Data Visualisation & Interpretation

The art of reading datasets

Devert Alexandre
School of Software Engineering of USTC

Table of Contents

Descriptive statistics

descriptive statistics helps to give a general summary of data

Mean

Example of descriptive statistics quantity

arithmetic mean

$$\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i$$

Mean

Example of descriptive statistics quantity

arithmetic mean

$$\bar{a} = \frac{1}{n}(a_1 + a_2 + \cdots + a_n)$$

Mean

The *mean* is defined in $\mathbb{R}^n \Rightarrow$ geometric center

Mean computation

You think, it is easy to compute the mean ?

$0.1+0.1+0.1+0.1+0.1+0.1+0.1+0.1+0.1$

Mean computation

A naive summation algorithm will return this

```
>>> 0.1+0.1+0.1+0.1+0.1+0.1+0.1+0.1+0.1
0.8999999999999999
```

Mean computation

An *accurate* summation algorithm will return this

```
>>> import math
>>> math.fsum(0.1+0.1+0.1+0.1+0.1+0.1+0.1+0.1+0.1)
0.9
```

Mean computation

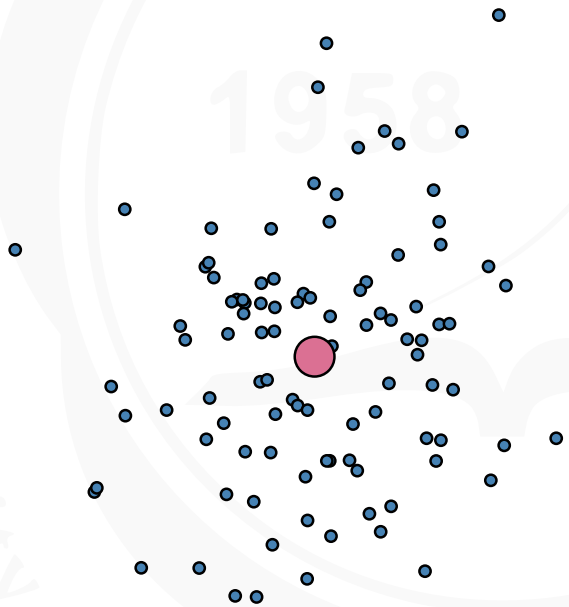
Algorithms like *Kahan summation algorithm* or *Shewchuk summation algorithm* reduces the numerical error

```
def KahanSum(data):
  s = 0.0
  c = 0.0
  for i in range(len(data)):
    y = data[i] - c
    t = s + y
    c = (t - s) - y
    s = t
  return s
```

Listing 1: Kahan summation

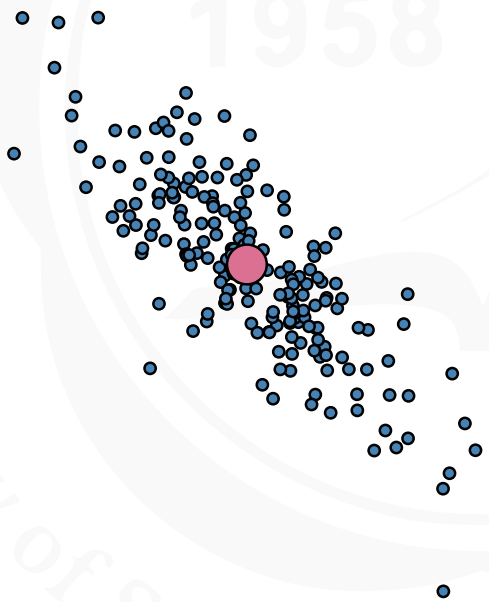
Central tendency

The mean is a measure of *central tendency* $\Rightarrow$ the main behaviour, the main value of some phenomenon



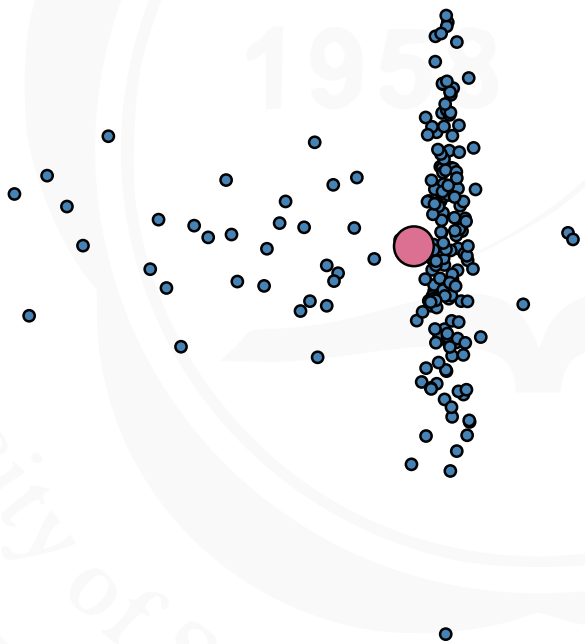
Central tendency

The mean is a measure of *central tendency* $\Rightarrow$ the main behaviour, the main value of some phenomenon



Mean robustness

The mean is not a *robust estimator* of the central tendency



Median

The *median* is the value such as 50% of the values are higher, 50% of the values are lower

$$a = [6, 1, 7, 9, 6, 3, 4, 5, 2]$$

$$a = [1, 2, 3, 4, \underline{5}, 6, 6, 7, 9]$$

$$\tilde{a} = 5$$

Median

The *median* is the value such as 50% of the values are higher, 50% of the values are lower

$$a = [6, 1, 7, 9, 6, 3, 4, 8, 5, 2]$$

$$a = [1, 2, 3, 4, \underline{5}, \underline{6}, 6, 7, 8, 9]$$

$$\tilde{a} = \frac{1}{2}(5 + 8) = 6.5$$

Median computation

To compute the median, you can

1. sort the list of samples
2.
   - if size is odd $\rightarrow \tilde{a} = a_{\frac{n+1}{2}}$
   - if size is even $\rightarrow \tilde{a} = \frac{1}{2}(a_{\frac{n}{2}} + a_{\frac{n+1}{2}})$

Note that it is for indexes starting from 1

Median computation

Let's code some python

```
def median(data):
  data.sort()
  if len(data) % 2 == 0:
    m = len(data) / 2
    return 0.5 * (data[m-1] + data[m])
  else:
    return data[(len(data) - 1) / 2]
```

