Ανάλυση δεδομένων στο περιβάλλον του SPSS

Λαβίδας Κωνσταντίνος
Μαθηματικός
lavidas@upatras.gr

Παρουσίαση των τιμών μιας μεταβλητής:

1. Κατανομή συχνοτήτων μια μεταβλητής
2. Αριθμητικά και Περιγραφικά Μέτρα για την παρουσίαση των τιμών μιας ποσοτικής μεταβλητής
3. Μέτρα Μορφής

1. Κατανομή συχνοτήτων μιας μεταβλητής

Κατανομή συχνοτήτων

- Απλός και αποτελεσματικός τρόπος περιγραφής της εικόνας των τιμών μιας μεταβλητής, είναι η παρουσίαση της συχνότητας εμφάνισης της κάθε τιμής (κατηγορίας) στα δεδομένα.
- Μπορεί να χρησιμοποιηθεί για την περιγραφή των δεδομένων οποιασδήποτε μεταβλητής.
  - Για τις ποσοτικές μεταβλητές προτείνεται η ομαδοποίηση των δεδομένων
- Η σειρά με την οποία κατατάσσονται και παρουσιάζονται οι κατηγορίες της μεταβλητής στην κατανομή συχνοτήτων είναι σημαντική μόνο σε ιεραρχικές κλίμακες:
  - Σε μεταβλητές με ιεραρχικές κλίμακες, οι τιμές πρέπει να παρουσιάζονται σύμφωνα με την ιεράρχηση τους.

Κατανομή ποιοτικών δεδομένων

q1_tr φύλο

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Άντρας	8	21,1	21,1	21,1
	Γυναίκα	30	78,9	78,9	100,0
	Total	38	100,0	100,0	

q10_tr βαθμό θεωρείτε ότι έχετε διδαχθεί παιδαγωγικά μαθήματα

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Καθόλου	5	13,2	13,2	13,2
	Λίγο	6	15,8	15,8	28,9
	Μέτρια	6	15,8	15,8	44,7
	Πολύ	13	34,2	34,2	78,9
	Πάρα πολύ	8	21,1	21,1	100,0
	Total	38	100,0	100,0	

Οι στήλες στον πίνακα κατανομής συχνοτήτων

- Η δεύτερη στήλη παρουσιάζει τις απόλυτες συχνότητες (**frequency**). Συμβολίζονται με το (**fi**). Το άθροισμα των απολύτων συχνοτήτων (Σfi) είναι ίσο με το μέγεθος του δείγματος (N).
- Η τρίτη στήλη παρουσιάζει τις συχνότητες υπό μορφή ποσοστών (**percent**) (σχετικές συχνότητες %), δηλαδή τις συχνότητες σε σχέση με το σύνολο. Το ποσοστό συμβολίζεται με (**Pi**).
  - $\mathbf{P}i = \frac{fi}{N} 100$
  - Η τέταρτη στήλη (**valid percent**) παρουσιάζει τα ποσοστά των απαντήσεων αποκλείοντας τις χαμένες τιμές.
    - Χαμένες τιμές (**missing values**), καταγράφονται στις περιπτώσεις εκείνες που δεν ήταν δυνατή η μέτρηση.
      - Για παράδειγμα σε μια ερώτηση του ερωτηματολογίου ο ερωτώμενος δεν απάντησε
- Οι επόμενη στήλη περιέχει, τα αθροιστικά εκατοστιαία ποσοστά (**cumulative percent**) (αθροιστικές σχετικές συχνότητες %). Αυτά εκφράζουν το ποσοστό των παρατηρήσεων που είναι μικρότερα ή ίσα με την τιμή (κατηγορία) που βρίσκεται στη γραμμή αυτή.
  - υπολογίζονται προσθέτοντας τα ποσοστά της γραμμής αυτής και τα προηγούμενα
  - δεν έχει νόημα σε κατηγορικές μεταβλητές που δεν διατάσσονται. Πχ. Το επάγγελμα ή το φύλο.

Κατανομή του φύλου των ερωτηθέντων

	Συχνότητα	Ποσοστό
Άντρας	8	21,1
Γυναίκα	30	78,9
Συνολικά	38	100,0

Κατανομή των απαντήσεων των ερωτηθέντων σχετικά με το βαθμό που θεωρούν ότι είναι ικανοποιημένοι από την επαφή που είχαν με τα παιδαγωγικά μαθήματα κατά την διάρκεια των σπουδών τους

	Frequency	Percent	Valid Percent	Cumulative Percent
Καθόλου	5	13,2	13,2	13,2
Λίγο	6	15,8	15,8	28,9
Μέτρια	6	15,8	15,8	44,7
Πολύ	13	34,2	34,2	78,9
Πάρα πολύ	8	21,1	21,1	100,0
Συνολικά	38	100,0	100,0	

Παρουσίαση πίνακα

- Συνήθως αναφερόμαστε στο σύνολο των υποκειμένων της έρευνας και μετά παρουσιάζουμε την κατανομή αξιοποιώντας τα αντίστοιχα ποσοστά των τιμών της μεταβλητής.
- Αν το σύνολο των χαμένων τιμών είναι μικρότερο από 5%, για τα ποσοστά αξιοποιούμε την τέταρτη στήλη (**valid percent**).
  - Προσοχή
    - σε διαφορετική περίπτωση (πολλές χαμένες τιμές) αναφερόμαστε στη τρίτη στήλη (**percent**) και παρουσιάζουμε επίσης και το ποσοστό των χαμένων τιμών.
    - στην περίπτωση αυτή θα πρέπει να μας προβληματίζει το μέγεθος των χαμένων τιμών και να διερευνηθούν ενδεχομένως και πιθανές εξηγήσεις.

Παράδειγμα πίνακα κατανομής συχνοτήτων ποσοτικής μεταβλητής

- Επίδοση μαθητών στα μαθηματικά

10,0	10,5	11,0	11,2	12,1	12,1	12,2	12,3
12,5	12,6	12,7	12,8	12,9	12,9	13,0	13,0
13,0	13,2	13,3	13,4	13,5	13,6	14,8	15,0
16,7	17,0	17,3	17,6	18,0	18,3	18,4	18,4
18,5	18,9	19,0	19,0	19,1	19,6	19,7	19,8

- Ας σκεφτούμε πως θα παρουσιάσουμε την κατανομή των τιμών, χρησιμοποιώντας πίνακα κατανομής συχνοτήτων;

Κάποιες διαπιστώσεις

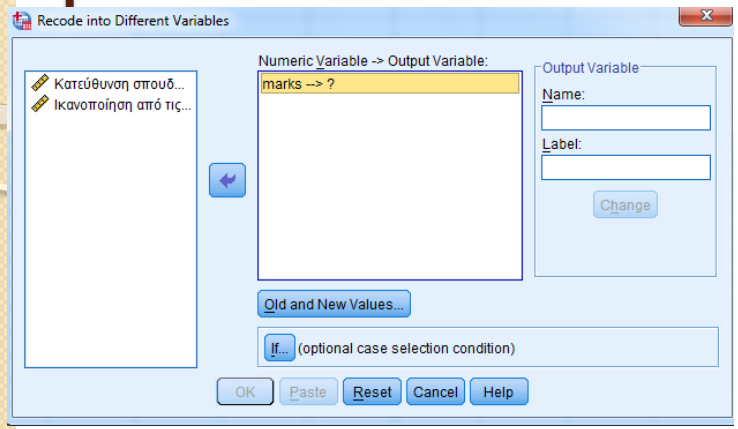
- Συνεχής ποσοτική μεταβλητή
  - είναι δυνατό να παίρνει οποιαδήποτε τιμή μεταξύ δύο ακραίων τιμών μιας δεδομένης κλίμακας.
- Δεν φαίνονται πολλές επαναλήψεις των τιμών της

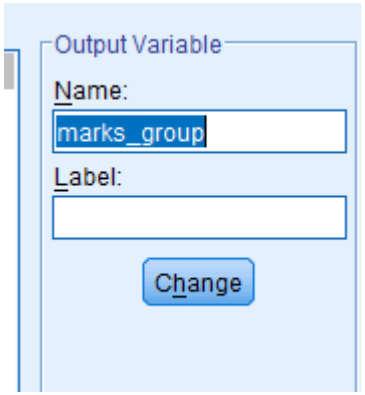
Πρόταση;

