A step-change in quantitative social science skills

Funded by the Nuffield Foundation, ESRC and HEFCE

SMART Skills Datacamp

Dr Hannah Bunting & Dr Kirils Makarovs

Introductions

Dr Hannah Bunting

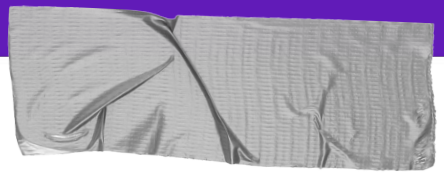
Q-Step & Politics

Researches electoral behaviour, British politics & trust in governments

Dr Kirils Makarovs

Q-Step & SPA

Researches public opinion of science & scientists



Course structure

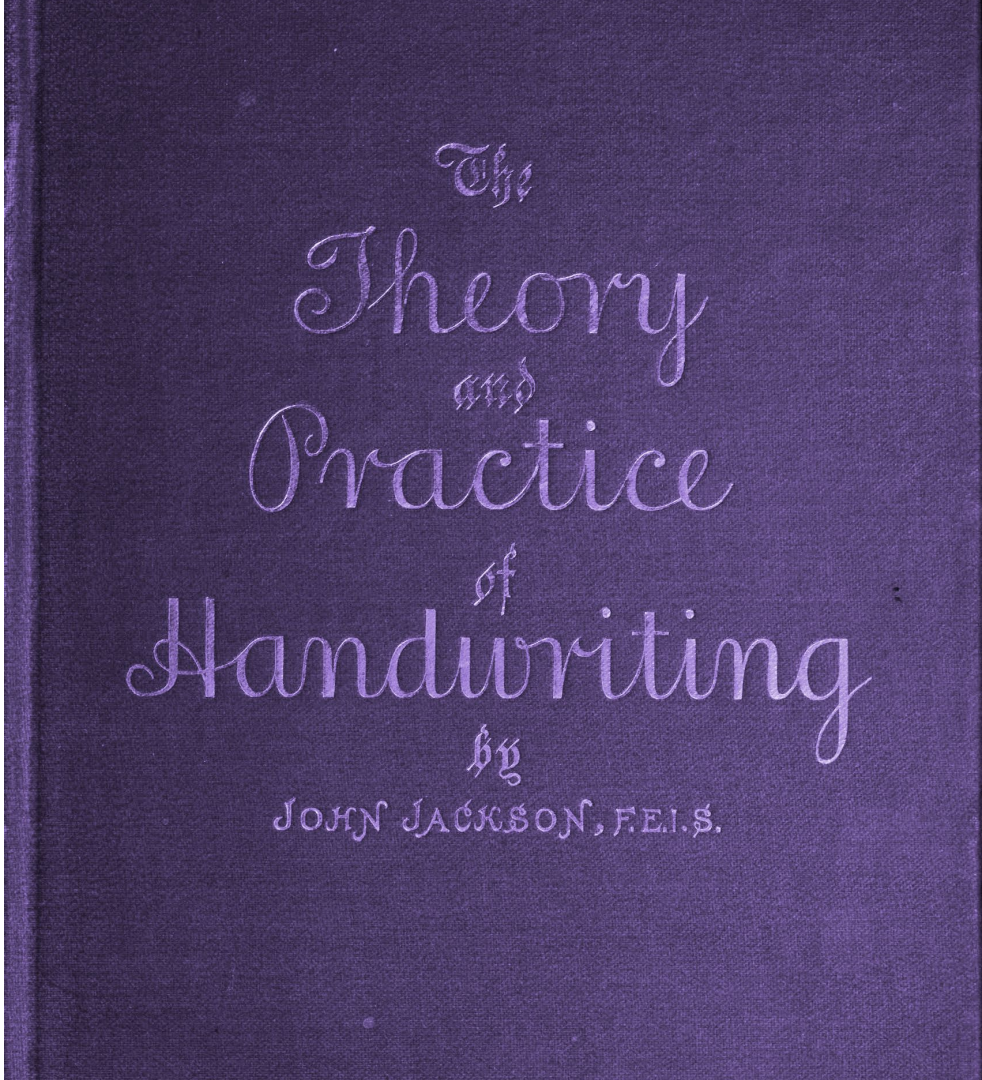
Intensive introduction to statistics and programming in two key languages.