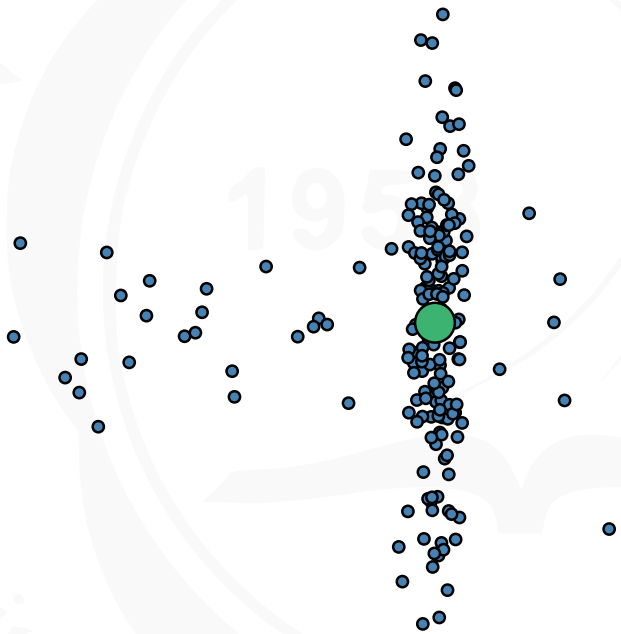
Median computation

Let's code some python

```
>>> a=[6,1,7,9,6,3,4,5,2]
>>> median(a)
5
```

Median computation

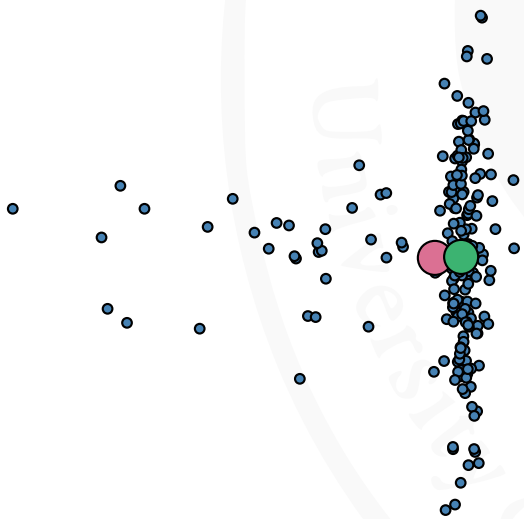
The *median* have an equivalent in $\mathbb{R}^n \Rightarrow$ median center



Compute the median for each dimension to get the median center

Median robustness

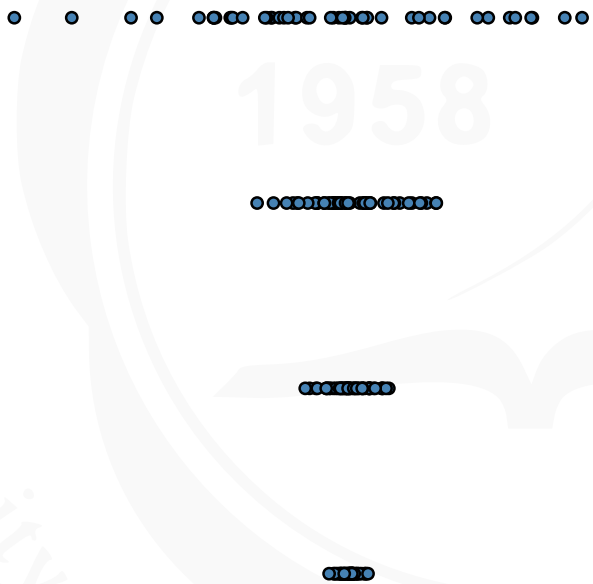
The median is a more *robust estimator* of the central tendency



- *green* is the median
- *pink* is the arithmetic mean

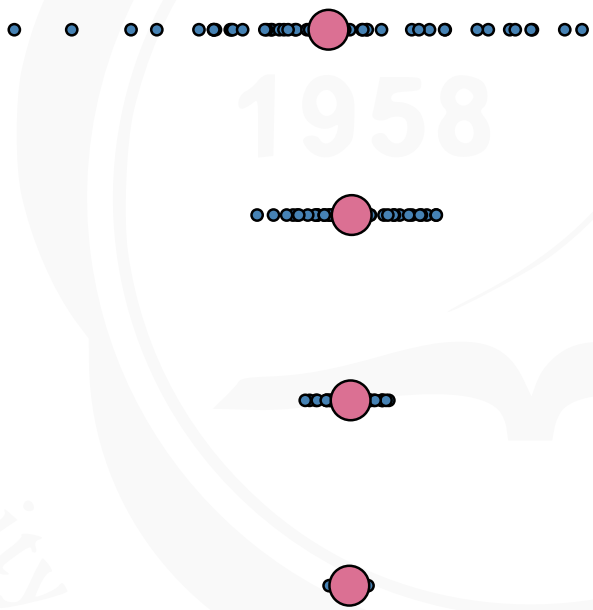
Statistical dispersion

The following datasets have the same central tendency



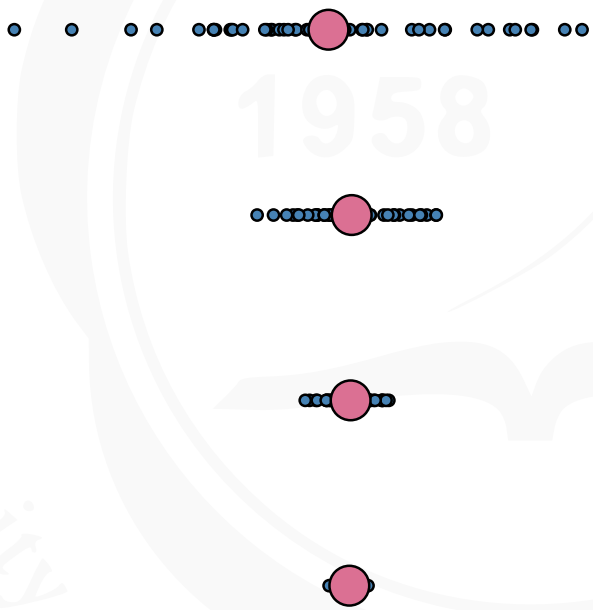
Statistical dispersion

The following datasets have the same central tendency



Statistical dispersion

But they have different *dispersions*



Standard deviation

A traditional measure of *dispersion* is the *standard deviation sigma*

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{N} (a_i - \bar{a})^2$$

Standard deviation computation

Robust computation of the standard deviation $\Rightarrow$
Knuth-Welford algorithm

```
def stdDev(data):
  n = 0
  mean = 0
  M2 = 0
  meanEstimate = math.fsum(data) / len(data)

  for x in data:
    y = x - meanEstimate
    n = n + 1
    delta = y - mean
    mean = mean + delta / n
    M2 = M2 + delta * (y - mean)

  return math.sqrt(M2 / (n - 1))
```

Standard deviation

Standard deviation suffers from the same robustness issues as mean. We will look why, later.

Quartiles

The *lower quartile* or *first quartile* is the value such as 75% of the values are higher, 25% of the values are lower

$$a = [6, 1, 2, 7, 9, 6, 3, 4, 5, 2, 6]$$

$$a = [1, 2, \underline{2}, 3, 4, 5, 6, 6, 6, 7, 9]$$

$$q_1 = 2$$

Quartiles

The *higher quartile* or *third quartile* is the value such as 25% of the values are higher, 75% of the values are lower

