Module 1. Basic concepts and sampling

Data Science

Sabine De Vreese Lieven Smits Pieter Van Der Helst Bert Van Vreckem

2024–2025

Contents

Basic Concepts in Data Science

Sample Testing

Learning Goals

- Variables and measurement levels
- Samples
- Basic concepts

Basic Concepts in Data Science

HO
GENT

Variables and Values

Variable General property of an object, allows to distinguish objects

Value Specific property, interpretation for that variable

Variable: gender
Value: male

Variable: hair
Value: bald

Variable: height
Value: 201cm

Variable: weight
Value: 95kg

Measurement Levels

- = Variable types
- Determine most suitable method for analysis
    - visualization methods
    - central tendency and dispersion
    - examine the relationship between variables

Measurement Levels

Qualitative vs quantitative

Qualitative	**Quantitative**
Not necessarily numeric	Number + unit of measurement
Limited number of values	Many values, often unique

Quantitative variables often contain the result of a **measurement**

Measurement Levels

Qualitative scales

Nominal Categories.
e.g. gender, race, country, shape, ...

Ordinal Order, rank.
e.g. military rank, level of education, ...

Measurement Levels

Quantitative scales

Interval No fixed zero point ⇒ no proportions
e.g. °C, °F

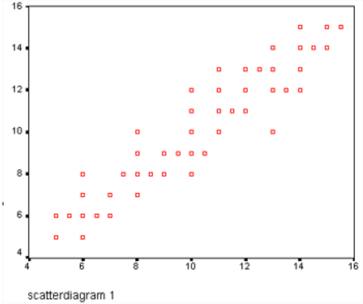
Ratio Absolute zero point ⇒ proportions
e.g. distance (m), energy (J), weight (kg) ...

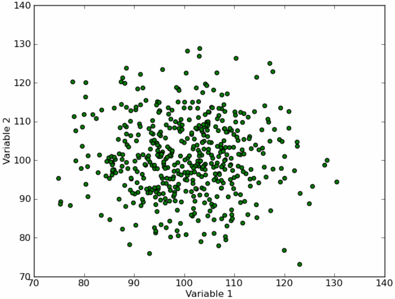
Proportions:

- 20 m is 1/3th or ~ 33% longer than 15 m
- 20 °C is **NOT** 1/3th warmer than 15 °C (convert to °F)

Relations between variables

Variables are related if their values change **systematically**.





Relations between variables: example

Is there a relationship between type of cola and taste appreciation?

	Pepsi	Coca Cola	Total
Like	56	24	80
Dislike	14	6	20
Total	70	30	100

HO GENT

Relations between variables: example

Is there a relationship between type of cola and taste appreciation?

	Pepsi	Coca Cola	Total
Like	56	24	80
Dislike	14	6	20
Total	70	30	100

Marginal totals

HO GENT

Causal Relationships

Researchers are often looking for **causal relationships**, e.g.

- Frustration leads to agression
- Alcohol leads to decreased alertness
- ...

Cause Independent variable

Consequence Dependent variable

Causal Relationships

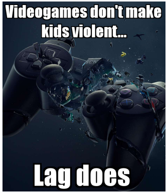
Fake correlations or “Spurious correlations”

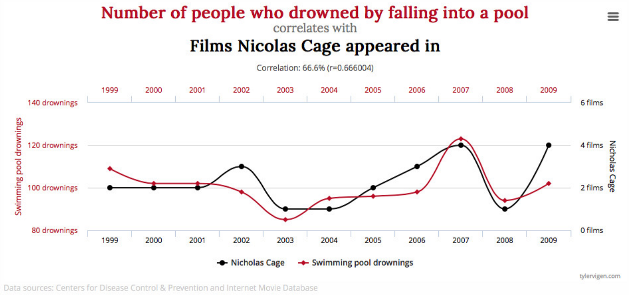
Warning!

A relationship between variables does not necessarily indicate a *causal* relation!

Examples:

- Violent video games lead to violent behaviour
- Vaccines can cause autism
- Relationship between drinking cola light and obesitas
- ...