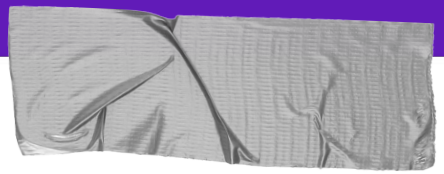
- **Stats concepts (today)**
  Data, probability, uncertainty, communicating analysis
- **RStudio (Tues & Weds)**
  Reading and writing code, summarising, visualising and analysing data
- **Python (Thurs & Fri)**
  Colab, Jupyter notebooks and learning a second programming language

Please head to Business Data Analytics Datacamp ELE page

- All course materials here
- Also details of how to download R & RStudio

Statistics is not just empirical.



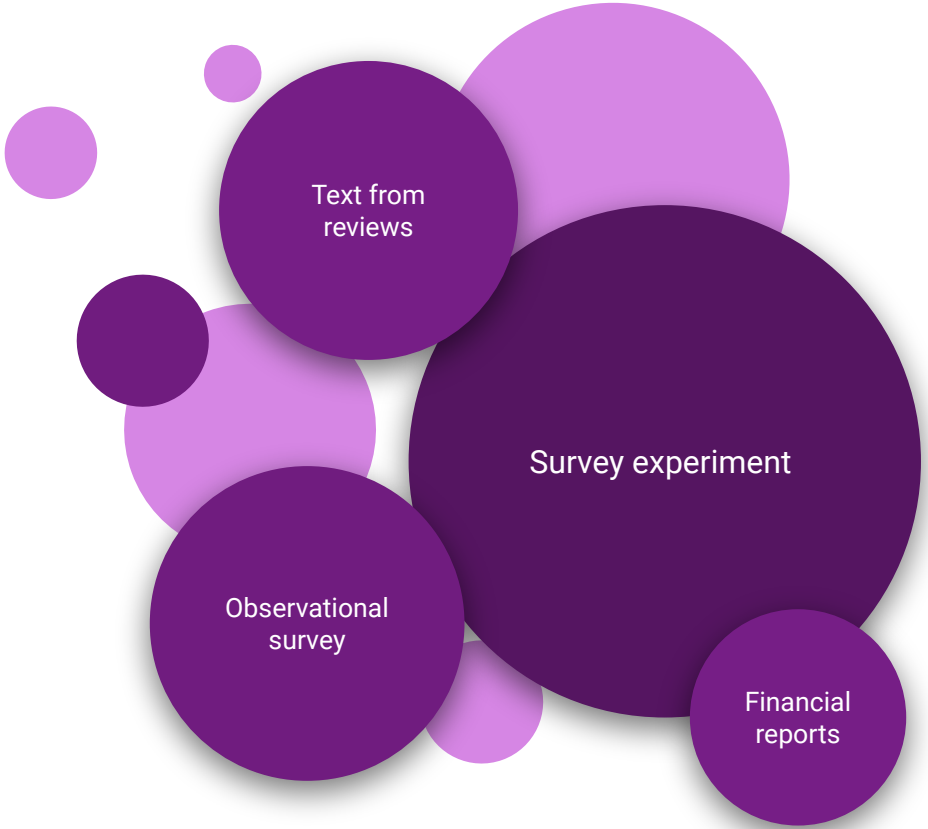


Topics today

- **Data**
  Where data comes from, types of data, approaches to data and analysis
- **Probability theory**
  Trees, birthday and Monty Hall problems
- **Uncertainty**
  Confidence, sample sizes
- **Hypothesis testing**
  Falsification, statistical significance
- **Communicating results**
  Communicating and visualising data

What is data?

Data are a collection of observations.



Where does data come from?

Observations can be gathered from almost anywhere. Usually we start with something we want to find out and search for or build our data from there.

Primary Data

“Data that are generated by a researcher who is responsible for the design of the study and their collection, analysis and reporting”

Secondary Data

“Raw data that have been collected by someone other than the researcher in question, either for some general information purpose, such as a government census, or for a specific research project.”

(Blaikie, 2003: 18)

Tip

There are lots of classifications for different types of data, but you should always know where your data came from.

Quantitative (Numeric) Data

"Data that are transformed into numbers immediately after they are collected or prior to the analysis, that remain in number during the analysis, and the findings from which are reported in numbers."

Qualitative (Non-numeric) Data

"Data that are recorded in words, that remain in words throughout the analysis, and the findings from which are reported in words."

