# Descriptive Statistics

Marco Torchiano
Version 1.3.1 - March 2020

## License 

# Introduction

## Descriptive Statistics

- The discipline of quantitatively describing the main features of a collection of information
    - Also used to signify the quantitative description itself
- Goal: summarize a sample of data
    - As opposed to inferential statistics that aims to *infer* characteristics

# Descriptive Statistics

- Type / Scale
- Summary
  - Central tendency
  - Dispersion
- Distribution
  - Shape



# Type

- Discrete
  - Nominal scale
    - Special case: Binomial/Dichotomous/Boolean
  - Ordinal scale
- Continuous
  - Interval scale
  - Ratio scale
  - Absolute scale

# Summary

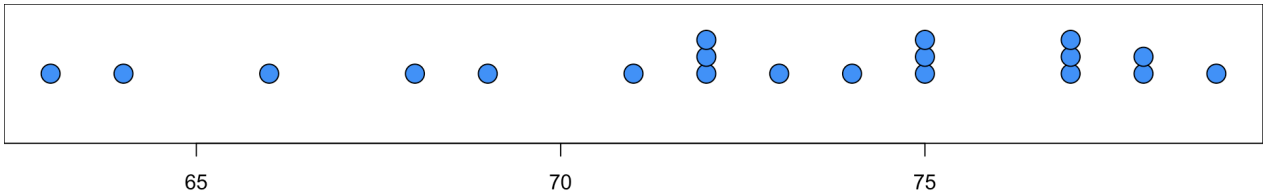
## Sample Data

Voter turn-out in all 20 Italian regions at 2018 political elections.

| Regione | turnout | Regione | turnout |
|---|---|---|---|
| **ABRUZZO** | 75 | MOLISE | 72 |
| **BASILICATA** | 71 | PIEMONTE | 75 |
| **CALABRIA** | 64 | PUGLIA | 69 |
| **CAMPANIA** | 68 | SARDEGNA | 66 |
| **EMILIA-ROMAGNA** | 78 | SICILIA | 63 |
| **FRIULI-VENEZIA GIULIA** | 75 | TOSCANA | 77 |
| **LAZIO** | 73 | TRENTINO-ALTO ADIGE | 74 |
| **LIGURIA** | 72 | UMBRIA | 78 |
| **LOMBARDIA** | 77 | VALLE D'AOSTA | 72 |
| **MARCHE** | 77 | VENETO | 79 |

# Example

Voter turn-out in italian regions




Represented as a strip-chart:

- Numbers are plotted along a line
- when multiple data share the same value they are stacked



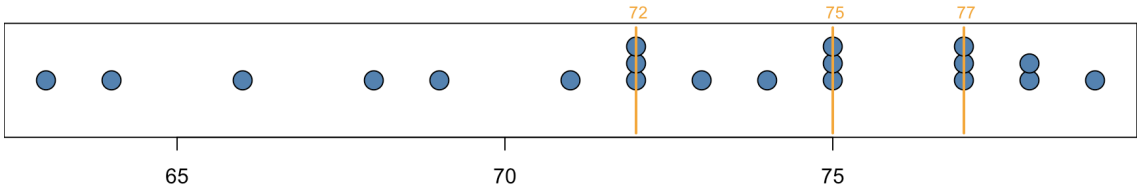
# Central tendency

## a.k.a. Location

- Mode
- Median
- Mean
- Mid-range

# Mode

- The most common value
  - Computed by counting the number of occurrences of each distinct value and taking the value(s) corresponding to the maximum count
- It is not unique!
  - We talk about *Unimodal* vs. *Multimodal* distributions



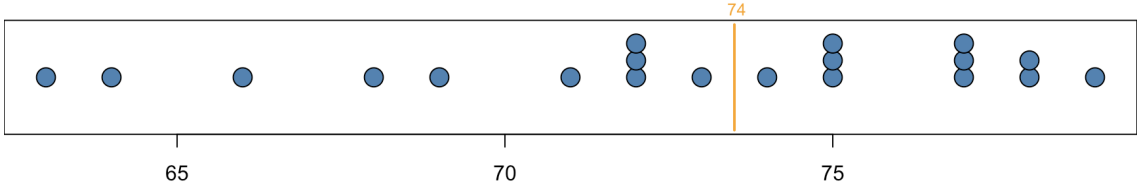



# Median

The value separating the higher half of values from the lower half.

