

Module 2. Univariate statistics

Data Science & AI

Sabine De Vreese Stijn Lievens Lieven Smits Bert Van Vreckem
2021–2022

**HO
GENT**

Contents

- Central Tendency and Dispersion
 - Measure of Central Tendency
 - Measures of Dispersion
 - Summary

- Data visualisation
 - Simple Graphs
 - Interpretation of Charts

Learning Goals

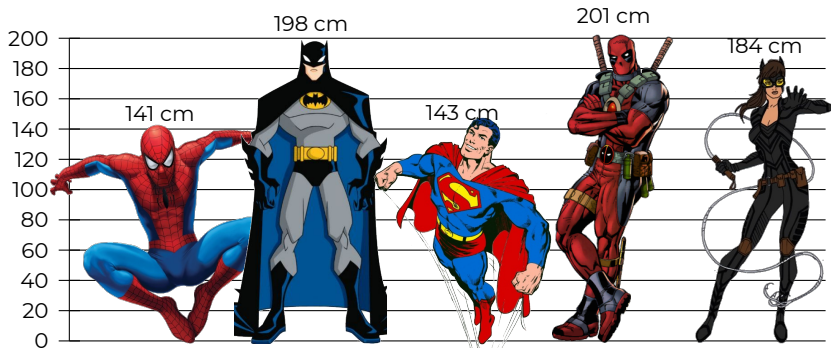
- Descriptive statistics
- Central tendency and dispersion for each measurement level
- Know formulas, being able to calculate
- Suitable visualization techniques for each measurement level

Central Tendency and Dispersion

**HO
GENT**

How tall are my friends?

Remember our superheroes:



**HO
GENT**

Measure of Central Tendency

What value is representative of the entire group?

**HO
GENT**

Mean or Average

Arithmetic mean

The **arithmetic mean** (notation: \bar{x}) is the sum of all values divided by the number of values

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

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184



Mean or Average

- Q1** What happens if Kabouter Wesley (10 cm) is added?
- Q2** The arithmetic mean of 15 numbers is 12. What number should be added to get a mean of 13?



**HO
GENT**

Median

Median

To find the **median**, sort all values and pick the middle number

- Odd number of values: no problem
- Even number of values: average of the middle two

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184



**HO
GENT**

Median

- Q1** What happens if Kabouter Wesley (10 cm) is added?
- Q2** What is the median of the number of people saved by Batman during the last eight years?

4	7	11	16	20	22	25	26
---	---	----	----	----	----	----	----



**HO
GENT**

Mode

Mode

The **mode** is the value that appears most often in a dataset.

Number of people saved by Superman during the last 15 years:

3	7	5	13	20	23	39	23	40	23	14	12	56	23	29
---	---	---	----	----	----	----	----	----	----	----	----	----	----	----



Number of people saved by Batman during the last 8 years:

4	7	11	16	20	22	25	26
---	---	----	----	----	----	----	----



**HO
GENT**

Measures of Dispersion

How large are the differences within the group?

Range

Range

The **range** of a dataset is the absolute value of the difference between the highest and the lowest value.

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184



**HO
GENT**

Quartiles

Quartiles

The **quartiles** of a sorted set of numbers are the three values that divide the set into 4 equally large subsets. Notation: Q_1, Q_2, Q_3

Number of people saved by Superman during the last 15 years:

3	7	5	13	20	23	39	23	40	23	14	12	56	23	29
---	---	---	----	----	----	----	----	----	----	----	----	----	----	----



**HO
GENT**

Calculating Quartiles

- Different software programs have slightly different ways of calculating quartiles.
- The following method is easy to perform by hand. Start by sorting the values.
 - When n is odd.
 - The median (Q_2) is the middle value (as before).
 - Leave out the median. Q_1 is the median of the first half, Q_3 is the median of the second half.
 - When n is even.
 - The median (Q_2) is the average of the two middle values.
 - Q_1 is the median of the first half, Q_3 is the median of the second half.

Variance and Standard Deviation

Variance

The **variance** (s^2 or σ^2) is the mean squared difference between the values of a data set and the arithmetic mean.

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

Standard deviation

The **standard deviation** (s or σ) is the square root of the variance

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184



Properties of the Standard Deviation

- Can the standard deviation be negative?

Properties of the Standard Deviation

- Can the standard deviation be negative?
- What is the smallest possible value? What does this imply?

Properties of the Standard Deviation

- Can the standard deviation be negative?
- What is the smallest possible value? What does this imply?
- What effect do outliers have on the standard deviation?

Properties of the Standard Deviation

- Can the standard deviation be negative?
- What is the smallest possible value? What does this imply?
- What effect do outliers have on the standard deviation?
- What is the unit of the standard deviation (in relation to the unit of the variable)?

Properties of the Standard Deviation

- Can the standard deviation be negative?
- What is the smallest possible value? What does this imply?
- What effect do outliers have on the standard deviation?
- What is the unit of the standard deviation (in relation to the unit of the variable)?
- How do you interpret the standard deviation combined with the average?

Properties of the Standard Deviation

Why $n - 1$ in the denominator and not n ?