(Blaikie,2003:20)

Tip

Quantitative and qualitative are also *approaches to research* where the latter focuses on methods such as interviews or ethnography.

Quantitative (numeric) data

Continuous

Discrete

Qualitative (non-numeric) data

Ordinal

Nominal

Traditional methods of collection

Anything that could've been done before the internet.
E.g. surveys, experiments, interviews

Newer methods of collection

Data generated thanks to the advent of the internet.
E.g. webscraping, tracking, 'smart' devices

Tip

There's no hierarchy for the 'best' type of data or collection. The best type always depends on what you're trying to find out.

Observed Data/Variables

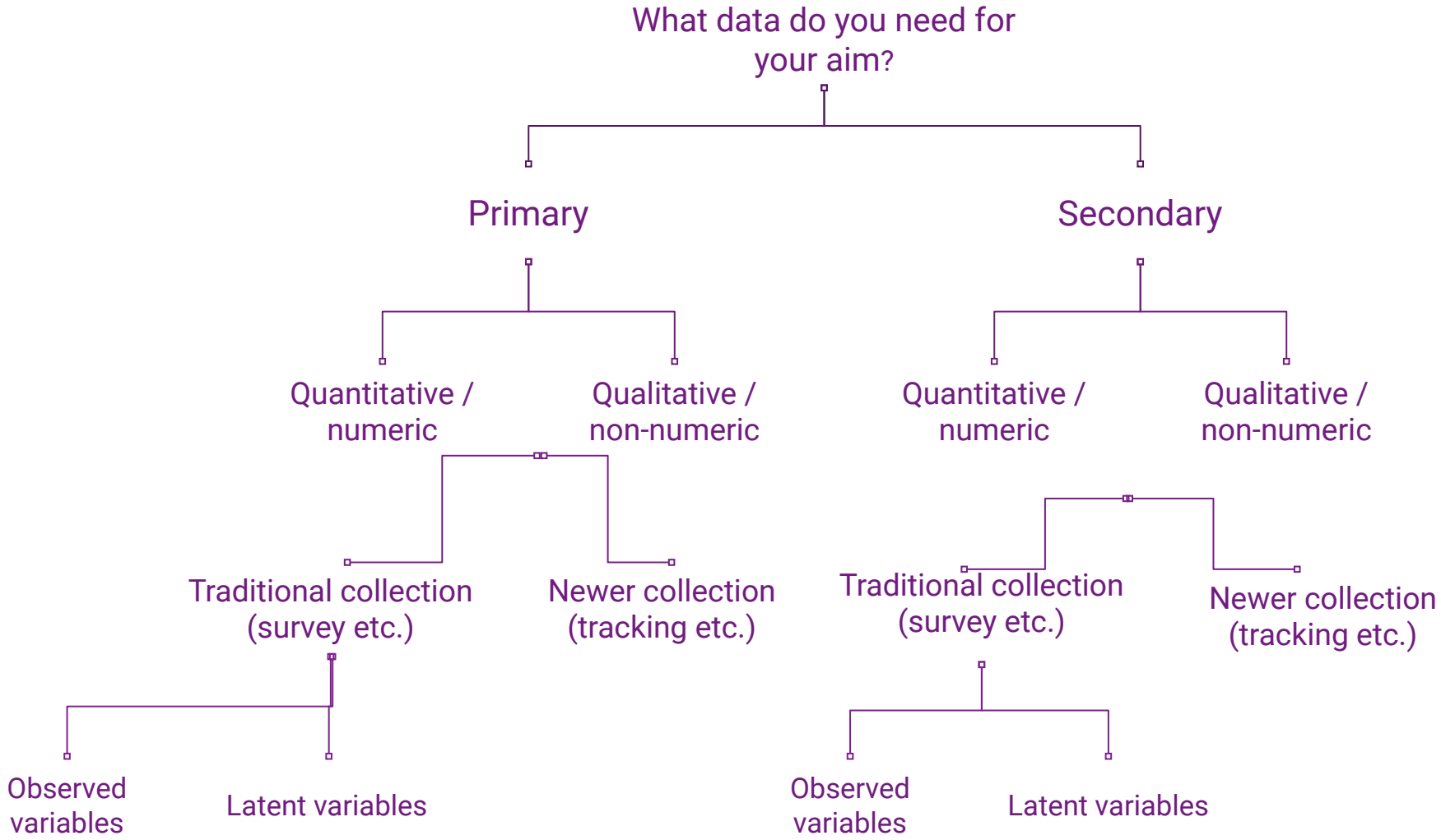
Something we can easily see and measure.
E.g. How many children do you have?
0, 1, 2, 3, 4, 5+

Latent Data/Variables

Generally a 'concept' or something not easily seen and/or measured.
E.g. *Social class* is a collection of economic and social factors. Or *attitudes to a company* may be made up of several components.

