ARISTOTLE UNIVERSITY
OF THESSALONIKI

FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE

MSc Health Statistics and Data Analytics

Graphs for quantitative and qualitative data and the Normal (Gaussian) distribution

Eirini Pagkalidou, PhDc

pagalidou@auth.gr

THESSALONIKI 2021-22

MSc Health Statistics & Data Analytics

Objectives

Upon completion of this lecture you will be able to:

- ❑ Choose appropriate graph(s) to describe variables
- ❑ Interpret basic graphs when you see them
- ❑ Recognize Normal distributed variables based on boxplots and histograms

Chart Suggestions/A Thought-Starter

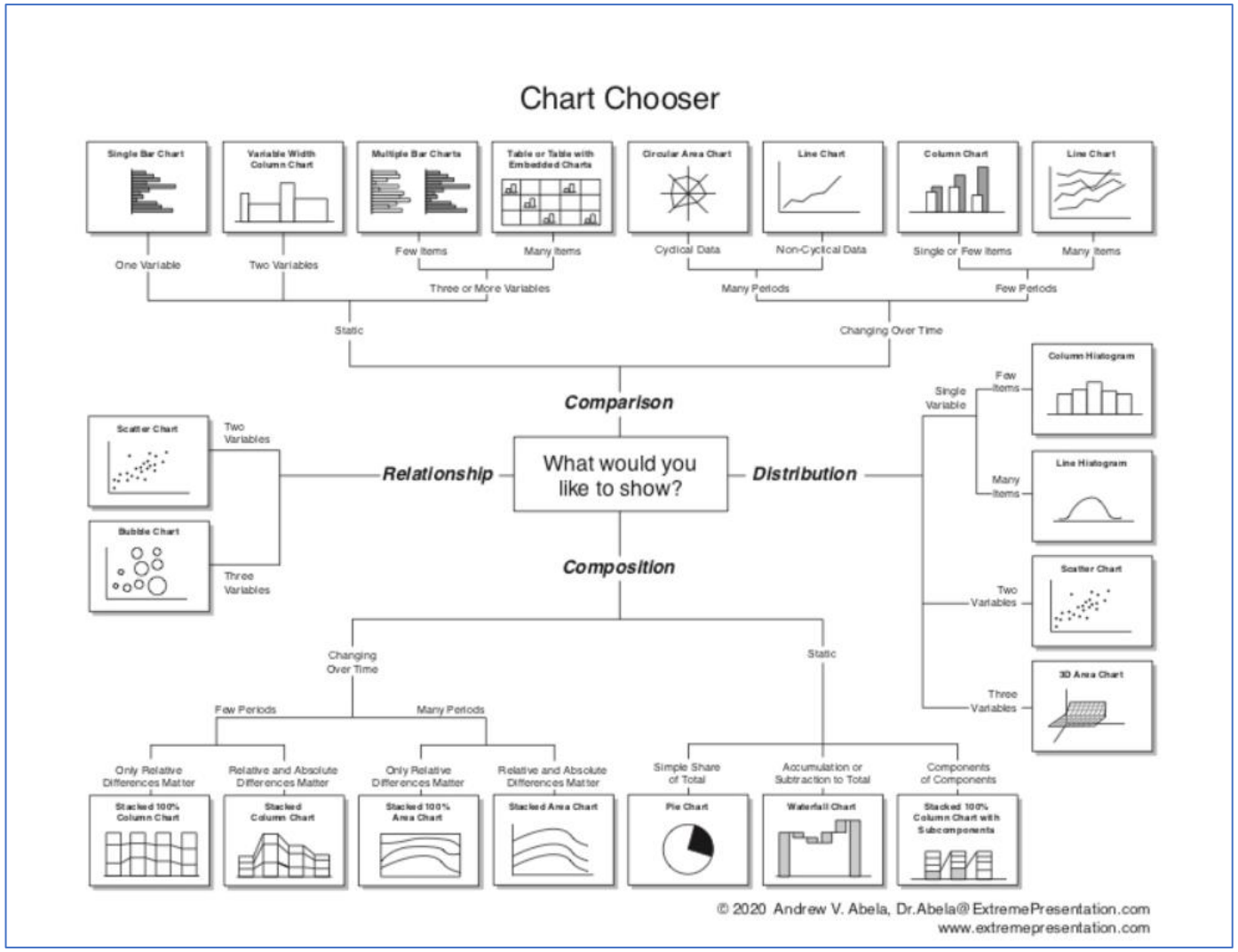
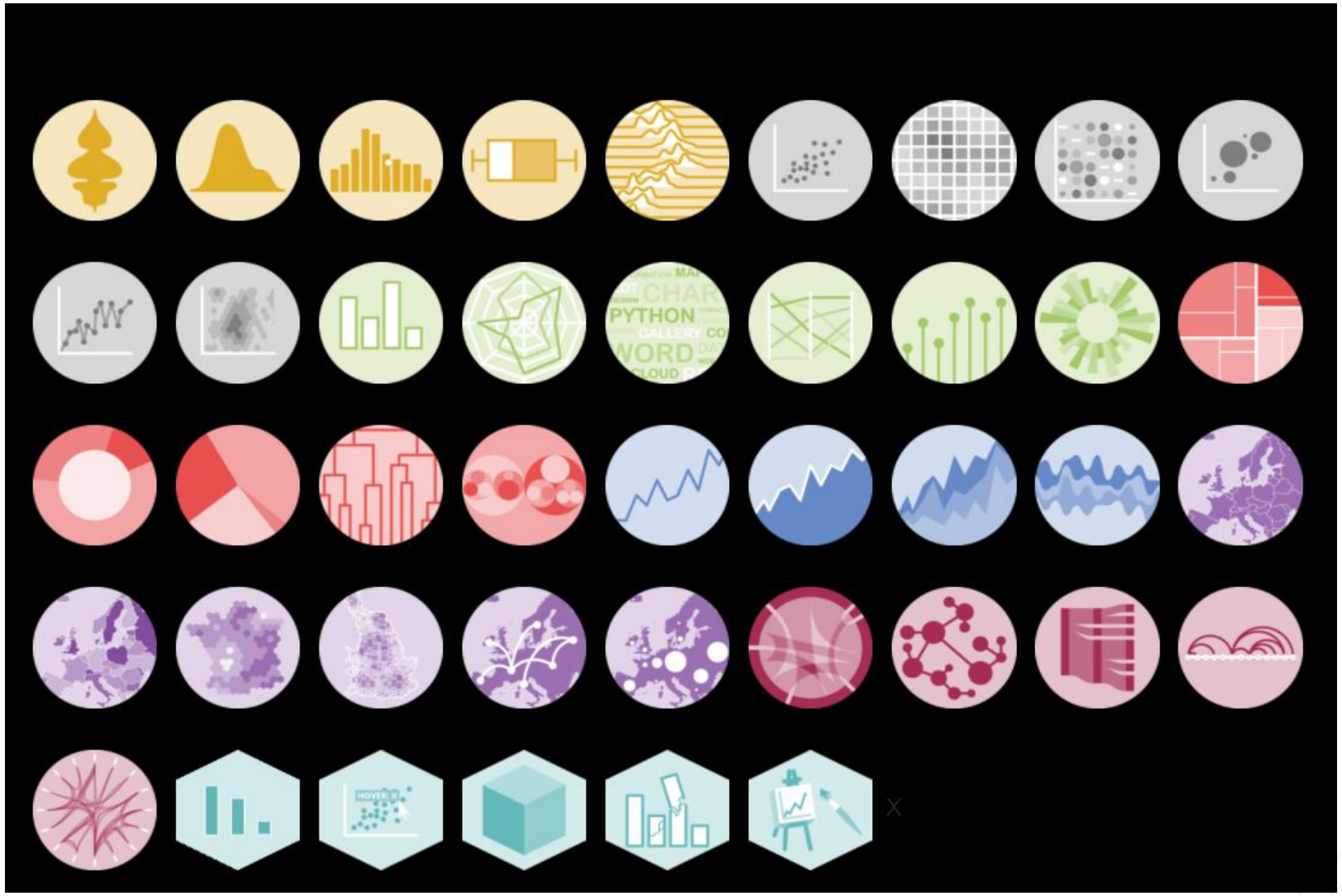
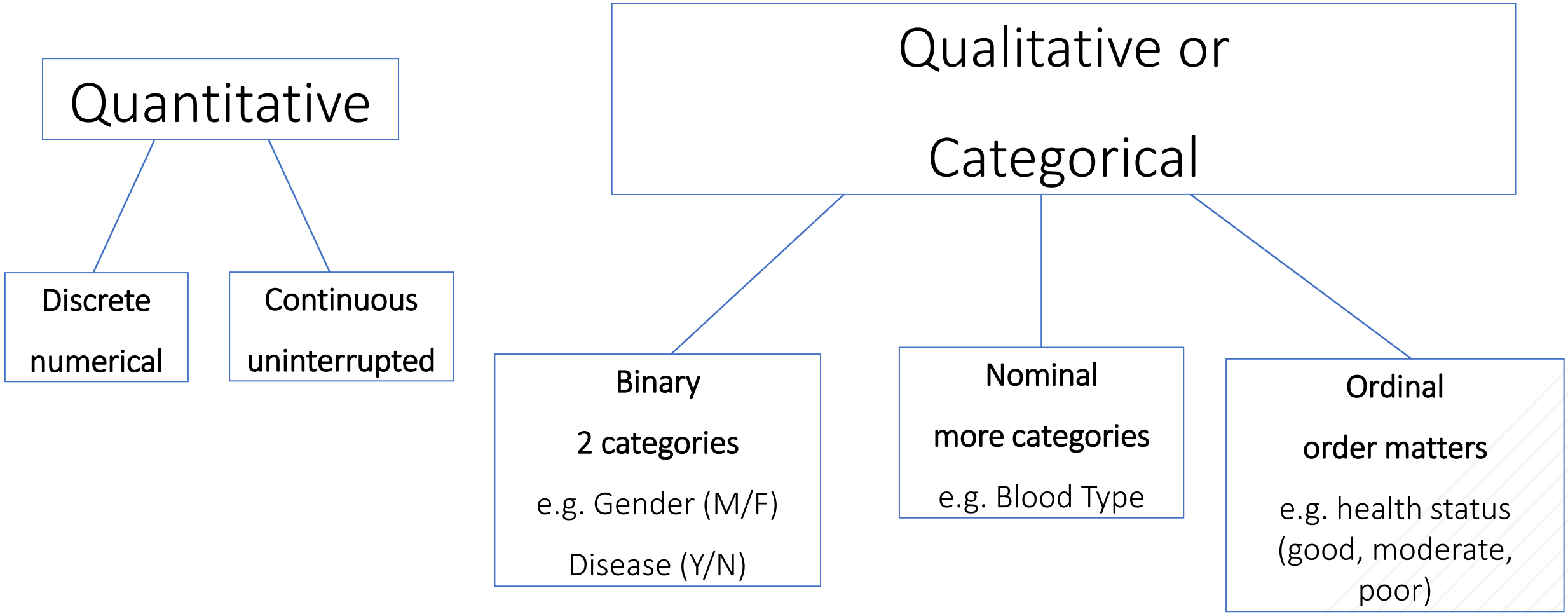


