Is my data normal?

Είναι τα δεδομένα κανονικά?

Ζιντζαράς Ηλίας, M.Sc., Ph.D.

Καθηγητής Βιομαθηματικών-Βιομετρίας
Εργαστήριο Βιομαθηματικών
Τμήμα Ιατρικής
Πανεπιστήμιο Θεσσαλίας

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Θεόδωρος Μπρότσης, MSc, PhD Candidate
Ακαδημαϊκός Υπότροφος
(http://biomath.med.uth.gr)
Πανεπιστήμιο Θεσσαλίας
Email: tmprotsis@uth.gr

ΕΞΕΤΑΣΤΕ ΤΑ ΔΕΔΟΜΕΝΑ ΣΑΣ ΓΡΑΦΙΚΑ

… πριν ξεκινήσετε την ανάλυσή σας.

Γνωρίστε τα δεδομένα. Κοιτάξτε για μοτίβα,
πιθανά προβλήματα, σχέσεις, κ.λπ.

Γραφική εξερεύνηση δεδομένων

- Κάνοντας χρήση διαγραμμάτων μπορούμε να αντλήσουμε πολλές πληροφορίες για τα δεδομένα μας
- Τα δεδομένα μας μπορεί να είναι λοξά (skewed), κυρτά (kurtosis, πλατιές άκρες) ή να ακολουθούν μία κατανομή που δεν είναι κανονική
- Σε αυτήν την παρουσίαση θα συζητήσουμε τα ακόλουθα διαγράμματα για να εξακριβώσουμε αν τα δεδομένα μας είναι «κανονικά»:
  - Ιστογράμματα
  - Φυλλογράμματα (Stem and leaf)
  - Θηκογράμματα (Box Plots)
  - P-P διαγράμματα
  - Q-Q διαγράμματα

Η καμπύλη κανονικής κατανομής

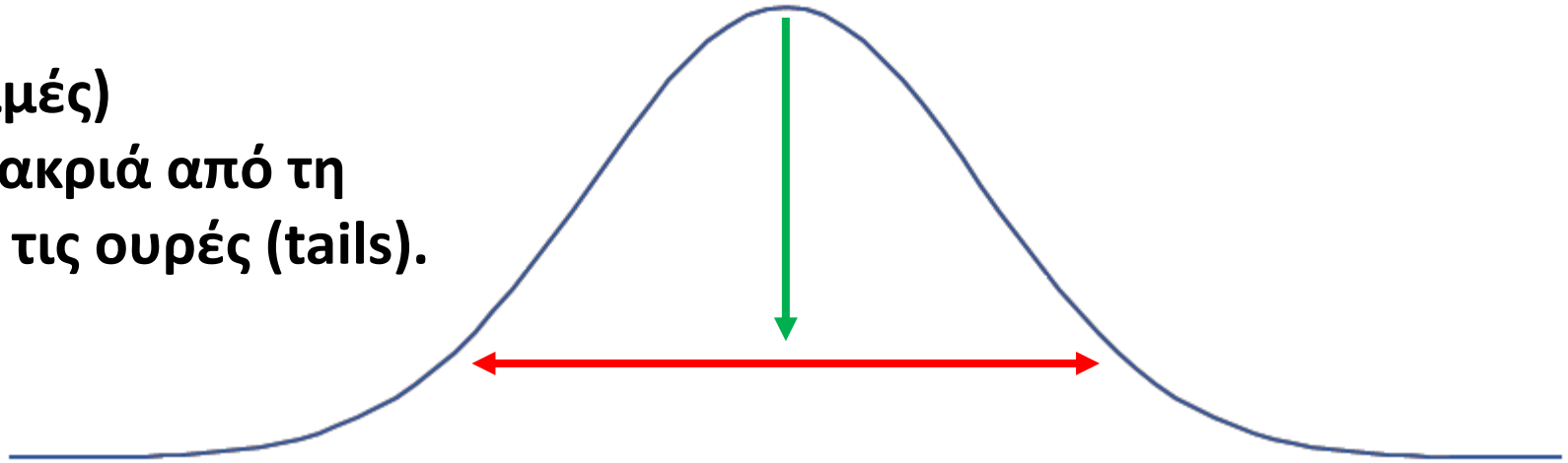
Πως μπορούμε να πούμε ότι τα δεδομένα μας ακολουθούν αυτό το σχήμα;

Πολλά στατιστικά τεστ προϋποθέτουν τα δεδομένα να ακολουθούν την κανονική κατανομή

