστημα

  Παράδειγμα.

  Φυσικό μέγεθος (ύψος), βιολογικό μέγεθος (χοληστερόλη)

ΓΡΑΦΗΜΑΤΑ

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 στόχος 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.

ΓΡΑΦΗΜΑΤΑ

Βασικά χαρακτηριστικά για την δημιουργία ενός γραφήματος είναι:

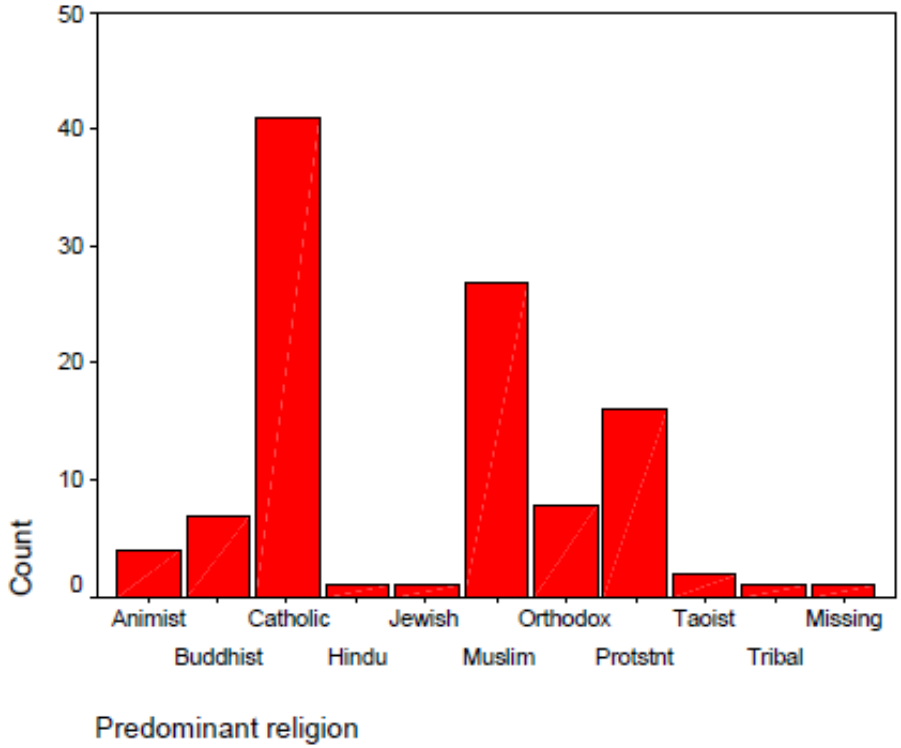
- Να ξεκαθαρίσουμε τους στόχους και τις προτεραιότητες σε ότι αφορά το μήνυμα που θέλουμε να δώσουμε.
- Να επιλέξουμε το κατάλληλο είδος γραφικής παράστασης.
- Να ενημερώσουμε τον αναγνώστη σχετικά με την φύση των απεικονιζόμενων πληροφοριών με σαφή τίτλο.
- Να κατασκευάσουμε ένα σχεδιάγραμμα το οποίο να είναι: παραστατικό, σαφές και ακριβές.

ΓΡΑΦΗΜΑΤΑ

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

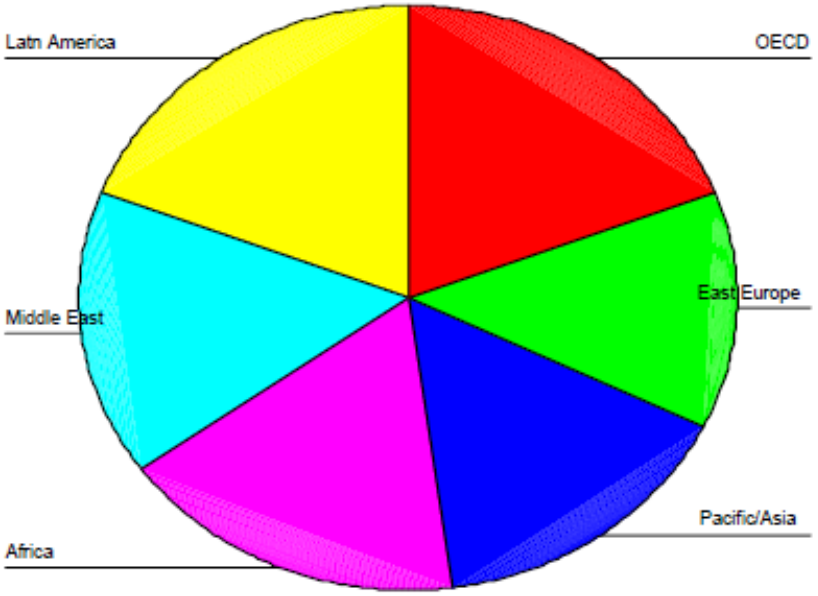
ΡΑΒΔΟΓΡΑΜΜΑ

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



ΚΥΚΛΙΚΟ ΔΙΑΓΡΑΜΜΑ

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.