Chart Suggestions- The R Graph Gallery

https://www.r-graph-gallery.com/

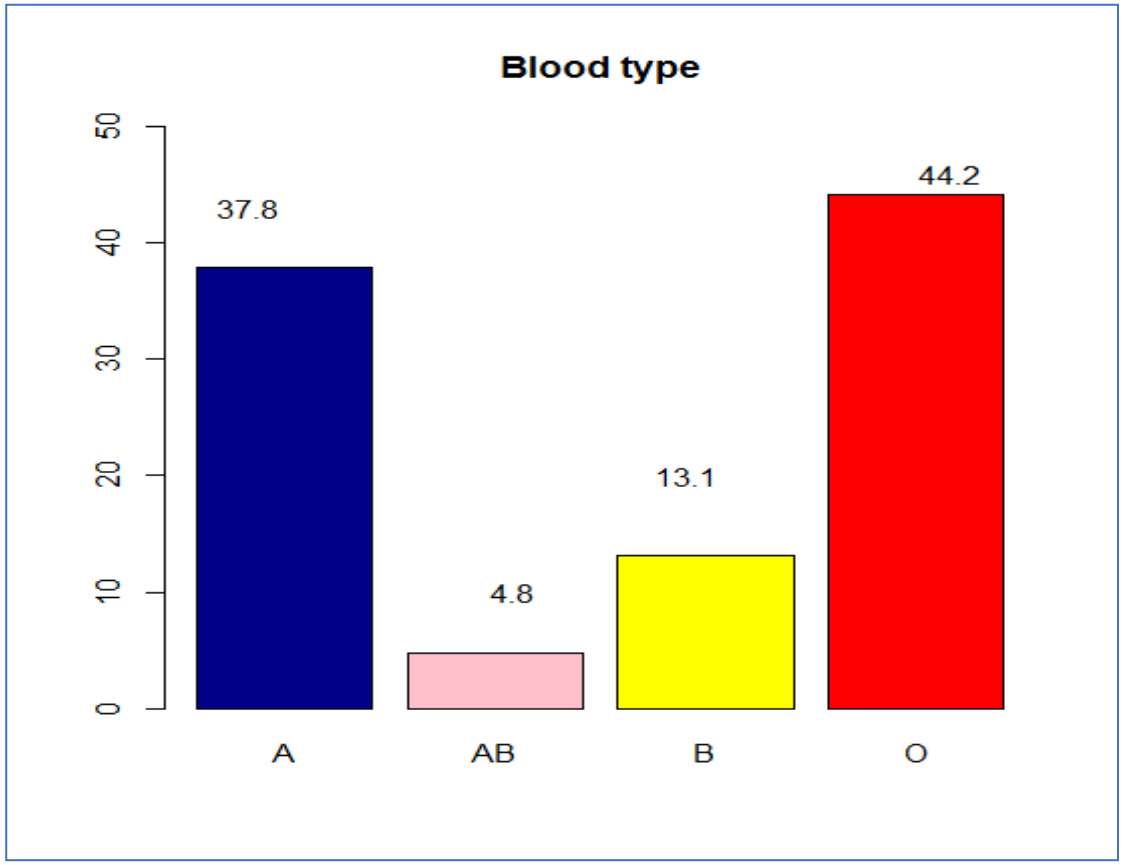


Types of Variables: Overview

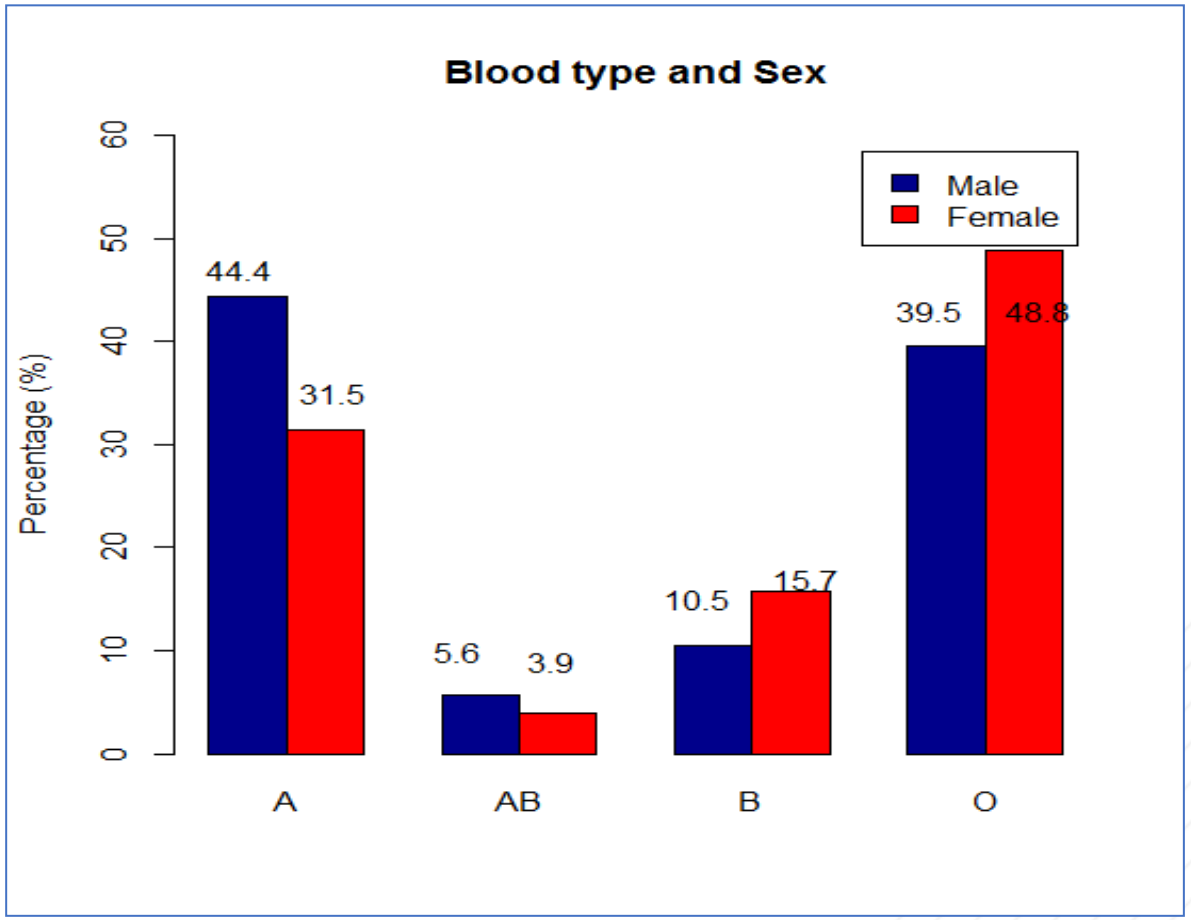


Bar Chart

- It is used to plot a categorical variable



- More than one variables

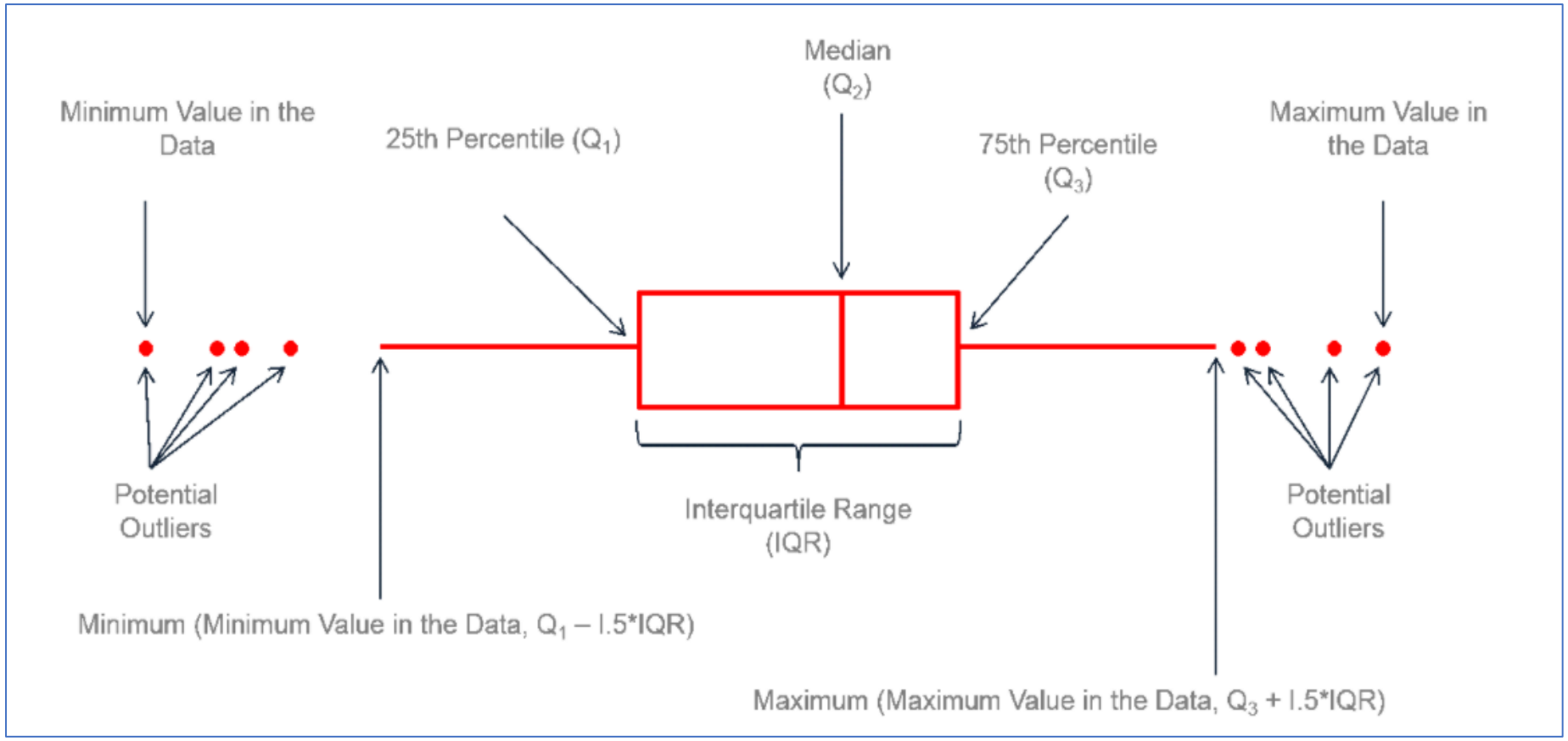


Separately present the blood type distribution in men and women

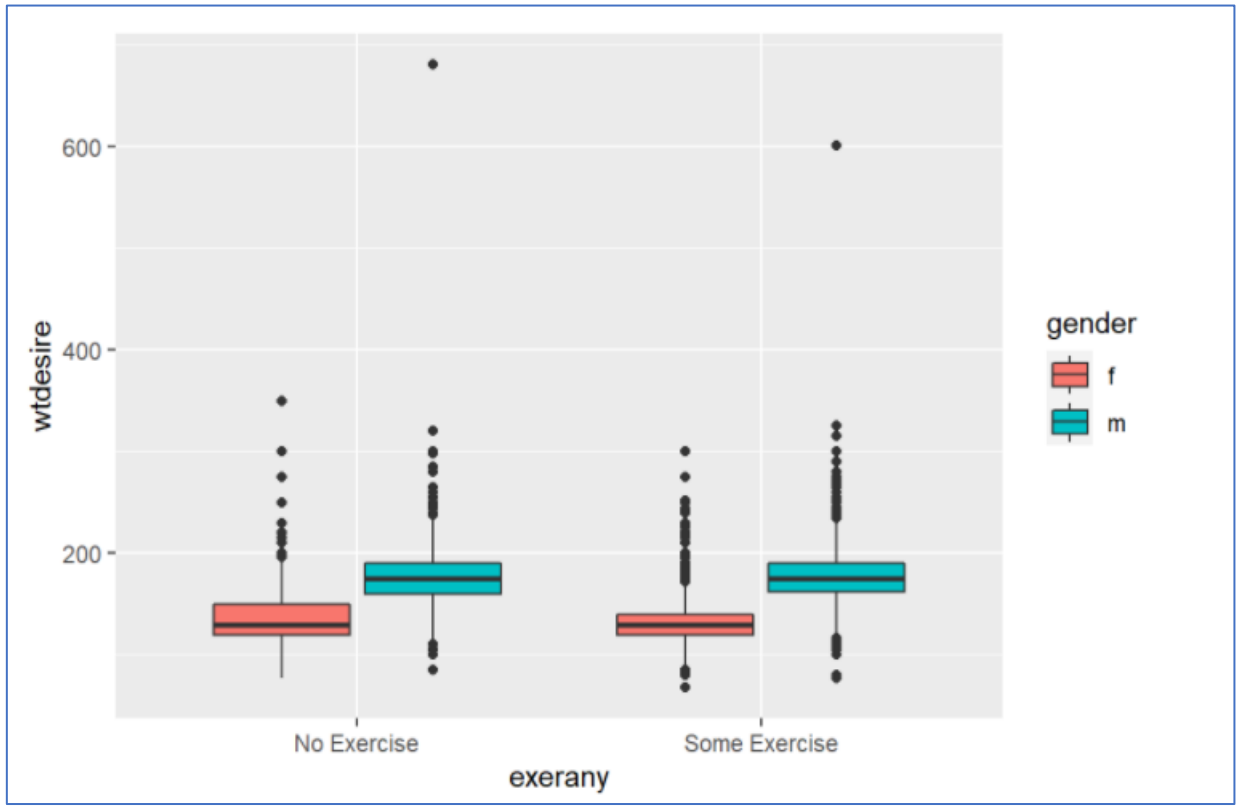
Box Plot

- It is used to plot a continuous variable or a combination of categorical and continuous variables.
- This plot is useful for visualizing the spread of the data and detect outliers.
- It shows five statistically significant numbers- the minimum, the 25th percentile, the median, the 75th percentile and the maximum.
- It shows the distribution (shape, center, range, variation) of continuous variables.