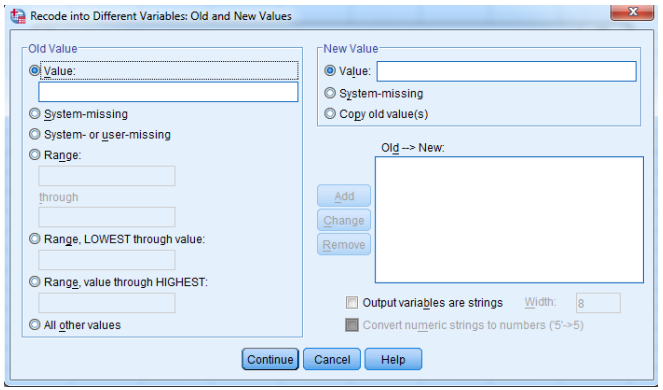
Ενδεικτική λύση στο πρόβλημα

- Ας κάνουμε αυτό που γίνεται στο σχολείο: ομαδοποίηση των βαθμών
- **Βαθμολογική Κλίμακα**
- Η βαθμολογική κλίμακα με βάση την οποία υπολογίζονται οι βαθμοί επίδοσης των μαθητών σε όλα τα μαθήματα, είναι 0 - 20 και προσδιορίζεται λεκτικά με τους παρακάτω χαρακτηρισμούς:
  - Κακώς 0 - 5.
  - Ανεπαρκώς 5,1 - 9,4.
  - Σχεδόν καλώς 9,5 - 13.
  - Καλώς 13,1 - 16.
  - Λίαν καλώς 16,1 - 18.
  - Άριστα 18,1 - 20.

Ομαδοποίηση τιμών μεταβλητής με το SPSS



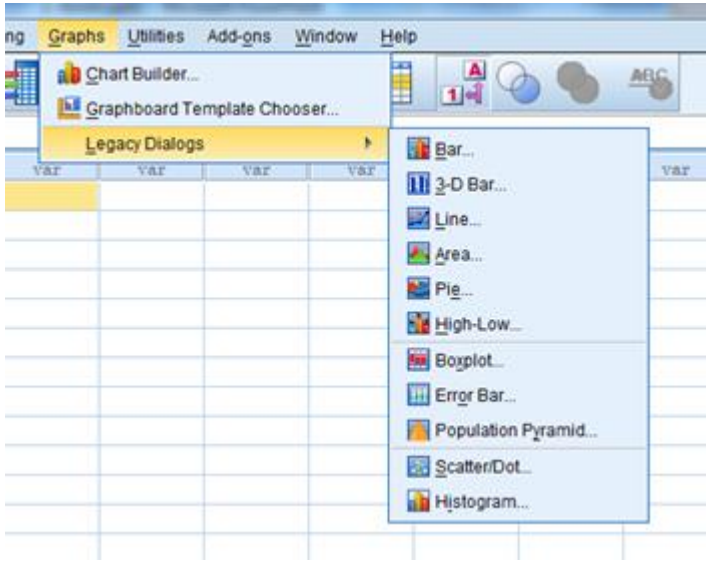


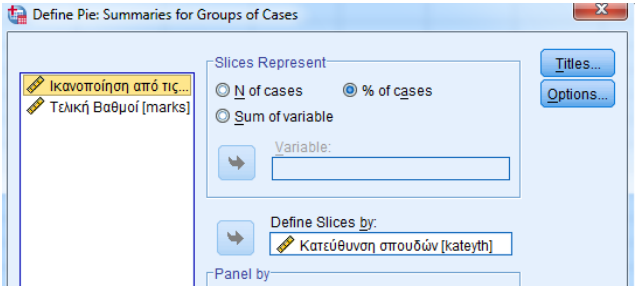


- Κατασκευάζω μια καινούργια μεταβλητή
  - Εντολές: Transform – Recode into different variable

- Εισάγω τις λεκτικές περιγραφές των τιμών
  - Σχεδόν καλώς: 9,5 – 13, Καλώς: 13,1 – 16, Λίαν καλώς: 16,1 – 18 και Άριστα: 18,1 - 20.
  - Αποθηκεύω
  - Εκτελώ την ανάλυση:
    - Analyze -Descriptive Statistic- frequencies
    - Μεταβλητή : «marks….» (καινούργια μεταβλητή)

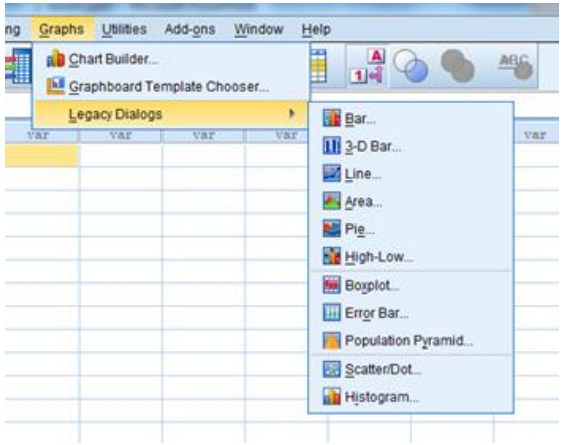
Γραφήματα παρουσίασης κατανομής ποιοτικών μεταβλητών (Ι)

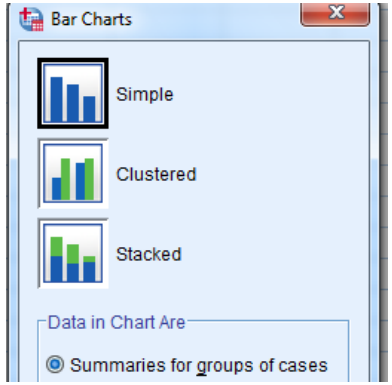


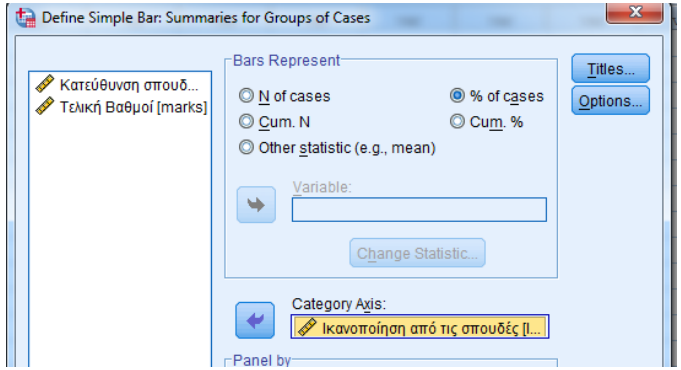


- Κυκλικό διάγραμμα (πίτα) – **Pie**
- **Πίτα για κατηγορικές μεταβλητές**: Graphs-Legacy Dialogs-Pie-Summaries for groups of cases - Define Slices by: μεταβλητή

Γραφήματα παρουσίασης κατανομής ποιοτικών μεταβλητών (2)



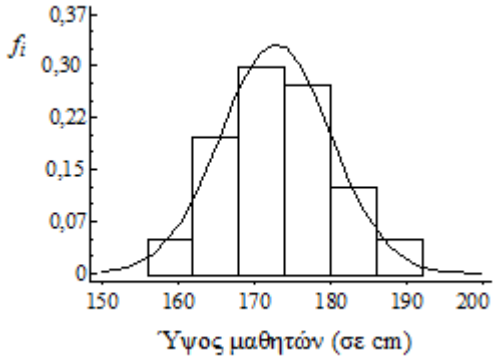




- Ραβδόγραμμα –Bar
- **Ραβδόγραμμα και για μεταβλητές διάταξης**: Graphs-Legacy Dialogs-Bar-Simple- Summaries for groups of cases – Category Axis: μεταβλητή