Computed by sorting the values and picking:

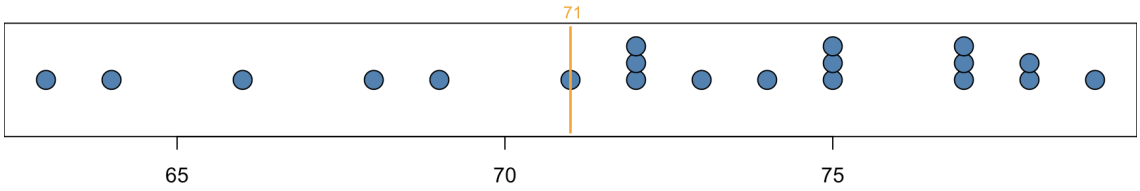
- value in the middle if odd number of values
- mid-value of two middle values if even number of values



# Mid-range

The value half-way between minimum and maximum values

$$MR = \frac{\min(x) + \max(x)}{2}$$



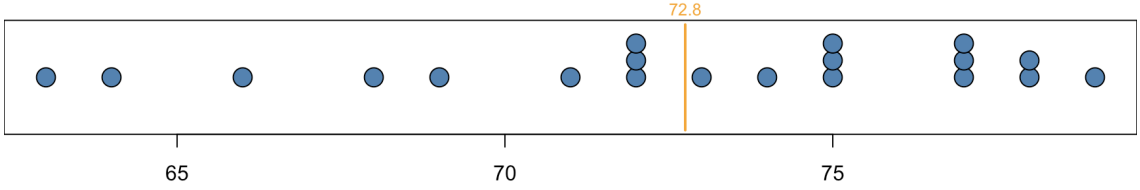



# (Arithmetic) Mean

Sum of values divided by number of values

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Divides the values in two sets such that the sum of distances are equal



# Other means

- Geometric mean

$$\bar{x} = \left( \prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

- Harmonic mean

$$\bar{x} = n \cdot \left( \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$



# Robustness

- What is the effect of an outlier on the central tendency measure?
- Outlier:
  - Observation (value) that is anomalous,
  - Distant from the others
- Typically due to an error

# Robustness example

Grades:

- John: 20, 21, 23, 23, 24, 25, 29
- Jane: 18, 25, 27, 28, 29, 30, 30

| Measure | John | Jane |
|---|---|---|
| **mode** | 23.00 | 30.00 |
| **median** | 23.00 | 28.00 |
| **mean** | 23.57 | 26.71 |
| **midrange** | 24.50 | 24.00 |



# Dispersion

# Dispersion

- Range
- Interquartile range (IQR)
- Median Absolute Deviation (MAD)
- Variance
- Standard Deviation

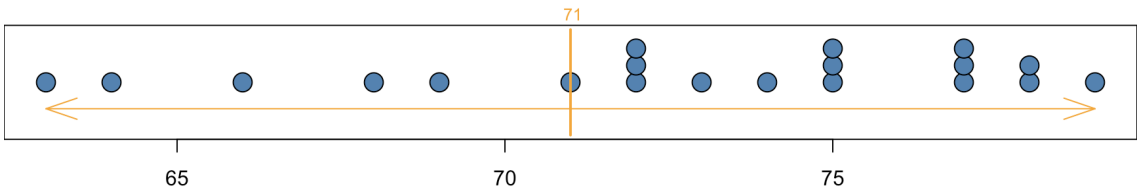


# Range

Distance between minimum and maximum.

$$\text{Range} = \max(x) - \min(x)$$

Range: 16



# Quantile

Split the values into groups according to their frequency (or probability).