Κυρτότητα

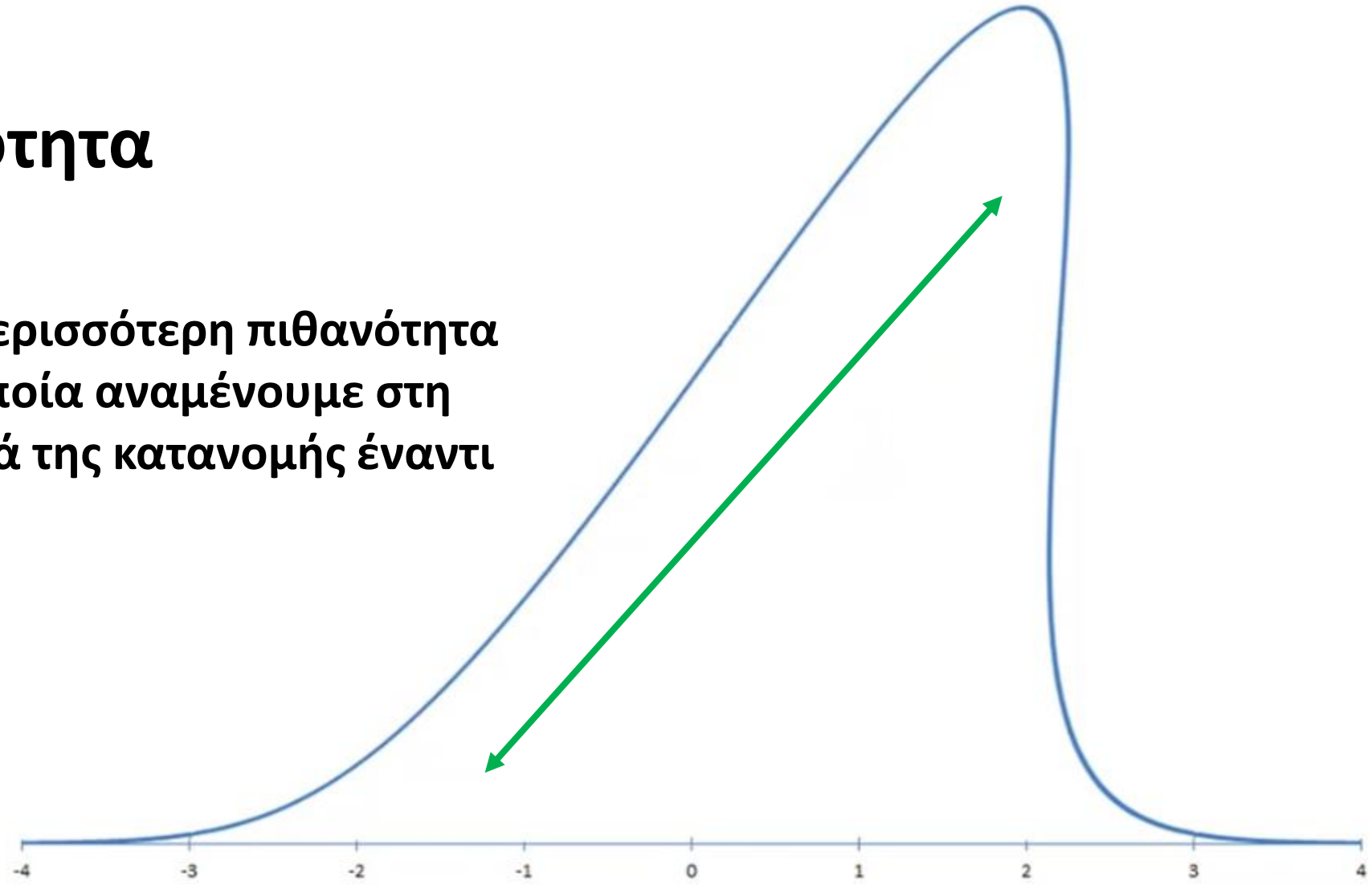
Περισσότερη πιθανότητα από την οποία αναμένουμε στις ουρές (tails) της κατανομής λόγω της ύπαρξης ακραίων τιμών μακριά από τη μέση τιμή.

Πιθανότητες (τιμές) σπρώχνονται μακριά από τη μέση τιμή προς τις ουρές (tails).



Λοξότητα

Υπάρχει περισσότερη πιθανότητα από την οποία αναμένουμε στη μία πλευρά της κατανομής έναντι της άλλης.



Άλλες κατανομές πιθανοτήτων

Πολλές φορές τα δεδομένα ακολουθούν άλλο τύπο κατανομής καλύτερα:

Exponential

Lognormal

μεταξύ άλλων…

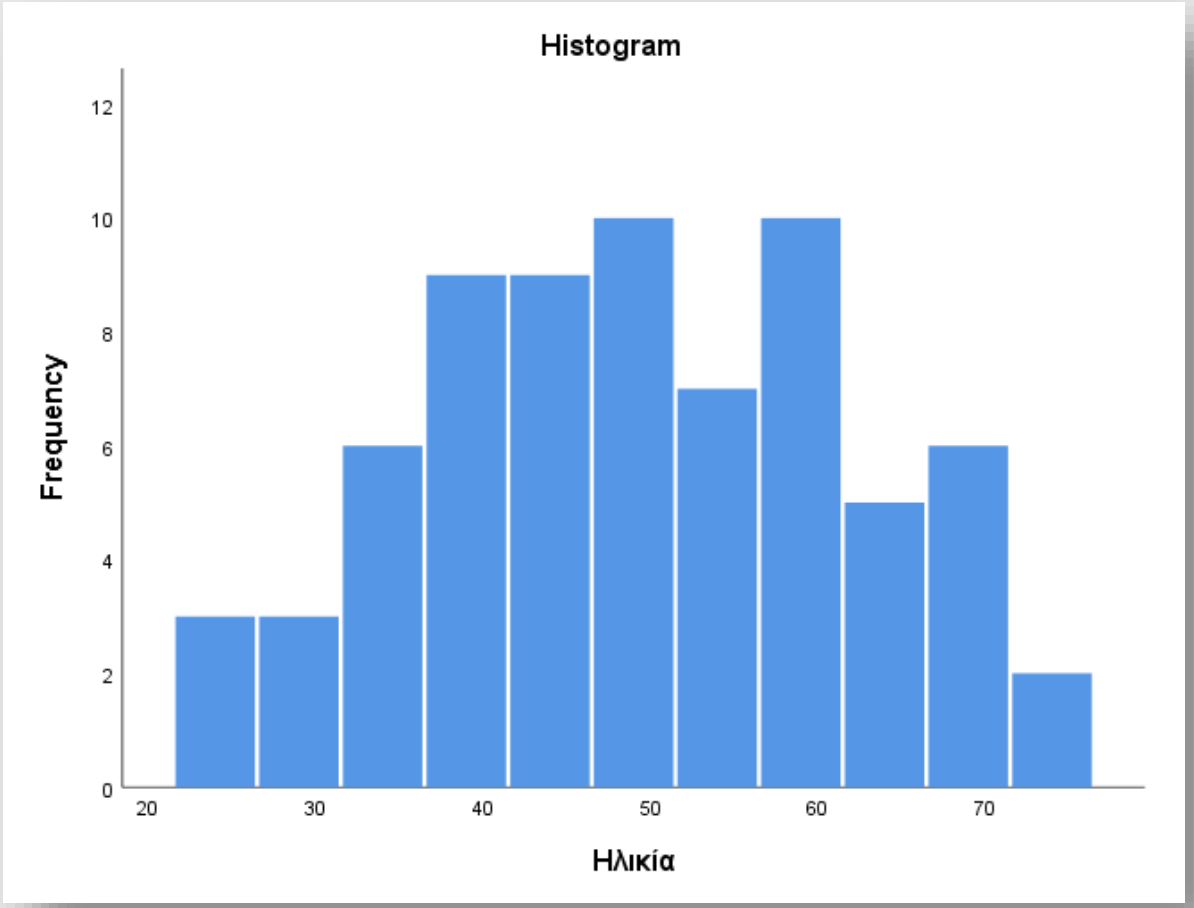
Weibull

Uniform

ΙΣΤΟΓΡΑΜΜΑ

Οι συχνότητες των τιμών για τα συγκεκριμένα διαστήματα ονομάζονται “bins”· πλάτος μπάρας

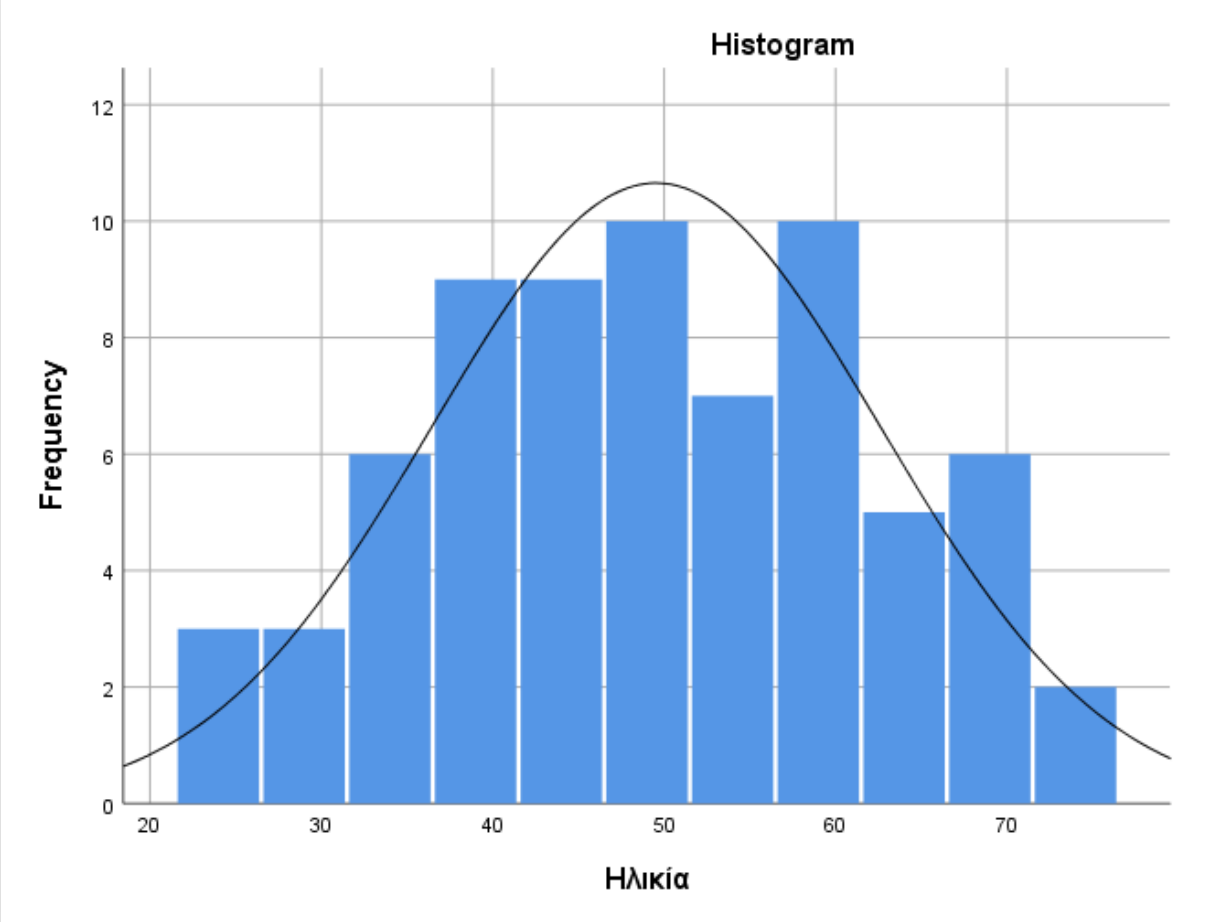
Μοιάζει αυτό το ιστόγραμμα στην κανονική κατανομή;



ΙΣΤΟΓΡΑΜΜΑ

Έτσι φαίνεται!

