Patrice Koehl

Data Exploration

Pre-processing data

With "help" from CS109, Harvard

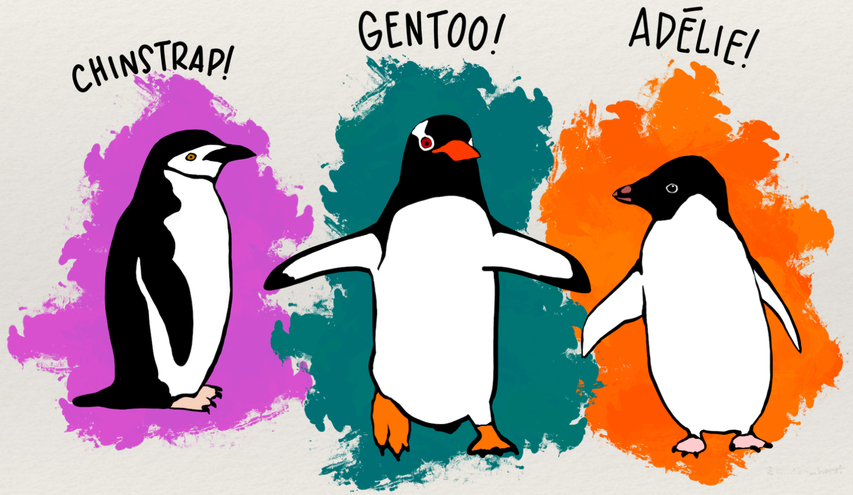
Why data exploration is important

- Ensure your data is as expected/valid/appropriate for the task
- Provides insights into a dataset
- Extract/determine important variables/attributes/features
- Detect outliers and anomalies
- Test underlying assumptions
- Make informed decisions in developing models

Example

The Palmer Archipelago (Antarctica) penguin dataset:

contains size measurements for three penguin species observed on three islands in the Palmer Archipelago, Antarctica.



Artwork by Allison Horst

Reference:
Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081.

The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Get The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Where?

Get The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Where? https://allisonhorst.github.io/palmerpenguins/reference/penguins.html

Get The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstra		50.8	19.0	0	Male

- Credible/Trustworthy?
- Original, or already preprocessed data?

Where?

https://allisonhorst.github.io/palmerpenguins/reference/penguins.html

Explore The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Explore The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Does it contain the necessary information?

Explore The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Missing data: What should we do?

Explore The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Are the data type OK?

Explore The Data

Penguin	Island	Beak length (mm)	Beak width (mm)	Mass (gr)	Sex
Adelie	Torgersen	39.1	18.7	3750	Male
Adelie	Biscoe	35.9	NaN	1800	Female
Gentoo	Biscoe	-45.2	14.8	5400	Female
Chinstrap	Dream	50.8	19.0	0	Male

Are the values reasonable?

Basic Data Analysis

➢Mean … the average value

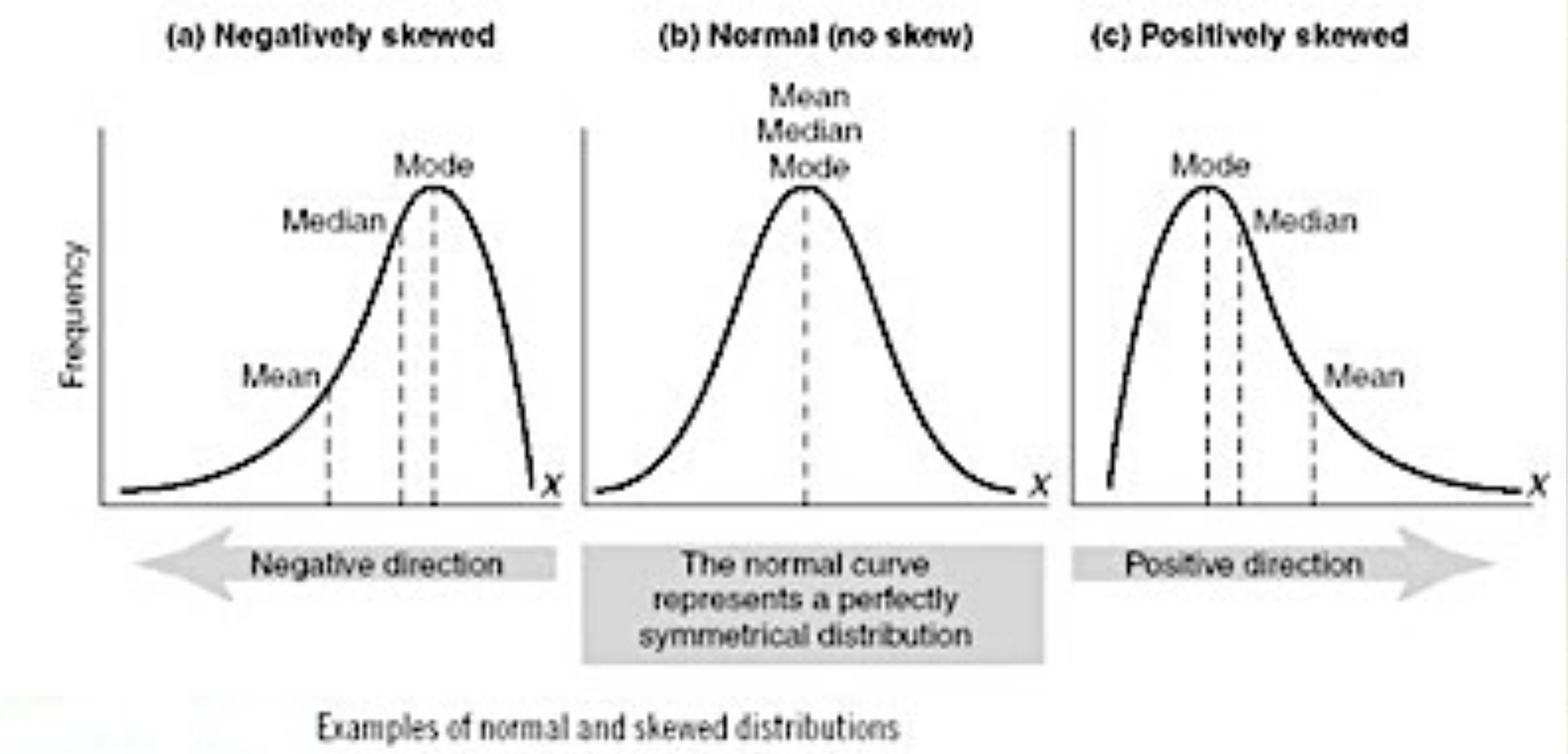
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