$$a = [6, 1, 7, 9, 6, 3, 4, 5, 2, 6]$$

$$a = [1, 2, 2, 3, 4, 5, 6, 6, \underline{6}, 7, 9]$$

$$q_3 = 6$$

Quartiles

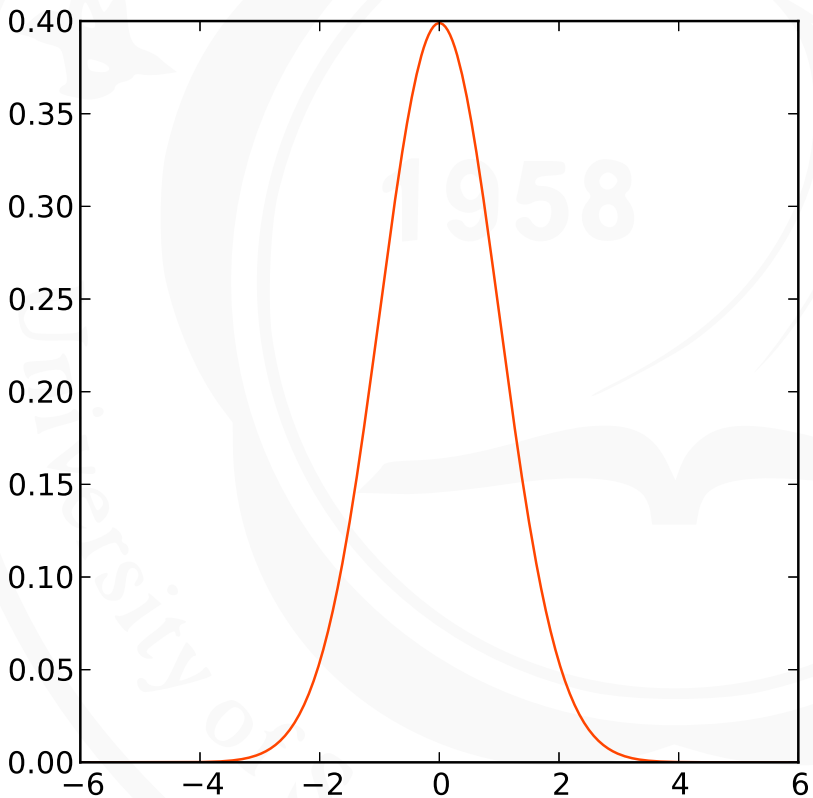
Where is the *second quartile* ? $\Rightarrow$ it's the median

Interquartile range

The difference $Q3 - Q1$ is the *interquartile range* or IQR
$\Rightarrow$ it's a more robust dispersion measure

normal distribution

A model for random variables, with 2 parameters μ and σ



normal distribution

The normal distributions have 2 parameters μ and σ.

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

This is the *probability density* of the normal distribution.

normal distribution

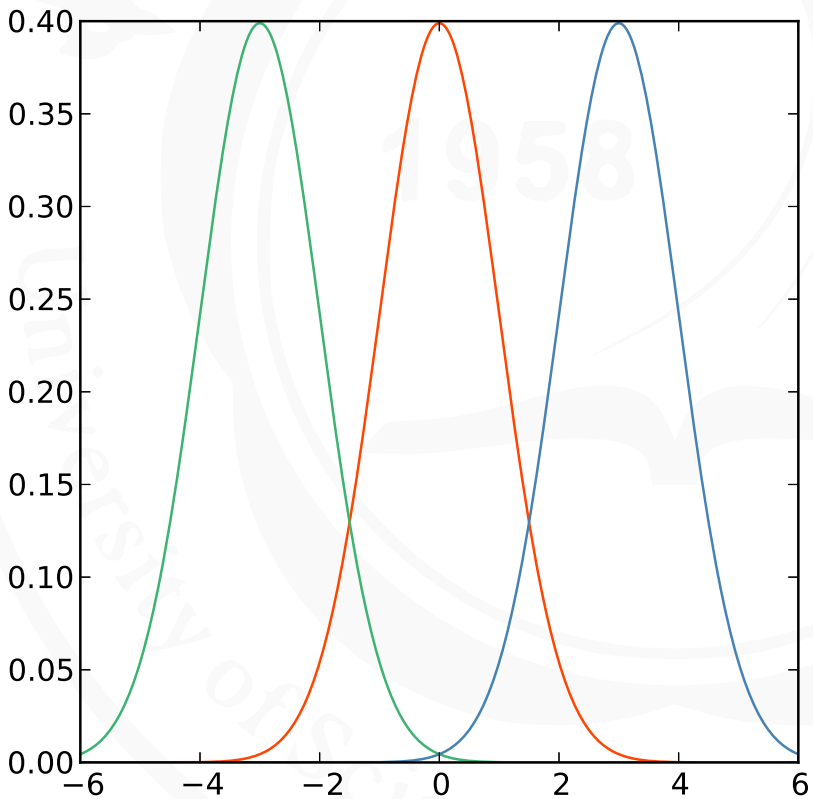
The normal distributions have 2 parameters μ and σ.

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

It tells the probability for x to appear, according to this distribution.

normal distribution

μ is the mode, the central tendency of the *normal* distribution



normal distribution

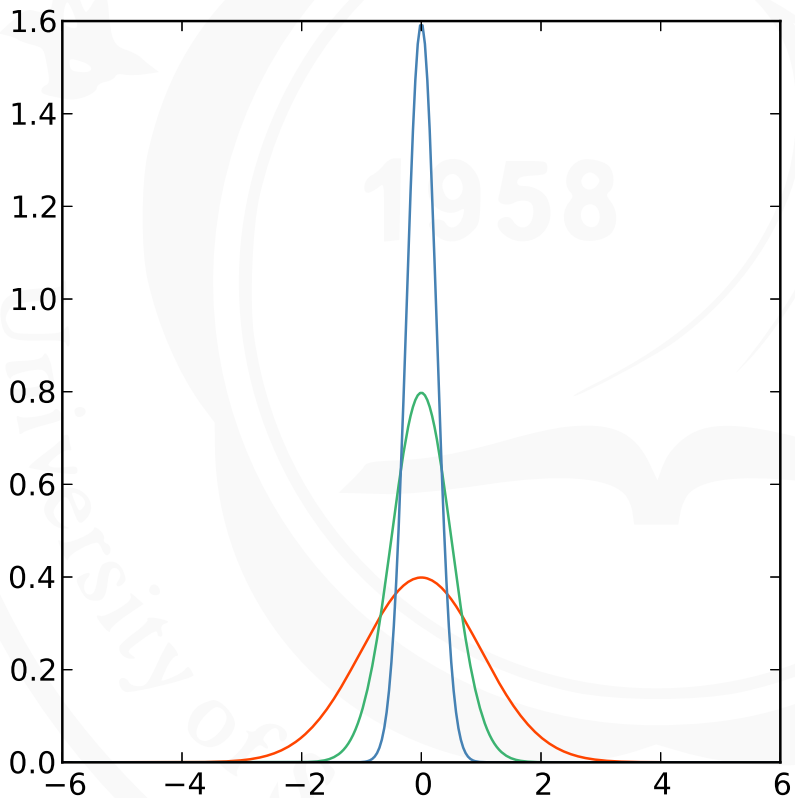
If some data are following a *normal* distribution, then