Προειδοποίηση:
Τα ιστογράμματα μπορεί να είναι παραπλανητικά. Η εμφάνιση του ιστογράμματος εξαρτάται από το πλάτος του bin.



Φυλλόγραμμα

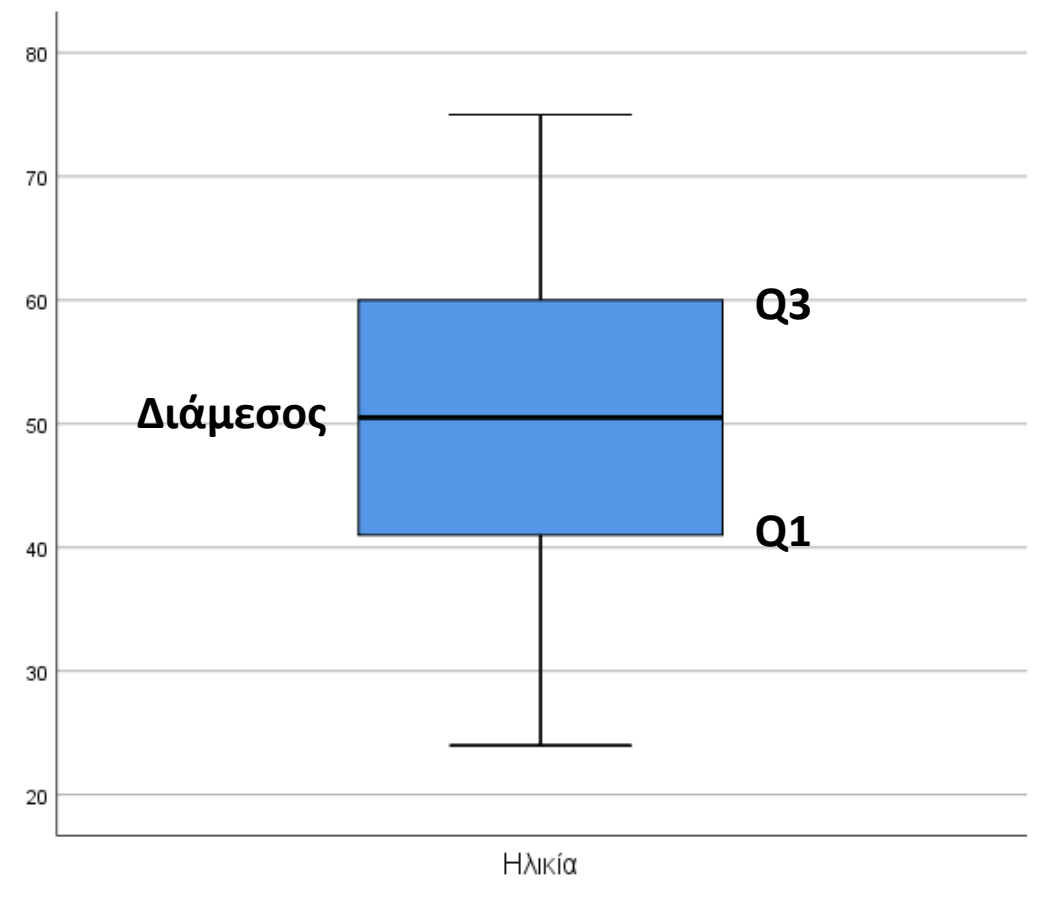
Ηλικία Stem-and-Leaf Plot

Frequency	Stem	&	Leaf
5.00	2	.	45689
11.00	3	.	02223359999
18.00	4	.	011112222244458999
15.00	5	.	011111345555578
16.00	6	.	000001112223389 9
5.00	7	.	11155

Stem width: 10
Each leaf: 1 case(s)