$$Q_p(X) = x \iff \frac{|\{y \in X | y < x\}|}{|X|} \leq p$$

$$\wedge$$

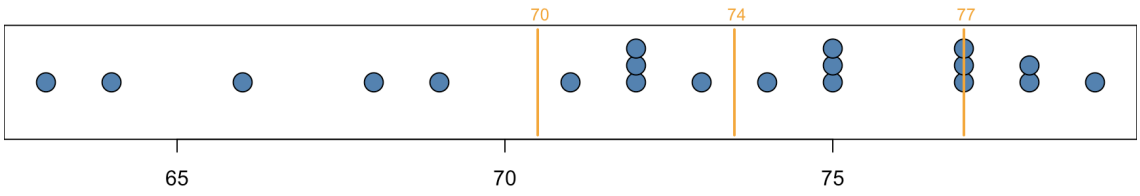
$$\frac{|\{y \in X | y \geq x\}|}{|X|} \geq p$$



# Quartiles

The quantiles that split the (ordered) values into four equally sized groups.

- Q1,Q2, Q3 corresponding to 25%, 50%, and 75% of values
- Q2 is the same as the median
- Sometimes you find Q0 (min) and Q4(max)

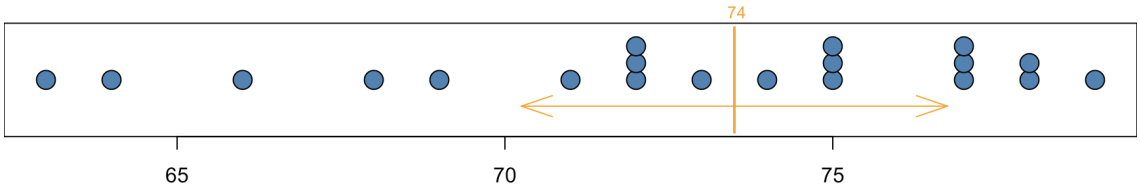


# IQR

The distance between Q1 and Q3

- encloses the central half of the values
- a.k.a. midspread, middle fifty, or middle 50%

$$IQR = Q3 - Q1$$

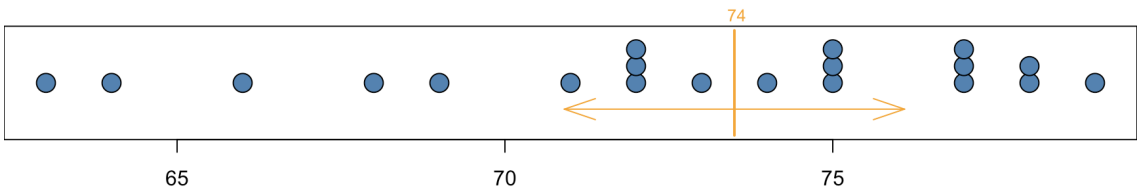





# MAD

Is the median of the absolute distance of values from the median

$$\text{MAD}(X) = \text{median}(\{\forall x_i \in X : |x_i - \text{median}(X)|\})$$



# Variance

Mean of the squared deviations from the mean.

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Always non-negative
- The distances are squared, therefore not directly comparable with the mean

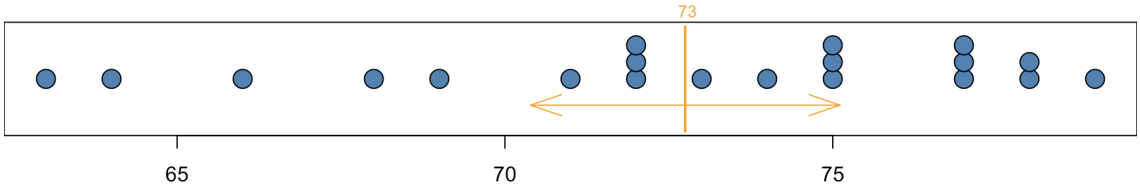


# Standard Deviation

Squared root of the variance

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Always non-negative
- Can be directly compared to the mean



