Is my data normal?

Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

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

Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

Our recommendation!

LOOK AT YOUR DATA GRAPHICALLY FIRST

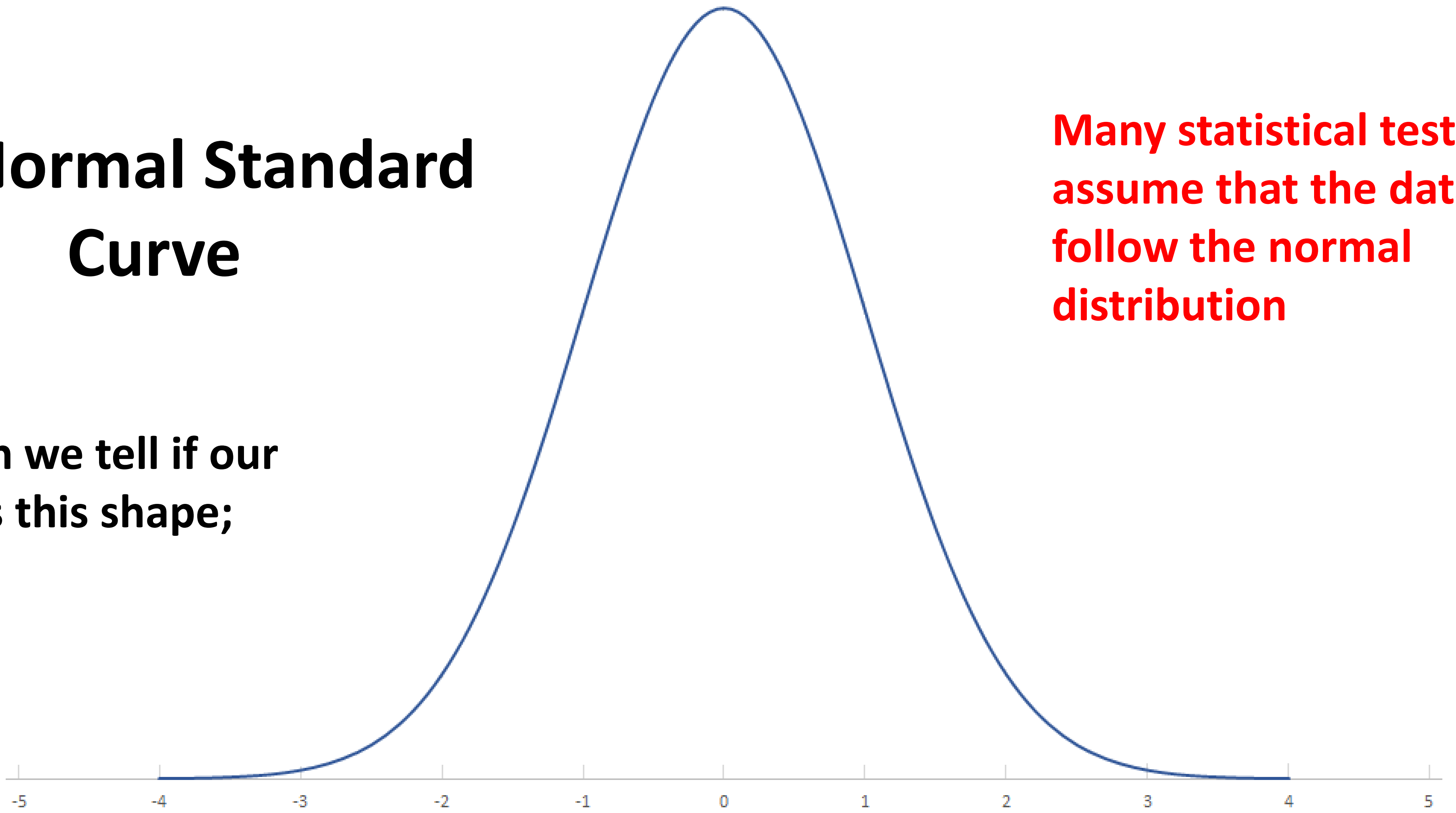
… before starting with the analysis.

Get to know the data. Look for patterns,
potentials problems, initials relationships, etc.

Graphical Data Exploration

- Charts allow us to extract meaningful information from our data
- Our data may be skewed, have high or low kurtosis (fat tails), or follow a non-normal distribution
- In this presentation, we will discuss the following charts to determine whether our data are **normally distributed**:
  - Histograms
  - Stem and leaf Plots
  - Box Plots
  - P-P Plots
  - Q-Q Plots

The Normal Standard Curve



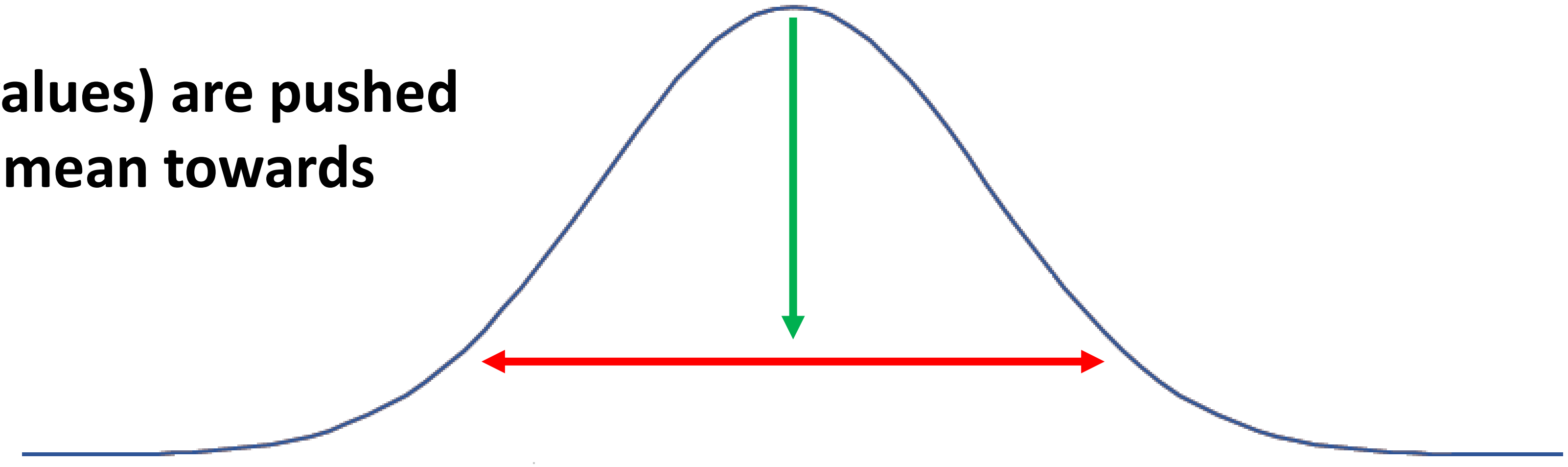
Many statistical tests assume that the data follow the normal distribution