➢Median … the value that lies in the middle after ranking all the values

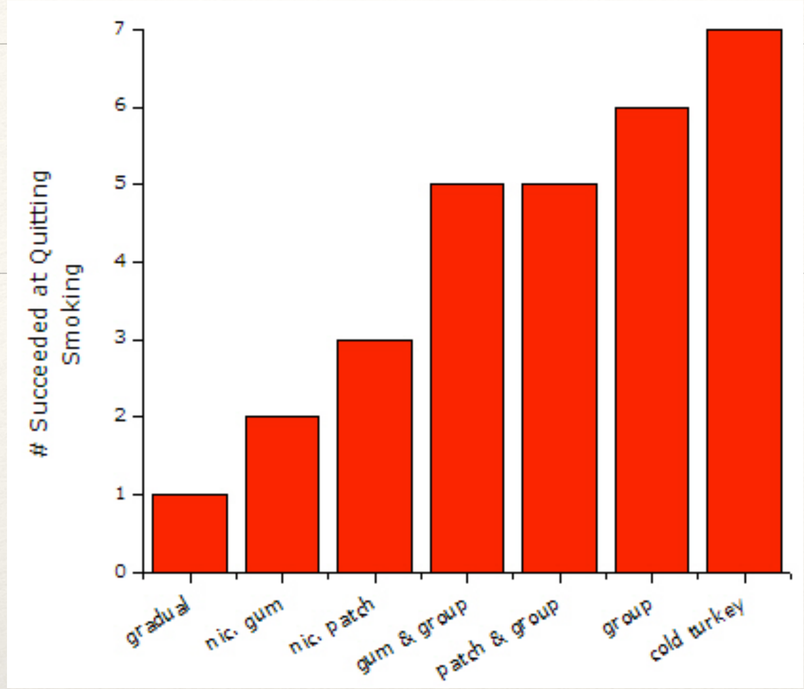
$$X_M = \begin{cases} X_{n/2+1} & \text{n odd} \\ \dfrac{X_{n/2} + X_{n/2+1}}{2} & \text{n even} \end{cases}$$

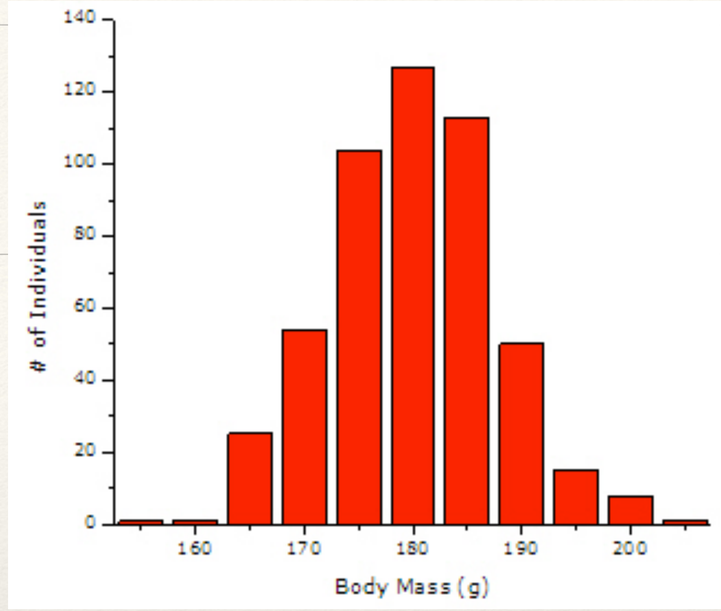
➢Mode … the most frequently occurring value(s)