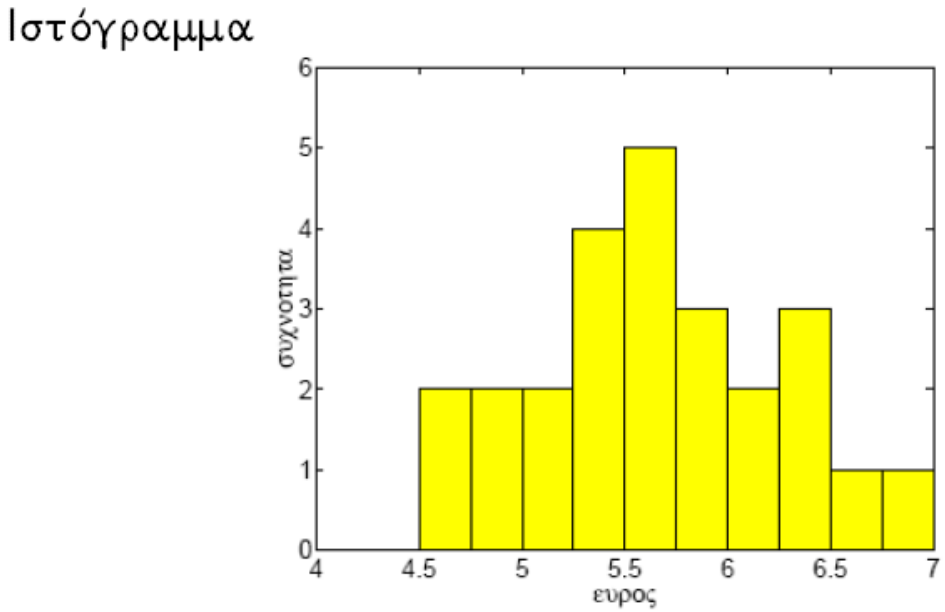


ΙΣΤΟΓΡΑΜΜΑ

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

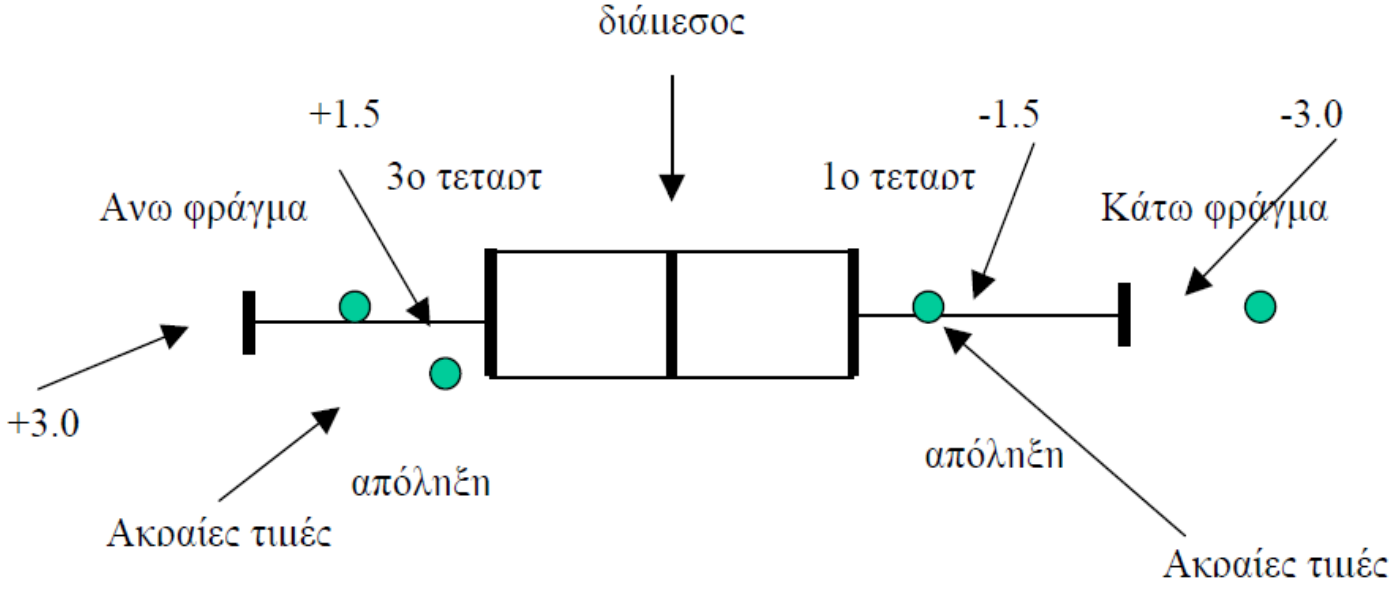
Το ύψος κάθε τμήματος αναπαριστά τη συχνότητα με την οποία εμφανίζεται αυτή η τιμή ή το διάστημα.

Επίσης χρησιμοποιείται για να δείξει το σχήμα μίας μεταβλητής.