How can we tell if our data fits this shape;

Kurtosis

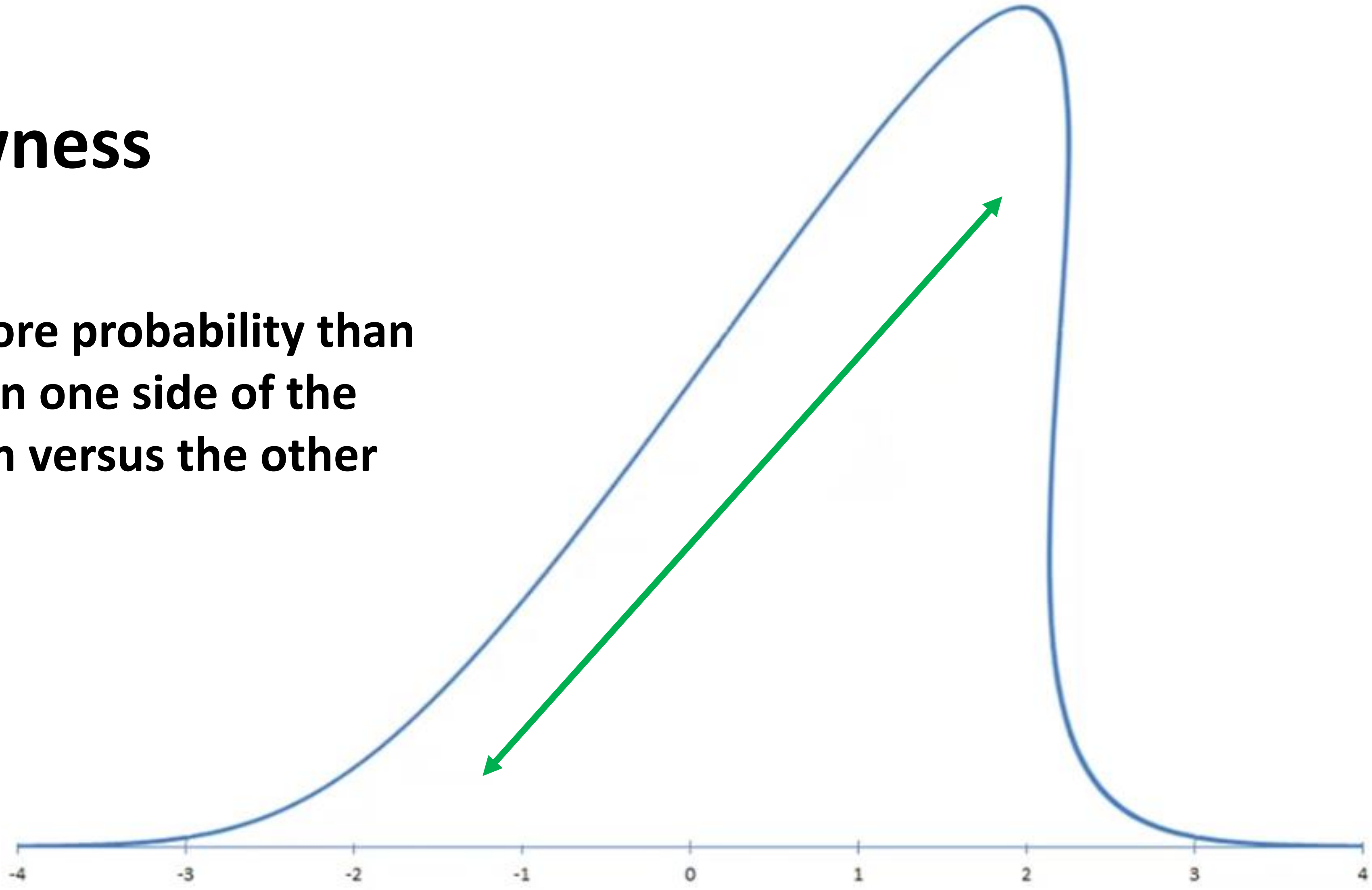
More probability than expected in the tails of the distribution due to extreme values away from the mean.

Probabilities (values) are pushed away from the mean towards the tails.



Skewness

There is more probability than excepted on one side of the distribution versus the other



Other probability distribution

Oftentimes data fits another type of distribution much better:

Exponential

Lognormal

among others…

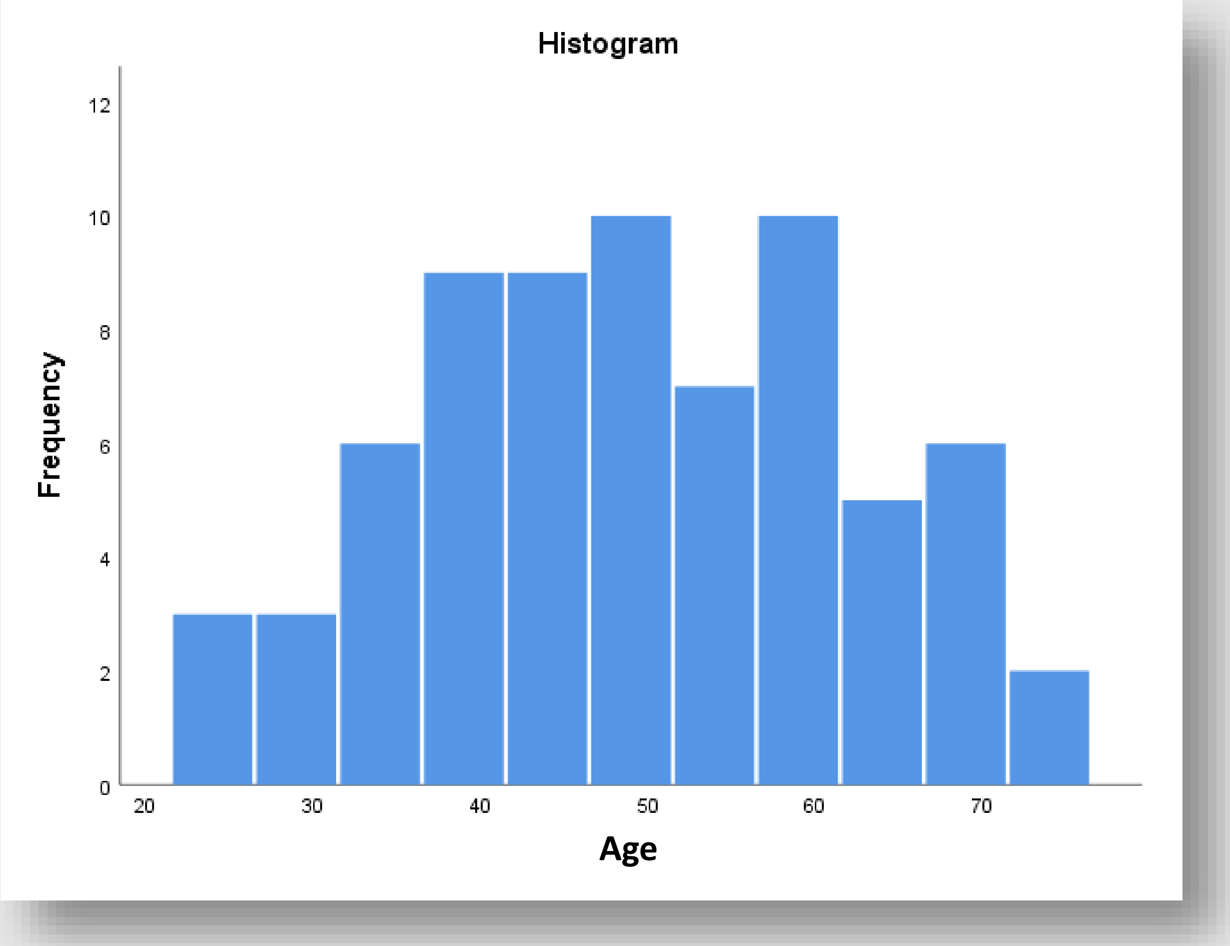
Weibull

Uniform

HISTOGRAM

The frequency of values over certain intervals is called bins

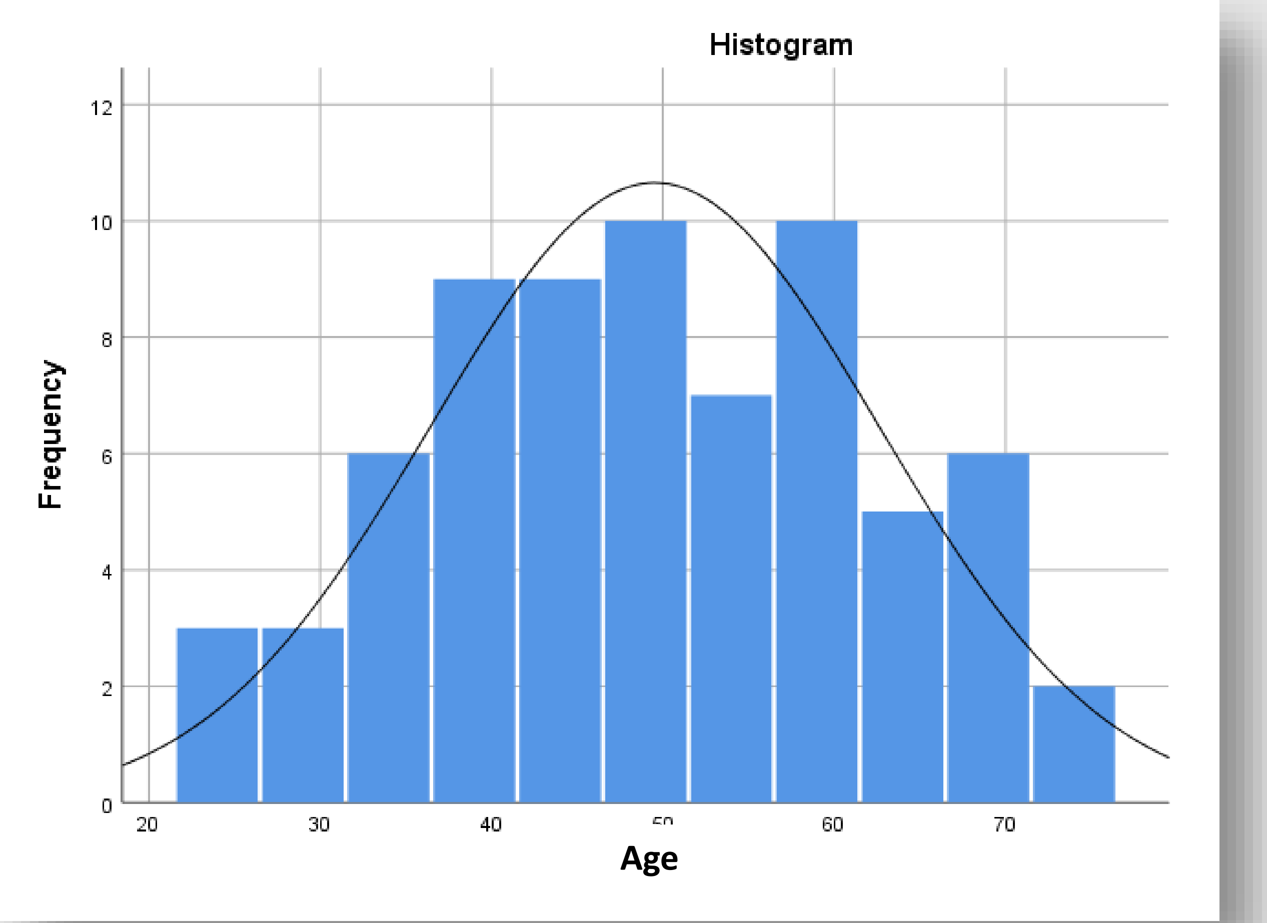
Does this histogram look like the normal curve?