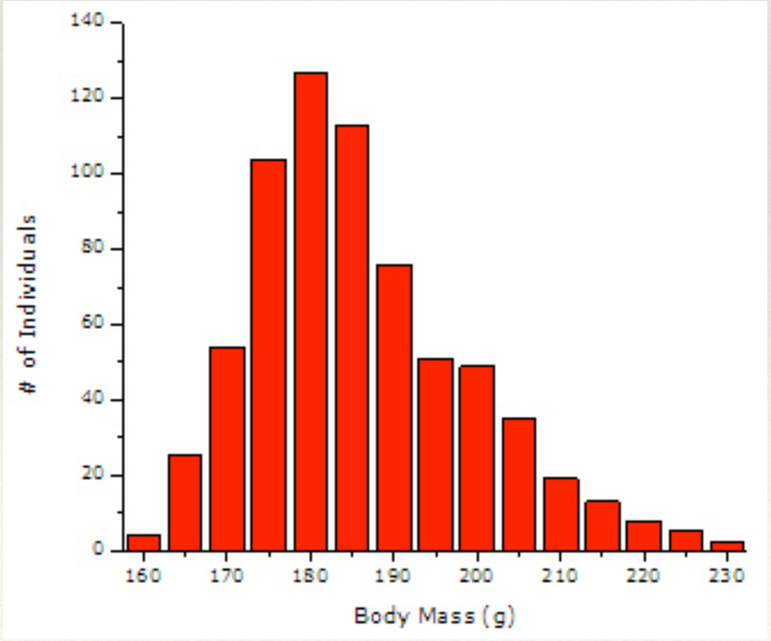
Basic Data Analysis

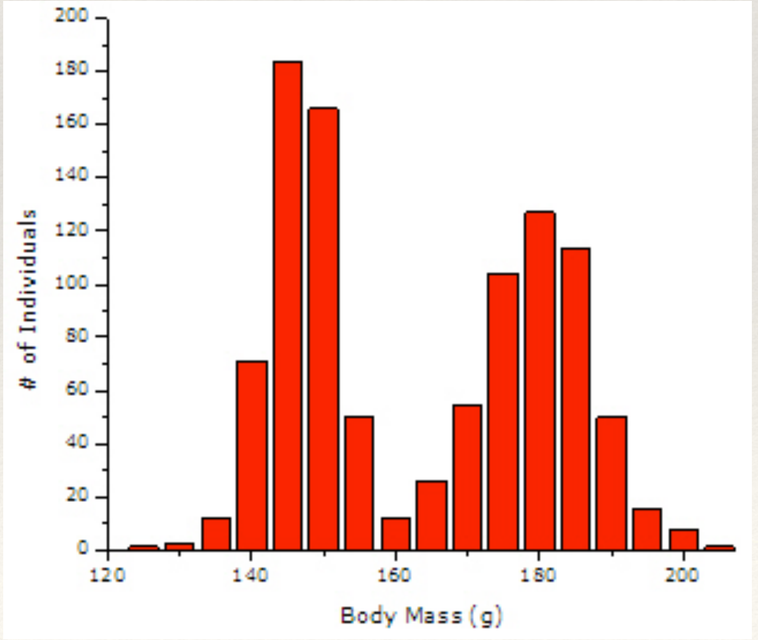


Examples of normal and skewed distributions

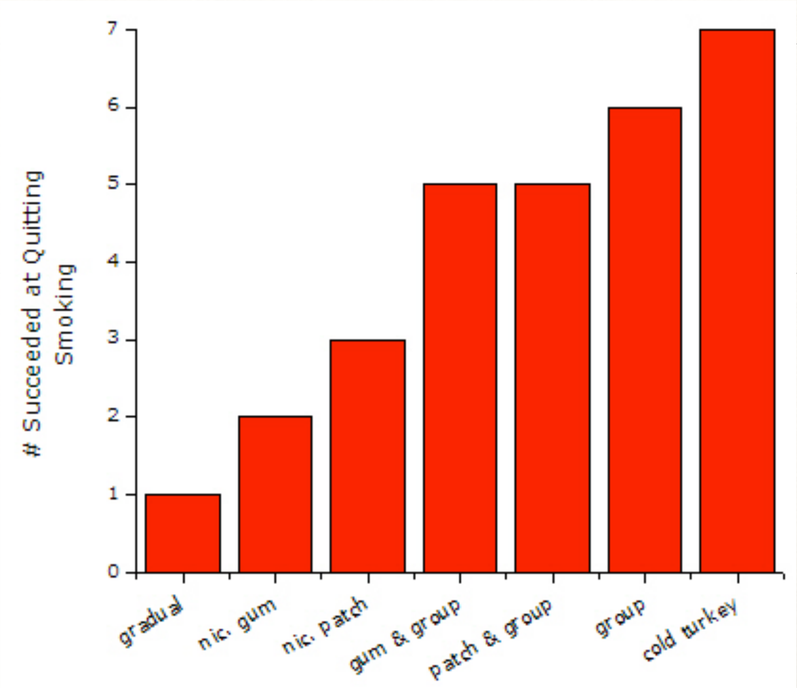

Succeeded at Quitting Smoking
0
1
2
3
4
5
6
7
gradual
nic. gum
nic. patch
gum & group
patch & group
group
cold turkey