# Normalized dispersion

Dispersion statistics are often compared to central location statistics

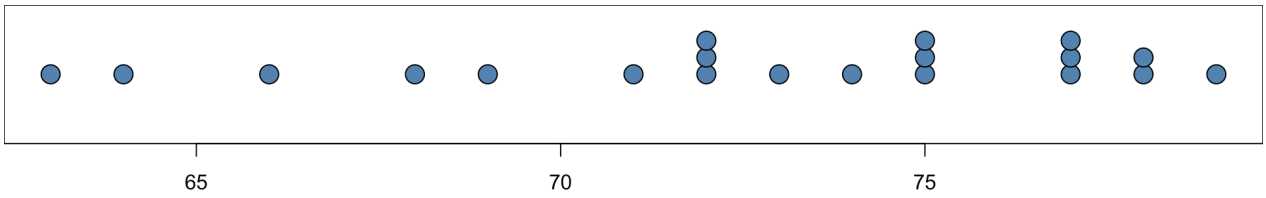
- To appreciate how much values are dispersed around the central location
- Typically a normalized version of the dispersions statistics is obtained dividing by the central statistics

In particular

$$\sigma_{Norm} = \frac{\sigma}{\overline{x}}$$



# Summary




| | Mean<br>SD | Median<br>MAD | Median<br>IQR | MidRange<br>Range |
|---|---|---|---|---|
| **Location** | 72.75 | 73.50 | 73.50 | 71.00 |
| **Dispersion** | 4.72 | 5.19 | 6.50 | 16.00 |
| **Normalized Disp.** | 0.07 | 0.07 | 0.09 | 0.23 |

# Anscombe's quartet

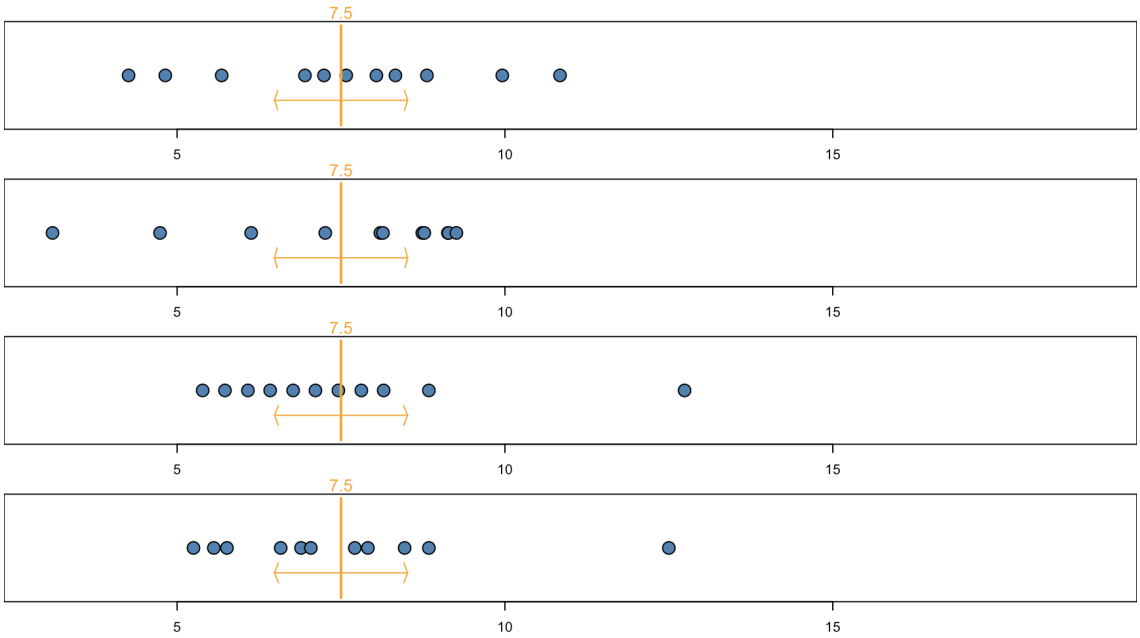
- Paired series (x,y) with 11 values
  - Mean x: 9
  - Variance x: 11
  - Mean y: 7.50
  - Variance y: 4.1
  - Correlation: 0.8
  - Linear regression: $y = 3 + 0.5x$