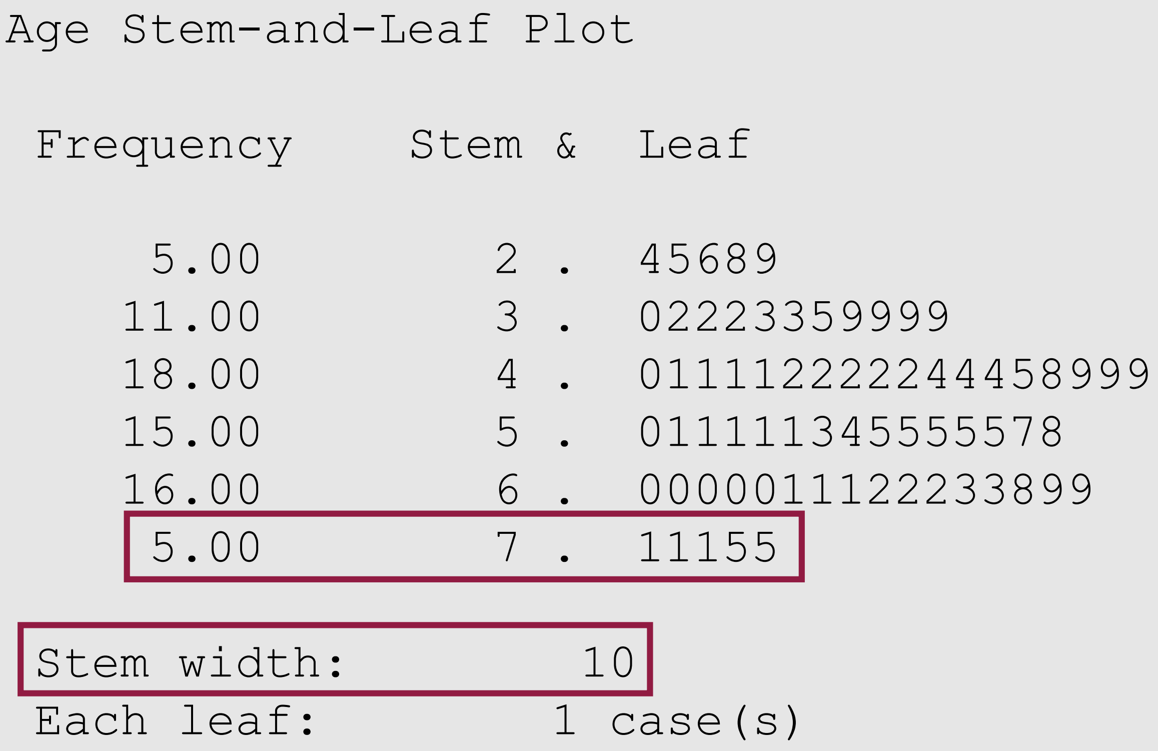
HISTOGRAM

So it seems!

Warning:
Histograms can sometimes be misleading due to their dependency on bin width.