Tip

We can use a series of observations to estimate a latent variable. This is generally considered quite advanced analysis.


What data do you need for your aim?
Primary
Secondary
Quantitative / numeric
Qualitative / non-numeric
Quantitative / numeric
Qualitative / non-numeric
Traditional collection (survey etc.)
Newer collection (tracking etc.)
Traditional collection (survey etc.)
Newer collection (tracking etc.)
Observed variables
Latent variables
Observed variables
Latent variables

Approaches to data

Just like with any science, there are different perspectives and approaches to statistics.

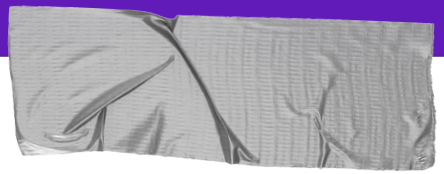
Frequentist

- More common
- P-values, confidence intervals
- Based on the data we have

Bayesian

- More intuitive
- Priors, credible intervals
- Based on probability

For more info:
https://cxl.com/blog/bayesian-frequentist-ab-testing/

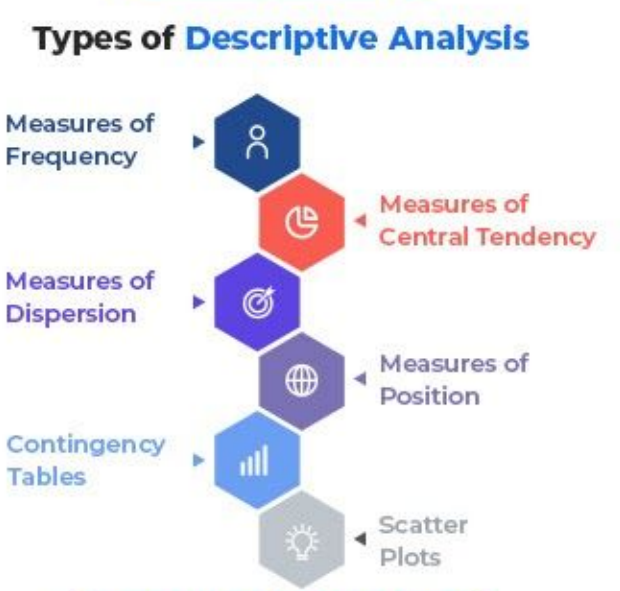


Types of analysis

- **Descriptive**
  As the name suggests, describing the data with percentages and trends

- **Inferential**
  Again self explanatory, making inferences about a whole population using the sample we have

- **Predictive**
  An extension of inferential that predicts how a population would behave using the sample we have