$$\mu = \bar{a}$$

The more sample, the more "true" it will be

normal distribution

σ controls the shape of the *normal* distribution



normal distribution

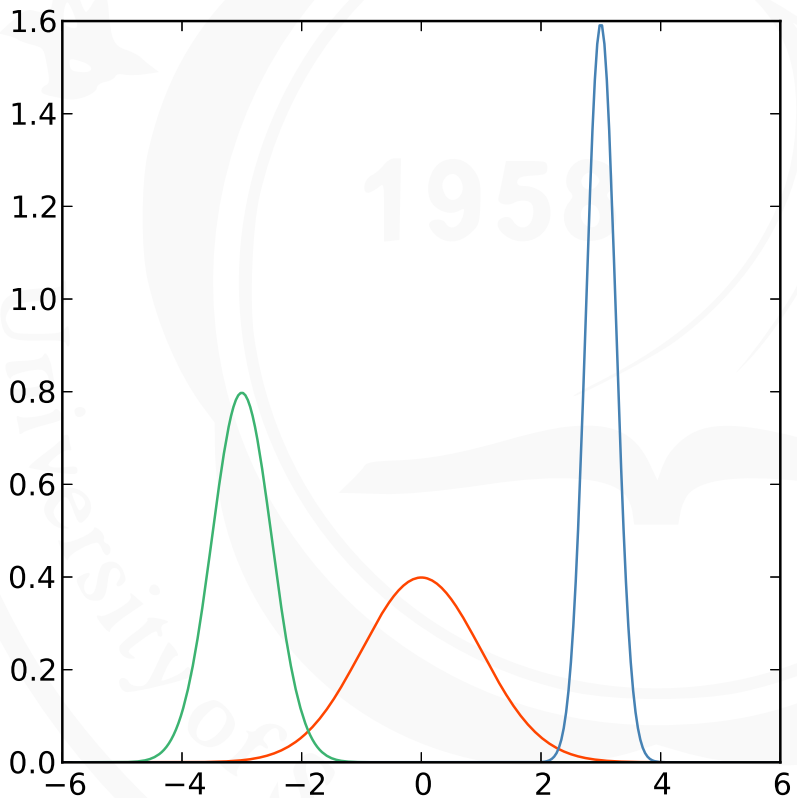
If some data are following a *normal* distribution

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{N}(a_i - \bar{a})^2$$

The standard deviation comes from here $\Rightarrow$ dispersion of a *normal* distribution

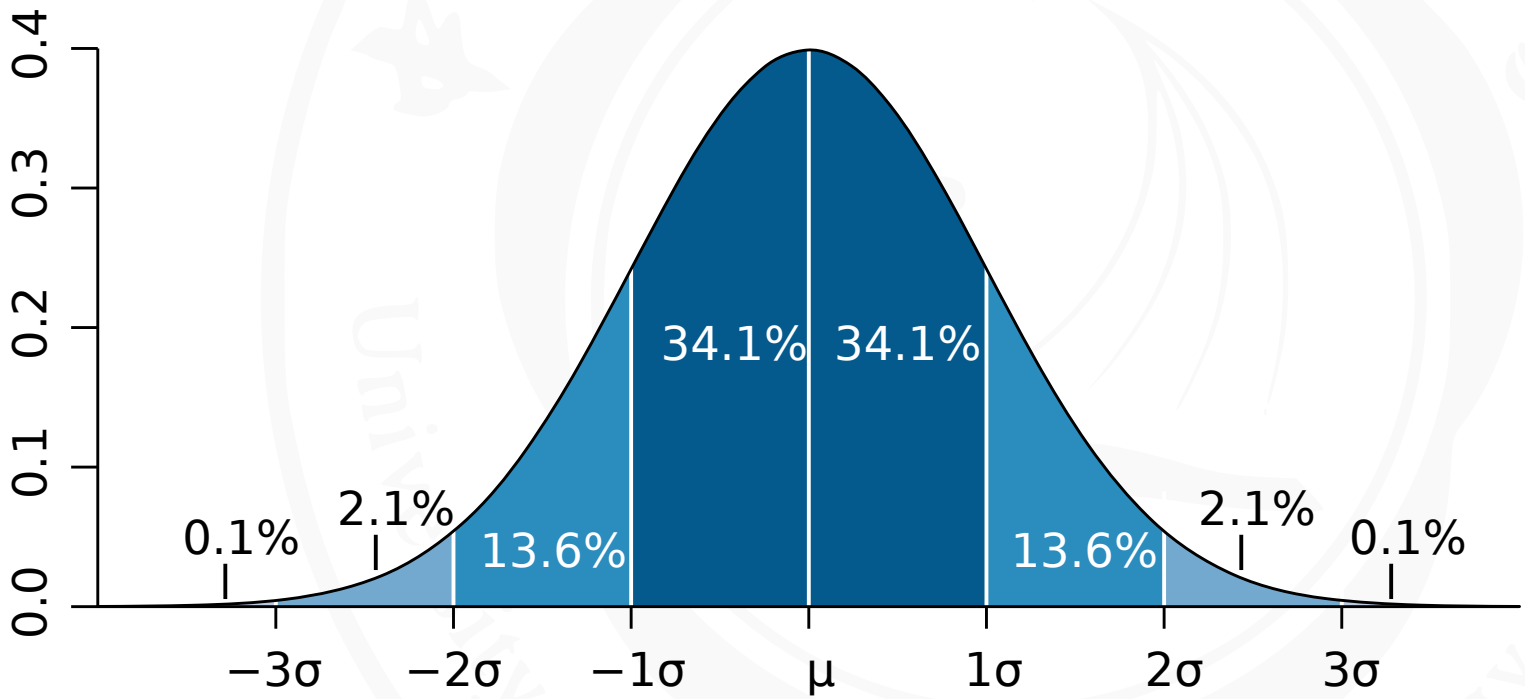
normal distribution

μ and σ are completely independent parameters



normal distribution

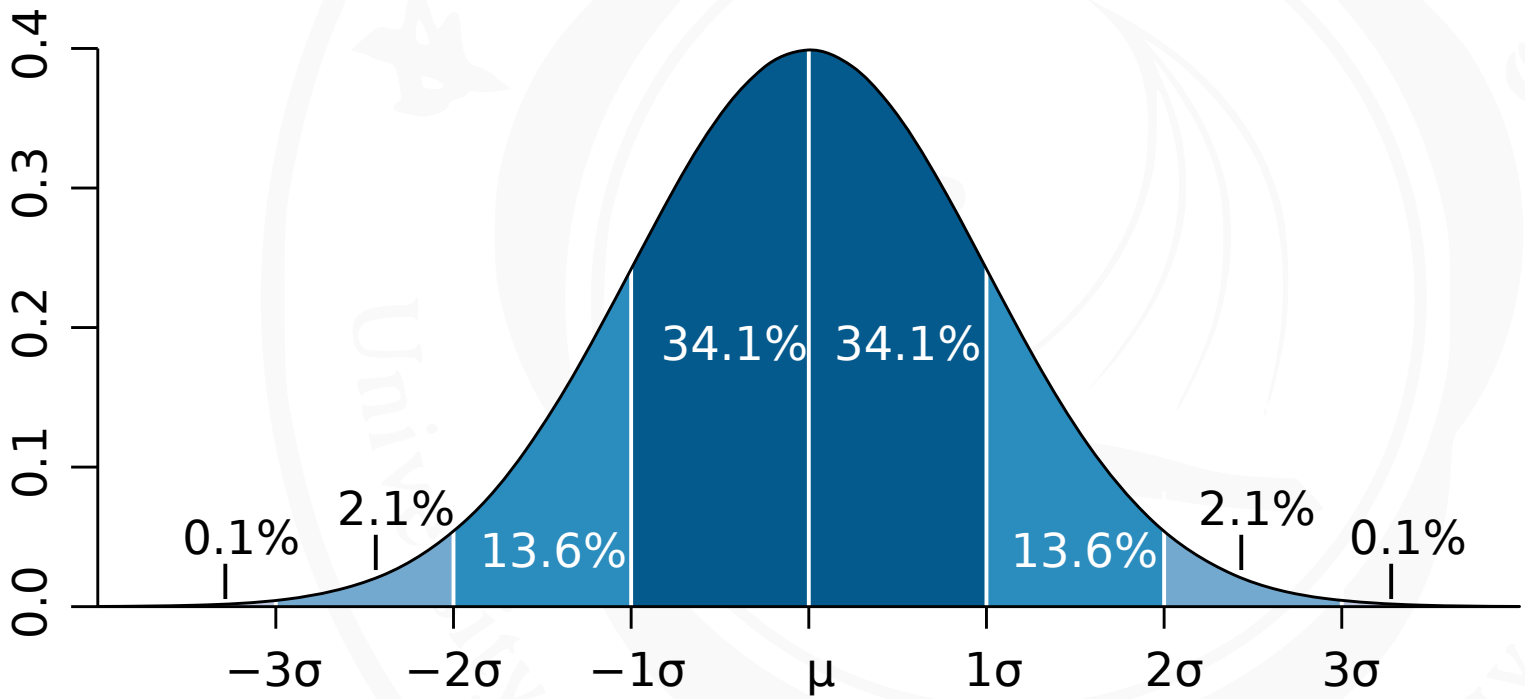
Practical interpretation of the normal distribution