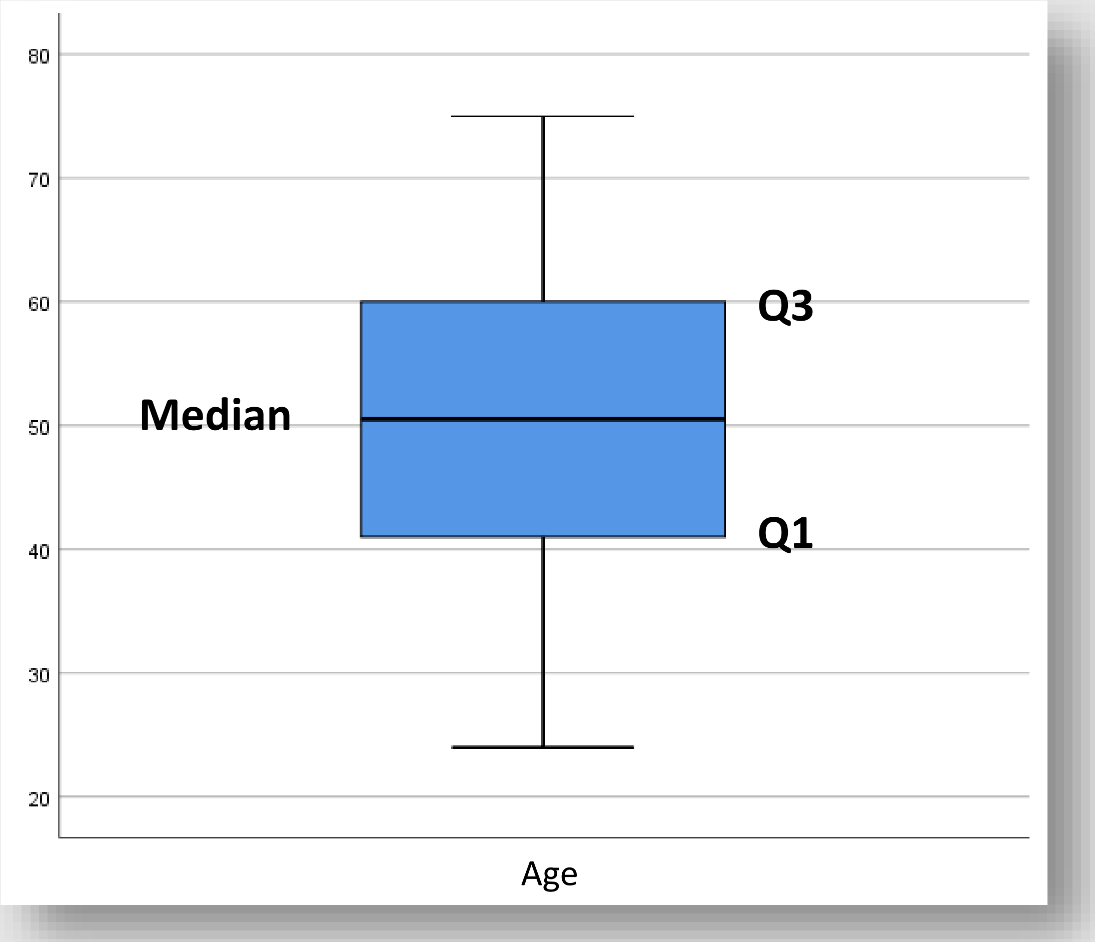
STEM AND LEAF



```
Age 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 .  0000011122233899
     5.00        7 .  11155

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

A “sideways” histogram

BOX PLOT

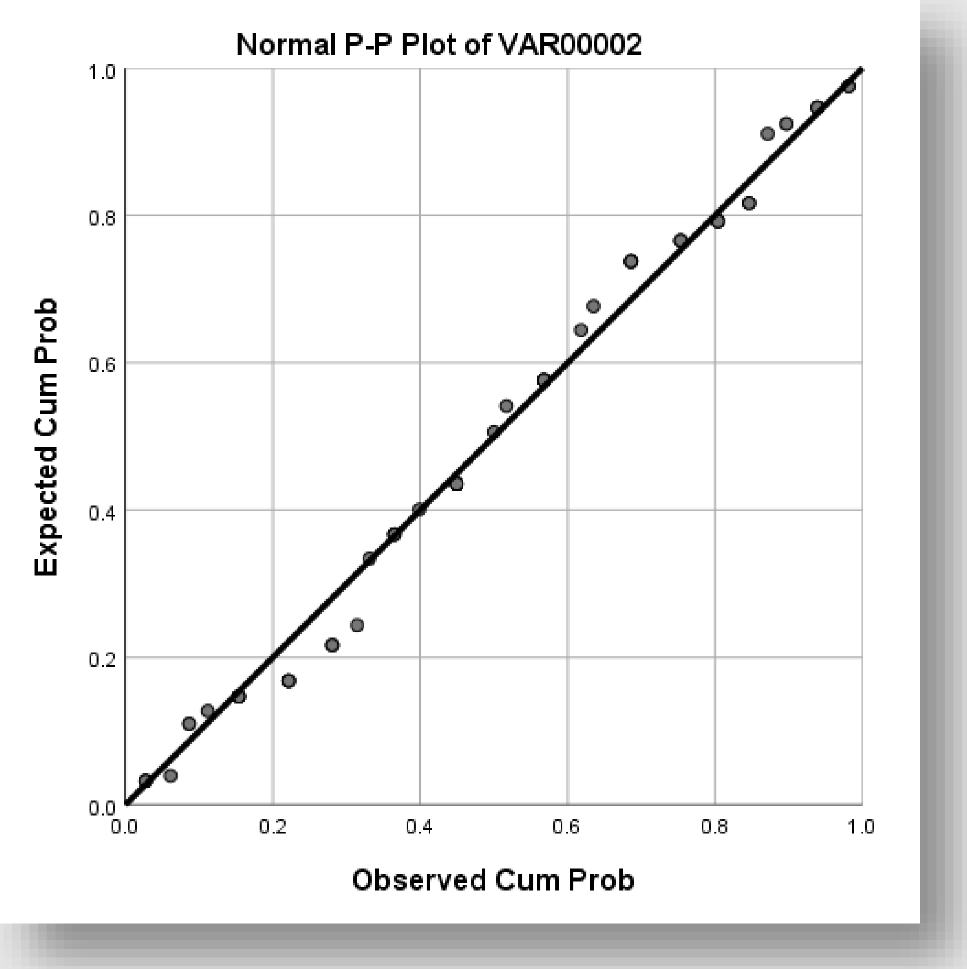


Box plots are simple graphs used to visualize the distribution of data

So, what should you look for?

1. **Is the box-plot symmetrical overall?**