Box Plot Anatomy



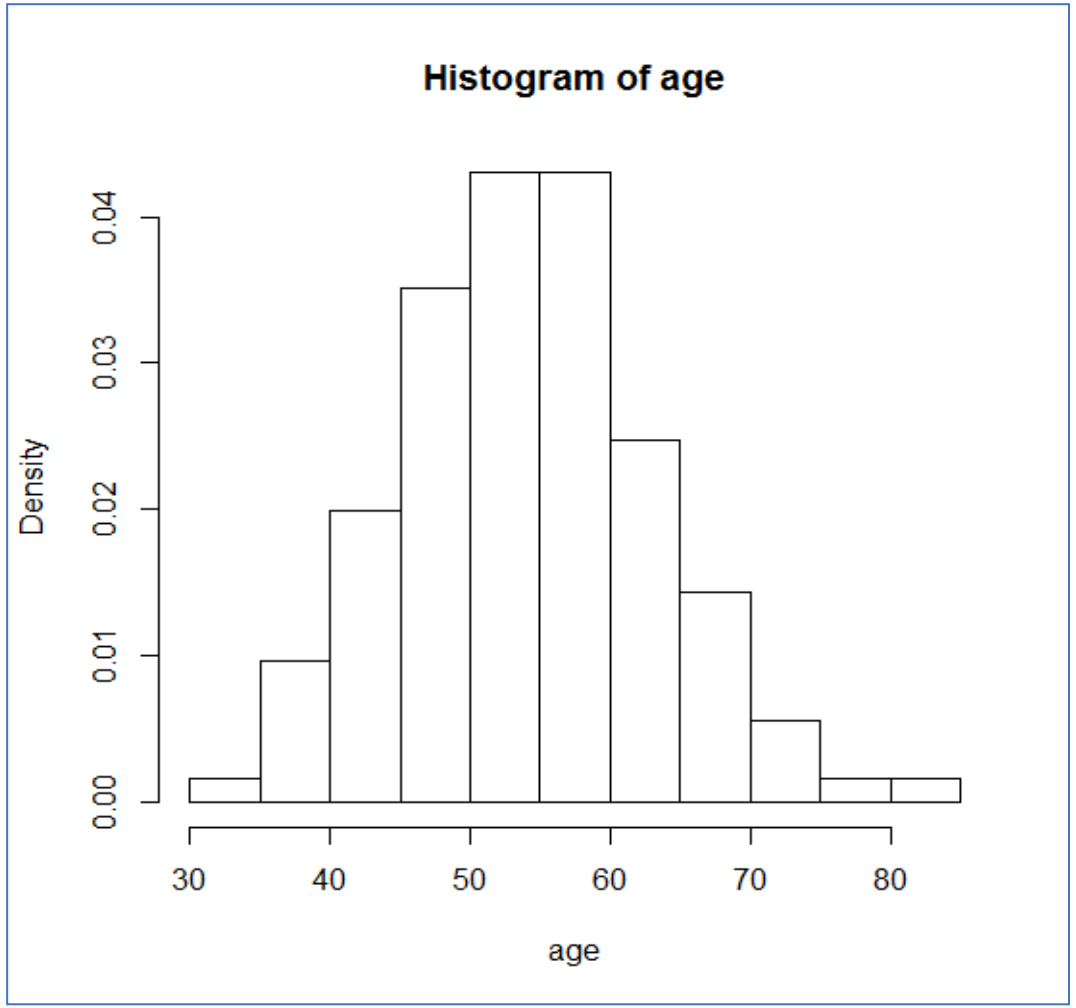
Grouped boxplots



Histogram

- Histogram is used to plot continuous variable.
- It breaks the data into bins and shows frequency distribution of these bins.
- We can always change the bin size and see the effect it has on visualization.

