1 - Fundamental concepts & sampling

Basic concepts in Data Science

Variables & Values

variable = property of an object
value = specific state of variable

Measurement Levels

= Variable types

Determine most suitable method for analysis

- visualization methods
- central tendency & dispersion
- examine relationship between variables

Qualitative vs quantative

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

Qualitative scales

- **Nominal** - Categories
  (gender, race, country, shape, ...)
- **Ordinal** - Order, rank
  (military rank, level of education, ...)

Quantitative Scales

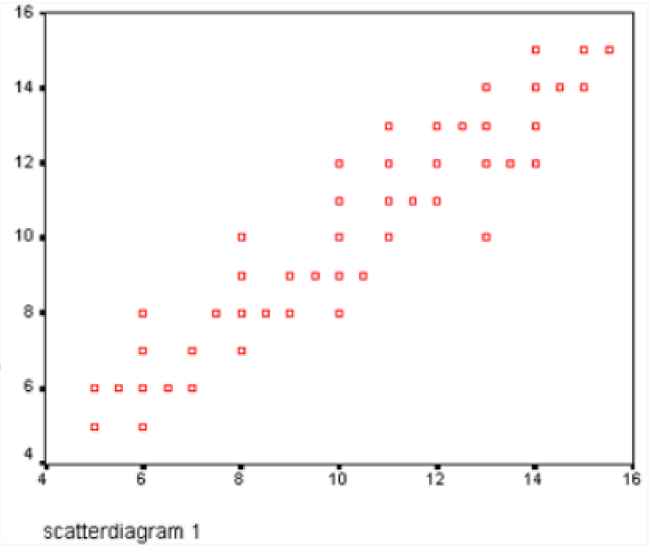
- **Interval** - No fixed zero point $\implies$ no proportions
  ($^\circ C$, $^\circ F$)
- **Ratio** - Absolute zero point $\implies$ proportions
  (distance (m), energy (J), weight (kg))

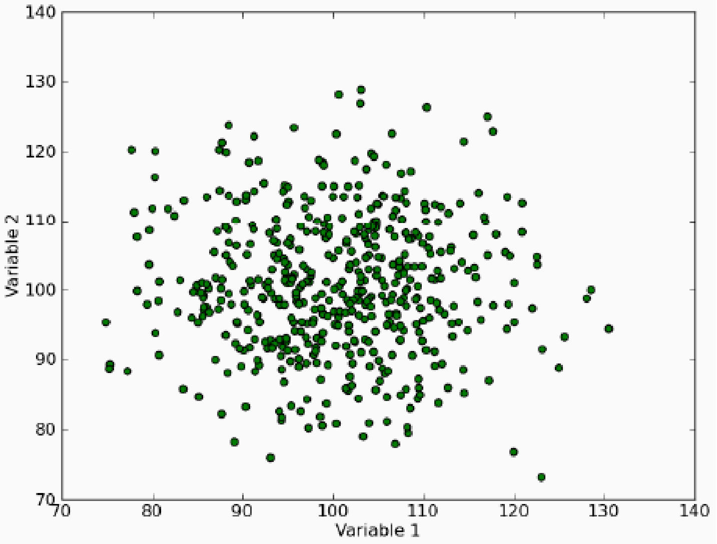
Proportions:

- $20m$ is $1/3$th longer than $15m$
- $20^\circ C$ **isn't** $1/3$th warmer than $15^\circ C$ (convert to $^\circ F$)

Relations between variables

Variables are related if values change **systematically**





Causal Relationships

- **Cause** - Independent variable
- **Consequence** - Dependent variable

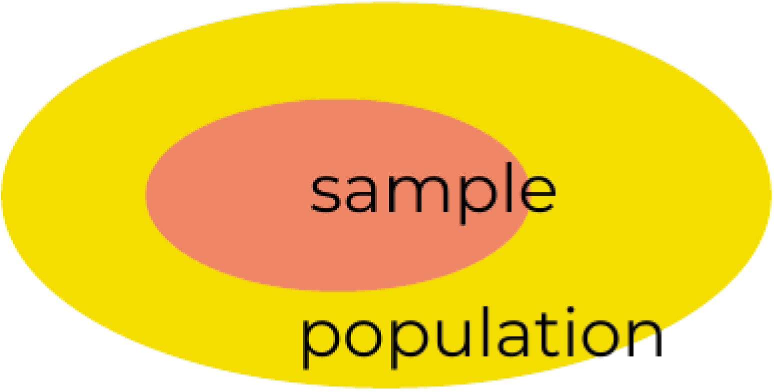
Fake correlations / "Spurious correlations"