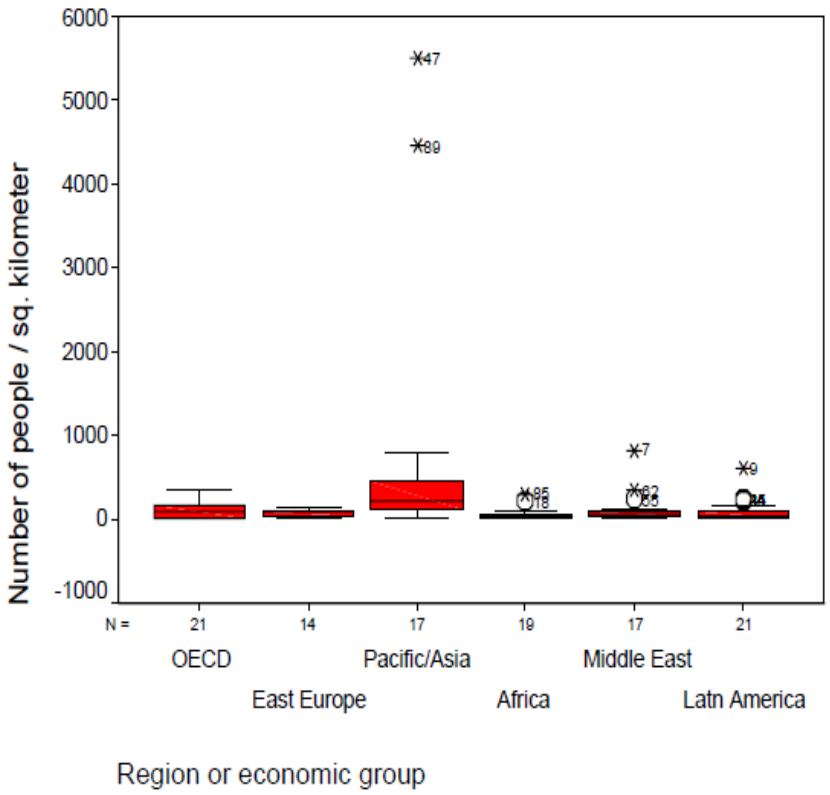


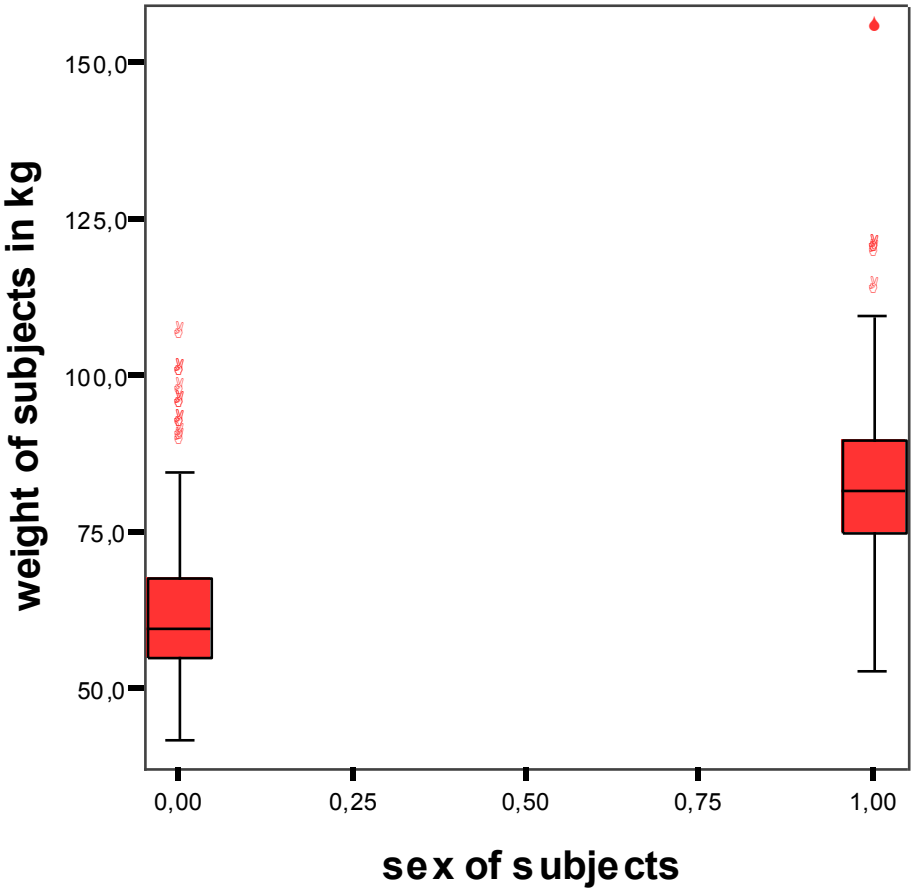
ΔΙΑΓΡΑΜΜΑΤΑ ΠΛΑΙΣΙΟΥ-ΑΠΟΛΗΞΕΩΝ

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



ΔΙΑΓΡΑΜΜΑΤΑ ΠΛΑΙΣΙΟΥ-ΑΠΟΛΗΞΕΩΝ

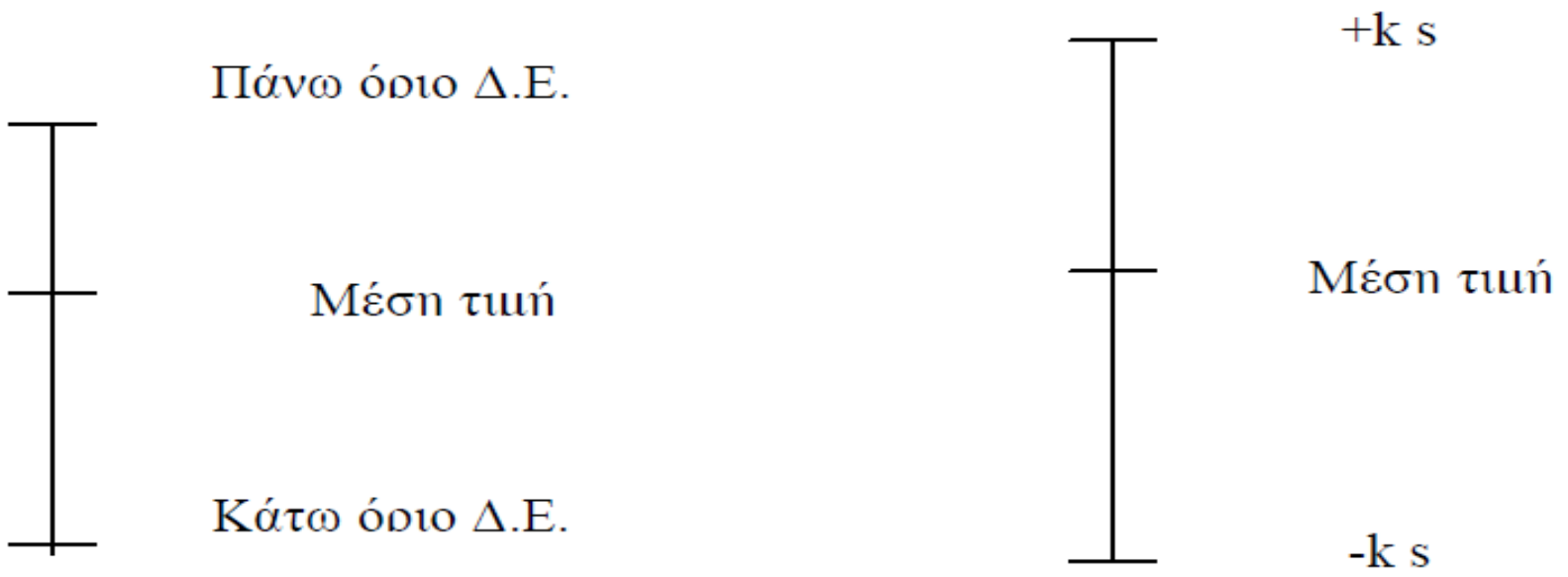




ΔΙΑΓΡΑΜΜΑΤΑ ΣΦΑΛΜΑΤΩΝ

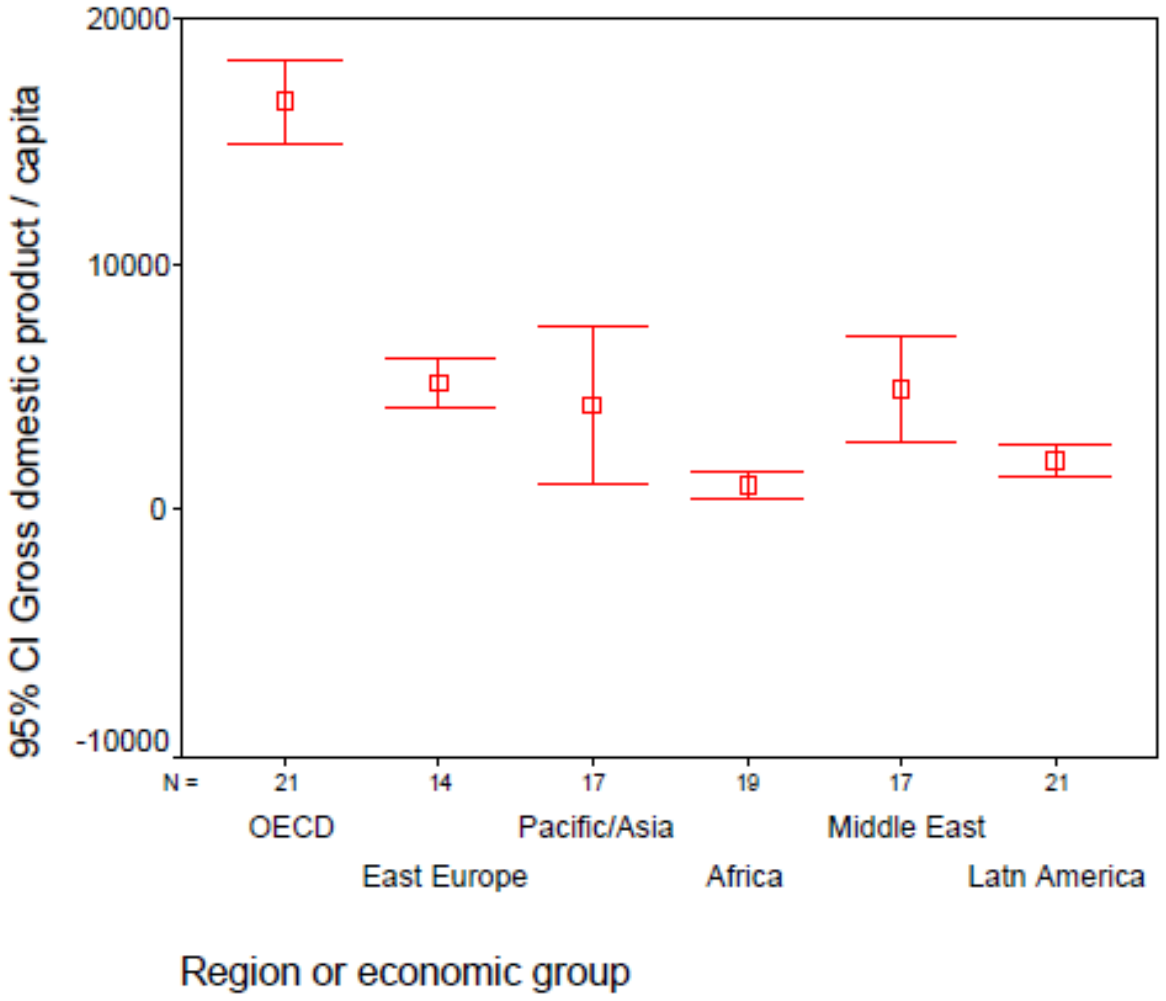
Είναι διαγράμματα που περιγράφουν περιληπτικά την κατανομή μίας ποσοτικής μεταβλητής σε διάφορα επίπεδα. Συνήθως αναπαριστά διαστήματα εμπιστοσύνης για το μέσο, αλλά εναλλακτικά μπορεί να χρησιμοποιηθεί και για την τυπική απόκλιση. Μοιάζουν με τα θηκογράμματα, αλλά συγκρίνουν διαστήματα εμπιστοσύνης και όχι κατανομές. Το σχήμα τους είναι μια ράβδος κατανεμημένη ισομερώς γύρο από το τη μέση τιμή των τιμών.