Four different variants with exactly the same summary characteristics

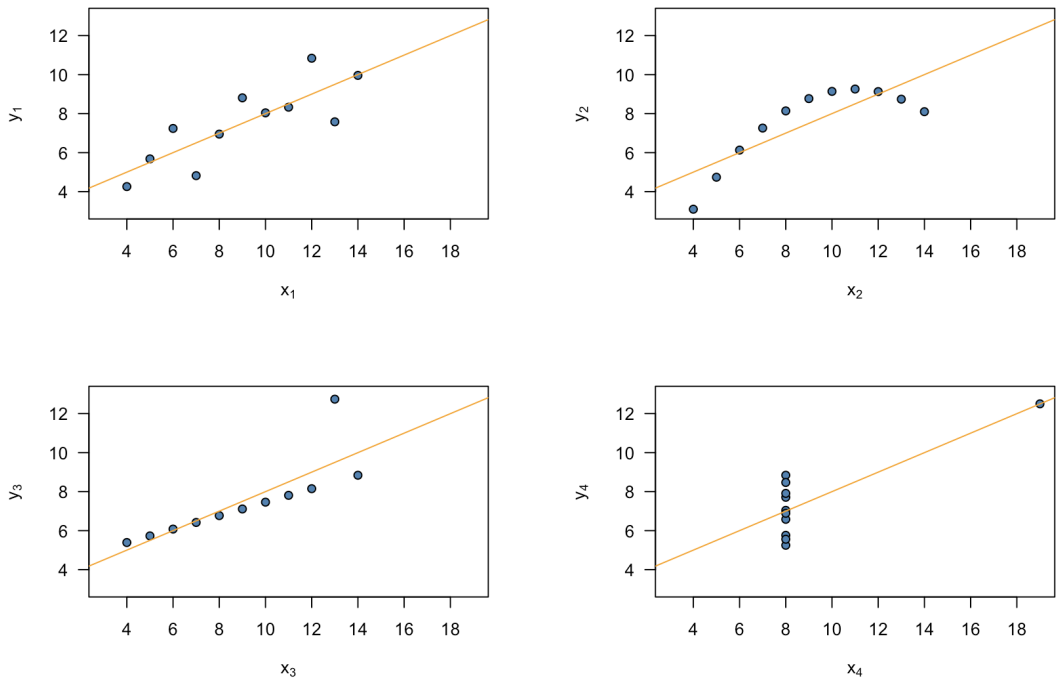


# Anscombe's quartets (y)

Mean and sd for the y components



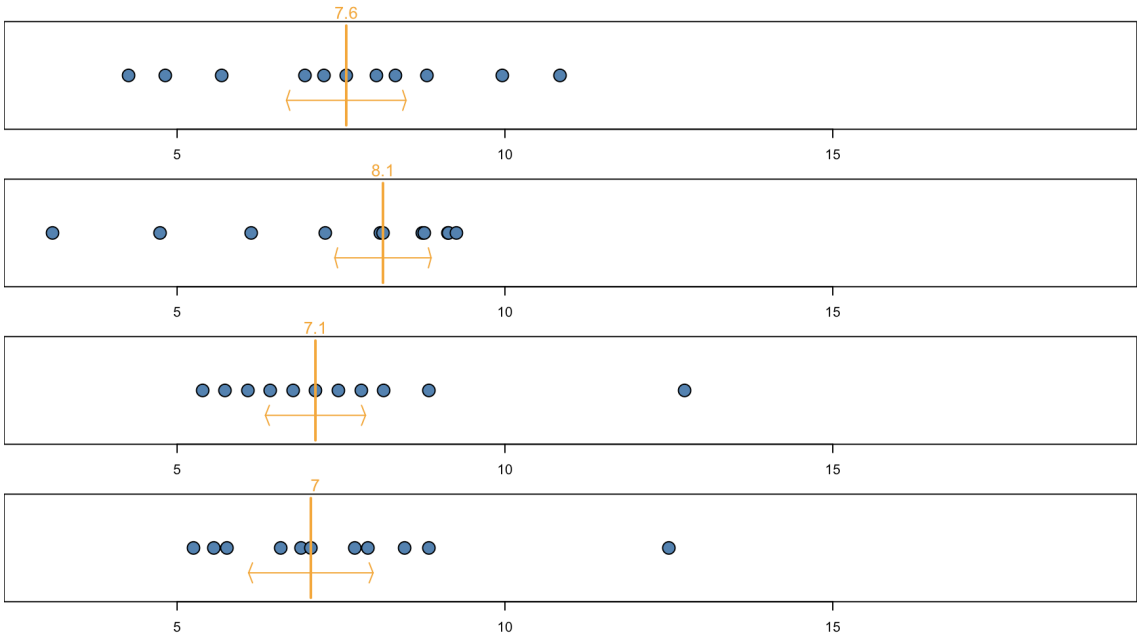
# Anscombe's quartets






# Anscombe's quartets (y)

## Medians and MADs



# Sampling

## Sampling

Often data from a sample are used to estimate characteristics of the whole population

- Population
  - The universe of entities (people, objects or any item)
- Sample
  - A subset of the population
- Census
  - Data from the entire population

# Parameter vs. Statistic

- Parameter: feature of the population
  - $\mu$: mean
  - $\sigma$: standard deviation
- Statistic: feature of the sample
  - $\bar{x}$: mean
  - s: standard deviation
  - Estimates of the population parameters