Ένα περιστραμμένο ιστόγραμμα

ΘΗΚΟΓΡΑΜΜΑ

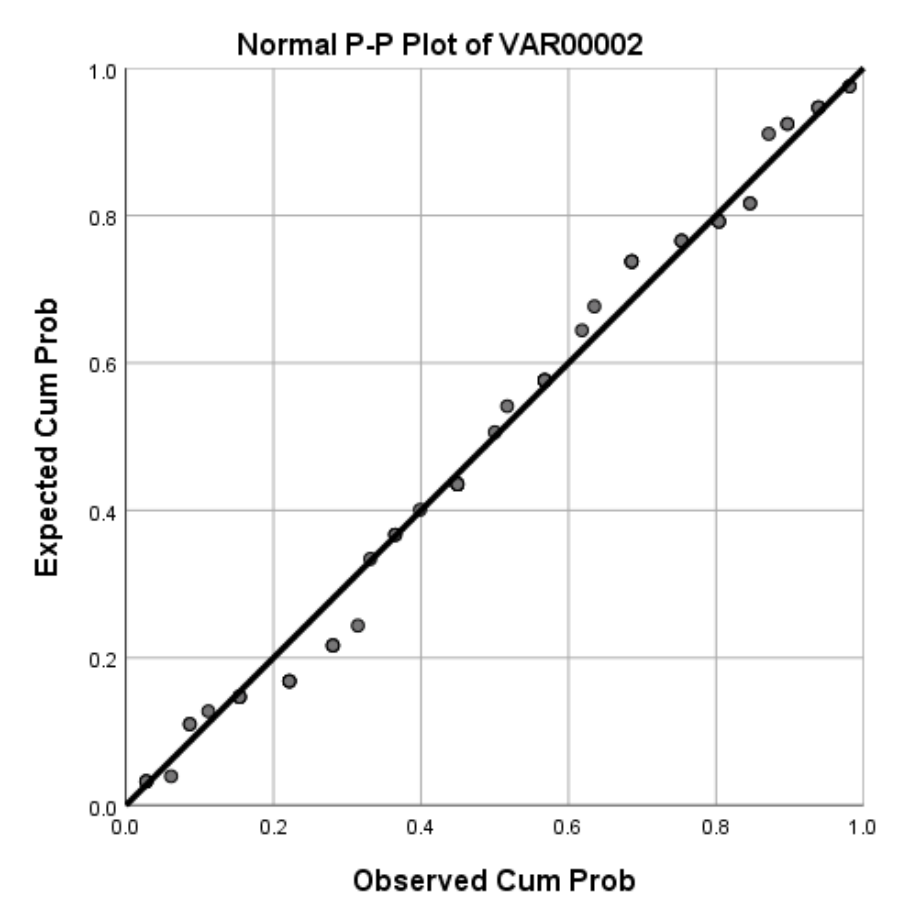


Τα θηκογράμματα είναι απλά γραφήματα που μας δείχνουν την κατανομή των δεδομένων.

Τι κοιτάμε λοιπόν;

1. **Είναι συμμετρικό;**
2. **Έχουν τα Q1 και Q3 κατά προσέγγιση την ίδια απόσταση από τη διάμεσο;**
3. **Έχουν τα μουστάκια την ίδια απόσταση;**

P-P διάγραμμα

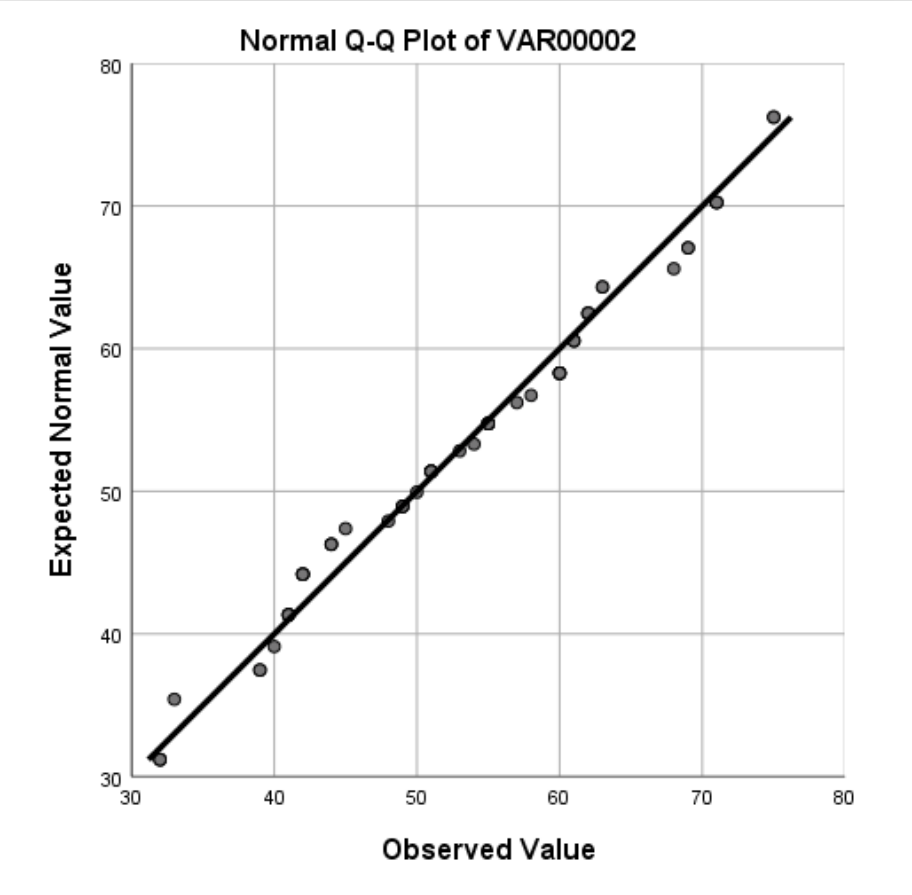


Σε ένα P-P διάγραμμα, συγκρίνουμε την αθροιστική πιθανότητα των δεδομένων μας με μία ιδανική «τεστ» κατανομή˙ στην περίπτωση αυτή με την κανονική κατανομή.

Η ερώτηση που κάνουμε:

«Πέφτουν» τα δεδομένα μας πάνω σε μία ευθεία γραμμή; Αν τα δεδομένα μας ακολουθούν την κανονική κατανομή τότε θα πρέπει.

Q-Q διάγραμμα

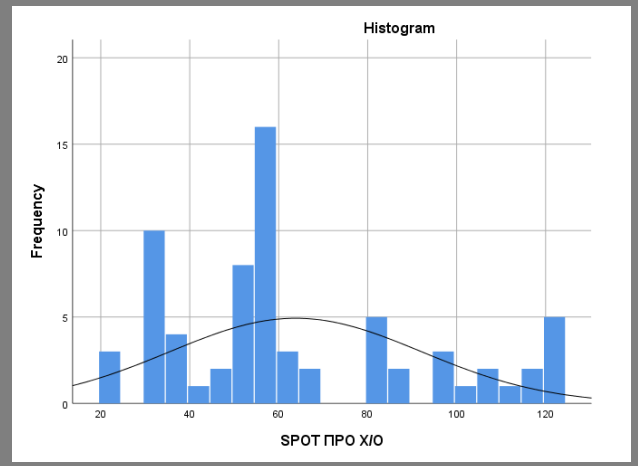


Σε ένα Q-Q διάγραμμα, συγκρίνουμε τα ποσοστιαία σημεία των δεδομένων μας με την ιδανική που στην περίπτωση αυτή είναι η κανονική κατανομή.