ΔΙΑΓΡΑΜΜΑΤΑ ΣΦΑΛΜΑΤΩΝ



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

ΔΙΑΓΡΑΜΜΑΤΑ ΣΦΑΛΜΑΤΩΝ



ΔΙΑΓΡΑΜΜΑΤΑ ΣΗΜΕΙΩΝ

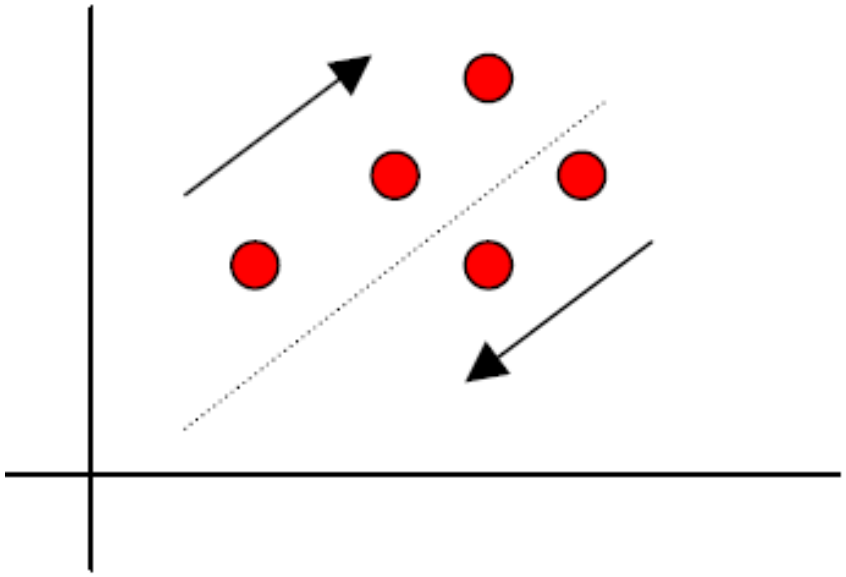
Ονομάζονται και διαγράμματα διασποράς. Αποτελούν το γραφικό τρόπο αναζήτησης σχέσεων μεταξύ μεταβλητών. Περιγράφουν τη δυσδιάστατοι κατανομή δύο ποσοτικών μεταβλητών. Κάθε σημείο απεικονίζει ένα ζευγάρι τιμών των υπό εξέταση μεταβλητών (συσχέτιση μεταβλητών). Εντοπίζονται εύκολα συσχετίσεις και ακραίες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

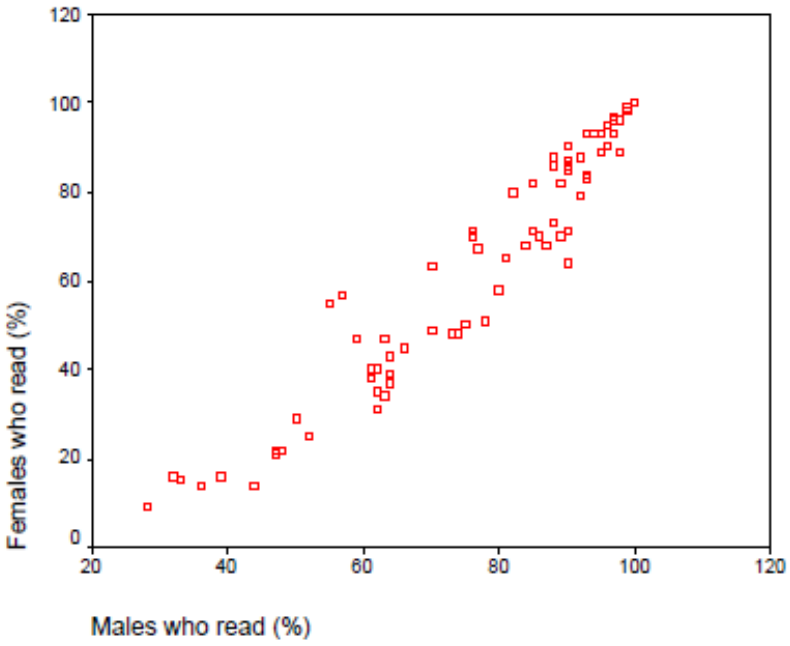
Συσχέτιση αναφέρεται ο βαθμός με τον οποίο σχετίζονται (συμμεταβάλλονται) δύο μεταβλητές

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

ΔΙΑΓΡΑΜΜΑΤΑ ΣΗΜΕΙΩΝ

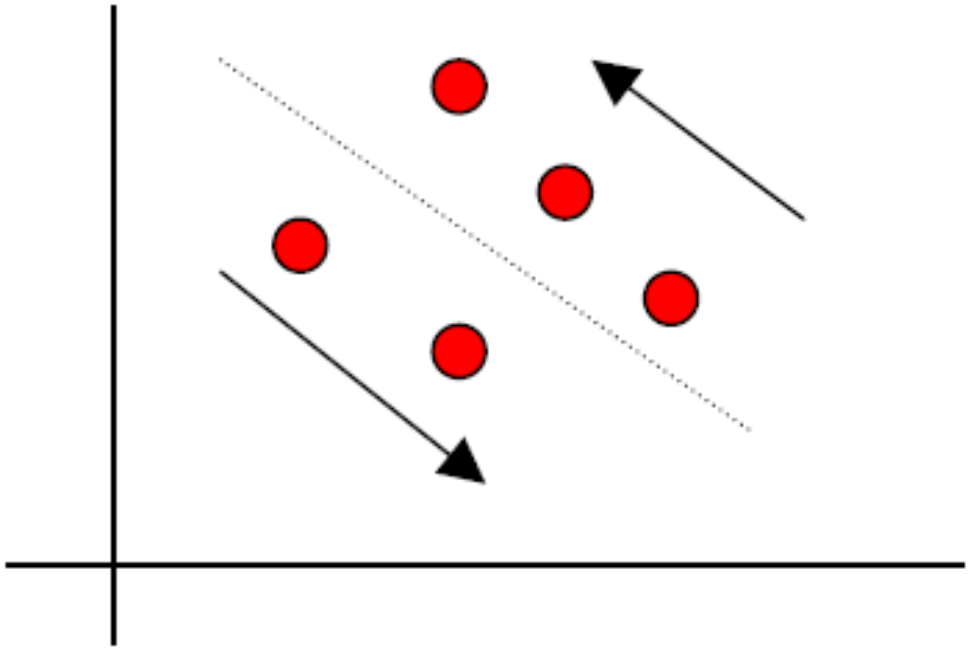
Θετική συσχέτιση όταν δύο μεταβλητές τείνουν να μεταβάλλονται προς την ίδια κατεύθυνση. Στην περίπτωση αυτή οι τιμές τείνουν να αυξάνονται ή να μειώνονται