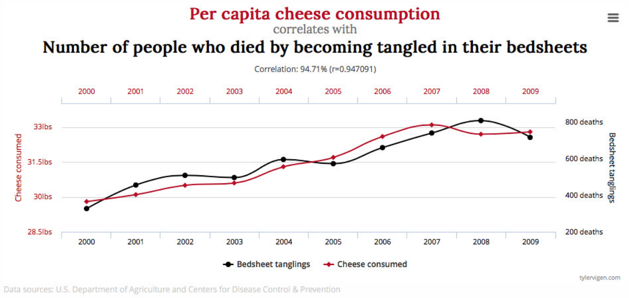
Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

Data sources: Centers for Disease Control & Prevention and Internet Movie Database



Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% (r=0.947091)

Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Sample Testing

HO
GENT

USA Today has come out with a new survey. Apparently, three out of every four people make up 75% of the population

—David Letterman

Suppose you want to analyze a group of friends

Questions you can ask:

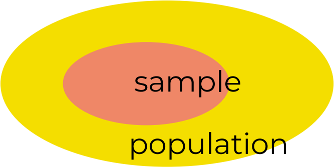
- How tall are my friends?
- What are their weights?
- How safe is their living environment?
- Do they have family?
- ...

Population

Sample and Population

Population the collection of all objects/people/…that you want to investigate

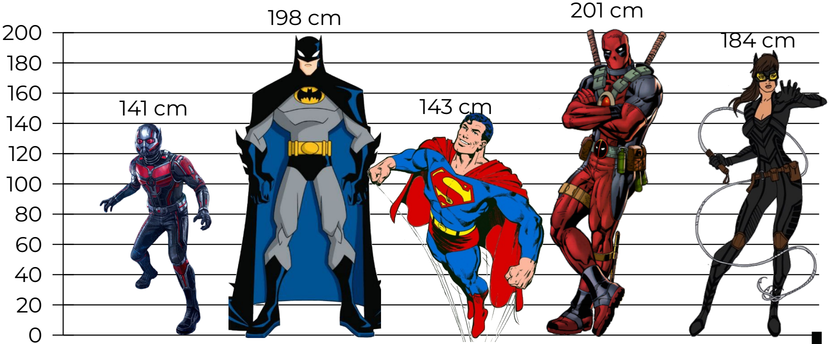
Sample a *subset* of the population from which measurements will be taken



Under certain circumstances, the results for a sample are representative for the population.

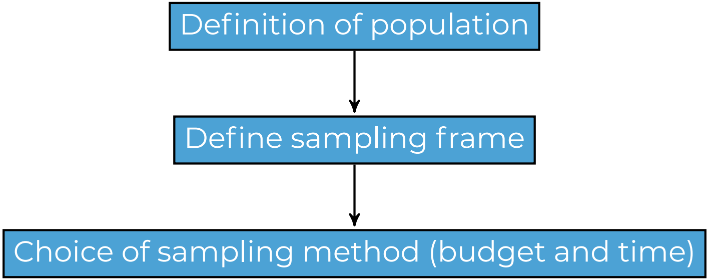
Sample and Population

A sample is easier to analyze than the entire population



HO GENT

Sampling Method



Definition of population

Define sampling frame

Choice of sampling method (budget and time)

How to pick elements for a sample?

Random sample : every element from the population has an equal chance of being included in the sample.

Non-random sample : the elements for the sample are *not* randomly selected. Objects that can be collected *easily* are more likely to be included (convenience sampling).

Stratified to variables

	Age				
Gender	≤ 18	$]18, 25]$	$]25, 40]$	> 40	Total
Woman	500	1500	1000	250	3250
Man	400	1200	800	160	2560
Total	900	2700	1800	410	5810

Stratified to variables

	Age				
Gender	≤ 18	$]18, 25]$	$]25, 40]$	> 40	Total
Woman	500	1500	1000	250	3250
Man	400	1200	800	160	2560
Total	900	2700	1800	410	5810

	Age				
Gender	≤ 18	$]18, 25]$	$]25, 40]$	> 40	Total
Woman	50	150	100	25	325
Man	40	120	80	16	256
Total	90	270	180	41	581

Possible Errors

Measurements in a sample will typically deviate from the value in the entire population ⇒ Errors!

- Accidental ↔ Systematic
- Sampling error ↔ Non-sampling error

Sampling Errors

- Accidental sampling errors
  - Pure coincidence

Sampling Errors

- Accidental sampling errors
  - Pure coincidence
- Systematic sampling errors
  - Online survey: people without internet are excluded
  - Street survey: only who is currently walking there
  - Voluntary survey: only interested parties participate

Non-sampling Errors

- Accidental non-sampling errors
    - Incorrectly ticked answers

Non-sampling Errors

- Accidental non-sampling errors
  - Incorrectly ticked answers
- Systematic non-sampling errors
  - Poor or non-calibrated measuring equipment
  - Value can be influenced by the fact that you measure
  - Respondents lie (number of cigarettes a day)