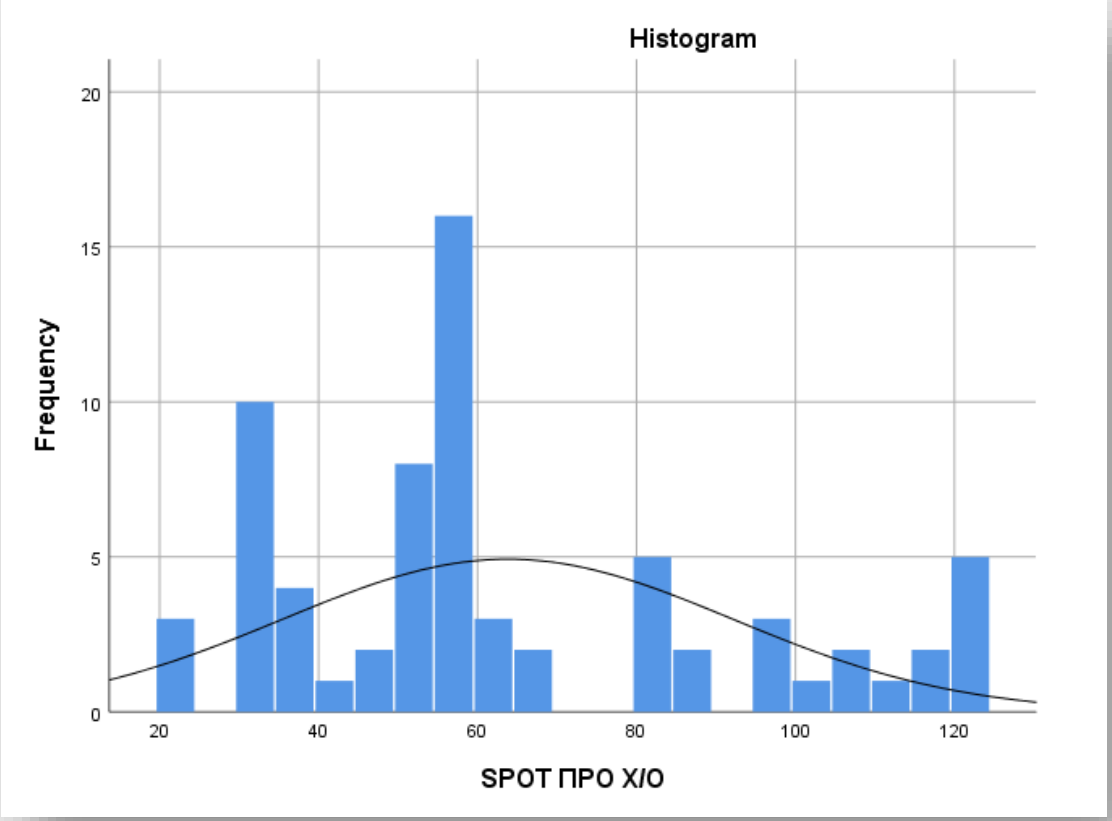
Η ερώτηση που κάνουμε:

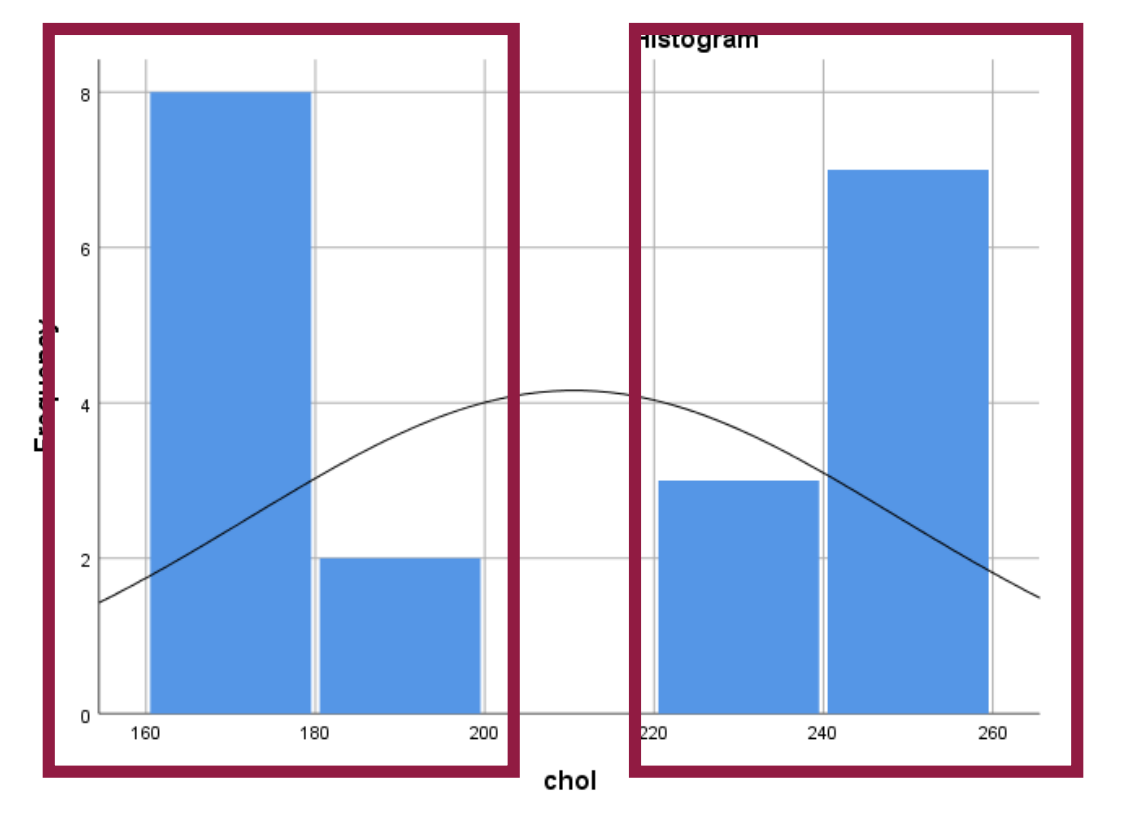
«Πέφτουν» τα δεδομένα μας πάνω σε μία ευθεία γραμμή;