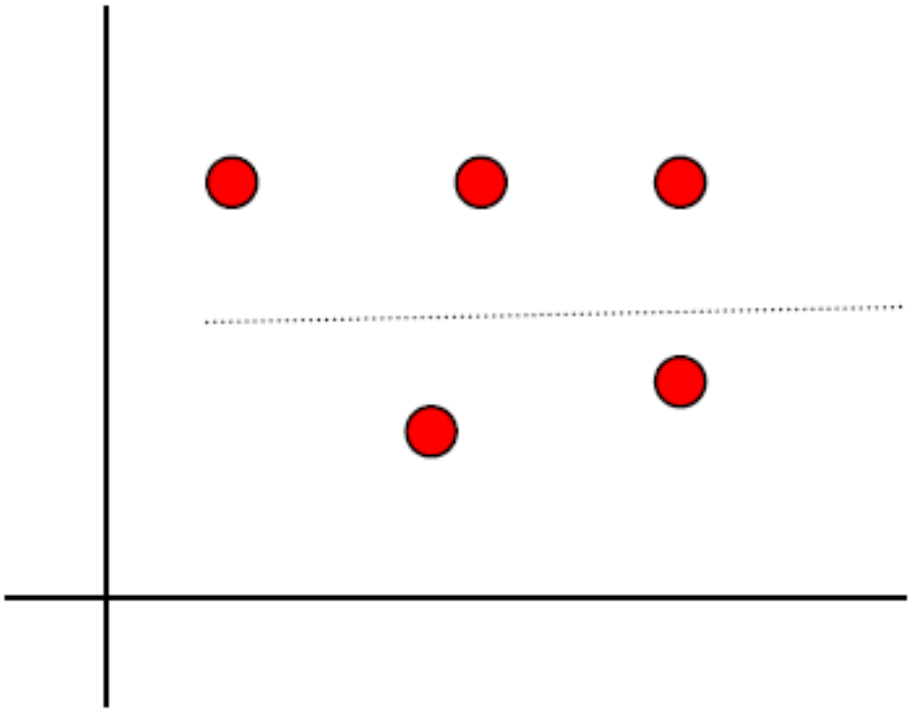
ΔΙΑΓΡΑΜΜΑΤΑ ΣΗΜΕΙΩΝ

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



ΔΙΑΓΡΑΜΜΑΤΑ ΣΗΜΕΙΩΝ

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



ΑΝΑΛΥΣΗ

- **Συχνότητα (f_i)**: Φυσικός αριθμός που δείχνει πόσες φορές εμφανίζεται η τιμή x_i της εξεταζόμενης μεταβλητής Χ στο σύνολο των παρατηρήσεων (n) του δείγματος
- **Σχετική συχνότητα (p_i)**: Ο λόγος της συχνότητας εμφάνισης (f_i) μίας τιμής (x_i) της εξεταζόμενης μεταβλητής Χ με το μέγεθος (n) του δείγματος

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

- **Αθροιστική (σχετική) συχνότητα (C_i)** μίας τιμής (x_i): Το πλήθος (το ποσοστό) των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i

ΑΝΑΛΥΣΗ

- Παράδειγμα

		Πλήθος (*fi*)
Άγαμοι	ΑΑΑΑΑΑΑΑ	8
Έγγαμοι	ΕΕΕΕΕΕΕΕΕΕ	11
Χήροι	ΧΧΧΧ	4
Διαζευγμένοι	ΔΔ	2
Συνολικός Αριθμός		n=25

- Οι αριθμητικές τιμές που προέκυψαν από την διαδικασία καταμέτρησης λέγονται **συχνότητες (frequencies).**
- Τα αποτελέσματα αυτά μας δίδουν την **εμπειρική ή πειραματική κατανομή (empirical distribution)** των μελών του δείγματος ανάλογα με την οικογενειακή τους κατάσταση.
- Οι σχετικές συχνότητες μπορούν να εκφραστούν
  - είτε ανά άτομο οπότε λαμβάνουν την μορφή της **αναλογίας** $\frac{f_i}{n}$ **(proportion)**
  - είτε ως **ποσοστό (percentage)** επί τοις 100 $\frac{f_i}{n}.100$.

ΑΝΑΛΥΣΗ

Δίνουν μια σαφή εικόνα των πραγματικών μεγεθών ενός φαινομένου.

	Πλήθος (*fi*)	$\frac{f_i}{n}$	$\frac{f_i}{n}.100$.
Άγαμοι	8	0,32	32
Έγγαμοι	11	0,44	44
Χήροι	4	0,16	16
Διαζευγμένοι	2	0,08	8
Συνολικός Αριθμός	n=25	1	100

ΙΣΤΟΓΡΑΜΜΑΤΑ

1	4	2	2	2	3	4	3	1	1	3	3
1	2	1	2	1	1	2	3	3	5	1	2
2	3	2	1	1	4	3	4	1	1	6	2
1	3	2	1	2	2	3	2	4	3	3	5
1	3	5	3	1	2	2	3	1	2	6	4
1	2	5	4	3	1	2	4	2	1	3	4
2	2	2	3	2	1	3	3	4	2	1	5
2	2	3	3	2	4	6	3	2	3	1	3
2	1	5	1	1	4	4	2	5	4	2	2
4	2	1	2	2	2	3	2	3	2	1	4

Δίνεται ο αριθμός των χρηστών που προσπαθούν να αποκτήσουν ευρυζωνική πρόσβαση

ΙΣΤΟΓΡΑΜΜΑΤΑ

Πίνακας συχνοτήτων

x_i	f_i	p_i	F_i	P_i
1	28	0.23	28	0.23
2	39	0.33	67	0.56
3	27	0.23	94	0.78
4	16	0.13	110	0.92
5	7	0.06	117	0.97
6	3	0.03	120	1.00
Άθροισμα	120	1.00		

Χωρίζουμε τα δεδομένα σε k ομάδες $\longrightarrow$ πίνακες / γραφήματα όπως πριν για τις k τιμές (ομάδες)

Χωρισμός σε ομάδες: (ίδιο εύρος τιμών r σε κάθε ομάδα)

Εύρος δεδομένων: $R = x_{\max} - x_{\min}$

$$R/k \quad \simeq \quad r$$

Το πρώτο διάστημα πρέπει να περιέχει το $x_{\min}$
Το τελευταίο διάστημα πρέπει να περιέχει το $x_{\max}$

ΙΣΤΟΓΡΑΜΜΑΤΑ

Χωρισμός σε ομάδες

$x_{\min} = 4.5 \qquad x_{\max} = 7.0$

$R = x_{\max} - x_{\min} = 7.0 - 4.5 = 2.5$

Διαλέγουμε να χωρίσουμε τα δεδομένα σε 10 ομάδες ($k = 10$)

$$r = \frac{R}{k} = \frac{2.5}{10} = 0.25$$

ομάδα 1: $4.50 - 4.75$

ομάδα 2: $4.75 - 5.00$

...