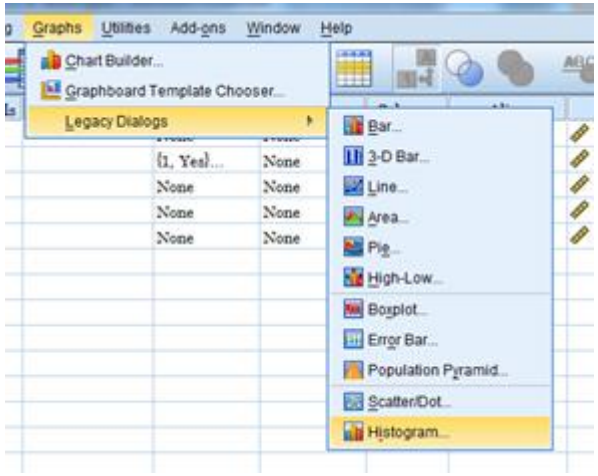
Καμπύλες Συχνοτήτων

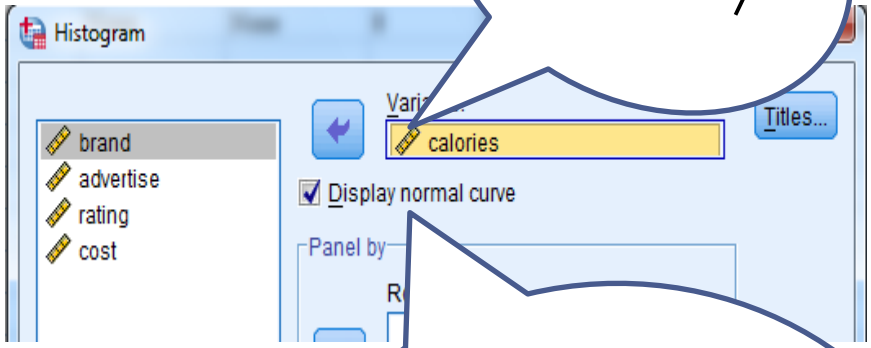
- Εάν ο αριθμός των κλάσεων για μια συνεχή μεταβλητή είναι αρκετά μεγάλος (θεωρητικά) τείνει στο άπειρο) και το πλάτος των κλάσεων είναι αρκετά μικρό (θεωρητικά τείνει στο μηδέν), τότε η πολυγωνική γραμμή συχνοτήτων τείνει να πάρει τη μορφή μιας ομαλής καμπύλης, η οποία ονομάζεται καμπύλη συχνοτήτων (frequency curve).
- Οι καμπύλες συχνοτήτων έχουν μεγάλη εφαρμογή στη Στατιστική: οι ιδιότητες τους μπορούν να χρησιμοποιηθούν για την εξαγωγή χρήσιμων συμπερασμάτων.
    - Η μορφή μιας κατανομής συχνοτήτων εξαρτάται από το πώς είναι κατανεμημένες οι παρατηρήσεις σε όλη την έκταση του εύρους τους.



Κατασκευή ιστογράμματος και καμπύλης (curve) συχνοτήτων με το SPSS

- Graphs- Legacy Dialogs- Histog



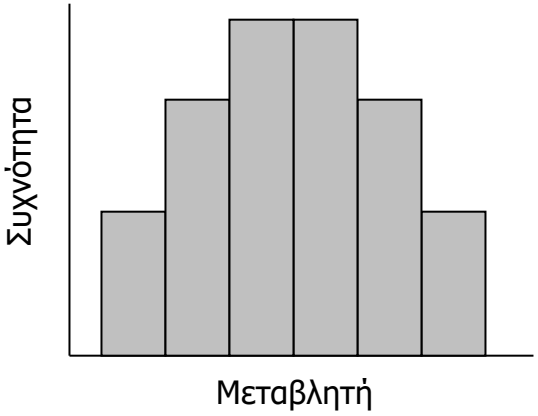


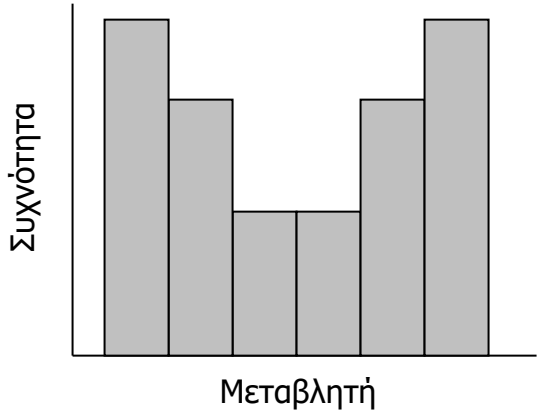
Μεταβλητή

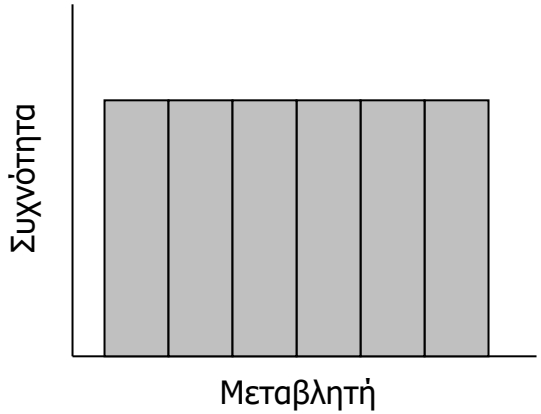
Για να εμφανιστεί η καμπύλη που αντιστοιχεί στην κατανομή των τιμών της μεταβλητής

Συμμετρικά Ιστογράμματα

Συμμετρικό: η κάθετη γραμμή στο κέντρο του ιστογράμματος, το χωρίζει ακριβώς σε δύο ίδια (μορφή και μέγεθος)

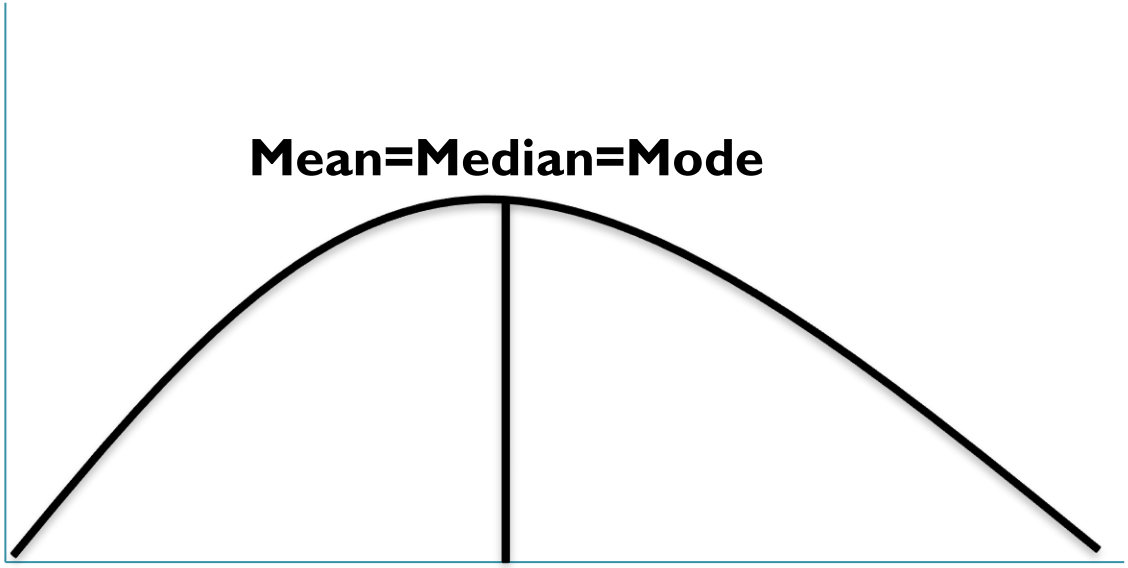






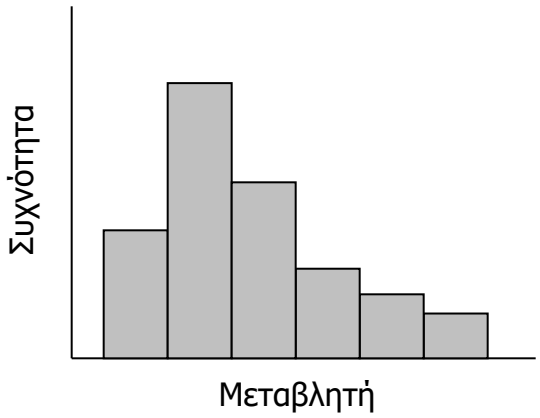
Συμμετρική κατανομή;

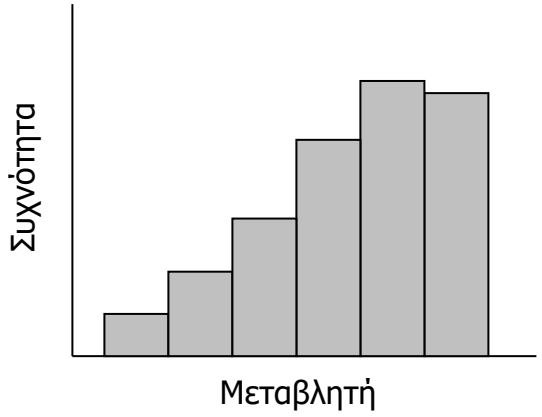
- Όταν η κατανομή των δεδομένων μας είναι συμμετρική, τότε οι τιμές και των τριών δεικτών είναι ίδιες. Προτιμάμε τον μέσο όρο