# Parameter vs. Statistic

Mean:

$$\mu \sim \bar{x} = \frac{1}{n} \sum_{i=0}^{n} x_i$$

Standard deviation:

$$\sigma \sim s \cdot \sqrt{\frac{n}{n-1}} = \sqrt{\frac{1}{n-1} \sum_{i=0}^{n} (x_i - \bar{x})^2}$$

$\frac{n}{n-1}$ is called Bessel's correction

# Distributions

## Distribution

A probability distribution describes the probability of a random variable to assume certain values

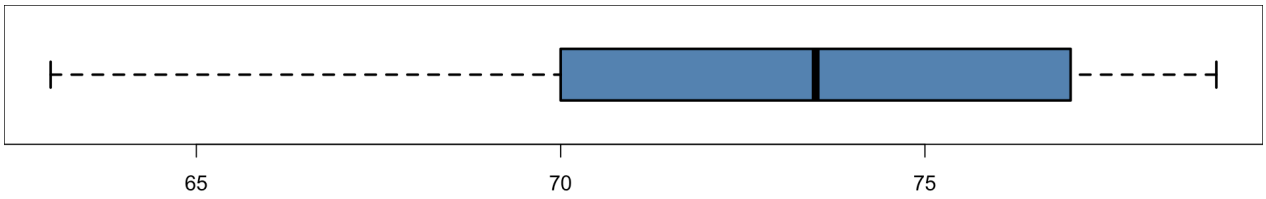
Distributions can be:

- Discrete
- Continuous

# Five Numbers Summary

| Minimum (Q0) | Lower Quartile (Q1) | Median (Q2) | Upper Quartile (Q3) | Maximum (Q4) |
|---|---|---|---|---|
| **63** | 70 | 73.5 | 77 | 79 |

Can be represented graphically as a *box-plot*.




Boxplot were devised by J.Tukey as a simple graphical summary, easy to draw – possibly by hand – with a few lines.