ομάδα 10: $6.75 - 7.00$

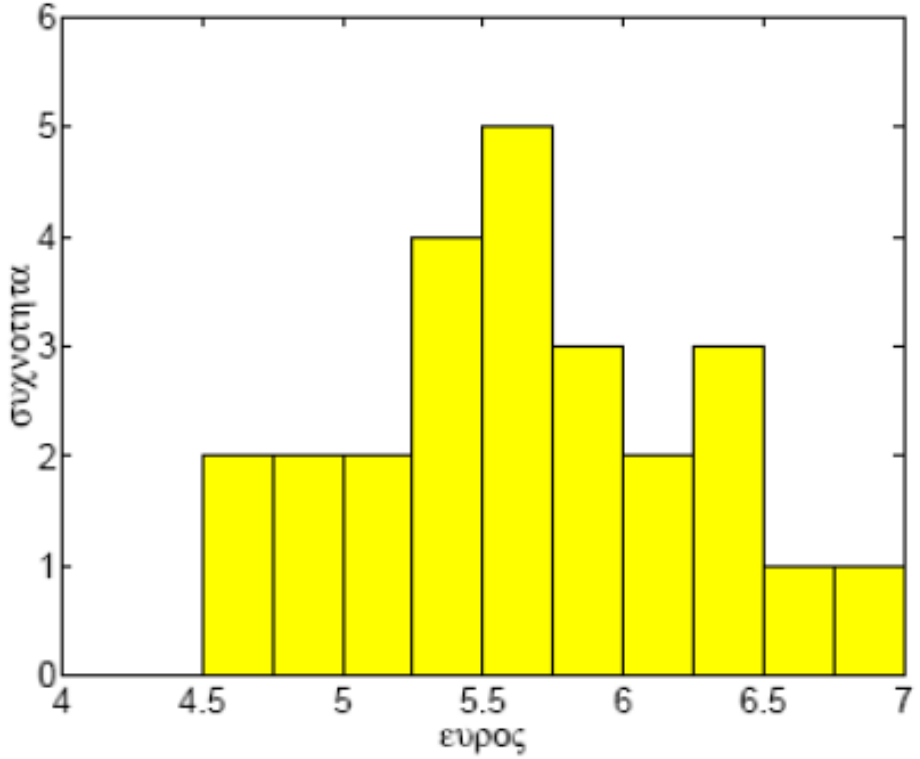
ΙΣΤΟΓΡΑΜΜΑΤΑ

Πίνακας συχνοτήτων

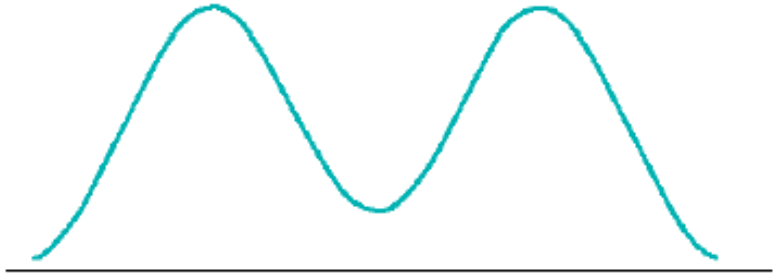
Διάστημα τιμών	f_i	p_i	F_i	P_i
4.50 – 4.75	2	0.08	2	0.08
4.75 – 5.00	2	0.08	4	0.16
5.00 – 5.25	2	0.08	6	0.24
5.25 – 5.50	4	0.16	10	0.40
5.50 – 5.75	5	0.20	15	0.60
5.75 – 6.00	3	0.12	18	0.72
6.00 – 6.25	2	0.08	20	0.80
6.25 – 6.50	3	0.12	23	0.92
6.50 – 6.75	1	0.04	24	0.96
6.75 – 7.00	1	0.04	25	1.00
Άθροισμα	25	1.00		

ΙΣΤΟΓΡΑΜΜΑΤΑ

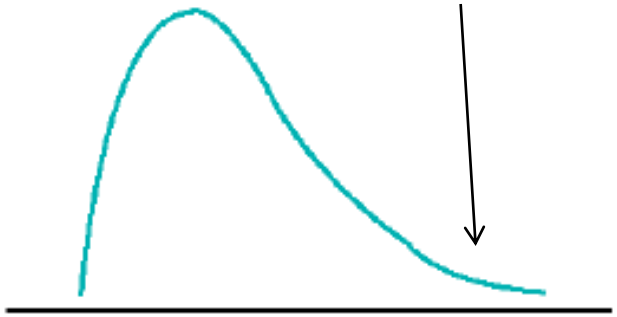
Ιστόγραμμα



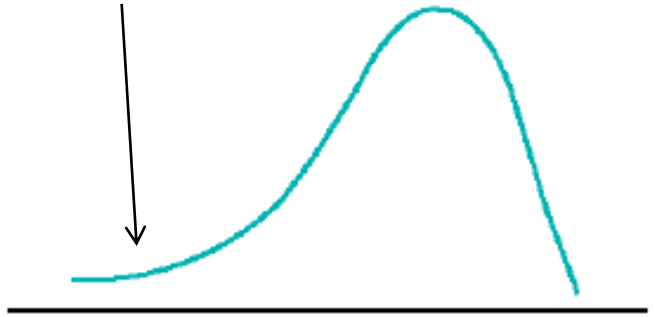
ΚΑΤΑΝΟΜΕΣ



Συμμετρική



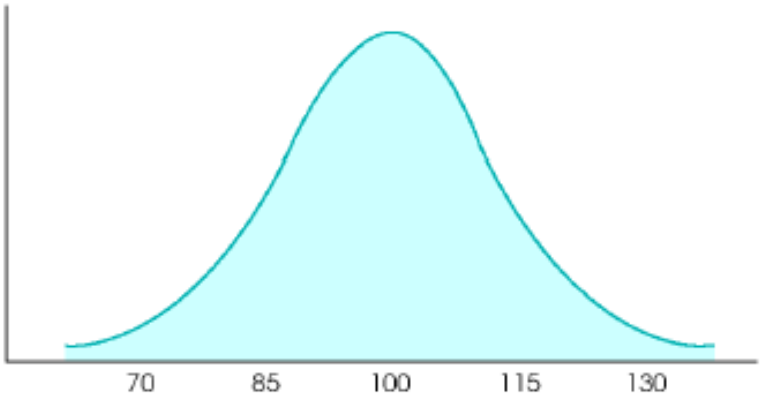
Θετικά
ασυμμετρική



Αρνητικά
ασυμμετρική

ΚΑΤΑΝΟΜΕΣ

- Κωδωνοειδής
- Συμμετρική γύρω από τη μέση τιμή, η οποία παρουσιάζει και την υψηλότερη συχνότητα



- Πλησιάζει ασυμπτωτικά τον οριζόντιο άξονα

ΚΑΤΑΝΟΜΕΣ

	Μέτρο	Μέτρα που δεν βασίζονται στη μέση τιμή
Κεντρική τάση	Μέση τιμή	Διάμεσος, Επικρατούσα τιμή
Μεταβλητότητα	Διασπορά (τυπική απόκλιση)	Εύρος, Ενδοτεταρτημοριακό εύρος
Λοξότητα	Συντελεστής λοξότητας	--
Κυρτότητα	Συντελεστής κυρτότητας	--

ΚΑΤΑΝΟΜΕΣ

– Μέτρα θέσης (κεντρικής τάσης): περιγραφή της θέσης της κατανομής από όπου προέρχονται τα δεδομένα

– Μέτρα διασποράς (μεταβλητότητας): εκφράζουν τις αποκλίσεις των τιμών μιας μεταβλητής γύρω από τα μέτρα θέσης

– Μέτρα λοξότητας: περιγράφουν τη συμμετρία της καμπύλης συχνοτήτων μίας μεταβλητής

– Συντελεστής κυρτότητας: περιγράφει τη μορφή της κορυφής της καμπύλης συχνοτήτων (συνήθως σε σχέση με την κανονική κατανομή)

ΜΕΤΡΑ ΘΕΣΗΣ

- Προσδιορίζουν ένα κεντρικό σημείο γύρω από το οποίο τείνουν να συγκεντρώνονται τα δεδομένα
- Τα κυριότερα μέτρα θέσης είναι:
  - Μέση τιμή (δειγματικός μέσος όρος, αριθμητικός μέσος)
  - Διάμεσος
  - Επικρατούσα τιμή
  - Ποσοστημόρια

ΜΕΤΡΑ ΘΕΣΗΣ

• **Μέση τιμή δείγματος** ορίζεται το άθροισμα των τιμών των παρατηρήσεων του δείγματος, προς το πλήθος των παρατηρήσεων