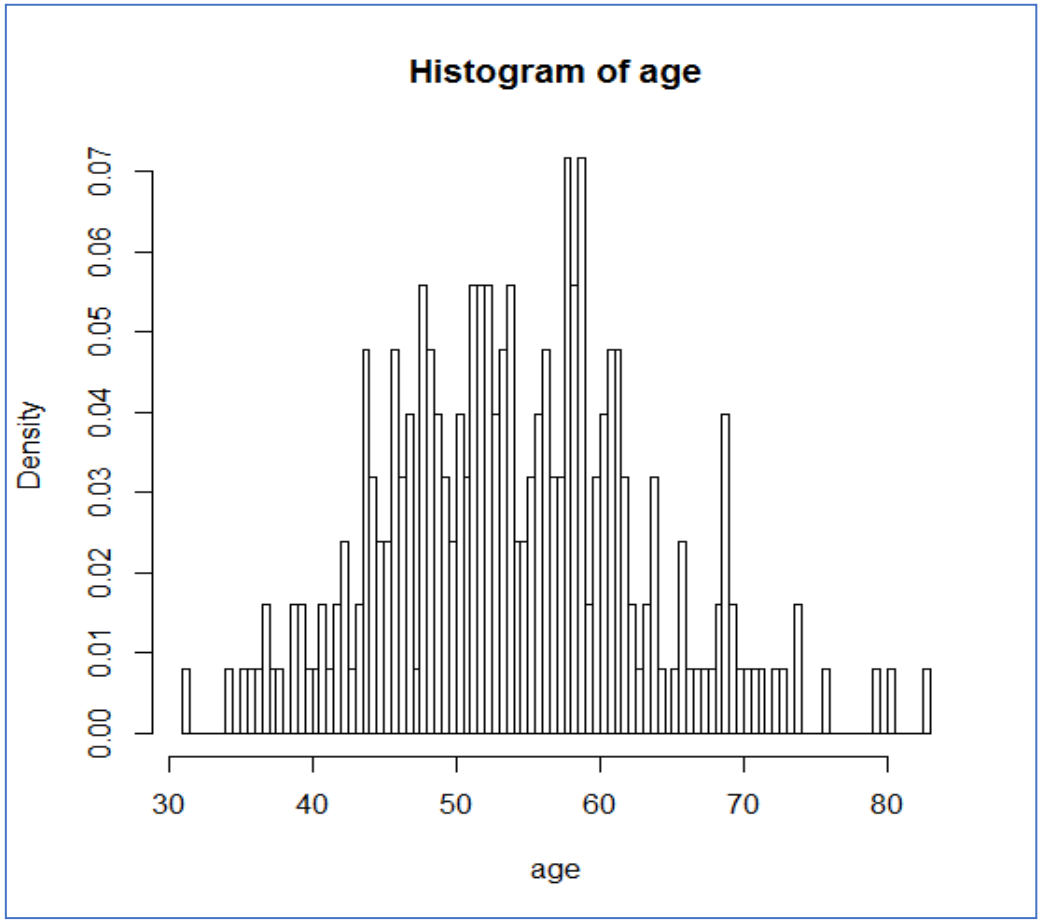
Histogram

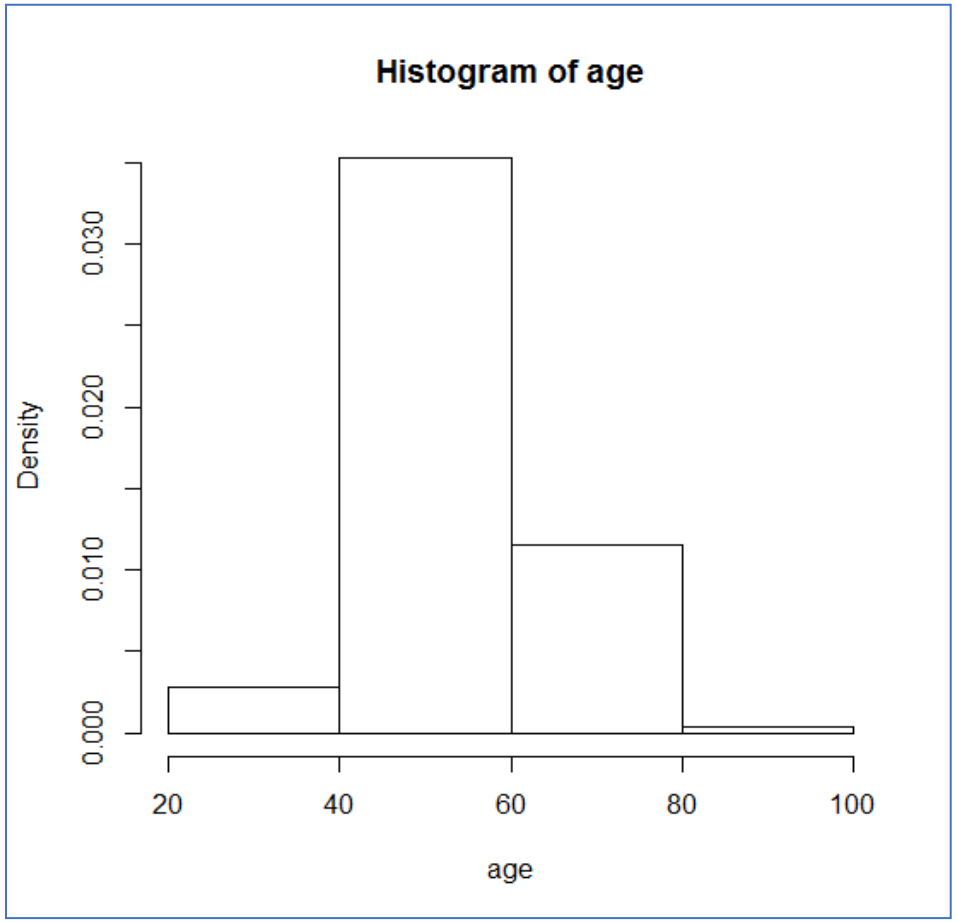


Note the shape: Although symmetric, slightly skewed to the right

10 “breaks”, age is categorized in 11 groups

Histogram



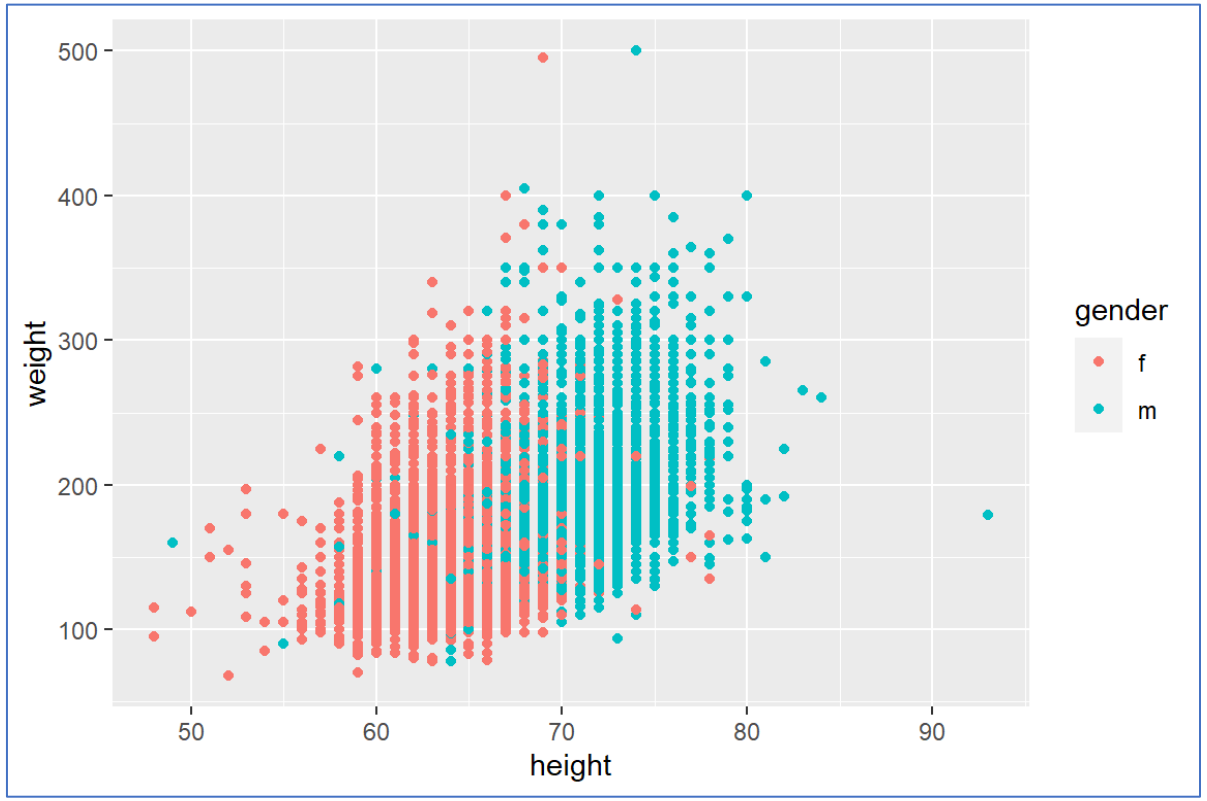


Use 100 “breaks”, instead

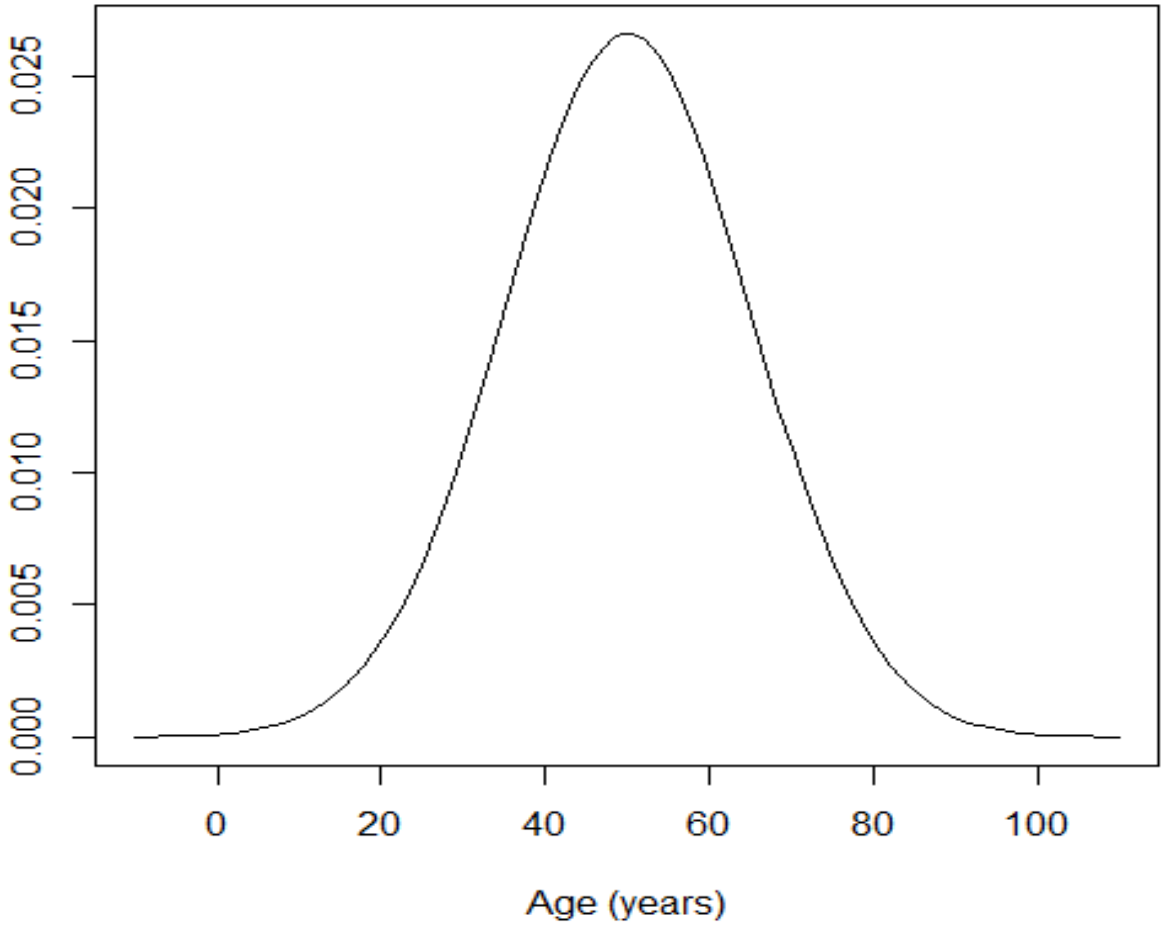
This is too much detail! We are only interested on the shape of the distribution…