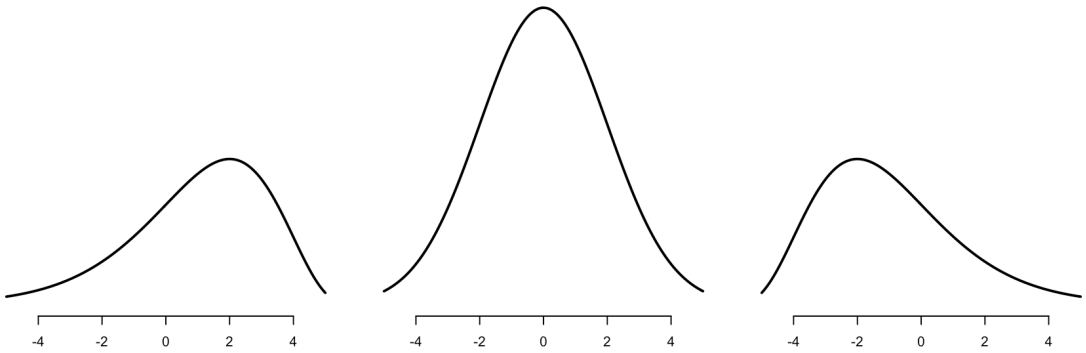


# Measures of Shape

- Skewness: absence of symmetry
  - Extreme values in one side of a distribution
- Kurtosis: peakedness of a distribution
  - Leptokurtic: high and thin
  - Mesokurtic: normal shape
  - Platykurtic: flat and spread out

# Skewness

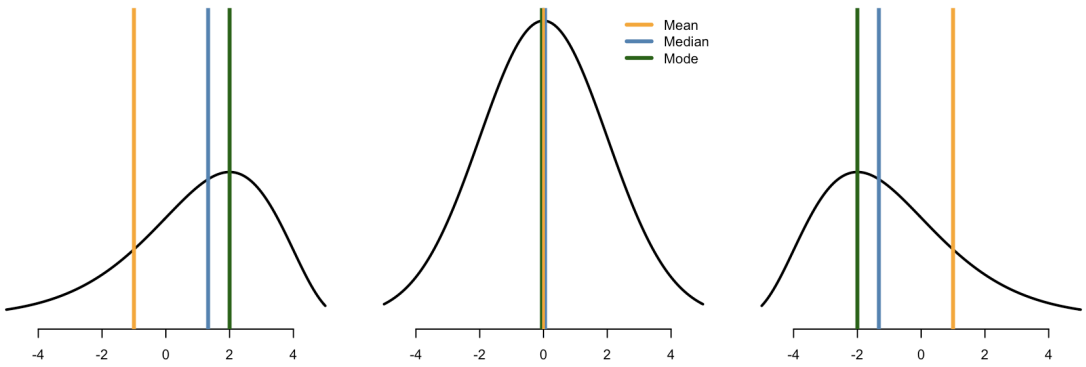
**Negatively skewed** **Not skewed** **Positively skewed**






# Skewness

**Negatively skewed** **Not skewed** **Positively skewed**



# Coefficient of Skewness

S: Summary measure for skewness

- If S < 0, the distribution is negatively skewed (skewed to the left).
- If S = 0, the distribution is symmetric (not skewed).
- If S > 0, the distribution is positively skewed (skewed to the right).