68% of the values within $[\mu - \sigma, \mu + \sigma]$

normal distribution

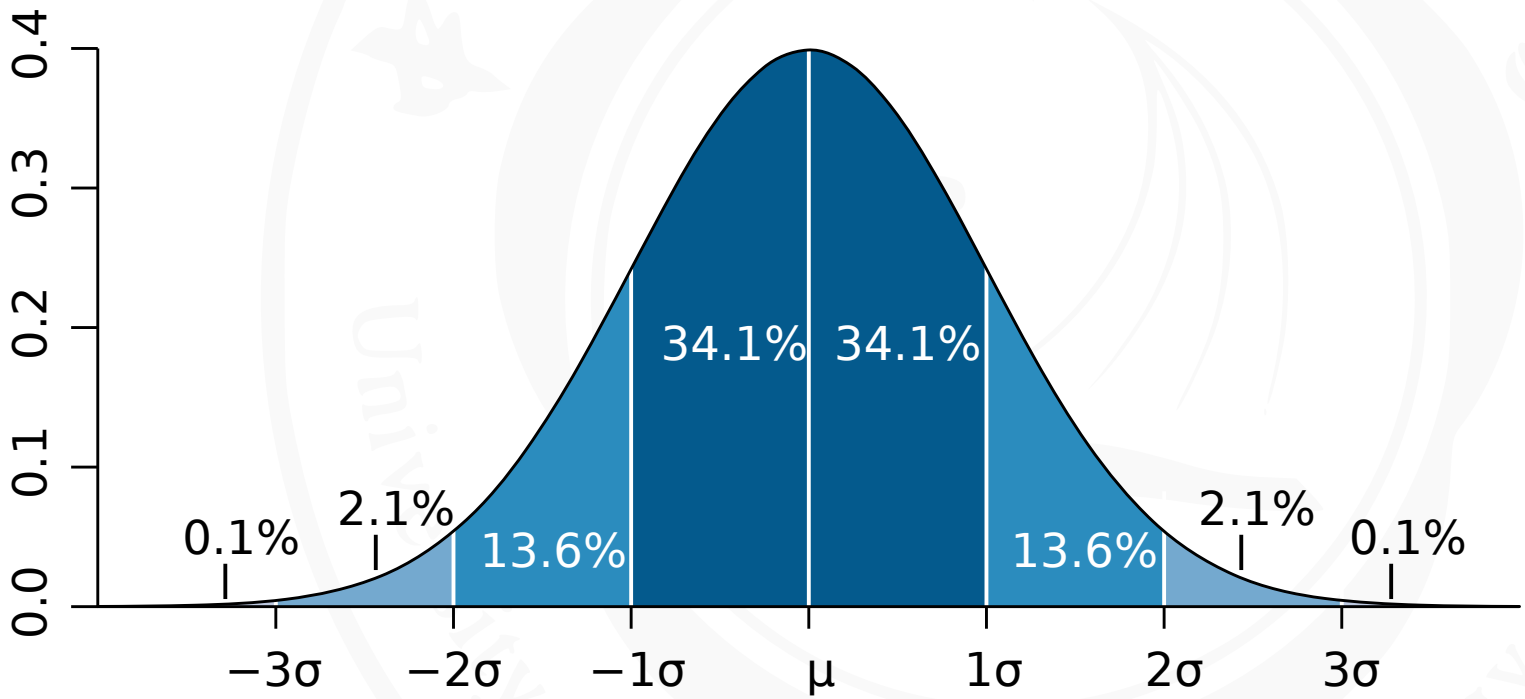
Practical interpretation of the normal distribution



95% of the values within $[\mu - 2\sigma, \mu + 2\sigma]$

normal distribution

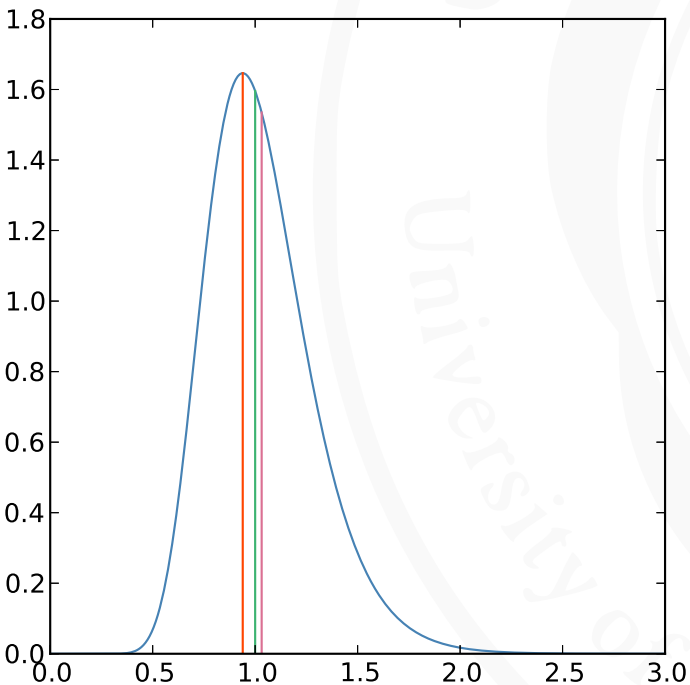
Practical interpretation of the normal distribution



99.7% of the values within $[\mu - 3\sigma, \mu + 3\sigma]$

skewed distributions

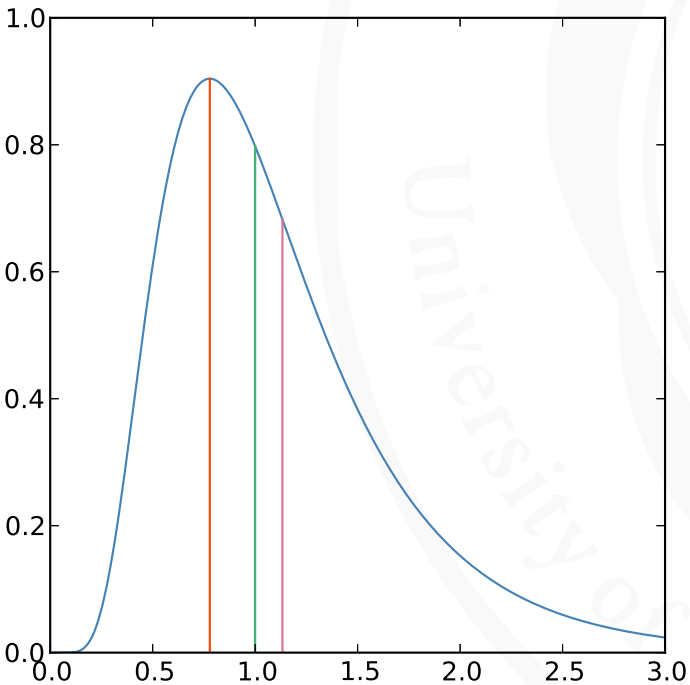
Your data might not have a symmetric distribution $\Rightarrow$ they might have a *skewed* distribution