Scatterplot

- Two continuous variables



The Normal Distribution



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

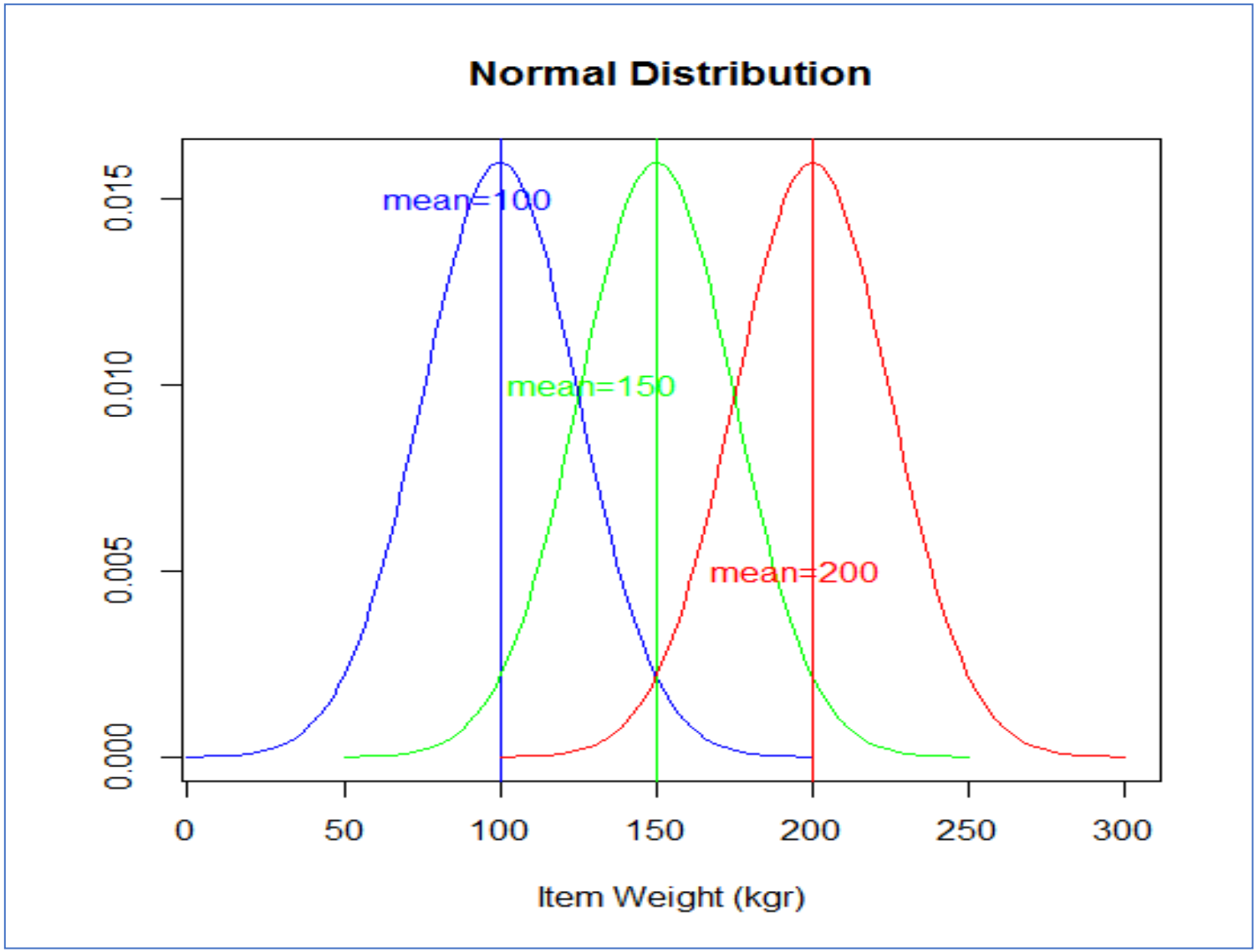
Note constants:
π=3.14159
e=2.71828

Properties of the Normal Distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center (around the mean).
- Half of the values are to the left of the mean and half of the values are to the right.
- The area under the curve is equal to 1.

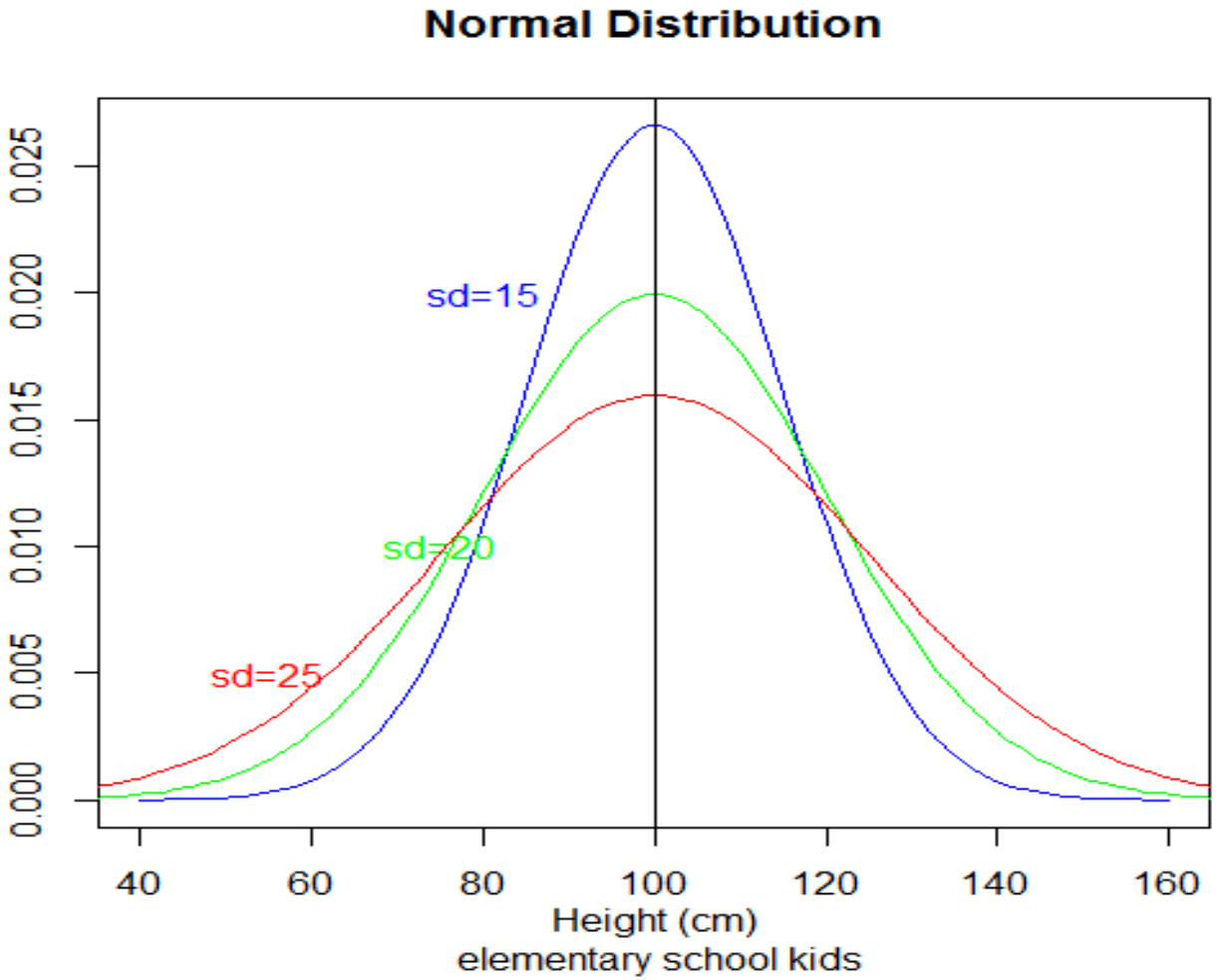
NOTE: We cannot use only these properties to declare that our data follow the Normal Distribution – we need to use a normality test!