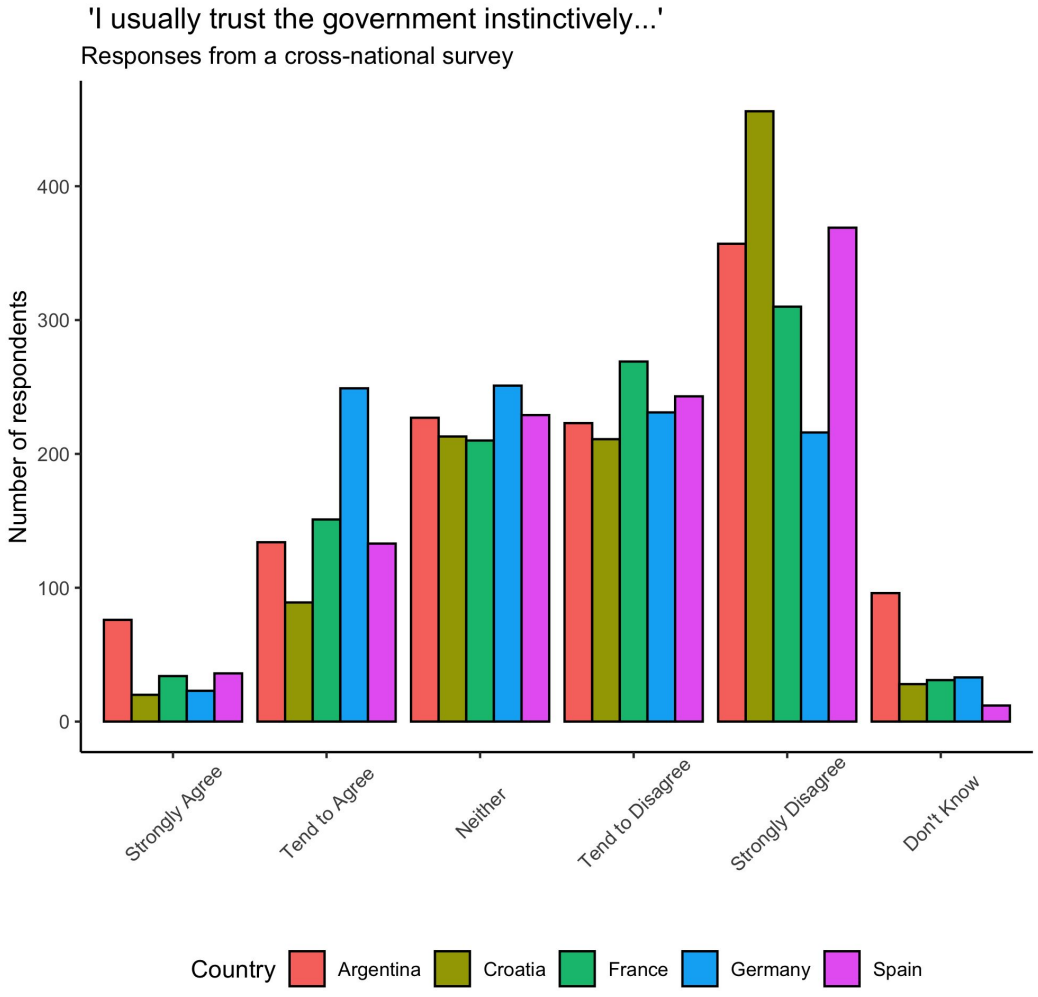
Descriptive

What do our data look like?

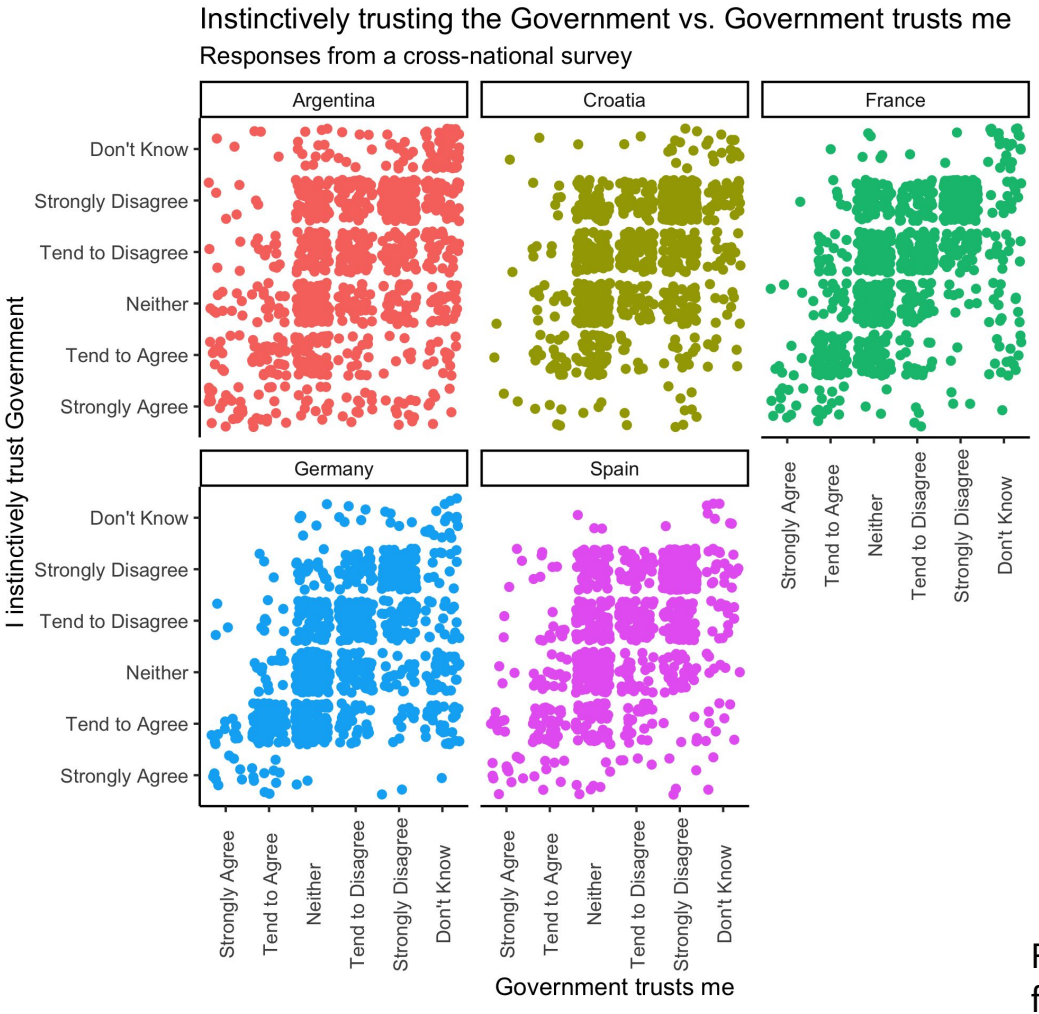
What are the minimum and maximum values?