Μη συμμετρικά ιστογράμματα

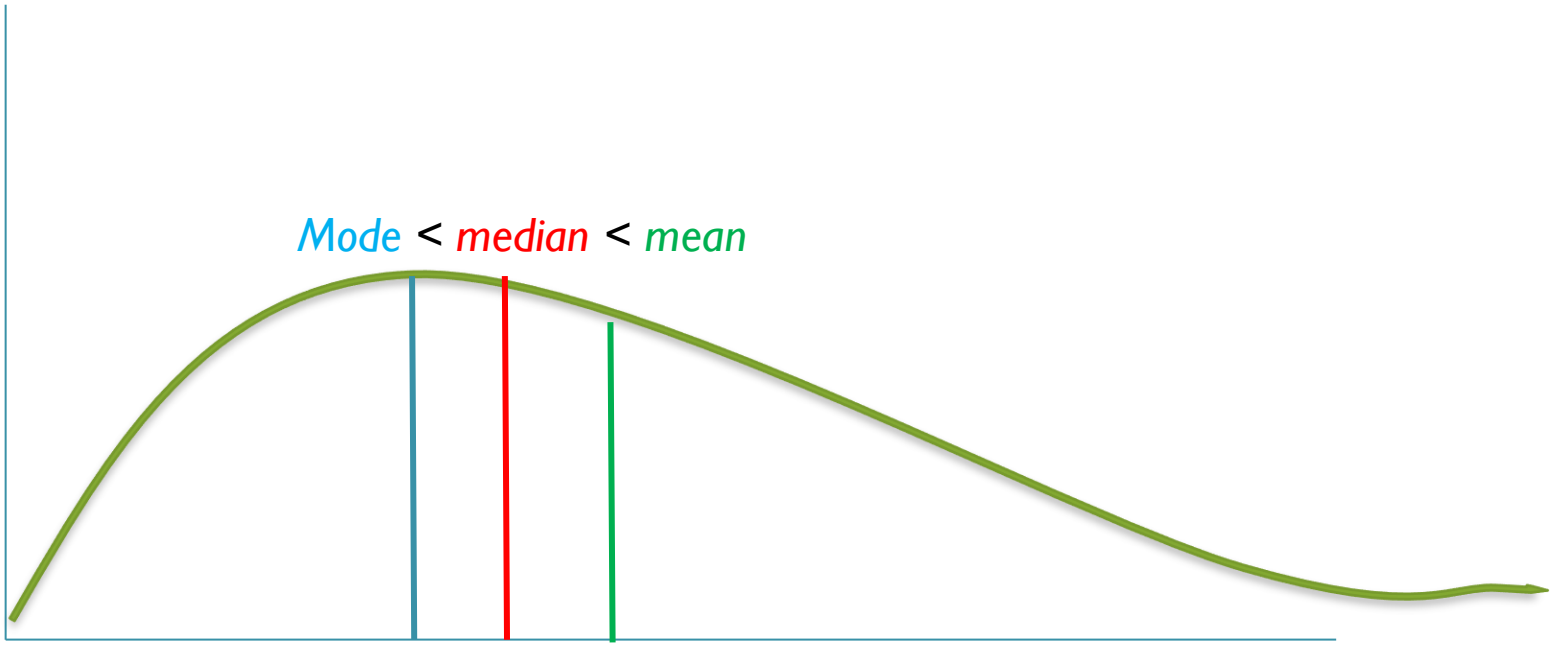
- Ασσύμετρο ή λοξό ιστόγραμμα: έχει μια εκτεταμένη μακριά ουρά προς τα δεξιά ή προς τα αριστερά .





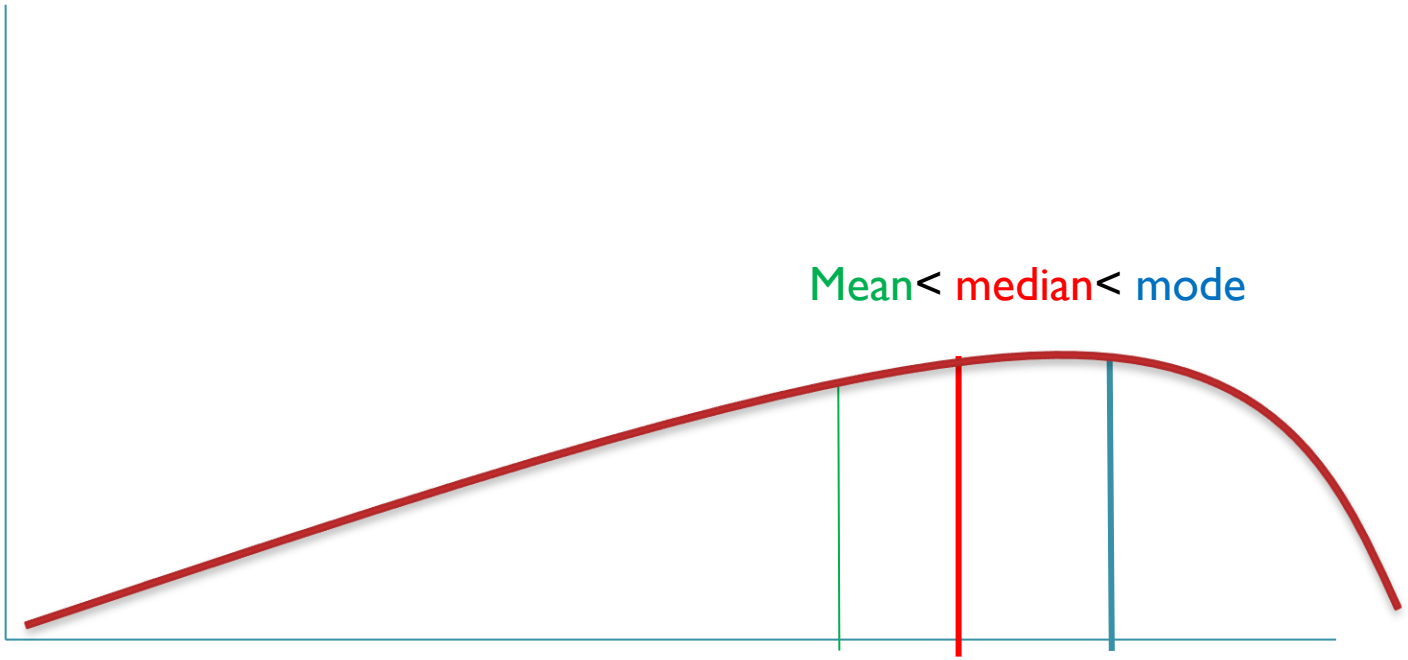
Θετικά ασύμμετρη –positive skew

- *ή ασύμμετρη δεξιά*
  - *Mode<median<mean*



Αρνητικά ασύμετρη –negative skew

- *ή ασύμμετρη αριστερά*
  - Mean< median< mode



2. Αριθμητικά και Περιγραφικά Μέτρα για την παρουσίαση των τιμών μιας ποσοτικής μεταβλητής

Μέτρα κεντρικής τάσης (Central Tendency) ή μέτρα θέσης

- Μια από τις ιδιότητες μιας ομάδας αριθμών είναι η κεντρική τάση.
  - Είναι η αριθμητική έκφραση της τάσης των τιμών για συγκέντρωση γύρω από μια τιμή της κλίμακας μέτρησής τους.
- Χρησιμοποιείται για να εκφράσει την **περιοχή μεγαλύτερης συγκέντρωσης των τιμών**, η οποία περιοχή κατά κάποιο τρόπο, αντιπροσωπεύει την ομάδα.
  - Μέση τιμή (Mean)
  - Διάμεσος (Median)
  - Επικρατούσα τιμή (Mode)

Δείκτες κεντρικής τάσης: Μέση τιμή

- **Μέσος Τιμή ή όρος:** Ορίζεται το άθροισμα των τιμών, διαιρουμένου από το πλήθος των τιμών της ομάδας.
- Υπολογίζεται σε ποσοτικές μεταβλητές
- Γενικά αν έχουμε Ν μετρήσεις ή τιμές που συμβολίζονται
- $X1, X2, ..., X_N$ ο μέσος όρος είναι:

π.χ. η μέση τιμή των τιμών, 8, 13, 9, 10, 20 είναι:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_v}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Μέση τιμή του δείγματος συμβολίζεται με: $\bar{x}$

Μέση τιμή του πληθυσμού συμβολίζεται με: μ

Δείκτες κεντρικής τάσης: Διάμεσος

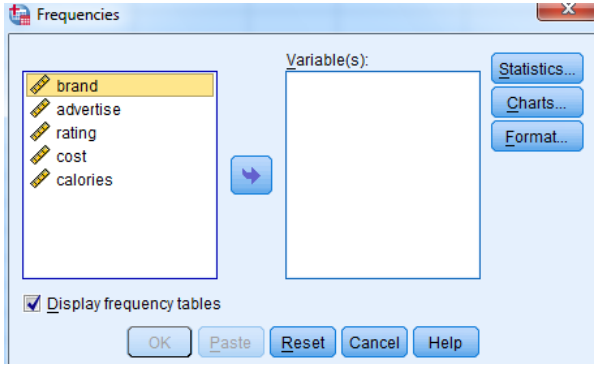
- **Διάμεσος:** Η διάμεσος είναι μια τιμή, τέτοια ώστε, ο αριθμός των παρατηρήσεων που είναι μεγαλύτερες απ' αυτήν, να είναι ίσος με τον αριθμό των παρατηρήσεων που είναι μικρότερές της. Το πολύ το 50% των παρατηρήσεων είναι πριν και μετά τη διάμεσο.
- Υπολογίζεται σε: Ποσοτικές μεταβλητές.
  - Και σε ποιοτικές μεταβλητές διάταξης
- Για να προσδιορίσουμε τη διάμεσο πρέπει να διατάξουμε τις τιμές της μεταβλητής σε αύξουσα σειρά.
  - Αν το πλήθος είναι άρτιο η διάμεσος είναι το ημιάθροισμα των δύο μεσαίων τιμών
  - Αν το πλήθος είναι περιττό η διάμεσος είναι η μεσαία τιμή.
- Παράδειγμα 1: Ας θεωρήσουμε τις παρακάτω τιμές της μεταβλητής Χ διατεταγμένες κατά αύξουσα σειρά : 5, 8, 12, 15, 20.
  - Διάμεσος είναι: 12

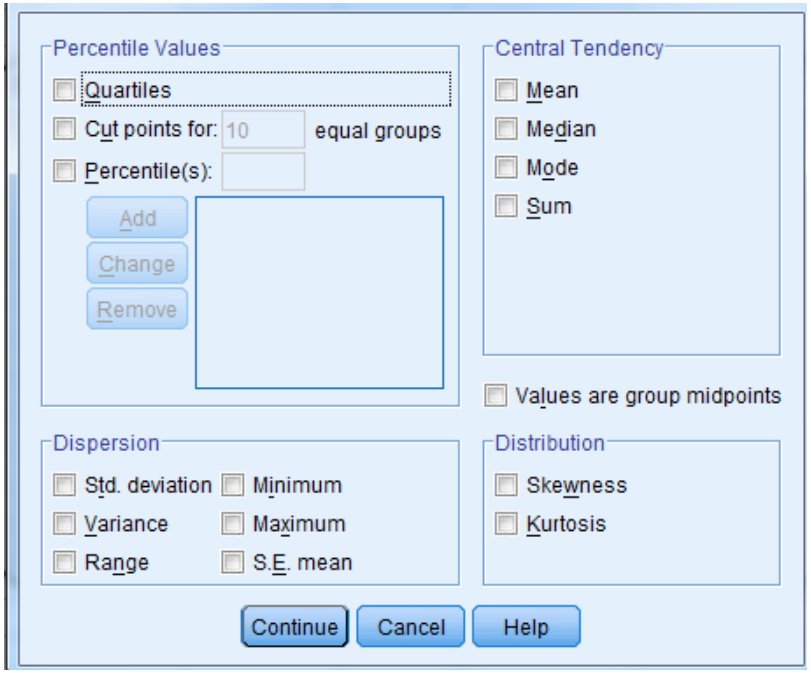
Δείκτες κεντρικής τάσης: Επικρατούσα τιμή

- **επικρατούσα ή δεσπόζουσα τιμή:** Είναι η συχνότερη τιμή (μεγαλύτερη συχνότητα).
- Υπολογίζεται σε: ποσοτικές και ποιοτικές μεταβλητές
- Παράδειγμα στα δεδομένα:
  - 7,7,7,8,8,8,8,10,10,10,10,10, 12,12,13.
    - Επικρατούσα τιμή είναι η 10

Υπολογισμός των Δεικτών Κεντρικής Τάσης με το SPSS

- Analyze - Descriptive Statistics - Frequencies
- Επιλογή Statistics





- Επιλέξτε τα Mean, Median και Mode.

Δείκτες διασποράς (Dispersion)

- Οι δείκτες κεντρικής τάσης δεν αρκούν για να περιγράψουν την εικόνα της κατανομής των τιμών.
  - Δεν απαντούν σε ερωτήματα, όπως: πόσο πολύ είναι οι παρατηρήσεις απλωμένες γύρω από κέντρο;
- Οι τιμές μιας μεταβλητής συνήθως βρίσκονται διασκορπισμένες μέσα σ' ένα διάστημα της κλίμακας, έχουν δηλαδή μια διακύμανση, που άλλες φορές είναι μικρή και άλλες μεγάλη.

Τα μέτρα θέσης δεν επαρκούν για την περιγραφή της κατανομής των τιμών μιας μεταβλητής.

- Παράδειγμα 1. Διαφοράς διακύμανσης
  - Στις παρακάτω ομάδες δεδομένων Α, Β Γ.
    - Α : 15,16,17,17,20
    - Β : 1, 13,17,17,37
    - Γ : 17,17,17,17,17
  - Είναι φανερή η διαφορά στην κατανομή των τιμών στις τρεις ομάδες.
  - Ας υπολογίσουμε τα μέτρα θέσης, οι ομάδες αυτές έχουν, μέση τιμή =…17….., διάμεσο=… 17 ….. και επικρατούσα τιμή=… 17 …..
  - Η διαφαινόμενη διαφορά στις κατανομές οφείλεται στην **διαφορετική διασπορά των τιμών** στις τρεις ομάδες;

Δείκτες μέτρησης της διασποράς:

- εύρος (range),
- ενδοτεταρτημοριακό εύρος (interquartile range),
- τυπική απόκλιση (standard deviation).

Εύρος (Range)

- Η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής της κατανομής
- Περιλαμβάνει και τις ακραίες τιμές της κατανομής
- Σε πολλές περιπτώσεις δεν παρουσιάζει μια αντιπροσωπευτική εικόνα της διασποράς της κατανομής
- Δεν μας λέει τίποτα για τη διασπορά των τιμών της κατανομής γύρω από το μέσο όρο
- Να υπολογίσουμε το εύρος της κατανομής των τιμών: 1, 2, 34, 56, 12

Ενδοτεταρτημοριακό εύρος (Interquartile Range)

- Το εύρος του μεσαίου τμήματος της κατανομής (50% των τιμών μιας κατανομής)
  - μετράει το άπλωμα των μεσαίων παρατηρήσεων.
- Προσδιορίζεται μέσω των τεταρτημορίων (**quartiles**): Τα σημεία που χωρίζουν την κατανομή σε τέσσερα ίσα μέρη:
  - 1ο τεταρτημόριο-**Q1**: κάτω από το οποίο βρίσκεται το πολύ το 25% των τιμών της μεταβλητής.
    - Το **Q1** είναι αντίστοιχα το 25-εκατοστιαίο σημείο ή εκατοστημόριο **P25**
  - 2ο τεταρτημόριο- **Q2**: κάτω από το οποίο βρίσκεται το πολύ το 50% των τιμών της μεταβλητής.
    - Το **Q2** είναι αντίστοιχα το **50**-εκατοστιαίο σημείο ή εκατοστημόριο **P50**
  - 3ο τεταρτημόριο-**Q3**: κάτω από το οποίο βρίσκεται το πολύ το 75% των τιμών της μεταβλητής.
    - Το **Q3** είναι αντίστοιχα το **7**5-εκατοστιαίο σημείο ή εκατοστημόριο **P75**
- **IQR= Q3-Q1**

Ενδοτεταρτομοριακό εύρος

- Μεγάλες τιμές IQR σημαίνει ότι το 1ο και 3ο τεταρτημόριο απέχουν υποδεικνύοντας υψηλό επίπεδο μεταβλητότητας.
- Είναι σχετικά εύκολο στον υπολογισμό του
- Είναι αντιπροσωπευτικό των κεντρικών τιμών της κατανομής
- Δεν λαμβάνει υπόψη τις ακραίες τιμές της κατανομής
- Δεν περιγράφει καμιά από τις παραμέτρους, οι οποίες είναι βασικές για την επαγωγική στατιστική.