- *red* is the true central tendency
- *green* is the median
- *pink* is the arithmetic mean

skewed distributions

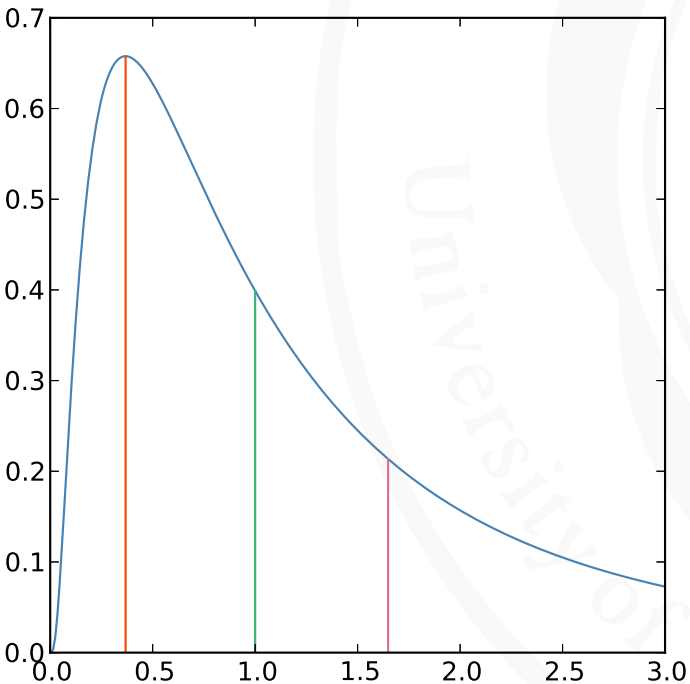
Your data might not have a symmetric distribution $\Rightarrow$ they might have a *skewed* distribution



- *red* is the true central tendency
- *green* is the median
- *pink* is the arithmetic mean

skewed distributions

Your data might not have a symmetric distribution $\Rightarrow$ they might have a *skewed* distribution



- *red* is the true central tendency
- *green* is the median
- *pink* is the arithmetic mean