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

Sample testing

Sample & population

- **Population** - collection of all objects/people/... that you want to investigate
- **Sample** - *subset* of population from which measurements will be taken



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

Sampling method

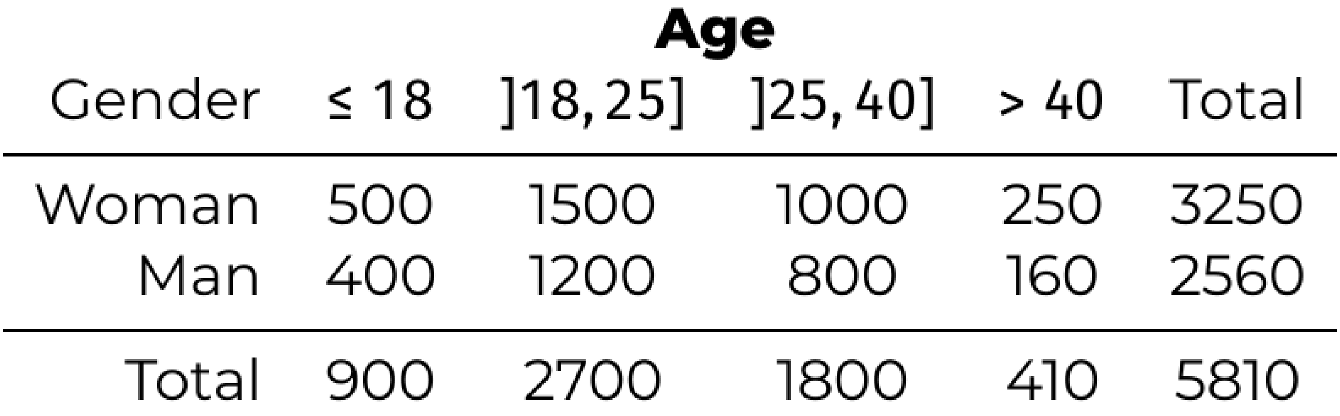
$$\text{Definition of population}$$
$$\downarrow$$

$$\begin{array}{c}\text{Define sampling frame}\\ \downarrow \\ \text{Choice of sampling method (budget \& time)}\end{array}$$

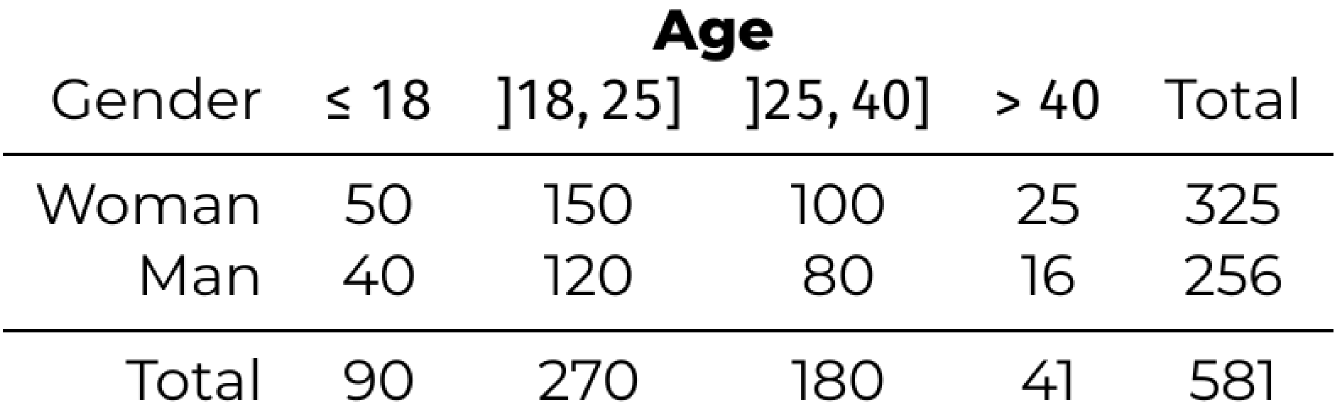
How select elements for sample

- **Random sample**
  - every element from population has equal chance of being included in sample
- **Non-random sample**
  - elements for sample are *not* randomly selected
  - objects that can be collected easily are more likey 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



		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 $\Longrightarrow$ Errors!

Sampling errors

- Accidental sampling errors
  - pure **coincidence**
- Systematic sampling errors

  - Online survey: people without internet are excluded
  - Street survey: only people who are walking there are included
  - Voluntary survey: only interested parties participate

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)