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

Πλεονεκτήματα

• Υπολογίζεται από όλες τις τιμές

• «Εκπρόσωπος» των παρατηρήσεων

• Εύκολος υπολογισμός / Απλή ερμηνεία

Μειονεκτήματα

• Ευαίσθητη σε ακραίες παρατηρήσεις. Η παρουσία ακραίων παρατηρήσεων καθιστά τη μέση τιμή μη αντιπροσωπευτική του δείγματος

• Δεν υπολογίζεται για ποιοτικά δεδομένα

ΔΙΑΜΕΣΟΣ

- Το σημείο εκείνο κάτω από το οποίο βρίσκεται το 50% των παρατηρήσεων και πάνω από το οποίο βρίσκεται το άλλο 50% των παρατηρήσεων
  – Η τιμή που διαιρεί το δείγμα (σε διατεταγμένες τιμές) σε δύο ακριβώς ίσα τμήματα

- Αν ν: περιττός $\delta = x_{(n+1)/2}$

- Αν ν: άρτιος $\delta = \frac{x_{n/2} + x_{(n/2)+1}}{2}$

Πλεονεκτήματα

- Δεν επηρεάζεται από ακραίες τιμές
- Είναι μοναδική σε κάθε σύνολο δεδομένων

Μειονεκτήματα

- Δεν χρησιμοποιούνται όλες οι τιμές
- Δεν υπολογίζεται για κατηγορικά δεδομένα

ΕΠΙΚΡΑΤΟΥΣΑ ΤΙΜΗ

- Η παρατήρηση με τη μεγαλύτερη συχνότητα (διακριτά δεδομένα)
- Η κεντρική τιμή της ομάδας (κλάσης) με τη μεγαλύτερη συχνότητα (ομαδοποιημένα δεδομένα)

Μειονέκτημα

- Δεν υπολογίζεται από όλες τις τιμές

Πλεονεκτήματα

- Δεν επηρεάζεται από ακραίες παρατηρήσεις
- Υπολογίζεται και για ποιοτικά δεδομένα

ΠΑΡΑΔΕΙΓΜΑ

- Για τα αποτελέσματα στη δοκιμασία μνήμης:

3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 9, 10, 10, 11

- Επικρατούσα τιμή = 6

- Διάμεσος = 6

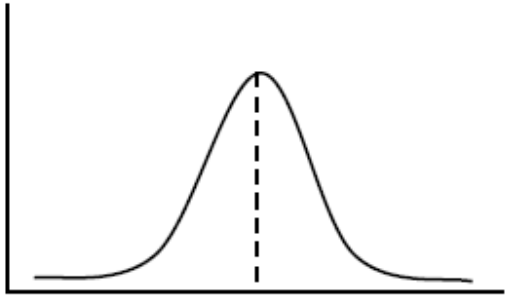
$$\delta = \frac{x_{n/2} + x_{(n/2)+1}}{2}$$

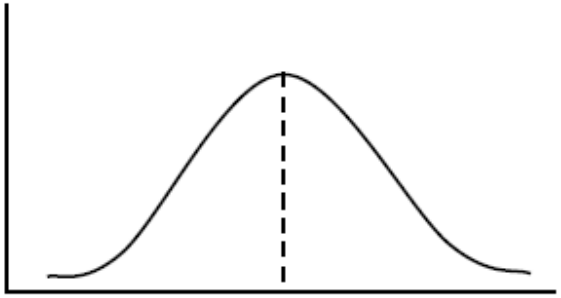
- Μέση τιμή = 6,35

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

ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ

- Όταν τα δεδομένα είναι πολύ σκορπισμένα, τα μέτρα θέσης δεν επαρκούν για να περιγράψουν την κατανομή
- Τα κυριότερα μέτρα διασποράς είναι
  - Εύρος
  - Διασπορά (διακύμανση)
  - Τυπική απόκλιση
  - Ενδοτεταρτημοριακό εύρος





ΔΙΑΚΥΜΑΝΣΗ

- Δηλώνει την απόσταση των παρατηρήσεων από τη μέση τιμή

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

  - Όταν οι τιμές απέχουν πολύ από τη μέση τιμή, η διακύμανση είναι μεγάλη
  - Όταν οι τιμές δεν διαφέρουν πολύ από τη μέση τιμή, η διακύμανση είναι μικρή

ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ

- Η τυπική απόκλιση αποτελεί τη θετική τετραγωνική ρίζα της διακύμανσης
- Αποτελεί μέτρο μεταβλητότητας
- Εκφράζεται στην ίδια μονάδα μέτρησης με το χαρακτηριστικό που μελετάμε

ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ

Η τυπική απόκλιση μπορεί να χρησιμοποιηθεί για την σύγκριση της μεταβλητότητας αρκετών κατανομών και τον χαρακτηρισμό της γενικής μορφής μιας κατανομής.

- Αν η καμπύλη κατανομής συχνοτήτων για την ποσοτική μεταβλητή που εξετάζουμε είναι κανονική ή περίπου κανονική, τότε η τυπική απόκλιση s έχει τις παρακάτω ιδιότητες :

ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ

- Περίπου το 68% των παρατηρήσεων βρίσκεται στο διάστημα :

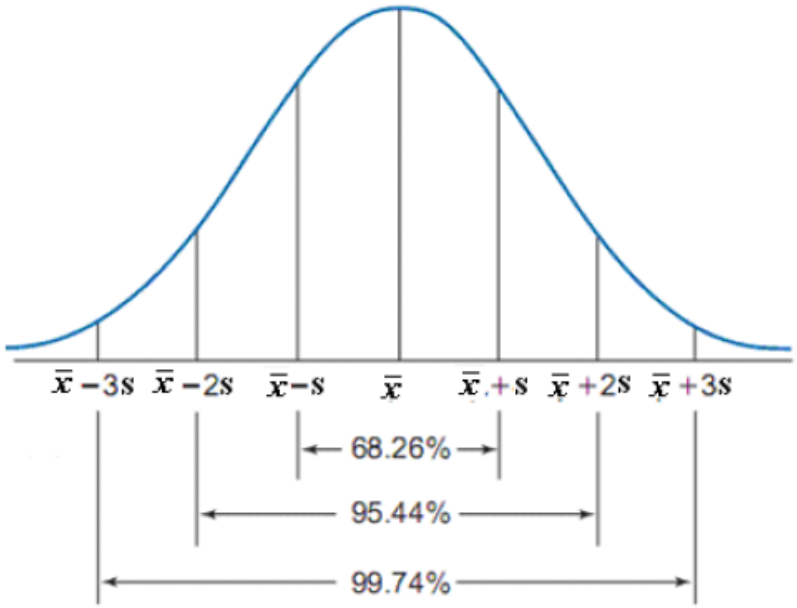
$$(\bar{x} - S \quad , \quad \bar{x} + S)$$

- Περίπου το 95% των παρατηρήσεων βρίσκεται στο διάστημα :

$$(\bar{x} - 2*S \quad , \quad \bar{x} + 2*S)$$

- Περίπου το 99% των παρατηρήσεων βρίσκεται στο διάστημα :

$$(\bar{x} - 3*S \quad , \quad \bar{x} + 3*S)$$



ΜΕΤΡΑ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ

- Ο συντελεστής μεταβλητότητας (CV) είναι ο λόγος της τυπικής απόκλισης προς τη μέση τιμή

$$CV = \frac{s}{\bar{x}}$$

- Μπορεί να χρησιμοποιηθεί για συγκρίσεις ομάδων τιμών οι οποίες εκφράζονται σε διαφορετικές μονάδες ή έχουν διαφορετικές μέσες τιμές
- Ένα δείγμα χαρακτηρίζεται ομοιογενές εάν CV $\leq 10\%$