# of Individuals
0
20
40
60
80
100
120
140
160
170
180
190
200
Body Mass (g)

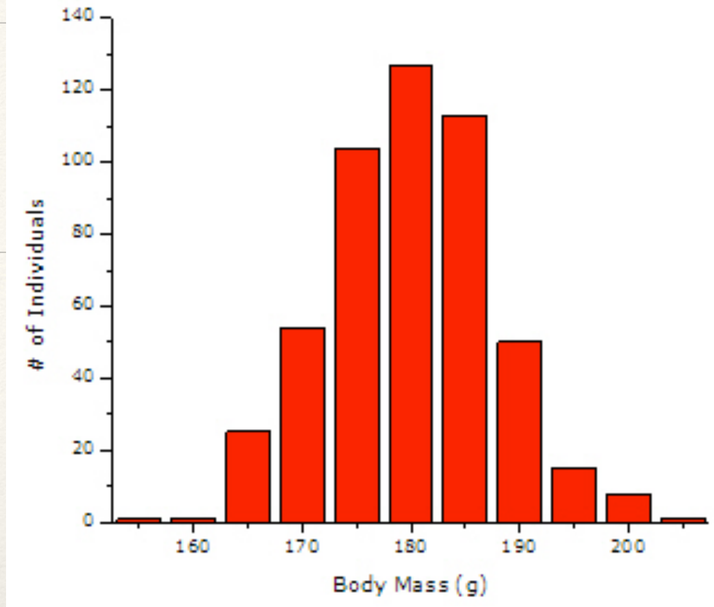

# of Individuals
0
20
40
60
80
100
120
140
160
170
180
190
200
210
220
230
Body Mass (g)


# of Individuals
0
20
40
60
80
100
120
140
160
180
200
120