Normal Distribution



For different means, the curve moves
to the right for larger means
to the left for smaller means

Normal Distribution



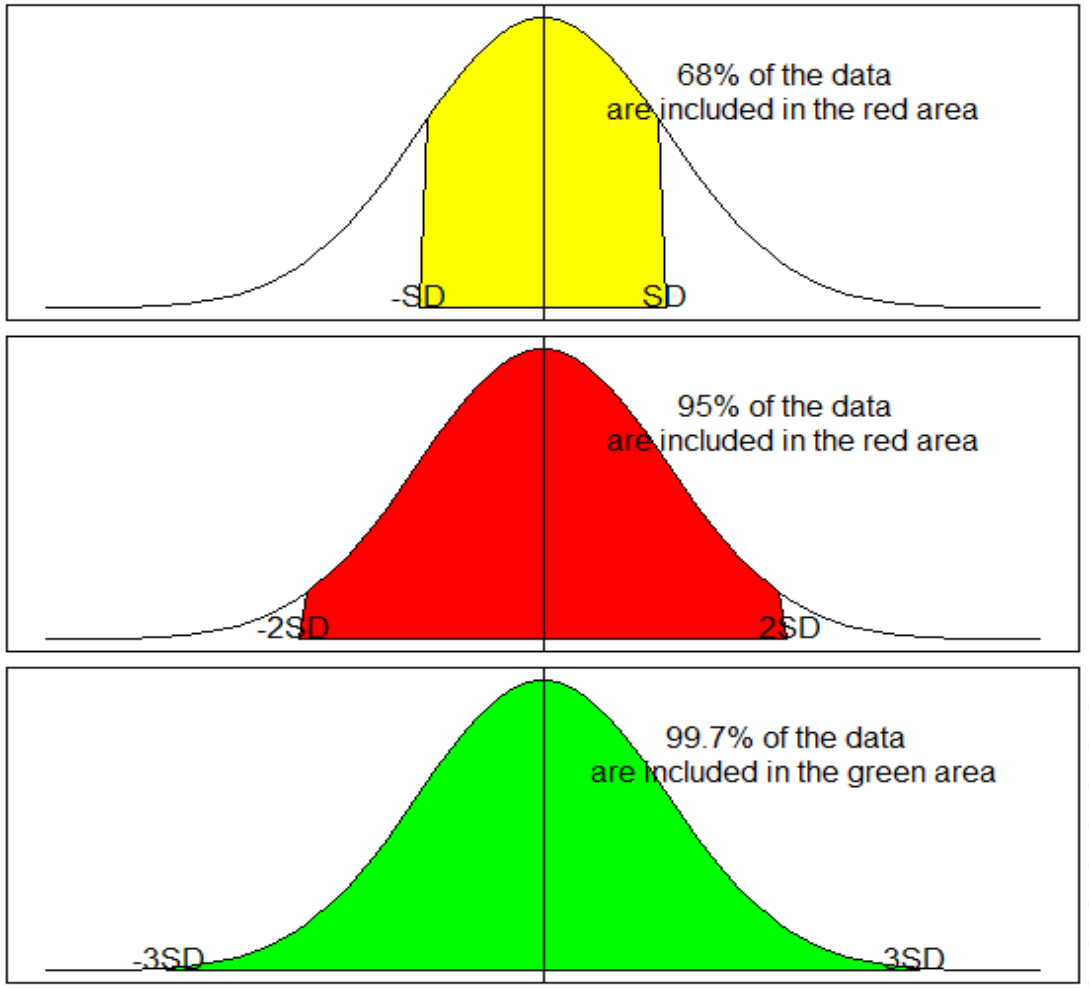
A smaller standard deviation indicates that the data is tightly clustered around the mean, the curve is taller.

A larger standard deviation indicates greater variability in our data, the curve is flatter and wider.

Empirical Rule

- The area between $\mu-\sigma$ and $\mu+\sigma$ is about 68%.
- The area between $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%.
- The area between $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%.

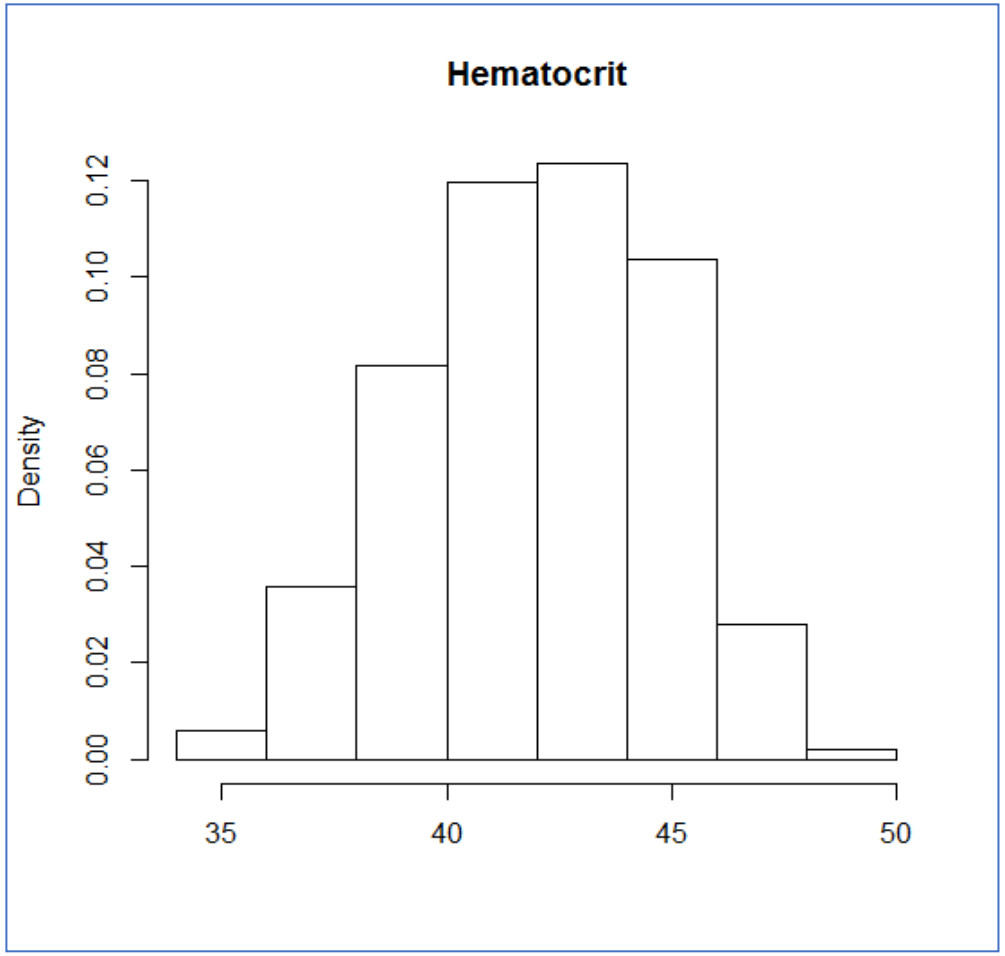
Almost all values fall within 3 standard deviations!