What is the average response?

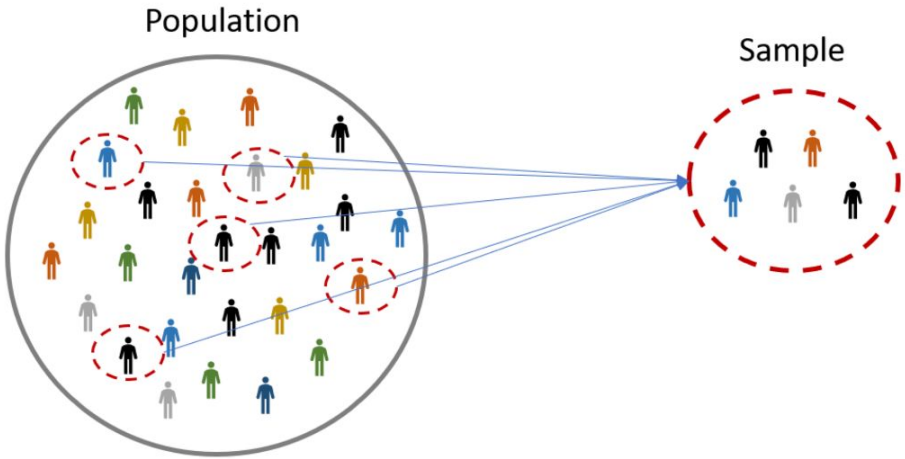
What % of people are in each category?