140
160
180
200
Body Mass (g)

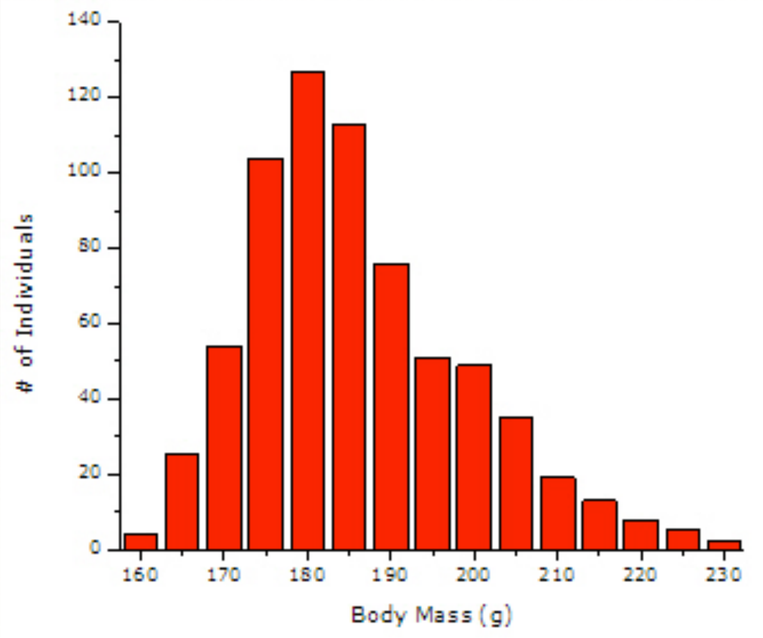
Mode



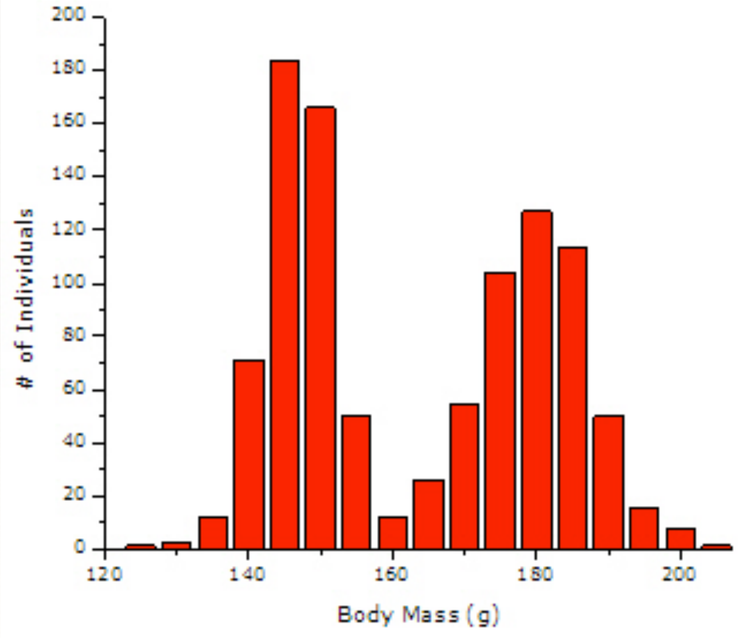
Mean



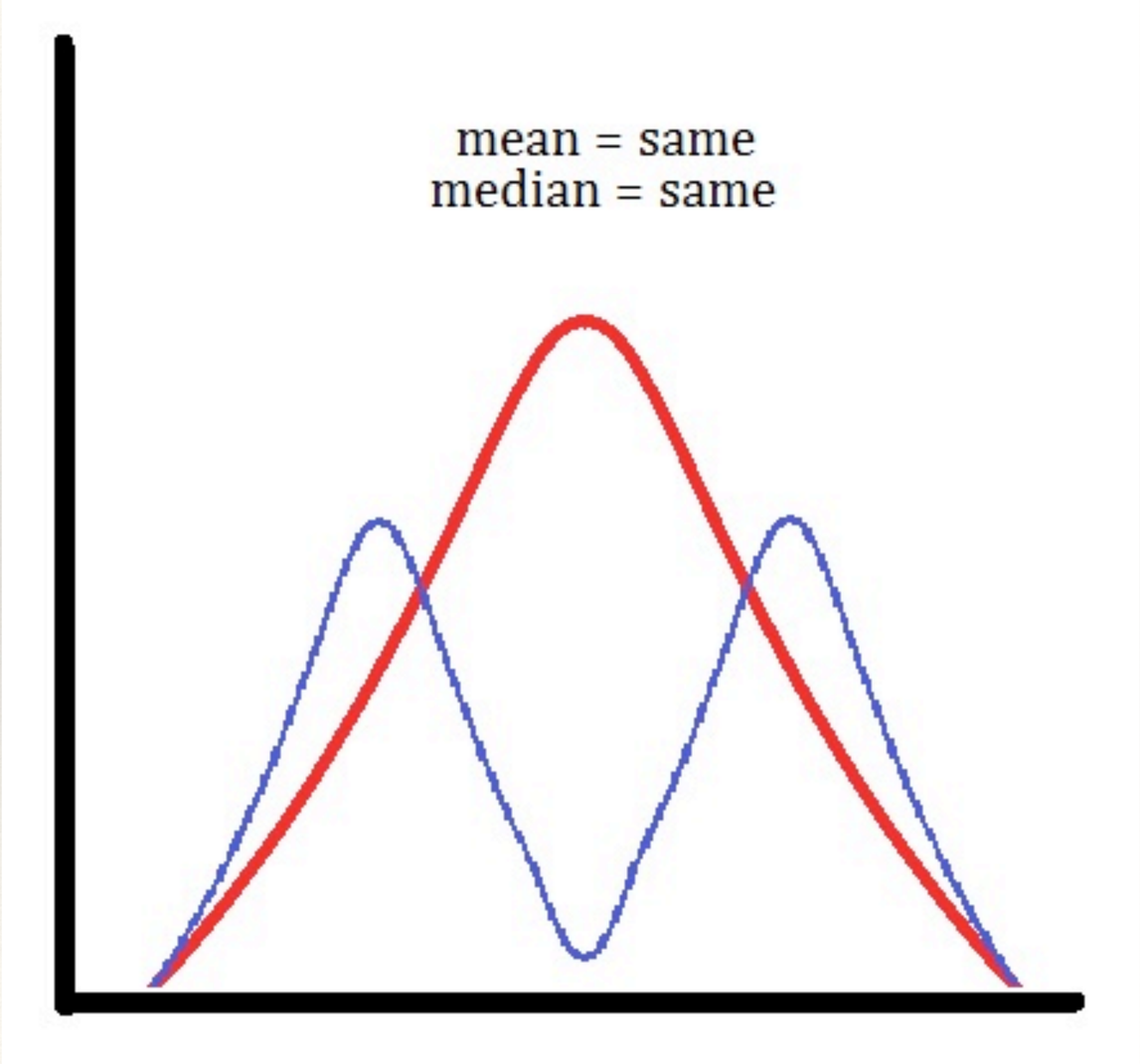
Median



None



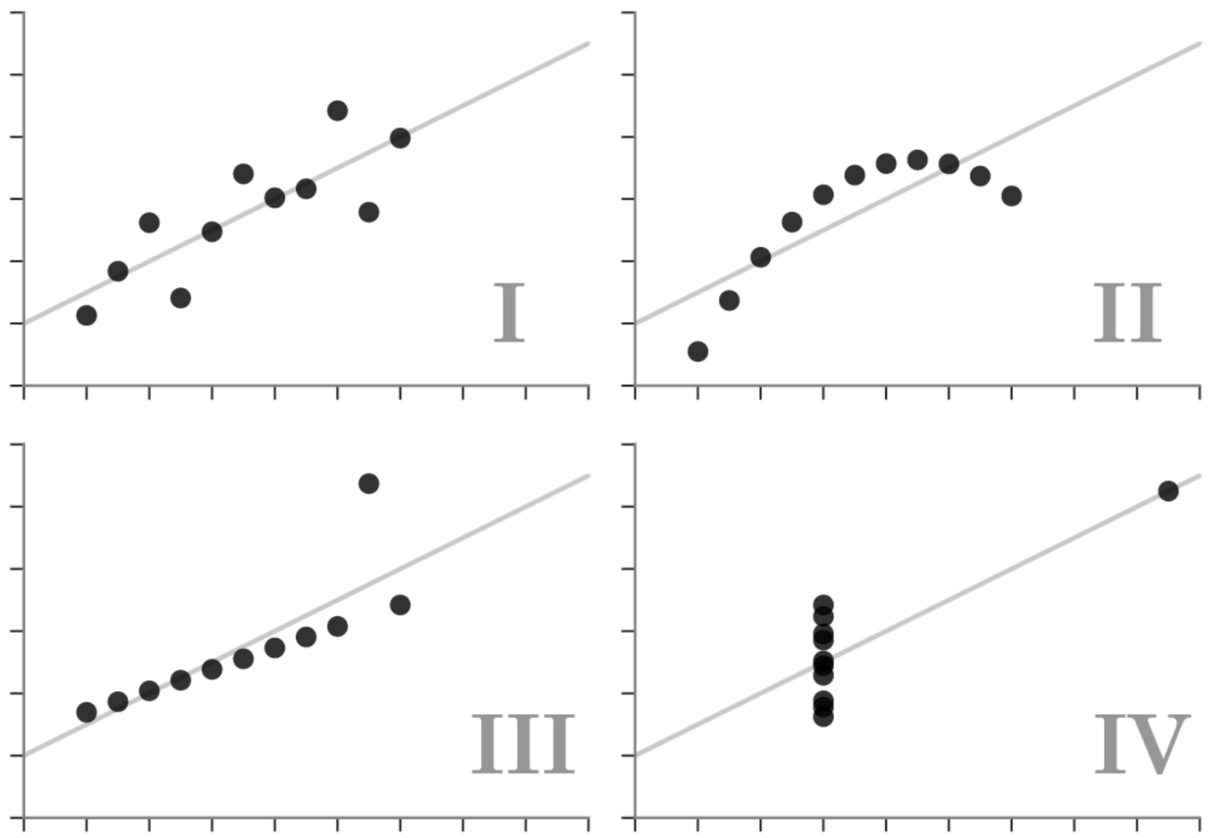
Attention: Danger!



Attention: Danger!

Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



Variation or Spread

- Range

$$Range = X_{Max} - X_{Min}$$

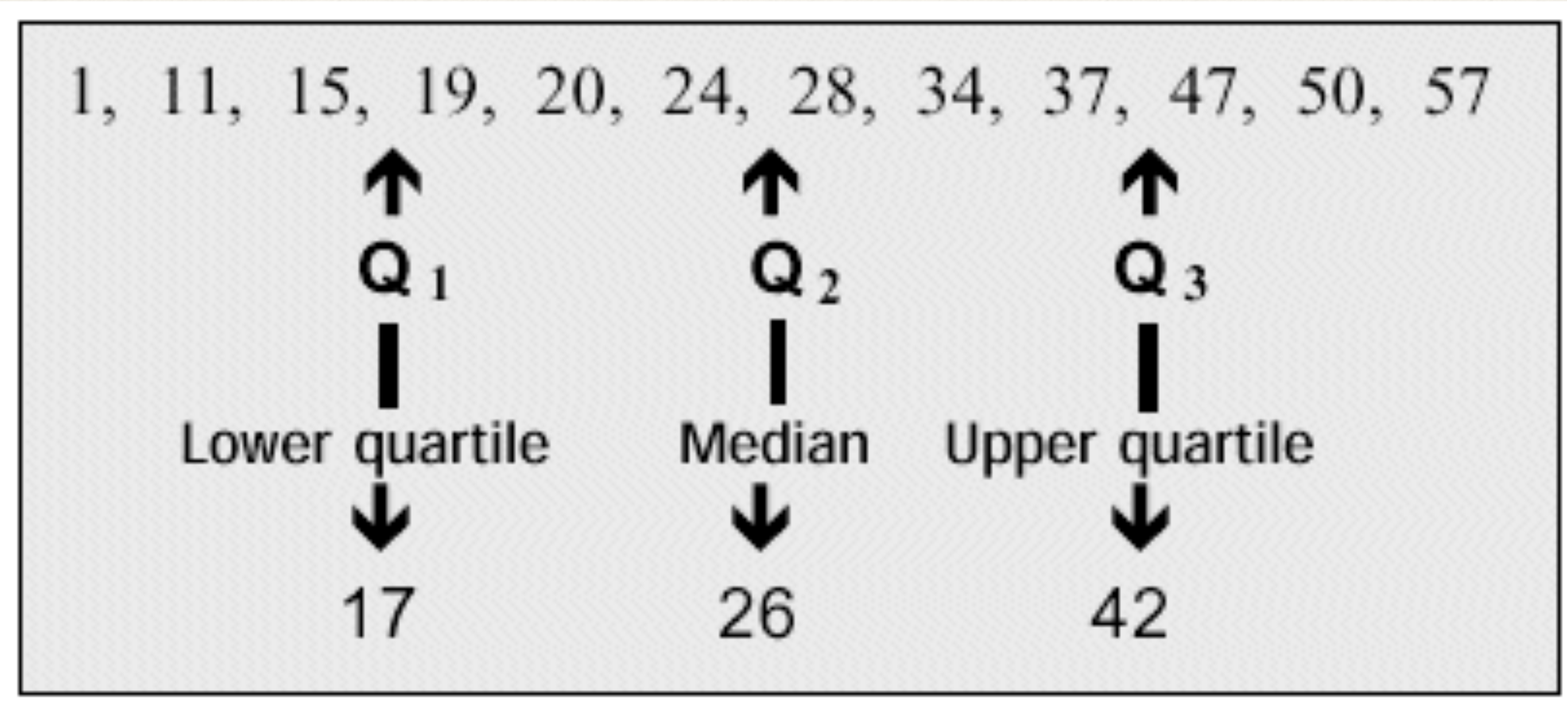
- Variance and Standard Deviation

$$Var(X) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

$$Std(X) = \sigma = \sqrt{Var(X)}$$

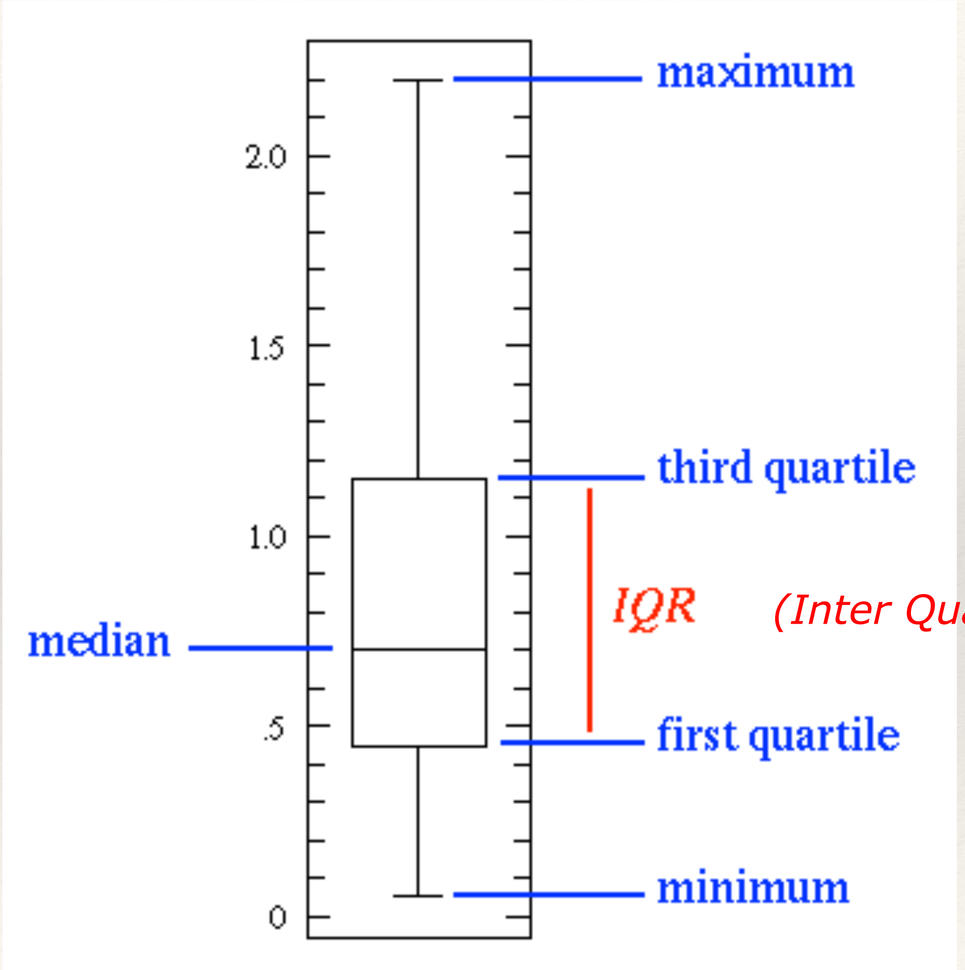
Variation or Spread

- Quartiles



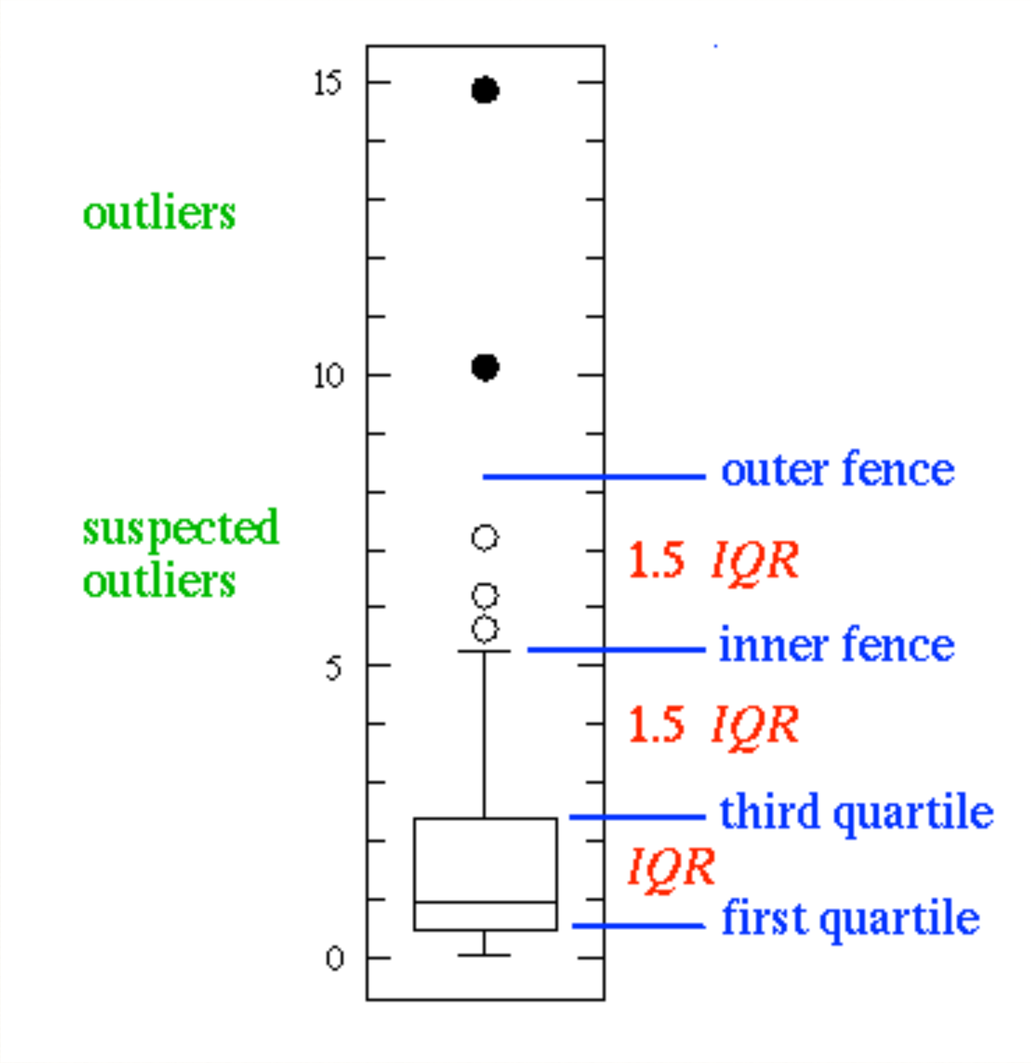
Variation or Spread

- Box plot



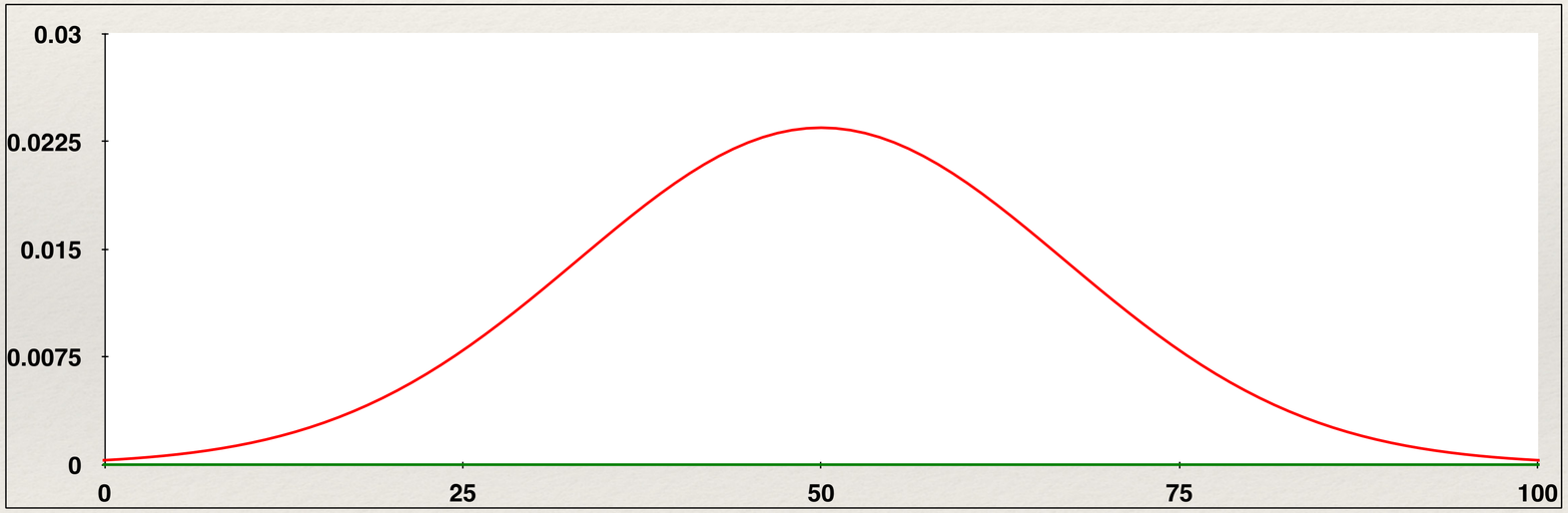
Detecting outliers

- Box plot



The normal distribution

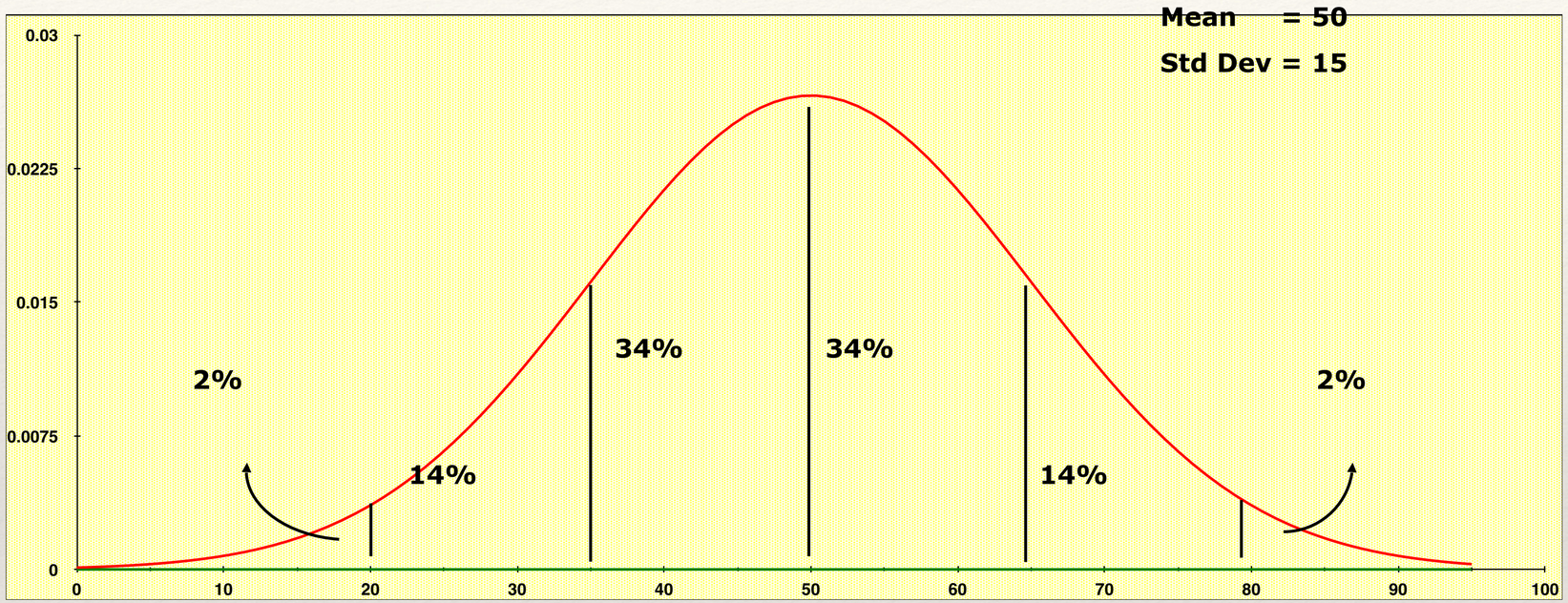