2. **Are Q1 and Q3 approximately the same distance from the median?**
3. **Are the whiskers of the plot approximately the same length?**

P-P PLOT

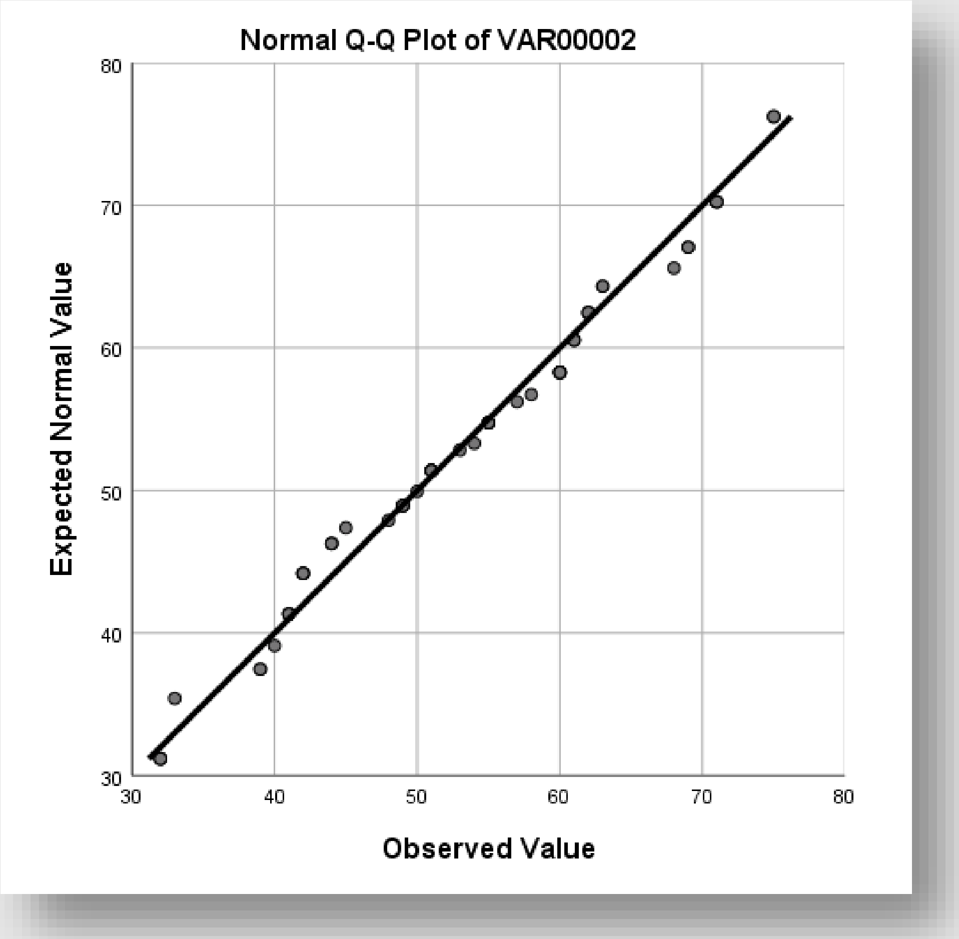


In a P-P plot, we compare the cumulative probability of our data with an ideal “test” distribution; in this case the normal distribution.

Question to Ask:

> **Do the data points fall in a straight line? If our data matches the test distribution they should.**

Q-Q PLOT

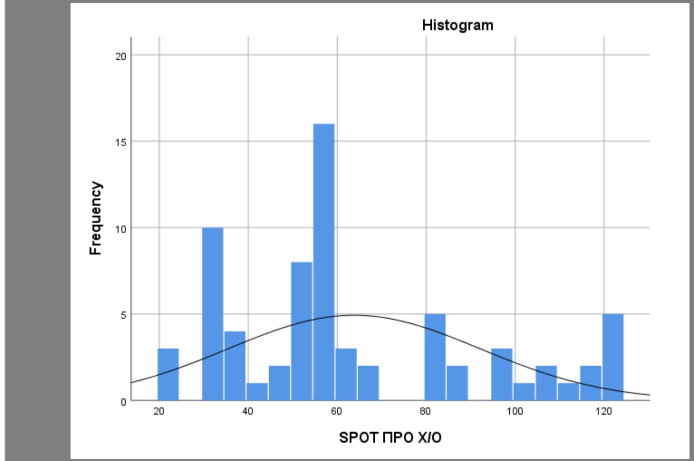