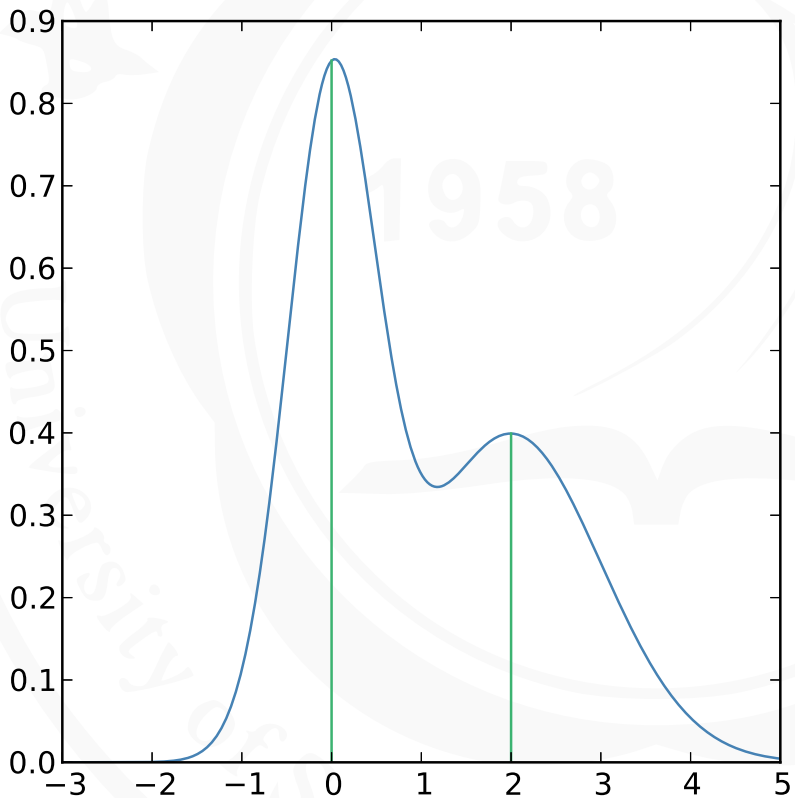
skewed distributions

You can compute the *skewness* of your data

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2\right)^{\frac{3}{2}}}$$

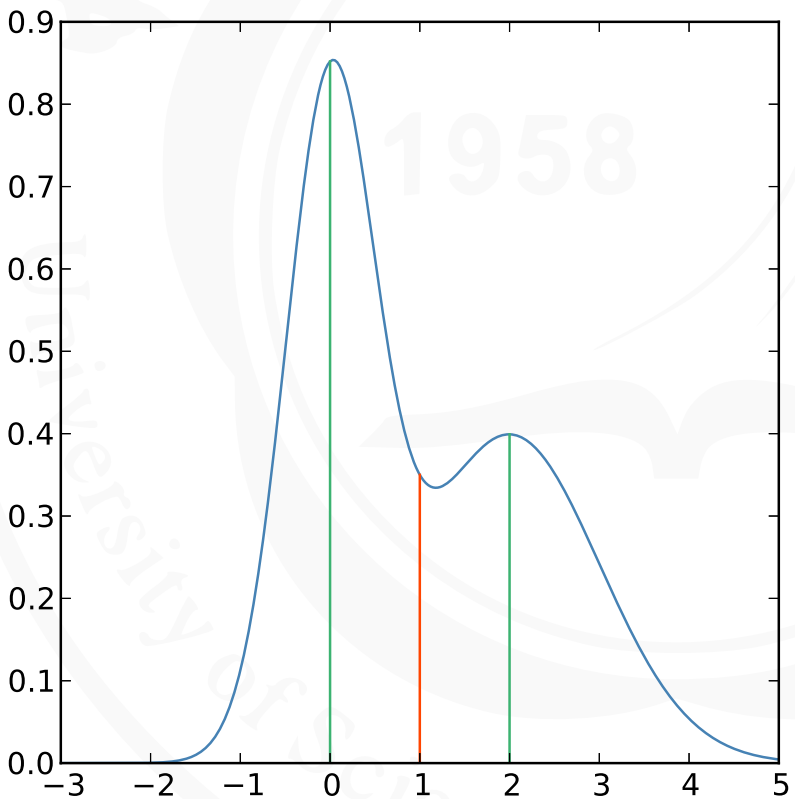
multimodal distributions

Your data might have multiple *m*odes



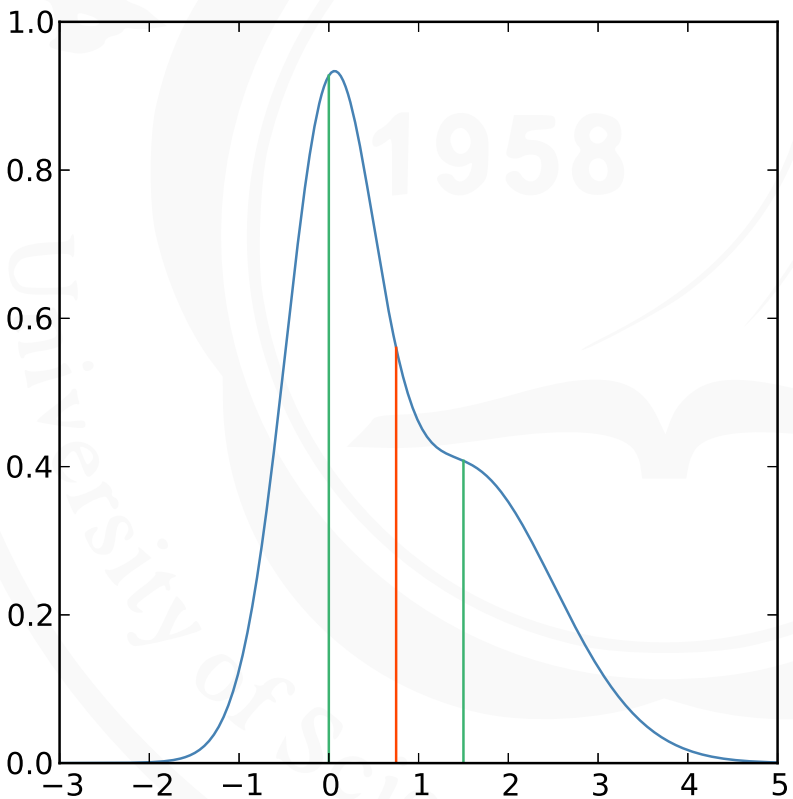
multimodal distributions

In such case, the mean, median and other descriptive quantities might have no reliable meaning



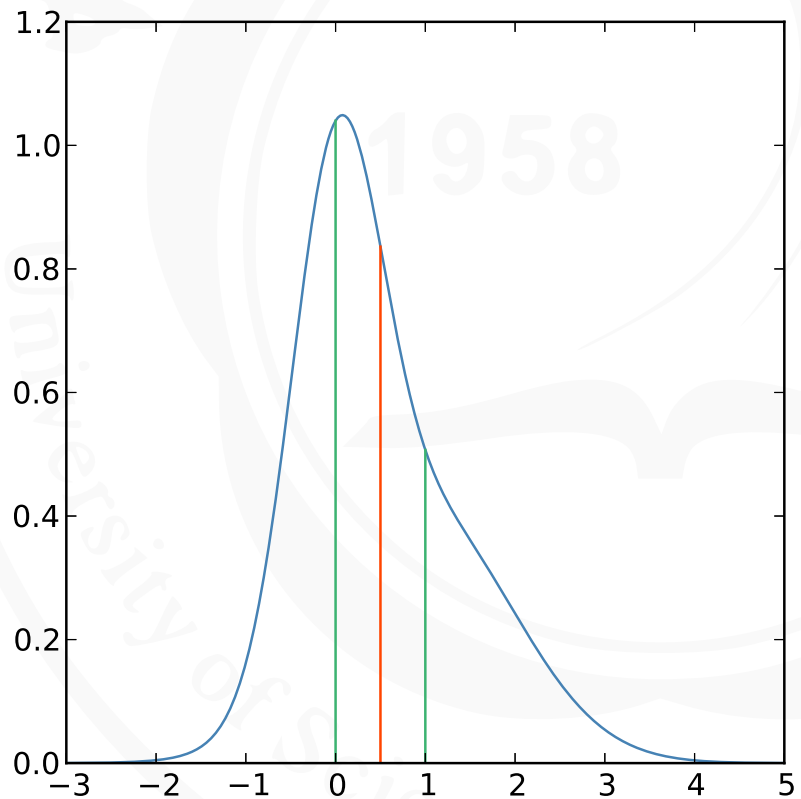
multimodal distributions

In such case, the mean, median and other descriptive quantities might have no reliable meaning



multimodal distributions

In such case, the mean, median and other descriptive quantities might have no reliable meaning



multimodal distributions

In such case, the mean, median and other descriptive quantities might have no reliable meaning