From Bunting et al., forthcoming



From Bunting et al., forthcoming



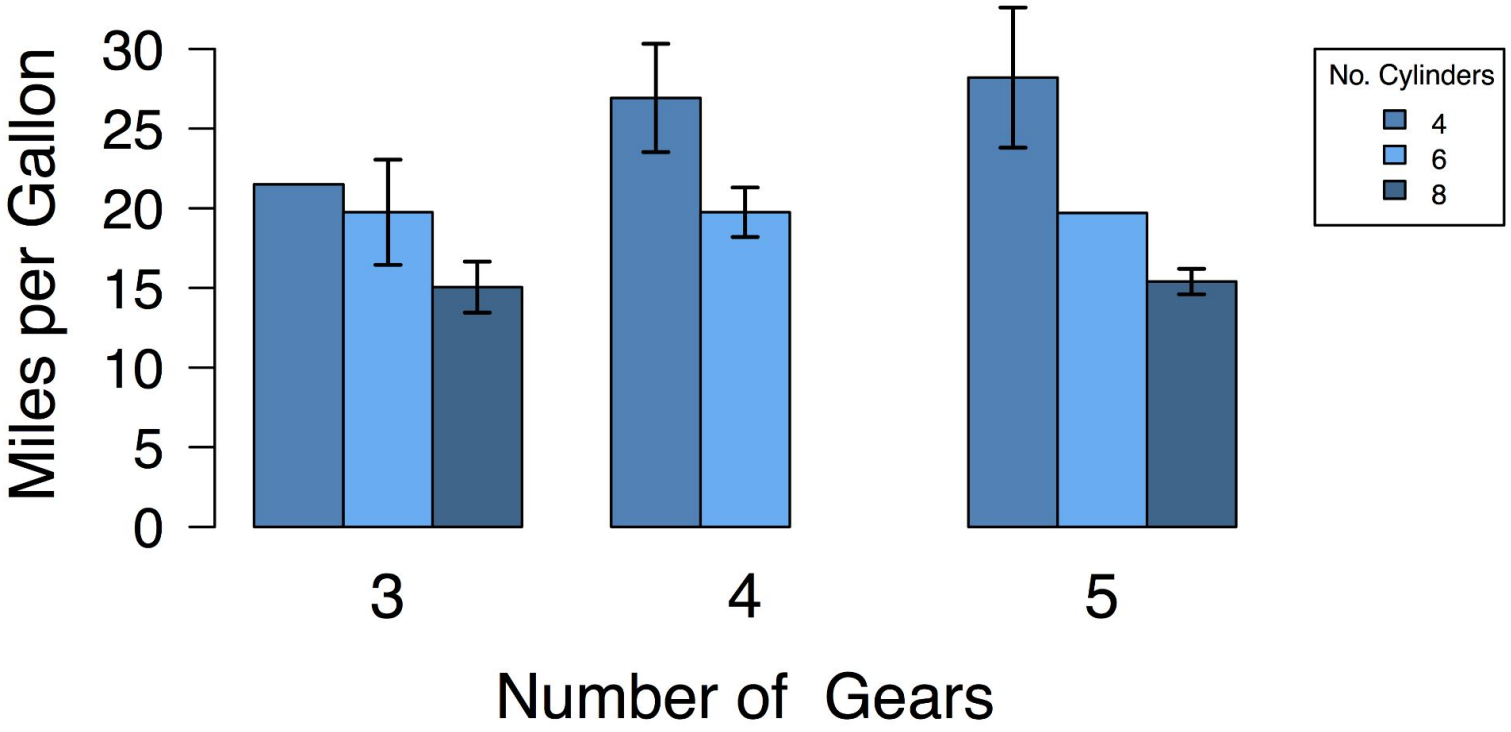
Inferential

What do customers think of our product?

Are we more successful at attracting men than other genders?

Favourite products by age group

Mileage by No. Cylinders and No. Gears



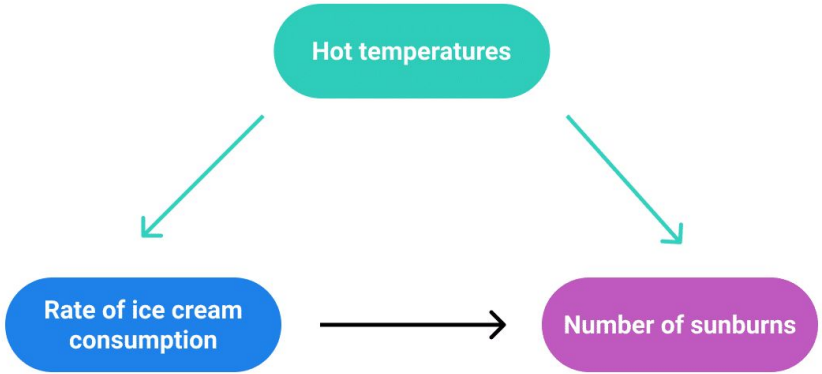
Predictive

"inferential analysis infers properties from tests and estimates, whereas predictive analytics focus on more the past and past behavior to better predict the future" - Barge, 2019

Who is the target for our latest product?

What factors make a customer more likely to buy?