In a Q-Q plot, we compare the quantiles of our data with the ideal.

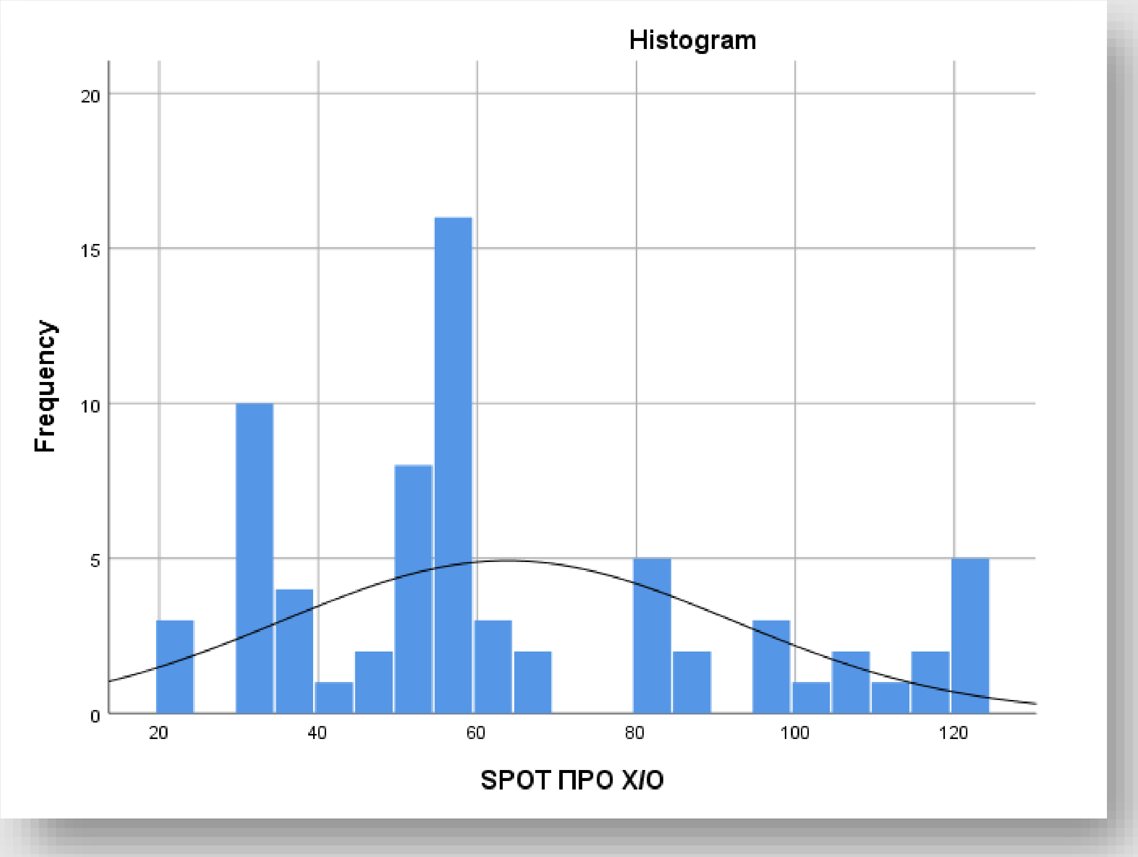
Question to ask:

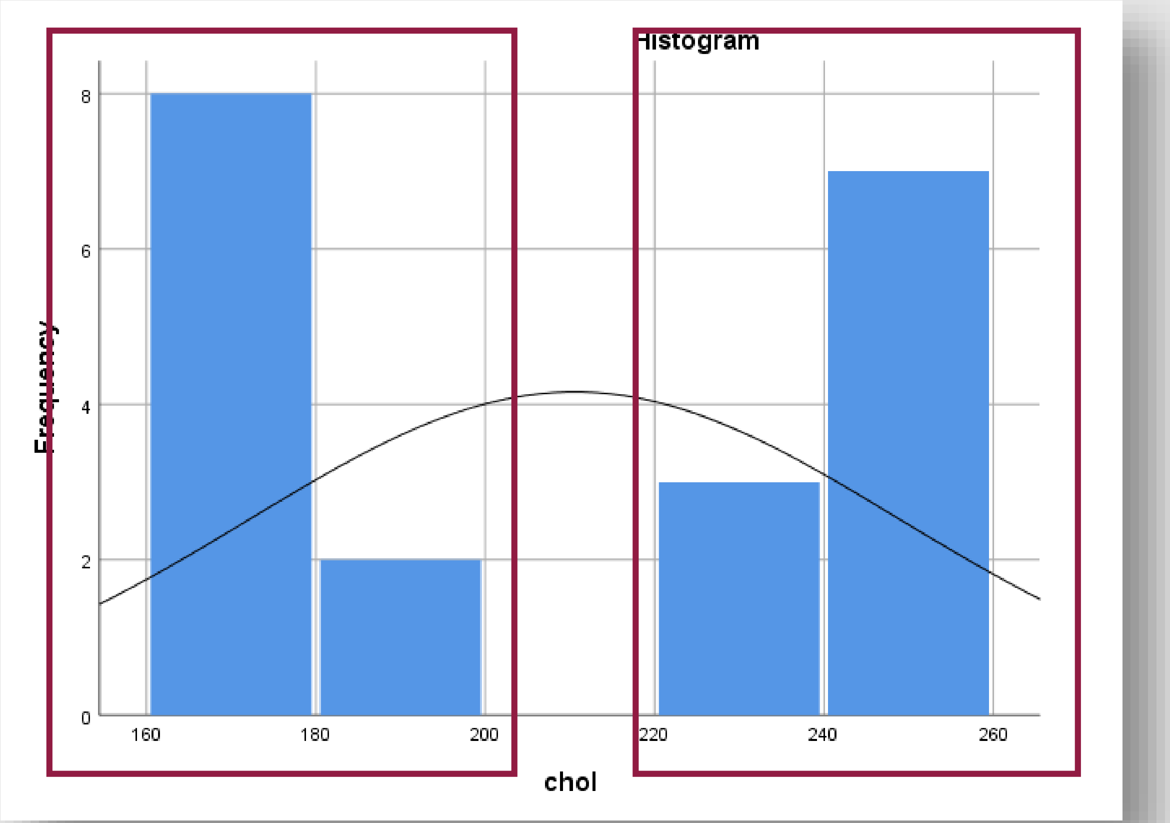
Do the data points fall in a straight line;

Is this data normal?



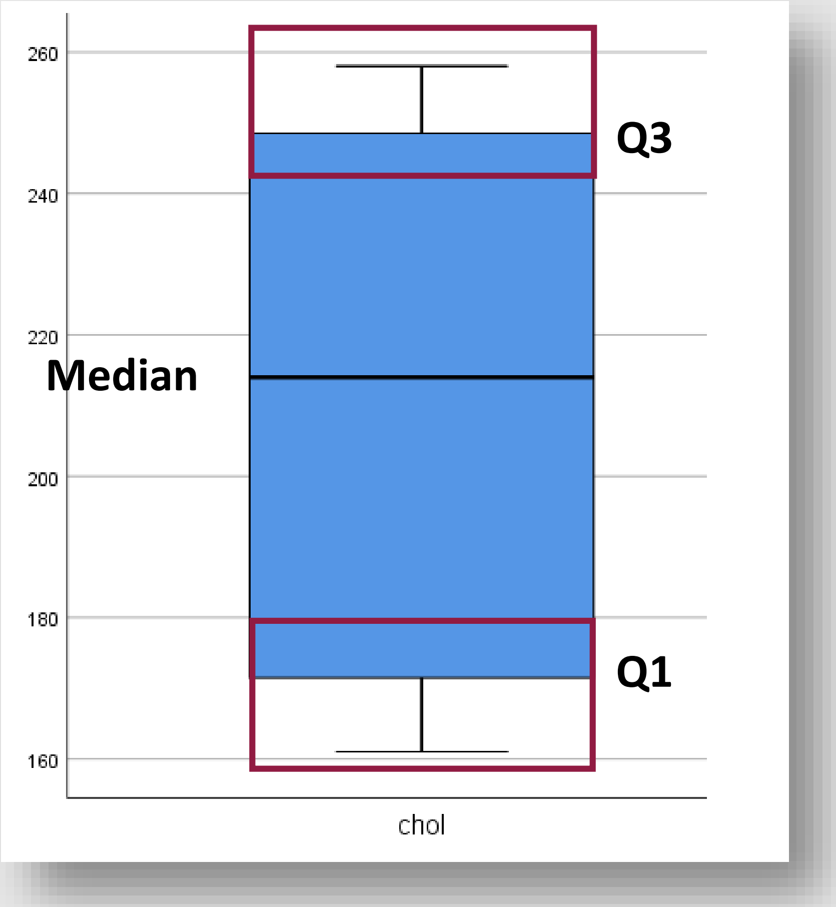
Histogram analysis



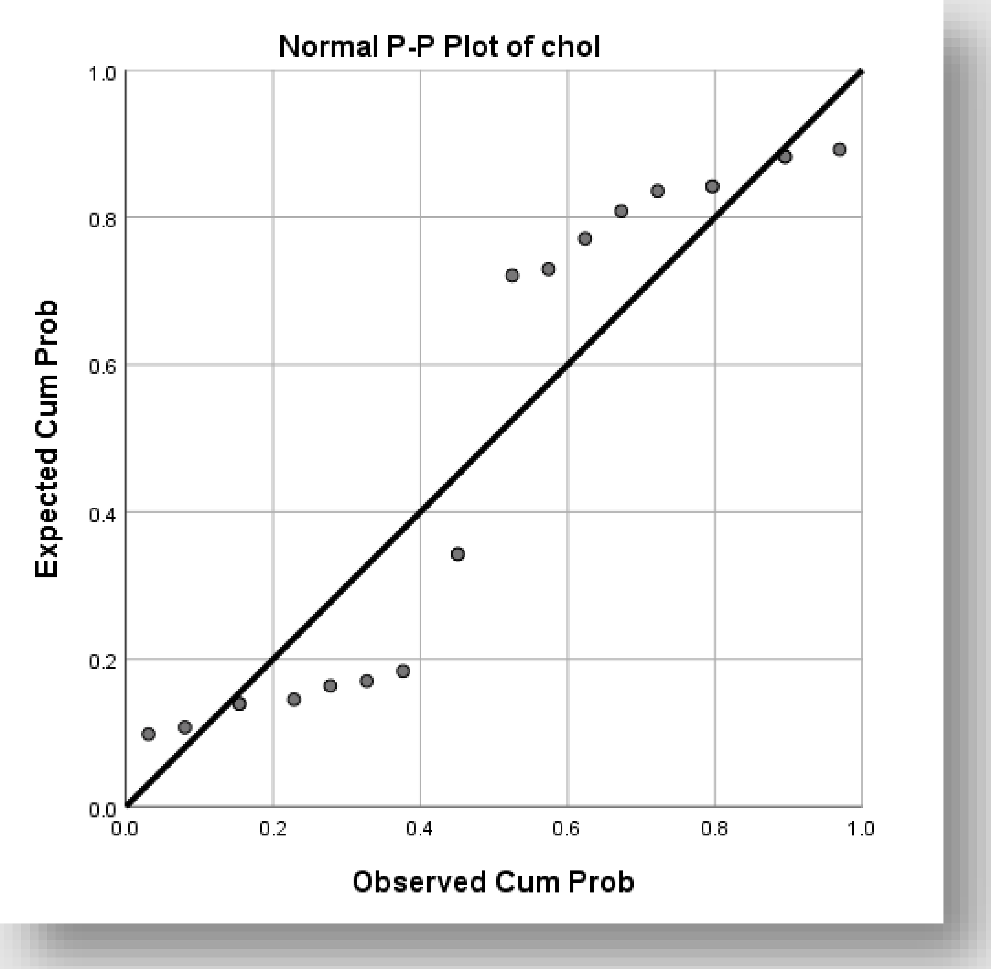


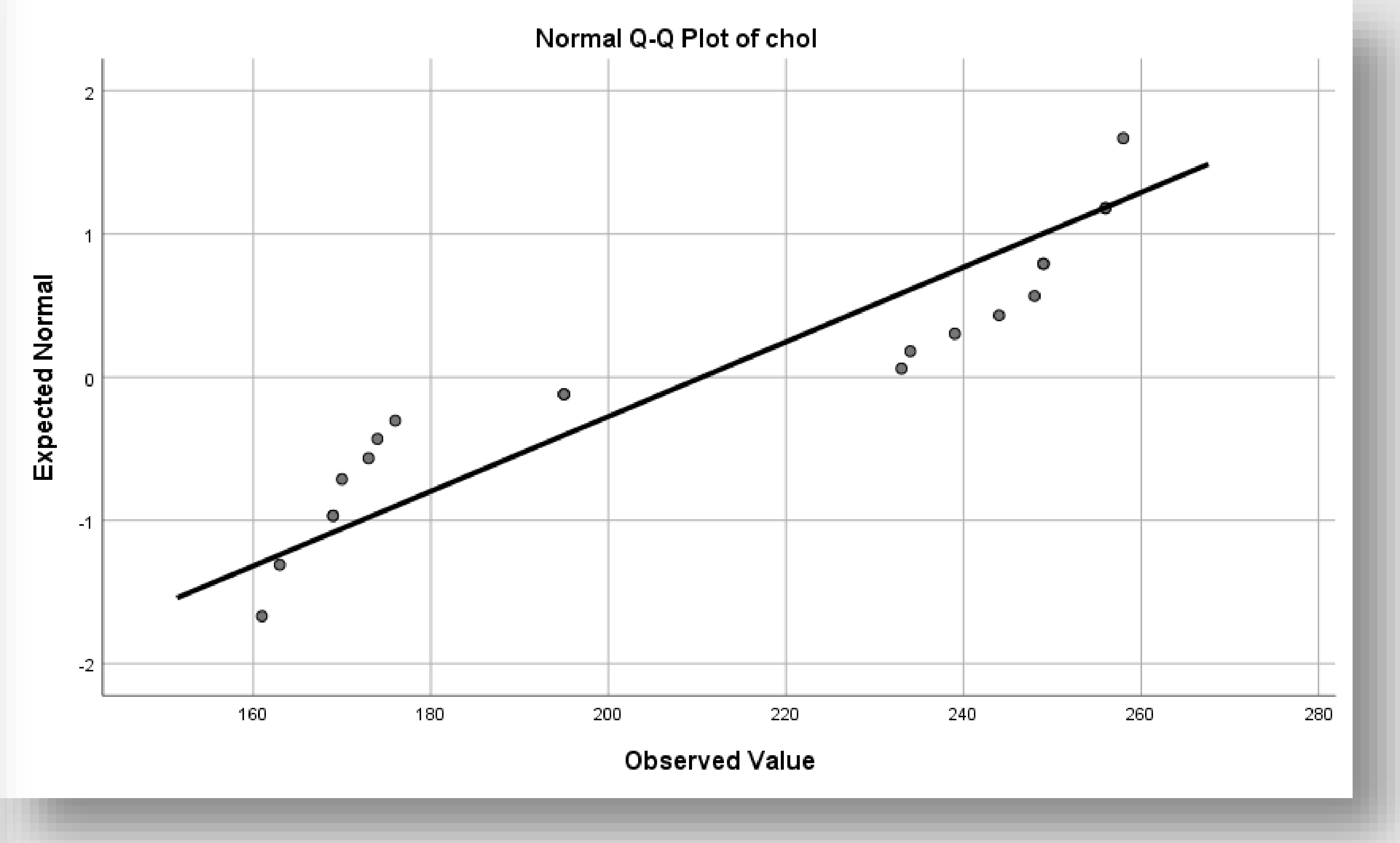
Stem and Leaf Plot and Box Plot

```
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 and Q-Q Plots





Conclusion?

This data does NOT fit the normal distribution

They follow a different distribution

A quick review

- Charts allow us to extract meaningful information from our data
- Our data may be skewed, have high or low kurtosis (fat tails), or follow a non-normal distribution
- In this presentation, we discussed the following charts to determine whether our data are **normally distributed**:
  - Histograms
  - Stem and leaf Plots
  - Box Plots
  - P-P Plots
  - Q-Q Plots