Υπολογισμός ενδοτεταρτομοριακού εύρους

- Υπολογίζω το 1ο και 3ο τεταρτημόριο
  - Τοποθετώ τα δεδομένα κατά αύξουσα σειρά
  - Q1 τεταρτημόριο είναι η τιμή στη θέση $\frac{N+1}{4}$,
  - Q2 ή διάμεσος είναι η τιμή στη θέση $\frac{N+1}{2}$,
  - Q3 τεταρτημόριο είναι η τιμή στη θέση $\frac{3(N+1)}{4}$
- Παράδειγμα 1ο
  - 4, 4, 5, 5, 7, 8, 8, 8, 9, 10, 11, 12, 12
  - Q1=….., Q3=……., IQR=
- Παράδειγμα 2ο
  - 4, 4, 5, 5, 7, 8, 8, 8, 9, 10, 11, 12
  - Q1=….., Q3=……., IQR=

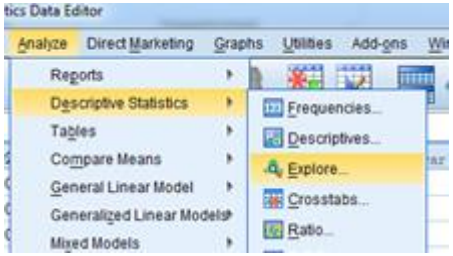
Τυπική απόκλιση (Standard Deviation)

- Τυπική απόκλιση και διακύμανση δύο σχεδόν ταυτόσημες έννοιες που περιγράφουν την διασπορά
- Ορίζουμε ως διακύμανση ενός πληθυσμού N τιμών με μέση τιμή μ τη μέση τετραγωνική απόκλιση- απόσταση των N μετρήσεων από τη μέση τιμή μ του πληθυσμού
  - $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
- Τυπική απόκλιση: είναι η τετραγωνική ρίζα της διασποράς-διακύμανσης: $\sigma = \sqrt{\sigma^2}$.
  - Ίδια μονάδα μέτρησης με τις τιμές.
- Την τυπική απόκλιση του δείγματος την συμβολίζουμε με **s**
- Μέση απόσταση των τιμών από τη μέση τιμή
  - Μπορεί να χρησιμοποιηθεί για τον υπολογισμό των παραμέτρων του πληθυσμού
  - Λαμβάνει υπόψη όλες τις τιμές της κατανομής
  - Είναι ο πιο ευαίσθητος από τους δείκτες διασποράς
  - Ο υπολογισμός της είναι σχετικά πιο περίπλοκος σε σχέση με τους υπόλοιπους δείκτες διασποράς
  - Είναι πολύ ευαίσθητη στις ακραίες τιμές της κατανομής

Διακύμανση του δείγματος ως μέτρο υπολογισμού παραμέτρων του πληθυσμού

- Ο συνηθέστερος τρόπος περιγραφής της διασποράς των τιμών μιας μεταβλητής είναι μέσω της τυπικής απόκλισης.
- Ο σημαντικότερος λόγος για τον οποίο προτιμάται η τυπική απόκλιση από τους υπόλοιπους δείκτες διασποράς είναι η δυνατότητα που προσφέρει να υπολογίσουμε παραμέτρους του πληθυσμού.

Υπολογισμός των Δεικτών Διασποράς με το SPSS



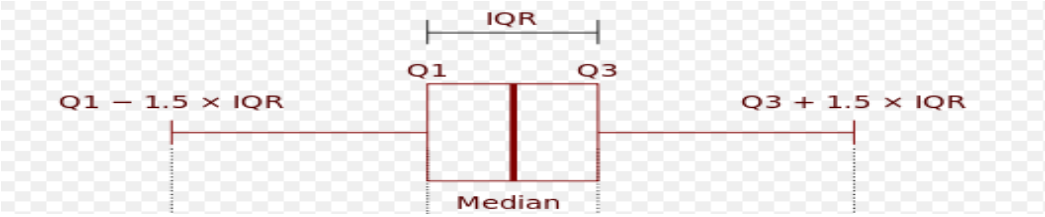
- Analyze - Descriptive Statistics - Explore

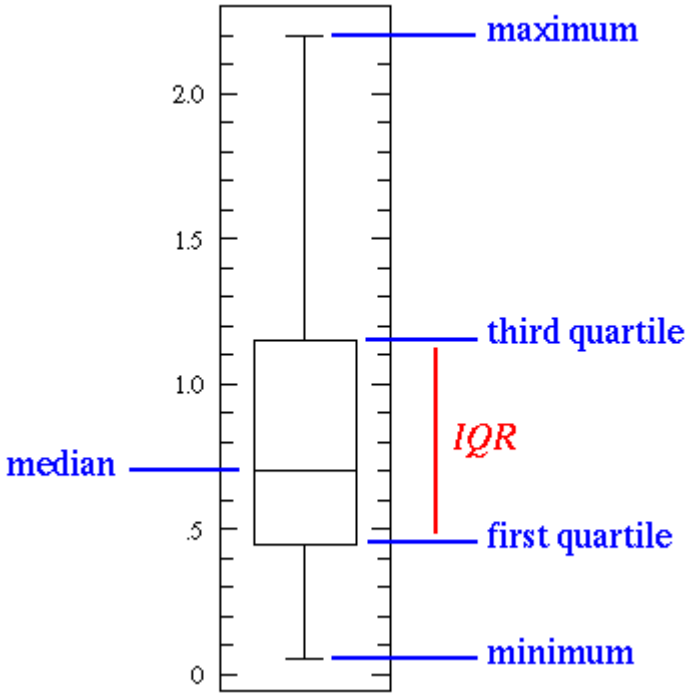
			Statistic	Std. Error
Ονομα Μεταβλητής	Mean		7,5833	,77321
	95% Confidence Interval for Mean	Lower Bound	5,8815	
		Upper Bound	9,2852	
	5% Trimmed Mean		7,5370	
	Median		8,0000	
	Variance		7,174	
	Std. Deviation		2,67848	
	Minimum		4,00	
	Maximum		12,00	
	Range		8,00	
	Interquartile Range		4,75	
	Skewness		,094	,637
	Kurtosis		-1,052	1,232

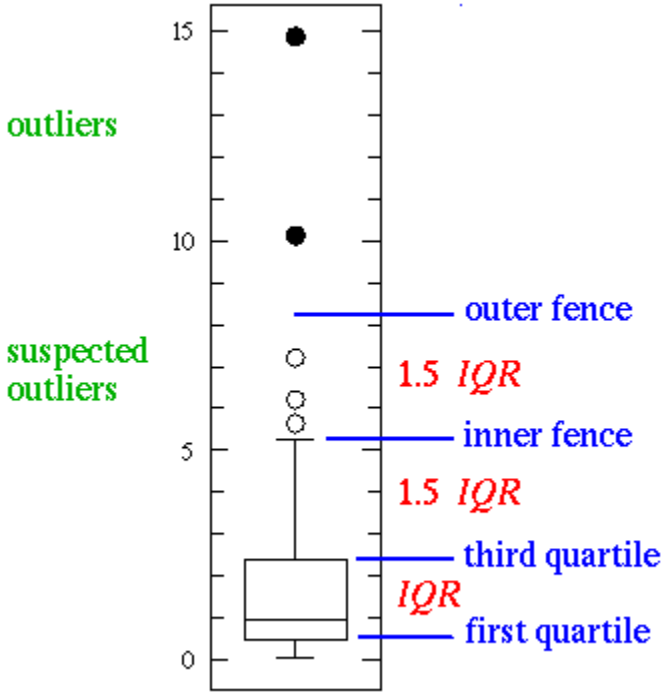
Θηκόγραμμα (boxplot) -Ι

- Το θηκόγραμμα ή θηκόγραμμα με απολήξεις (box-and whisker plot) είναι η γραφική παράσταση της διασποράς και των ακραίων τιμών ενός δείγματος
  - Αποτελείται από ένα ορθογώνιο με βάσεις το πρώτο και το τρίτο τεταρτημόριο
    - Ενδιάμεσα τοποθετείται η διάμεσος.
    - Από τα μέσα των βάσεων αναπτύσσονται γραμμές οι οποίες συνδέουν τις οριακές τιμές της μεταβλητής.