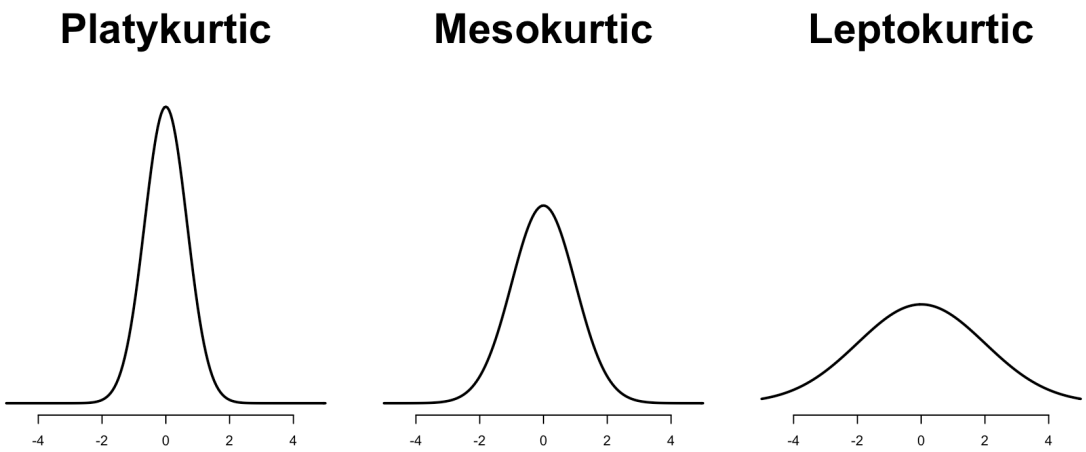
# Kurtosis

Peakedness of a distribution

K: measure of kurtosis:

- K > 0: Leptokurtic: high and thin
- K = 0: mesokurtic: normal in shape
- K < 0: platykurtic: flat and spread out

# Kurtosis






# Discrete measures

- Absolute frequency (count/tally)
  - Sorted by descending count (if nominal)
  - Sorted by measure order (if ordinal)
- Relative frequency (proportion)
  - as percentage
- Cumulate relative frequency (Cumulative proportion)
  - as percentage

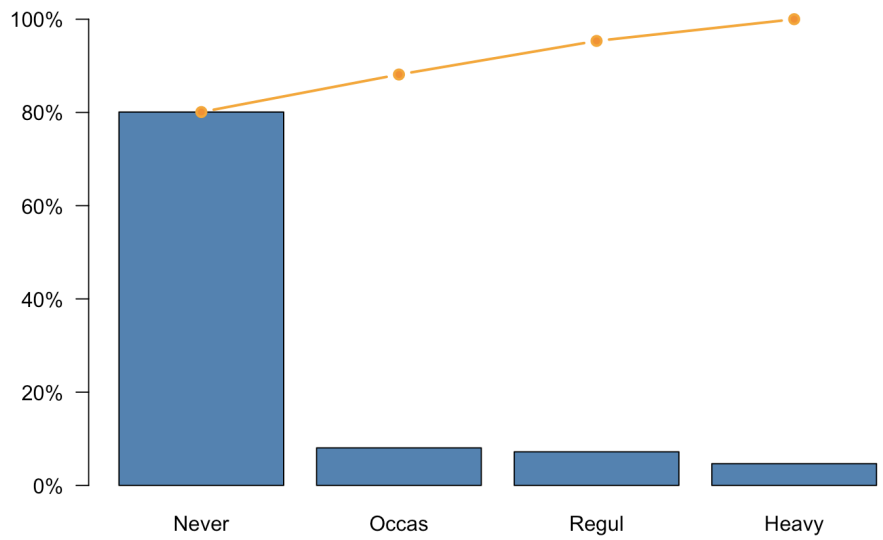
# Discrete measures example

Sample data from a student survey, here reported smoking habits.

| | Frequency | Proportion | Cum. Prop. |
|---|---|---|---|
| **Never** | 189 | 0.801 | 0.801 |
| **Occas** | 19 | 0.081 | 0.881 |
| **Regul** | 17 | 0.072 | 0.953 |
| **Heavy** | 11 | 0.047 | 1.000 |



# Discrete measures example



# References

## References


- Tukey, John Wilder (1977). Exploratory Data Analysis. Addison-Wesley
- Anscombe, F. J. (1973). “Graphs in Statistical Analysis”. In American Statistician 27 (1): 17–21
- Stephen Few (2015). Signal: Understanding What Matters in a World of Noise, Analytics Press
- Ministero dell’Interno. “Archivio storico delle elezioni” https://elezionistorico.interno.gov.it/index.php?tpel=C&dtel=04/03/2018