You can prove the reason for the change mathematically, but we will investigate it empirically

See Python example code in `demo-analysis-1-var.ipynb`



Het journaal 1 - 21/02/14



Het weer 13.30u



100'' Journaal 15u

HOME

VIDEOZONE

LIVE CENTER

PROGRAMMA'S

NIUWS

WAUTERS VS. WAES

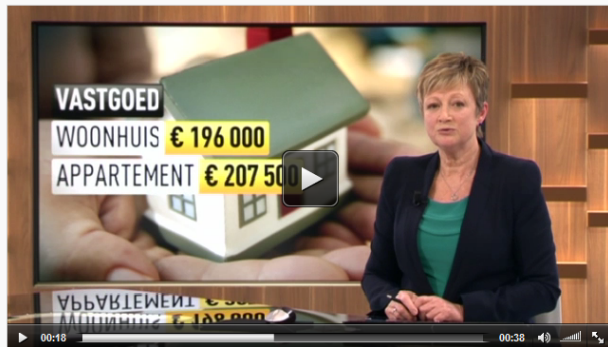
OOK DAT NOG

60 SECONDEN

VK14 OP TV

LIVESTREAM

MEEST BEKEKEN



Eigen dak boven hoofd wordt steeds duurder

Andere afleveringen



Het journaal 1 - 21/02/14



Het journaal L - 20/02/14



Het journaal 1 - 20/02/14



Het journaal L - 19/02/14

This news item reports on high prices for houses and flats. Do the numbers give a good idea of the situation?

Remember!

Providing only a center value is never sufficient!

- What is the dispersion?
- How is the data distributed? Normal distribution?
- Is the group sufficiently homogeneous?

**HO
GENT**

Central Tendency and Dispersion: Summary

Measurement Level	Center	Spread Distribution
Qualitative	Mode	—
Quantitative	Average/Mean Median	Variance, Standard Deviation Range, Interquartile Range

Summary of Symbols

	Population	Sample
number of elements	N	n
average or mean	μ	\bar{x}
variance	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
standard deviation	σ	s

Data visualisation

**HO
GENT**

Chart type overview

Measurement level	Chart type
Qualitative	Bar chart
Quantitative	Boxplot Histogram Density plot

See Python-example code in `demo-analysis-1-var.ipynb`

Pie Chart

Attention!

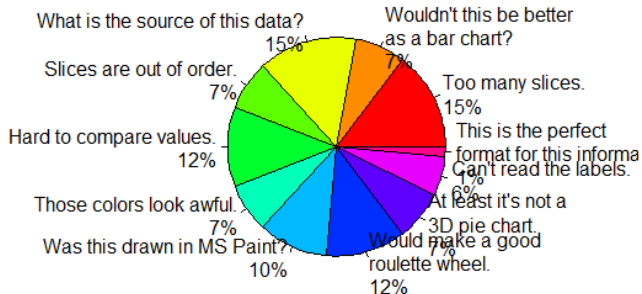
Avoid using a pie chart!

Disadvantages:

- Comparing angles is harder than comparing length
- Unusable for data with many categories

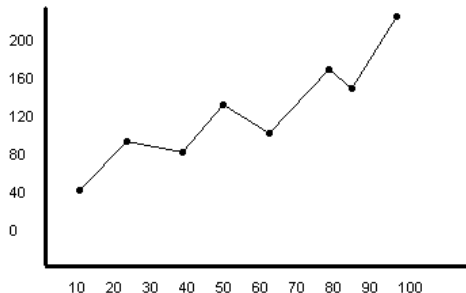
Pie Chart

What people are saying about your pie chart



**HO
GENT**

Interpretation of Charts



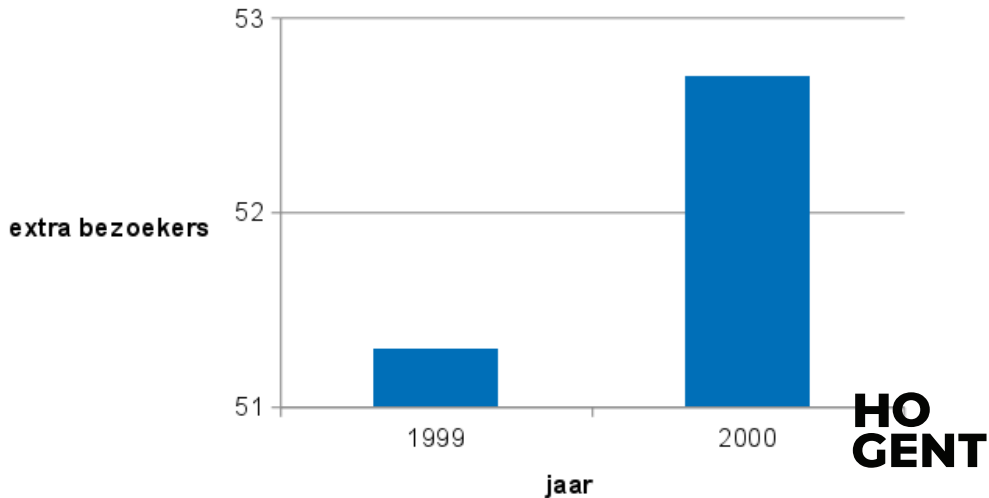
Tips:

- Label the axes
- Add a clear title
- Name the unit (and, if necessary, order of magnitude)
- Add a label that clarifies the chart

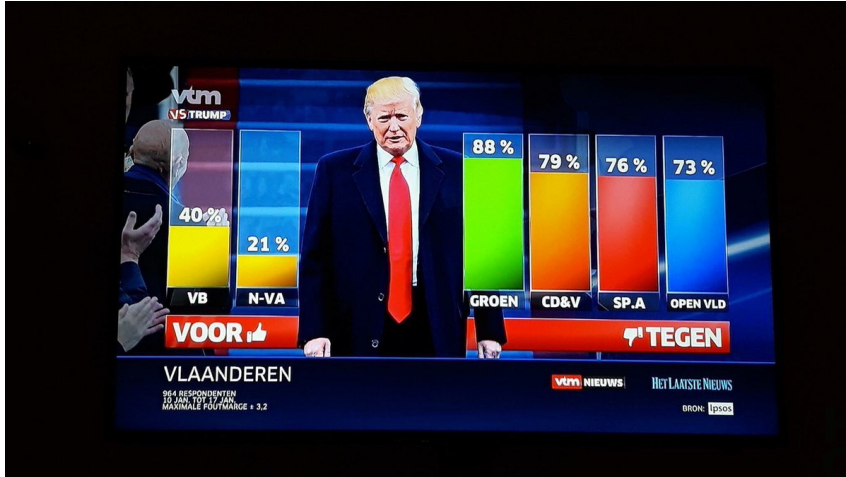
**HO
GENT**

Data distortion

= misrepresenting data so that invalid conclusions are drawn



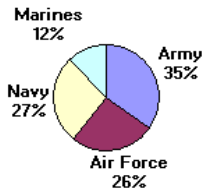
Data distortion



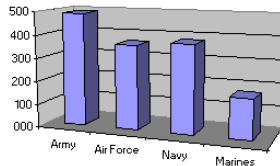
Data distraction

- Avoid bells and whistles
- Minimize “ink to data” ratio

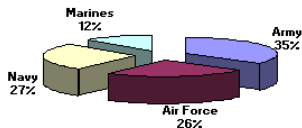
Active Duty Personnel, 1998



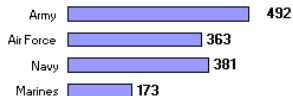
Active Duty Personnel, 1998
(millions)



Active Duty Personnel, 1998



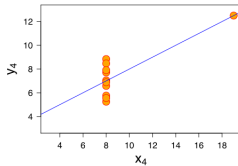
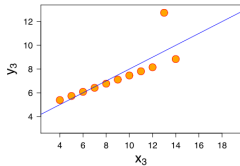
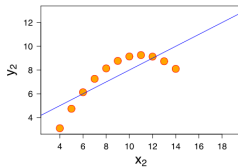
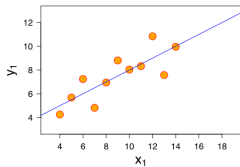
Active Duty Personnel, 1998
(millions)

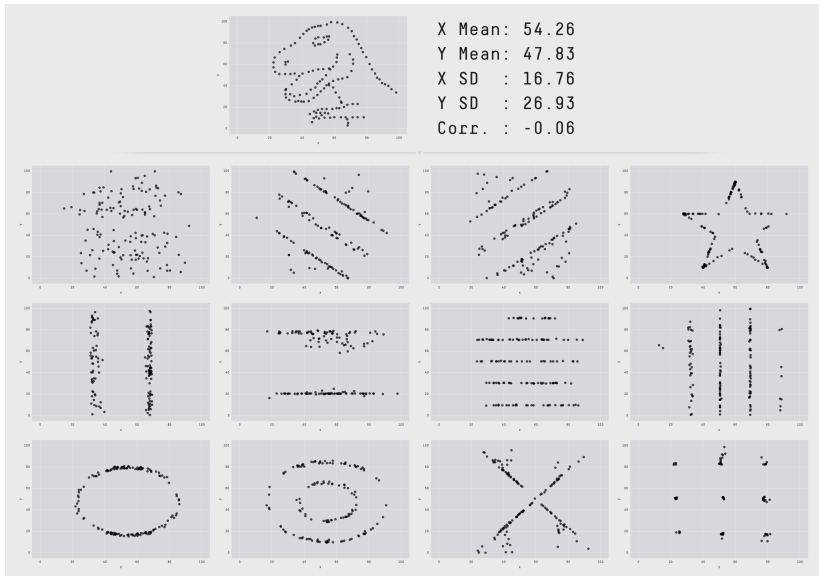


**HO
GENT**

The importance of data visualization

Anscombe's Quartet are four completely different datasets with the same measurements of central tendency and dispersion.





“The Datasaurus Dozen” (Source:
<https://www.autodeskresearch.com/publications/samestats>)