Θηκόγραμμα (boxplot) -2





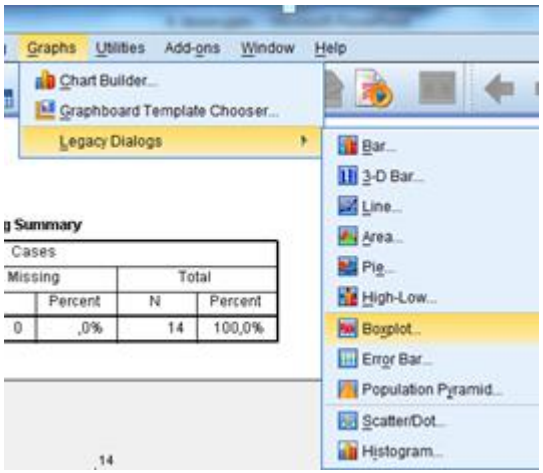


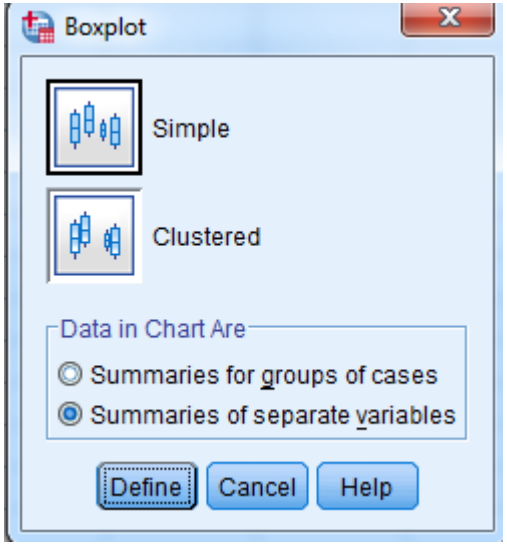
Ακραίες (extreme) ή παράτυπες (outliers) παρατηρήσεις

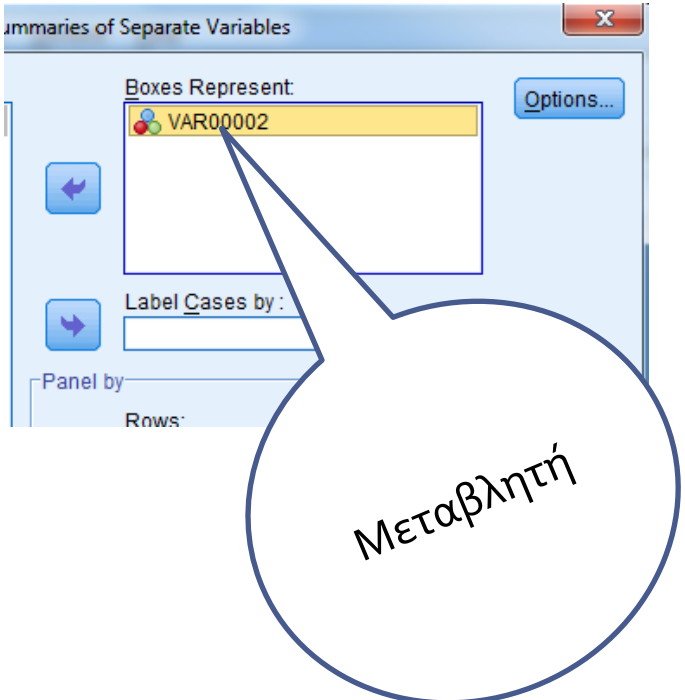
- Οι παρατηρήσεις που απέχουν από 1,5 μέχρι 3 φορές IQR πάνω και κάτω από το κουτί ονομάζονται ακραίες τιμές και συμβολίζονται με κύκλο (o).
  - Μεγάλη προσοχή στις ακραίες τιμές κατά την ανάλυση
- Οι παρατηρήσεις που απέχουν περισσότερες από 3 φορές IQR πάνω και κάτω από το κουτί ονομάζονται εξαιρετικά ακραίες ή παράτυπες τιμές και συμβολίζονται με αστεράκι (*).
  - Ίσως θα πρέπει να επαναληφθεί η ανάλυση χωρίς αυτές τις τιμές

Κατασκευή boxplot με το SPSS

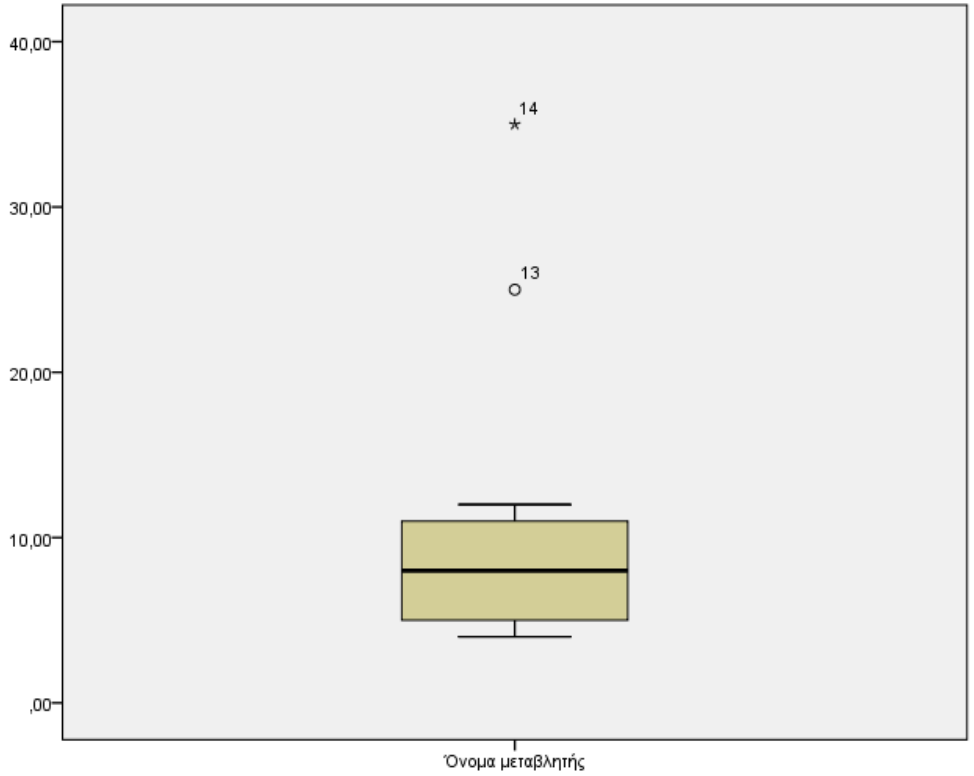
- Graphs- Legacy Dialogs- boxplot
- Simple – summaries of separate variables