In everyday life many variables such as height, weight, shoe size and exam marks all tend to be normally distributed, that is, they all tend to look like:

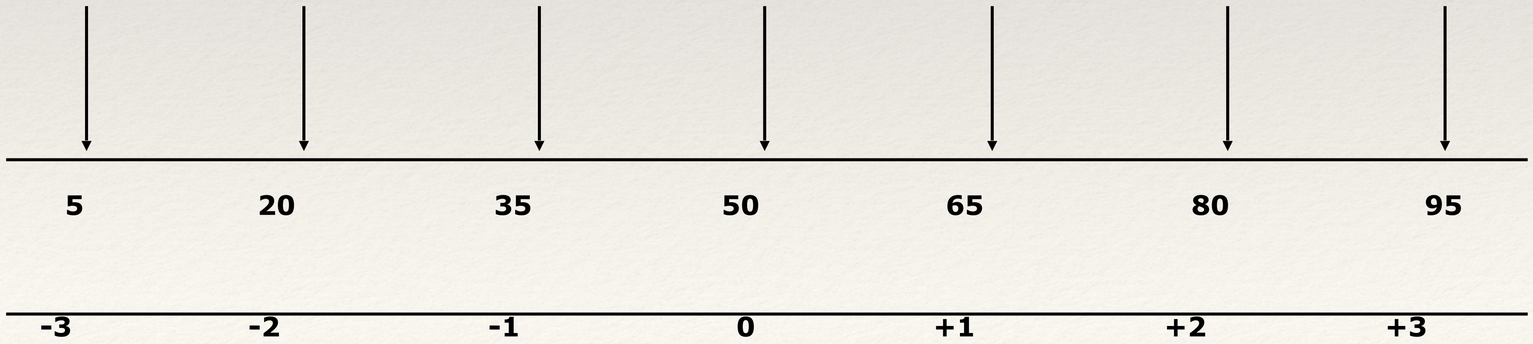


It is bell-shaped and symmetrical about the mean

The mean, median and mode are equal

The normal distribution





Beware!

A real example from a medical study* comparing the success rates of two treatments of kidney stones:

	Treatment A	Treatment B
Patients	78% (273/350)	**83% (289/350)**

*Charig et al, Br Med J, 292, 879 (1986)

Beware!

A real example from a medical study* comparing the success rates of two treatments of kidney stones:

	Treatment A	Treatment B
Small Stones	**93% (81/87)**	87% (234/270)
Large Stones	**73% (192/263)**	69% (55/80)
Patients	78% (273/350)	**83% (289/350)**

**Charig et al, Br Med J, 292, 879 (1986)*

Beware!

A real example from a medical study* comparing the success rates of two treatments of kidney stones:

	Treatment A	Treatment B
Small Stones	**93% (81/87)**	87% (234/270)
Large Stones	**73% (192/263)**	69% (55/80)
Patients	78% (273/350)	**83% (289/350)**

What is happening here?

**Charig et al, Br Med J, 292, 879 (1986)*