ΛΟΞΟΤΗΤΑ

- Η κατανομή ενός πληθυσμού μπορεί να είναι συμμετρική ή ασύμμετρη
- Ασύμμετρη: η κορυφή χωρίζει την κατανομή σε δύο μέρη, τα οποία δεν περιέχουν ίσο αριθμό παρατηρήσεων
- **Θετική ασυμμετρία**: Συγκέντρωση τιμών αριστερά της μέσης τιμής και ακραίες τιμές (ουρά) προς τα δεξιά της κορυφής της καμπύλης συχνοτήτων
- **Αρνητική ασυμμετρία**: Συγκέντρωση των παρατηρήσεων δεξιά της μέσης τιμής και ακραίες τιμές (ουρά) προς τα αριστερά της κορυφής της καμπύλης συχνοτήτων

ΛΟΞΟΤΗΤΑ

'Εστω $x_1, x_2, \ldots x_n$ n παρατηρήσεις. Τότε,

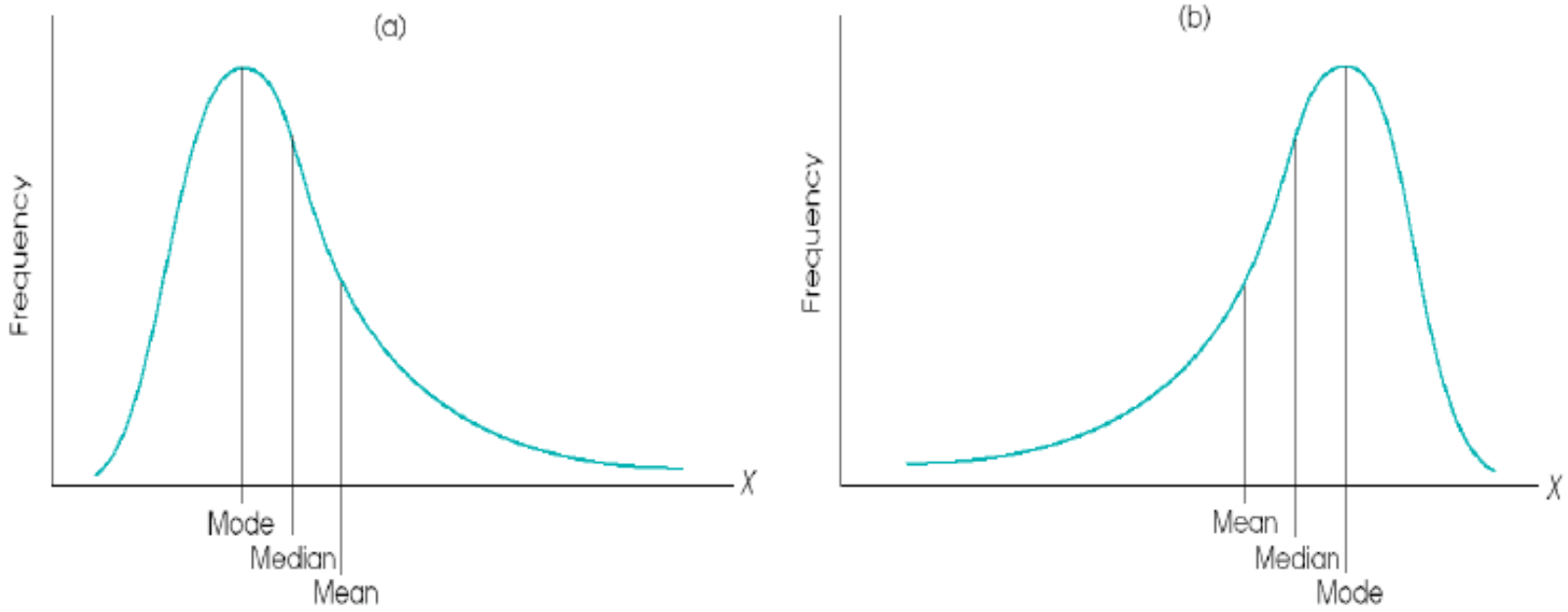
$$\text{Συντελεστής} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sum_{i=1}^{n}(x_i - \bar{x})^2}^{3/2}$$

Συντελεστής λοξότητας=0, συμμετρική

Συντελεστής λοξότητας >0, θετικώς λοξή

Συντελεστής λοξότητας <0, αρνητικώς λοξή

ΛΟΞΟΤΗΤΑ


(a)
Frequency
X
Mode
Median
Mean
(b)
Frequency
X
Mean
Median
Mode

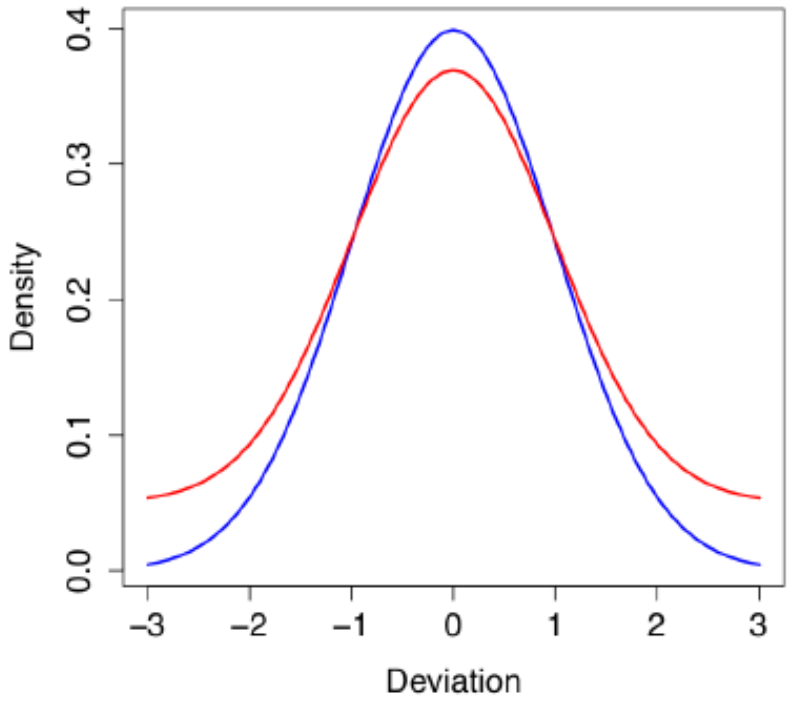
ΚΥΡΤΟΤΗΤΑ

Έστω, $x_2, \ldots x_n$ η παρατηρήσεις ενός δείγματος. Τότε,

$$\text{Συντελεστής κύρτωσης} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} - 3$$

$$\beta_2 = \begin{cases} > & 3 & \lambda\varepsilon\pi\tau o\kappa\upsilon\rho\tau\eta \\ = & 3 & \mu\varepsilon\sigma o\kappa\upsilon\rho\tau\eta \\ < & 3 & \pi\lambda\alpha\tau\upsilon\kappa\upsilon\rho\tau\eta \end{cases}$$

ΚΥΡΤΟΤΗΤΑ



- Ο συντελεστής κύρτωσης αποτελεί ένα μέτρο αποτύπωσης της κορυφής της κατανομής και της διασποράς
- Η τυπική κανονική κατανομή (μπλε γραμμή: $\mu = 0$; $\sigma = 1$) έχει συντελεστή κύρτωσης 0
- Η καμπύλη συχνοτήτων με την κόκκινη γραμμή έχει συντελεστή κύρτωσης <0 με χαμηλότερη κορυφή σε σχέση με τα άκρα της κατανομής