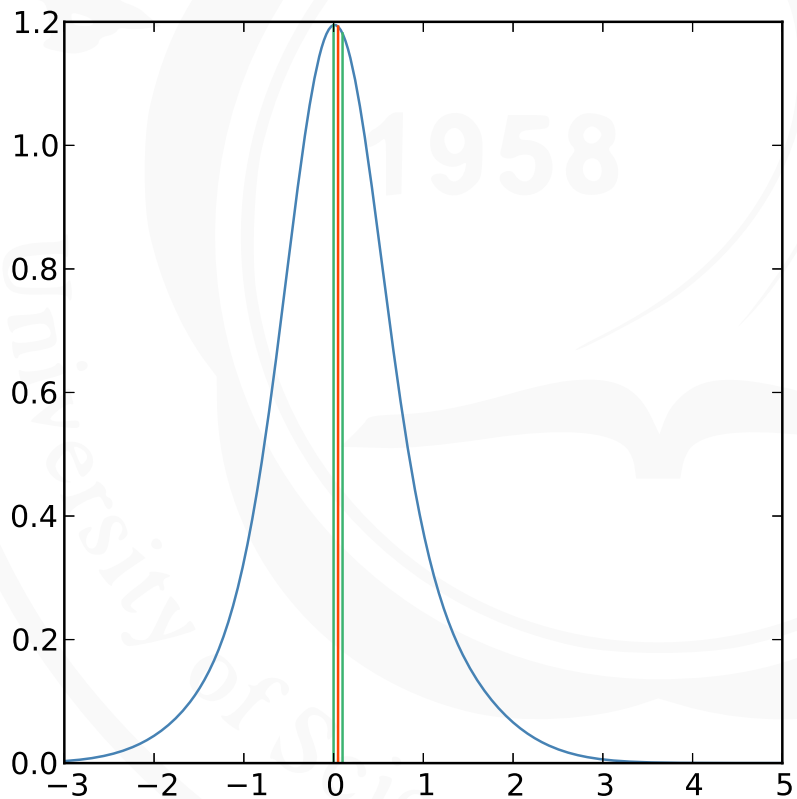


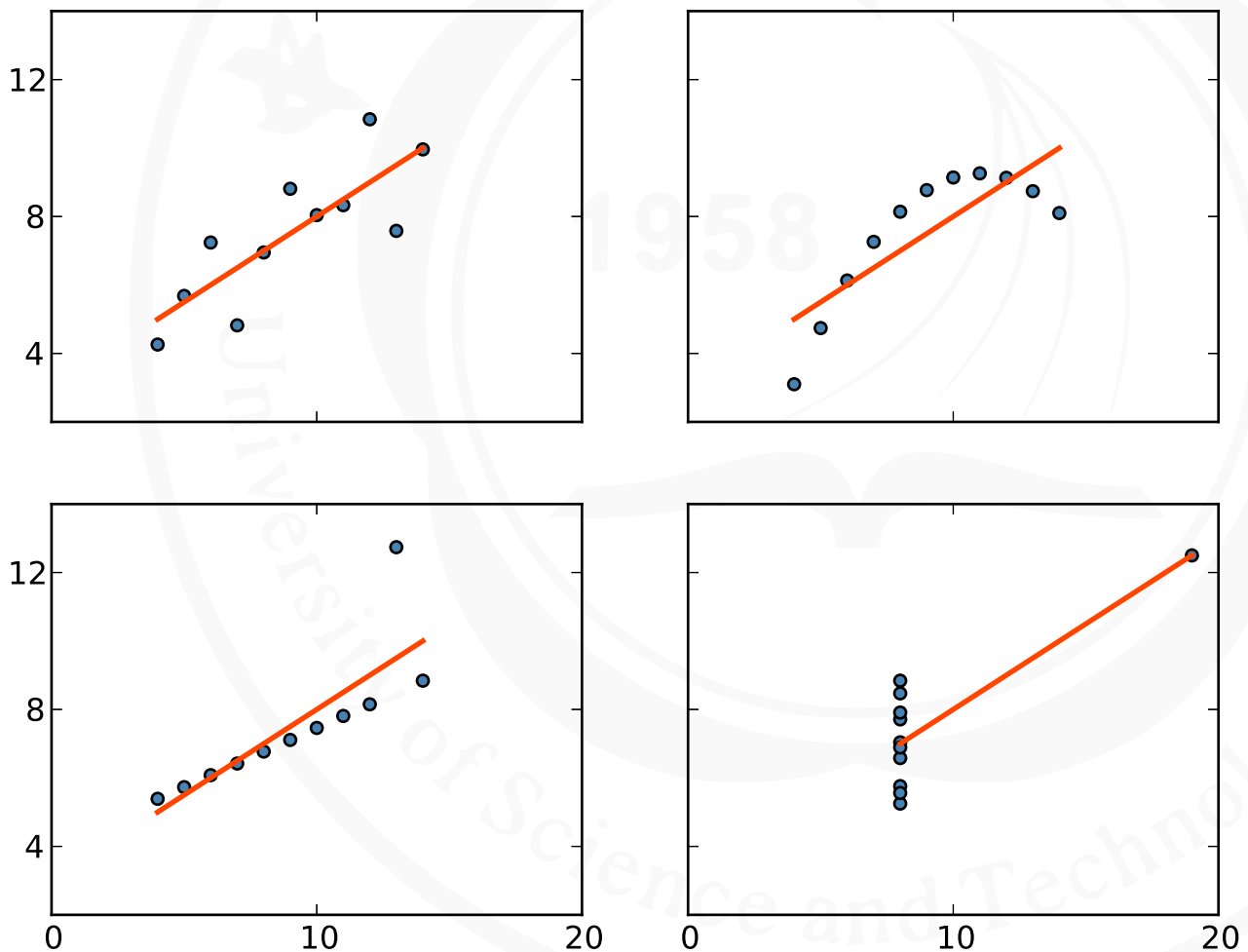
Table of Contents

Observe your data

Descriptive statistics can completely miss important informations from your data !

Observe your data

The *Anscombe's quartet*



Observe your data

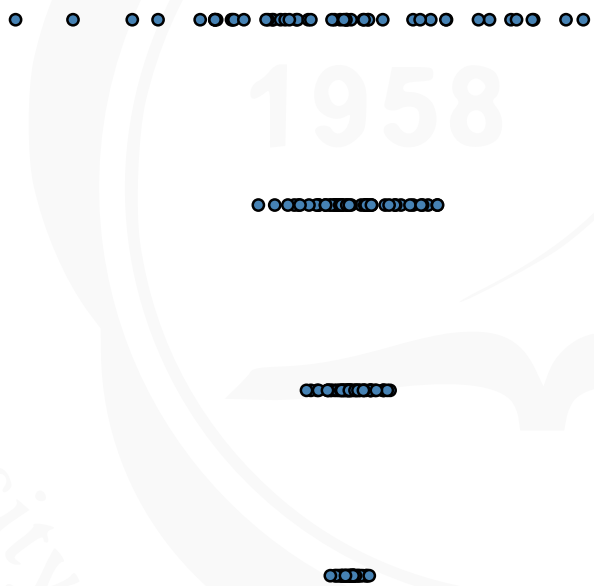
Those 4 datasets have *exactly* the same

- mean
- variance
- regression line

But they are not quite the same things !

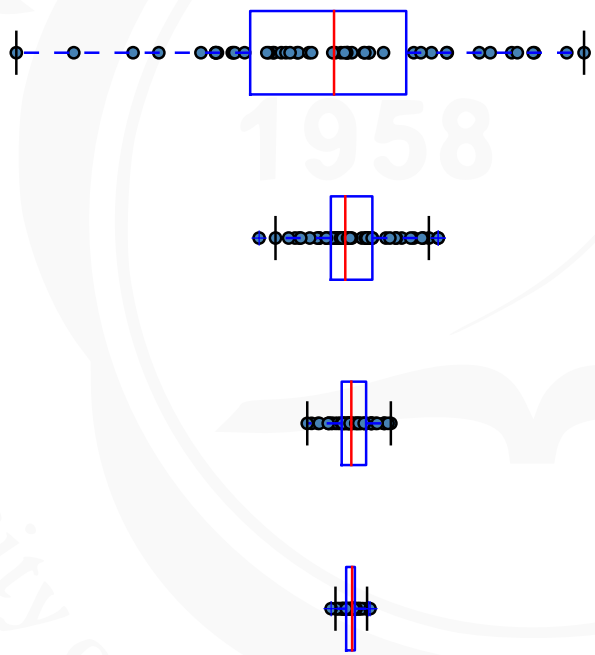
Boxplot

A nice way to summarize data distribution is the *boxplot*



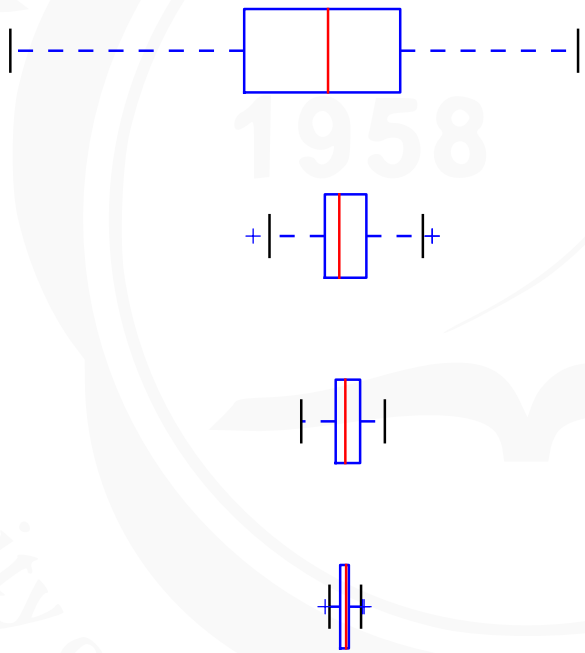
Boxplot

A nice way to summarize data distribution is the *boxplot*



Boxplot

A nice way to summarize data distribution is the *boxplot*



Boxplot

The red mark shows the mean

Boxplot

The box goes from the lower quartile to the upper quartile

Boxplot

The box is thus centred on the median

Boxplot

The whiskers are the minimum and maximum values

Boxplot

Outliers values are shown as blue crosses

Outliers are values which are beyond $1.5\times$ IQR from the quartiles

Scatter plot

A *scatter* plot is simply a plot with the data as points along 2 dimensions