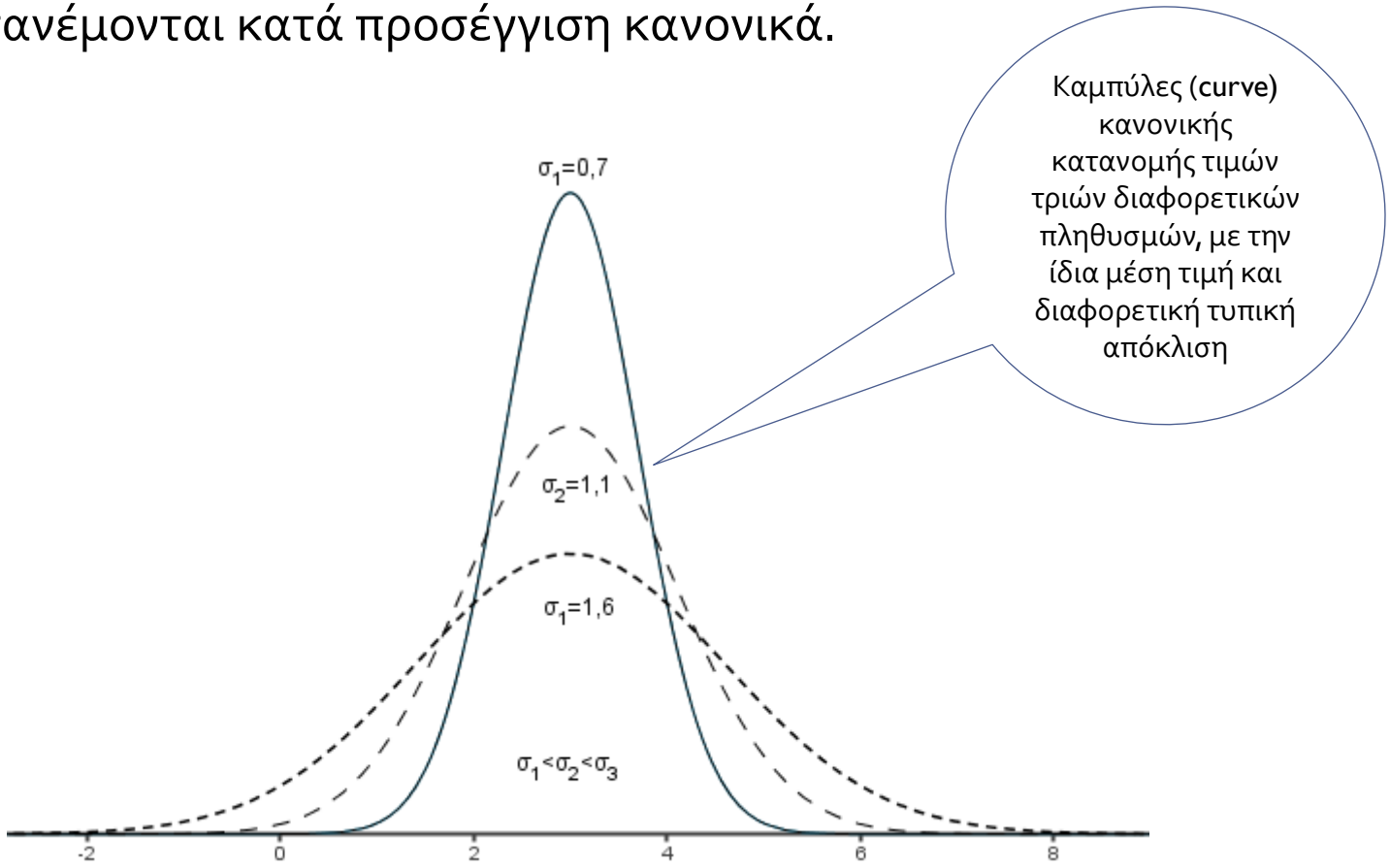
Παράδειγμα Boxplot



Ποιες τιμές είναι ακραίες ή παράτυπες και πρέπει να ελεγχθούν;

Κανονική Κατανομή

- Η κανονική κατανομή (normal distribution) είναι πολύ σημαντική αφού, πολλά φαινόμενα που συμβαίνουν γύρω μας κατανέμονται κατά προσέγγιση κανονικά.



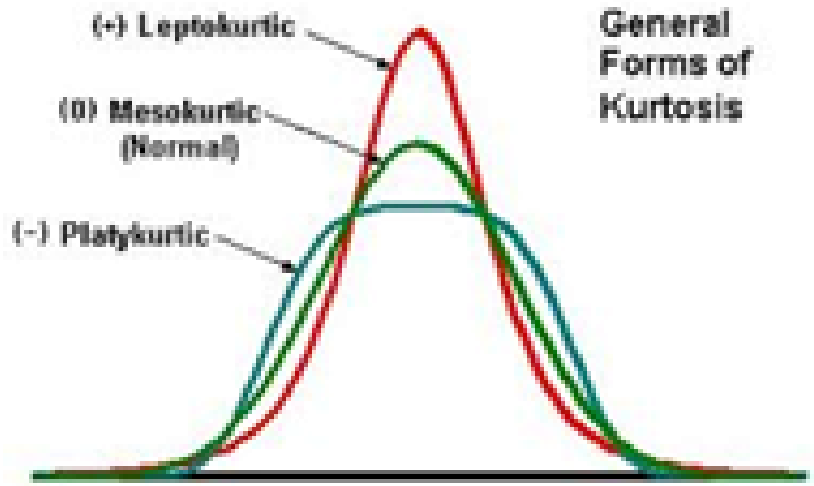
3. Μέτρα Μορφής

Μέτρα Μορφής

- Δύο συντελεστές που μας βοηθούν να κατανοούμε κατά πόσο αποκλίνει η καμπύλη των τιμών της μεταβλητής μας από την καμπύλη της κανονικής κατανομής, είναι οι συντελεστές:
  - κύρτωσης και
  - ασυμμετρίας.

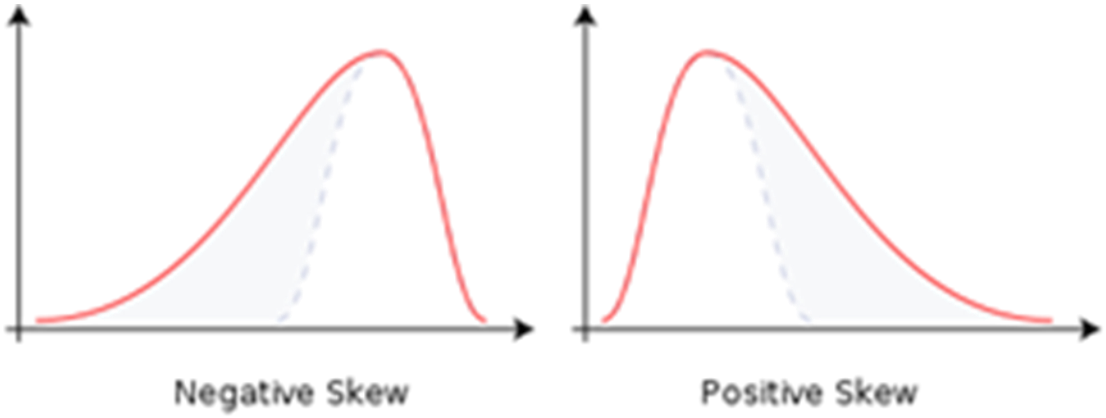
Συντελεστή **κύρτωσης** (kurtosis)

- Θετικός τότε τα δεδομένα είναι συγκεντρωμένα γύρω από μια κεντρική τιμή και η κατανομή λέγεται λεπτόκυρτη: *Προσοχή στις ακραίες τιμές.*
- Αρνητικός τότε οι τιμές είναι περισσότερο διάσπαρτες και η κατανομή λέγεται πλατύκυρτη.



Ο συντελεστής ασυμμετρίας (scewness):

- συντελεστής ασυμμετρίας:
  - αρνητικός, τότε τα δεδομένα είναι τραβηγμένα «αριστερά»
  - θετικός, τότε τα δεδομένα είναι «τραβηγμένα» δεξιά.



Έλεγχος της κανονικότητας μιας κατανομής

- Δεν υπάρχουν ασφαλείς μέθοδοι που εξασφαλίζουν τον έλεγχο της κανονικότητας. Κυρίως οι μέθοδοι μας βοηθούν να στηρίξουμε την συμμετρία, η οποία είναι προϋπόθεση της κανονικότητας:
- Ελέγχουμε κατά πόσο το ιστόγραμμα των τιμών της μεταβλητής είναι συμμετρικό.
  - Η αυτόματη εμφάνιση της καμπύλης της κανονικής κατανομής (**normal curve**) των τιμών της μεταβλητής: Γραφική παράσταση της εκθετικής συνάρτησης, λαμβάνοντας υπόψη τη μέση τιμή και την τυπική απόκλιση των τιμών.
  - Συμβάλει στον κατά προσέγγιση έλεγχο: αν και κατά πόσο η κατανομή των τιμών αποκλίνει από την καμπύλη της κανονική κατανομής.
- *Θα πρέπει οι συντελεστές ασσυμετρίας και κύρτωσης να είναι εντός [-1, 1]: λιγότερο αυστηρό*
- Θα πρέπει οι λόγοι ασσυμετρίας προς το αντίστοιχο τυπικό της σφάλμα και κυρτότητας προς το αντίστοιχο τυπικό της σφάλμα, είναι μεταξύ [-2, 2]
- Q-Q plot:
  - *Normal Q-Q Plot:* Τα σημεία πρέπει να βρίσκονται στην διχοτόμο της γωνίας
  - Detrended Normal Q-Q Plot: Τα σημεία εκατέρωθεν της ευθείας, πρέπει να μην σχηματίζουν κάποιο συγκεκριμένο πρότυπο.
- Έλεγχοι: **Kolmogorov – Smirnov, (Lillefors) και Shapiro – Wilk**. *Μη στατιστικά σημαντικό αποτέλεσμα.*