Είναι τα δεδομένα μου κανονικά;



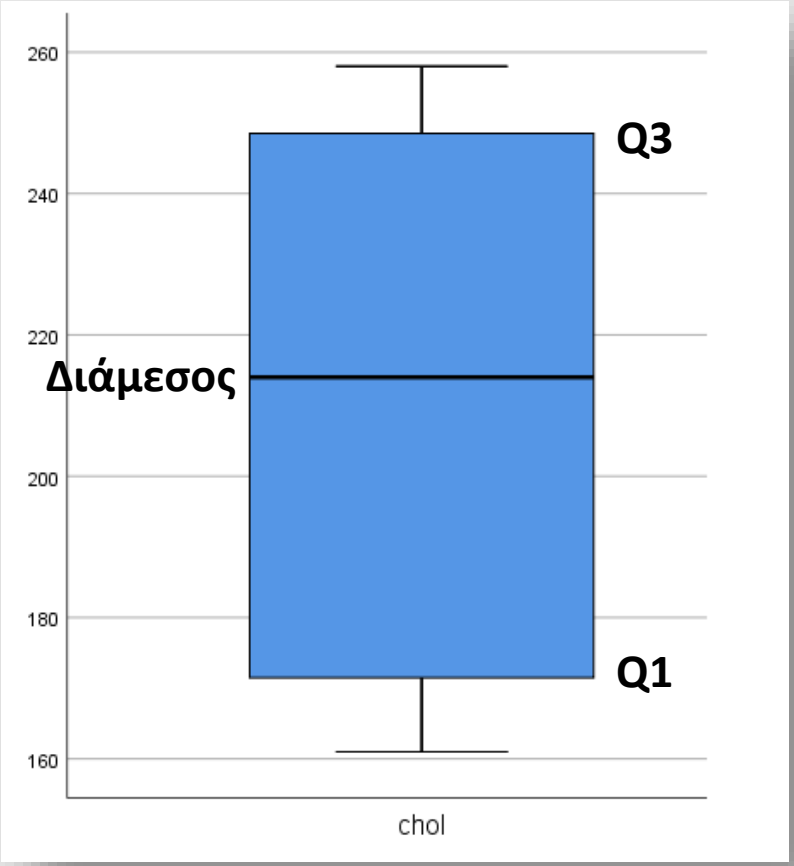
Ανάλυση ιστογράμματος



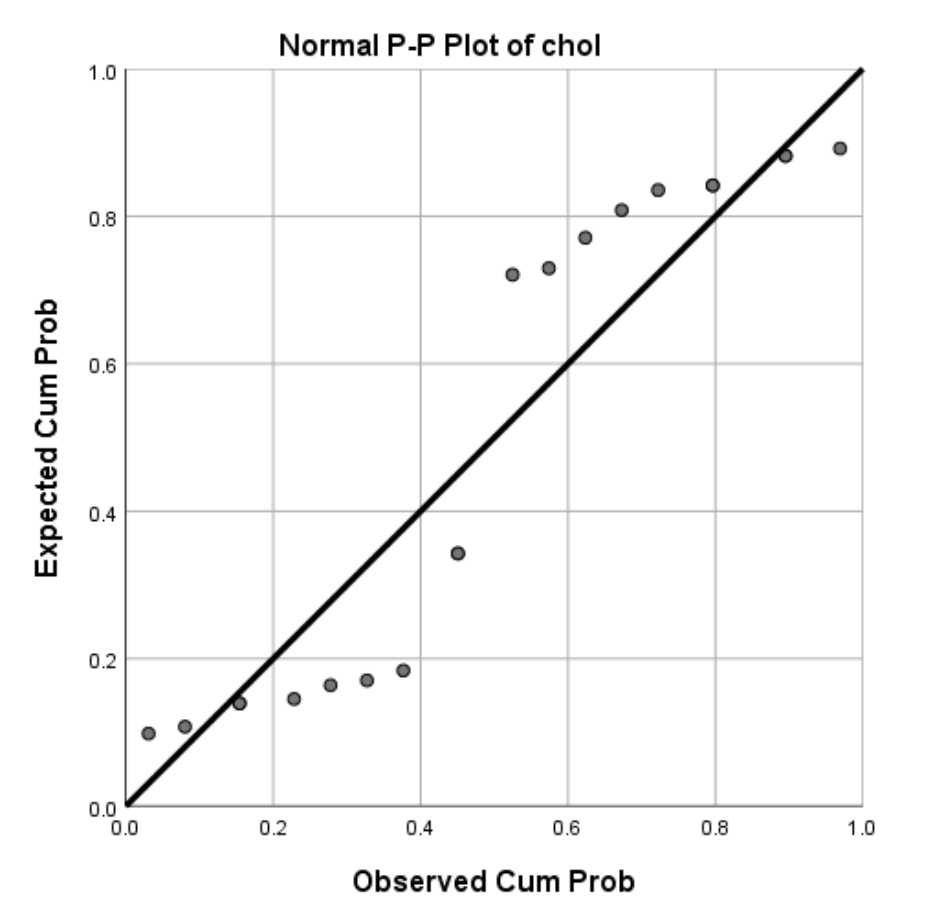


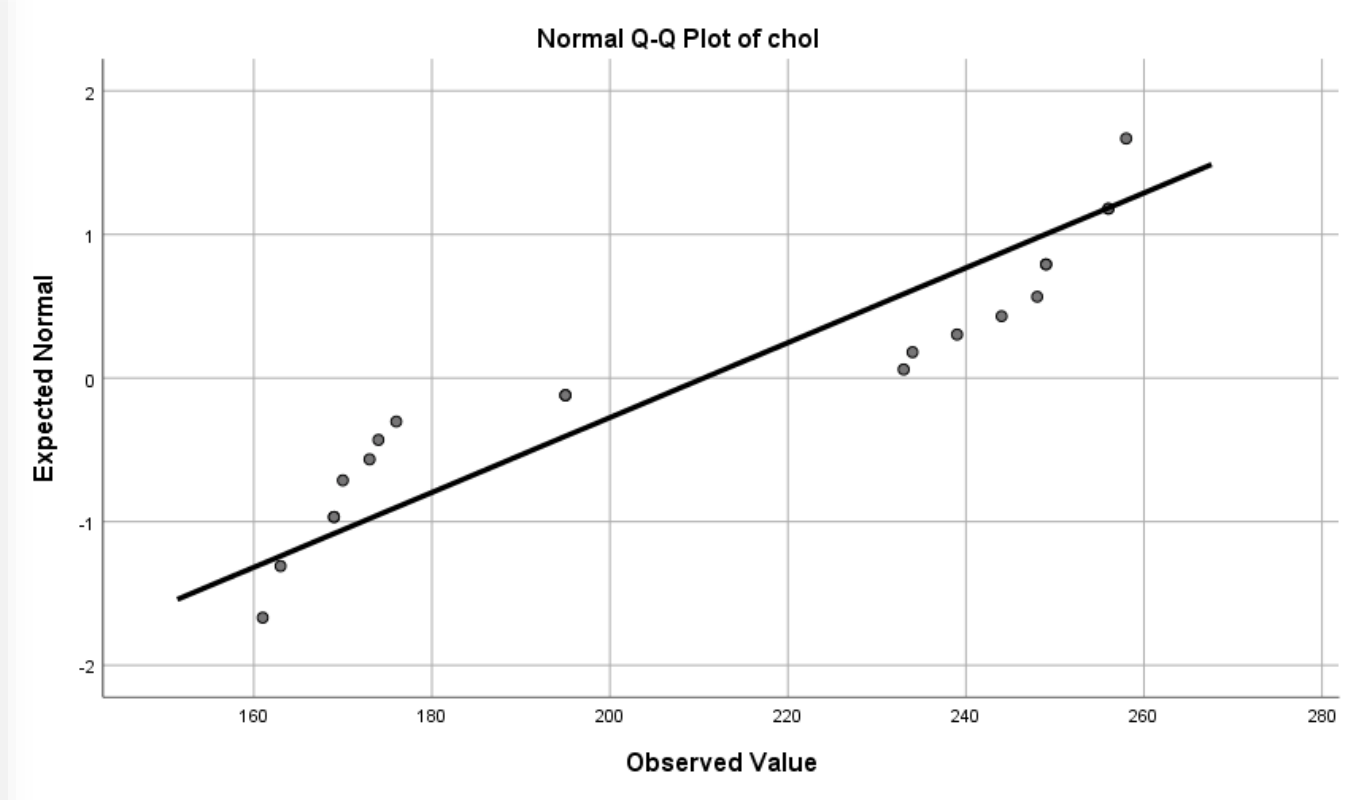
Φυλλόγραμμα και θηκόγραμμα

```
chol Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    10.00        1 .  6666777799
     7.00        2 .  3334444
     3.00        2 .  555

 Stem width:       100
 Each leaf:        1 case(s
```



P-P και Q-Q διάγραμμα





Συμπέρασμα;

Τα δεδομένα μας δεν ακολουθούν την κανονική κατανομή

Ακολουθούν κάποια άλλη κατανομή

Ανασκόπηση

- Κάνοντας χρήση διαγραμμάτων μπορούμε να αντλήσουμε πολλές πληροφορίες για τα δεδομένα μας
- Τα δεδομένα μας μπορεί να είναι λοξά (skewed), κυρτά (kurtosis, πλατιές άκρες) ή να ακολουθούν μία κατανομή που δεν είναι κανονική
- Σε αυτήν την παρουσίαση συζητήσαμε για τα παρακάτω διαγράμματα για να εξακριβώσουμε αν τα δεδομένα μας είναι «κανονικά»:
  - Ιστογράμματα
  - Φυλλόγραμμα (Stem and leaf)
  - Θηκογράμματα (Box Plots)
  - P-P διαγράμματα
  - Q-Q διαγράμματα