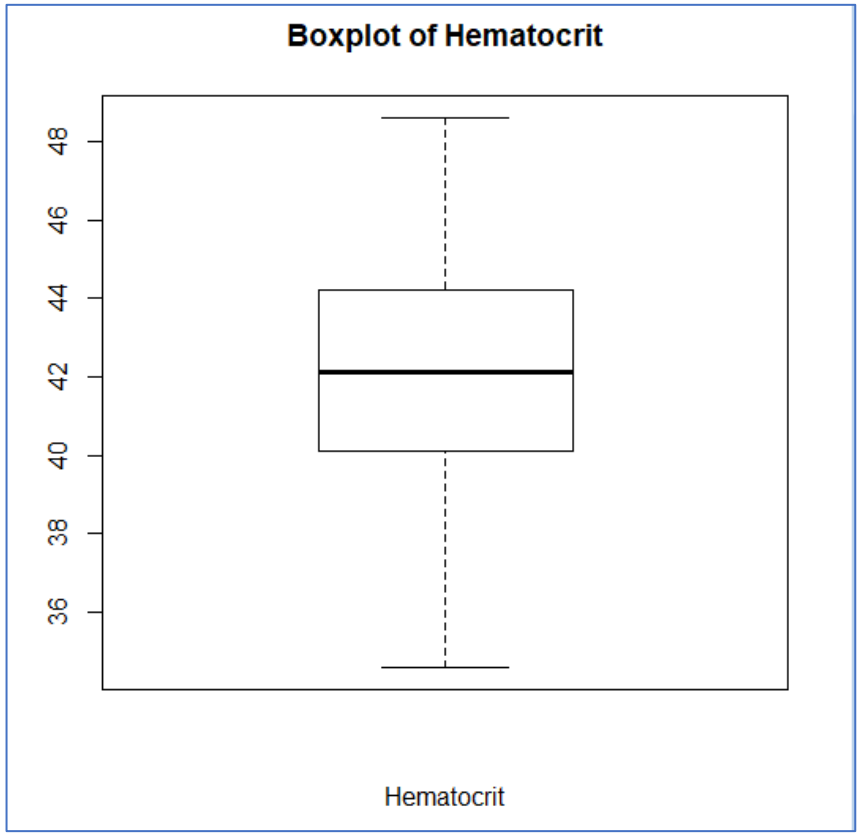


Are my data normally distributed?

- Look at the histogram! Does it appear bell shaped?
- Compute descriptive summary measures—are mean, median, and mode similar?
- Run tests of normality (such as Shapiro-Wilk). But, be cautious, highly influenced by sample size!

Are my data normally distributed (I)?



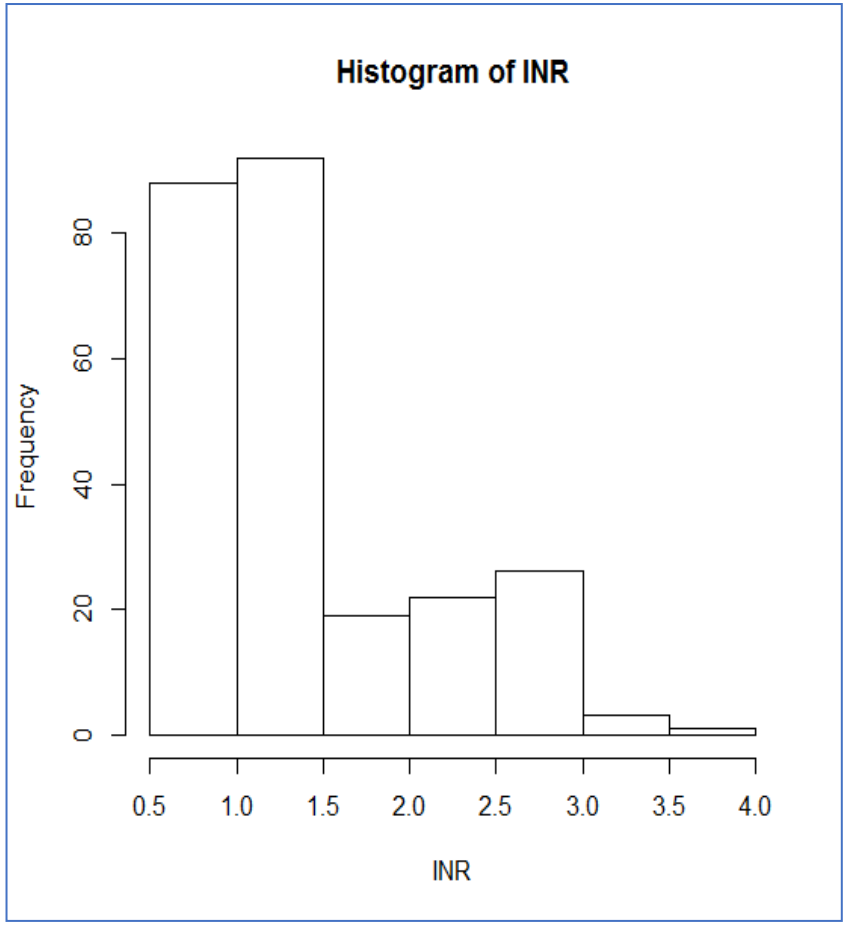


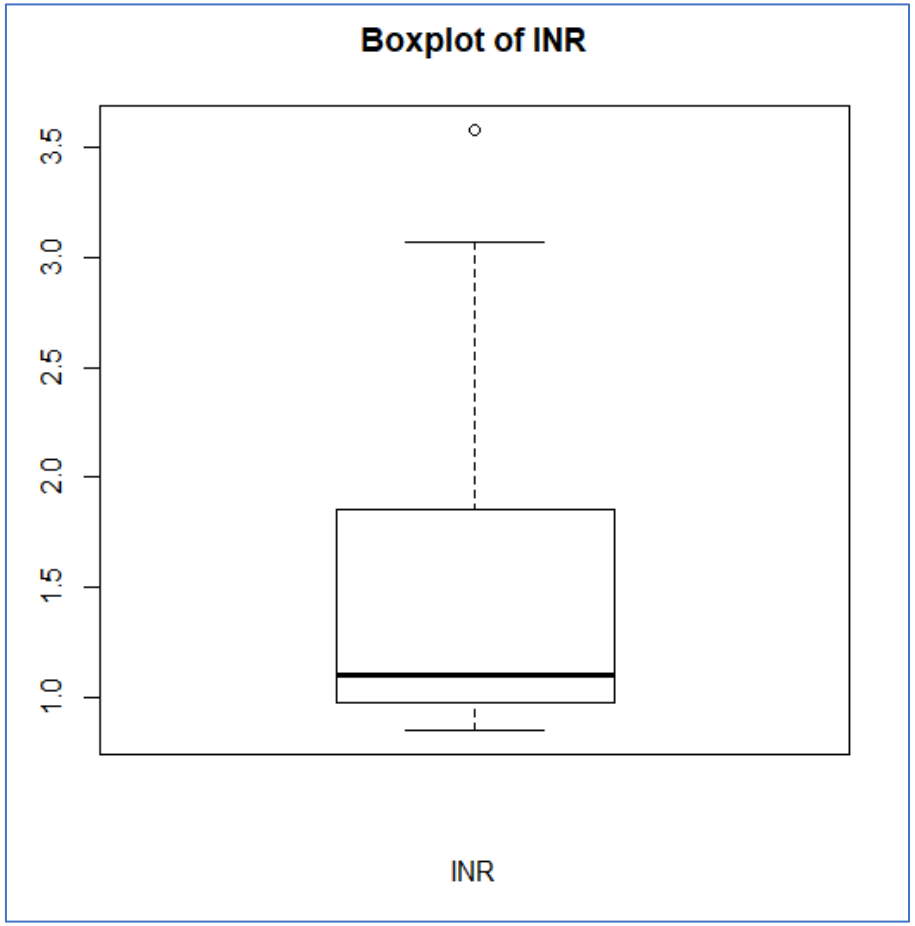
Median = 42.10

Mean = 42.03

Mode = 42

Are my data normally distributed (II)?





Median = 1.1

Mean = 1.4

Mode = 0

Formal tests for normality

- For a formal test for normality, we can perform a Shapiro-Wilk test.

  H_0: normal

  H_a: not normal

- Results: (Shapiro-Wilk)

Hematocrit: No evidence of non-normality (p=0.136 s-w)

INR: Strong evidence for non-normality (p<0.001)

- All indication converge to the conclusion that Hematocrit **can** be assumed to be normally distributed, while INR **cannot** be assumed to be normally distributed