How can we summarise this? We're looking for the relationship between x and y



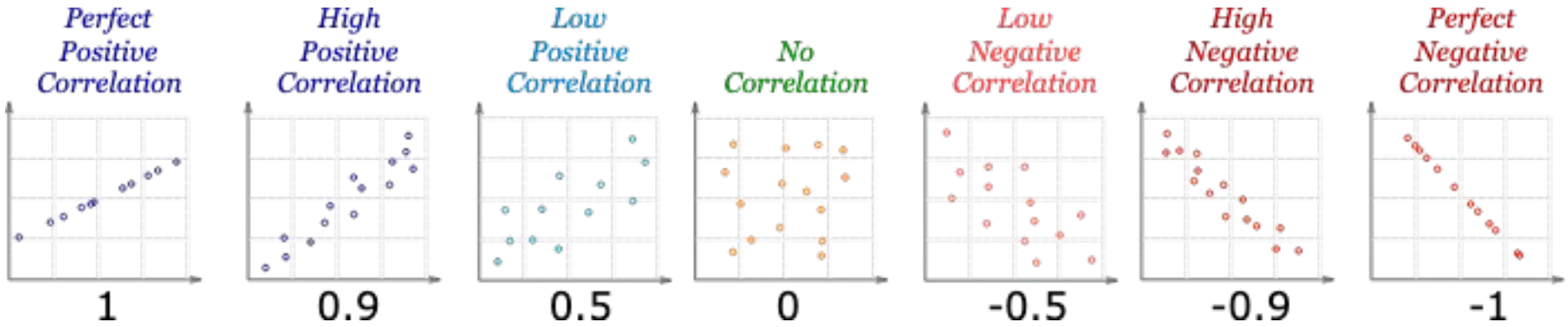
Relationships between variables

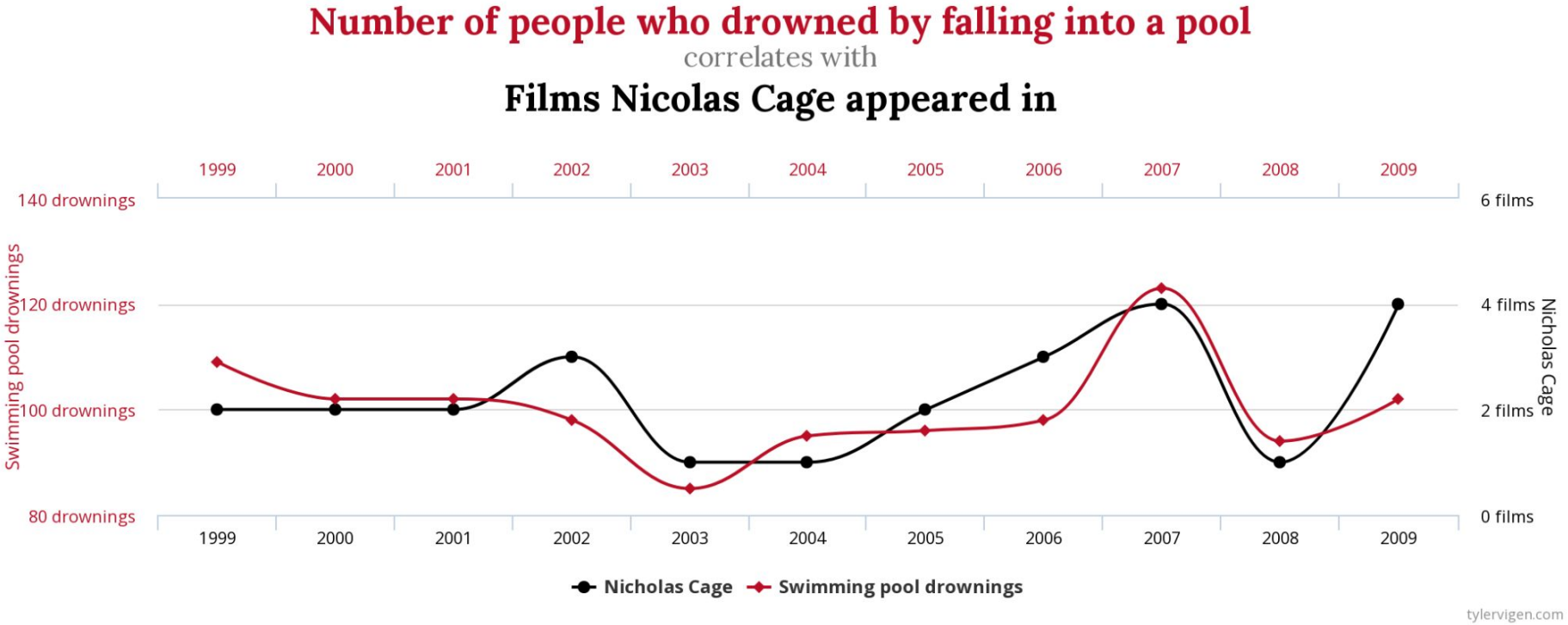
Does being a member of group *x* mean you're more likely to have *y* outcome?

Does an increase in *x* mean an increase in *y*?

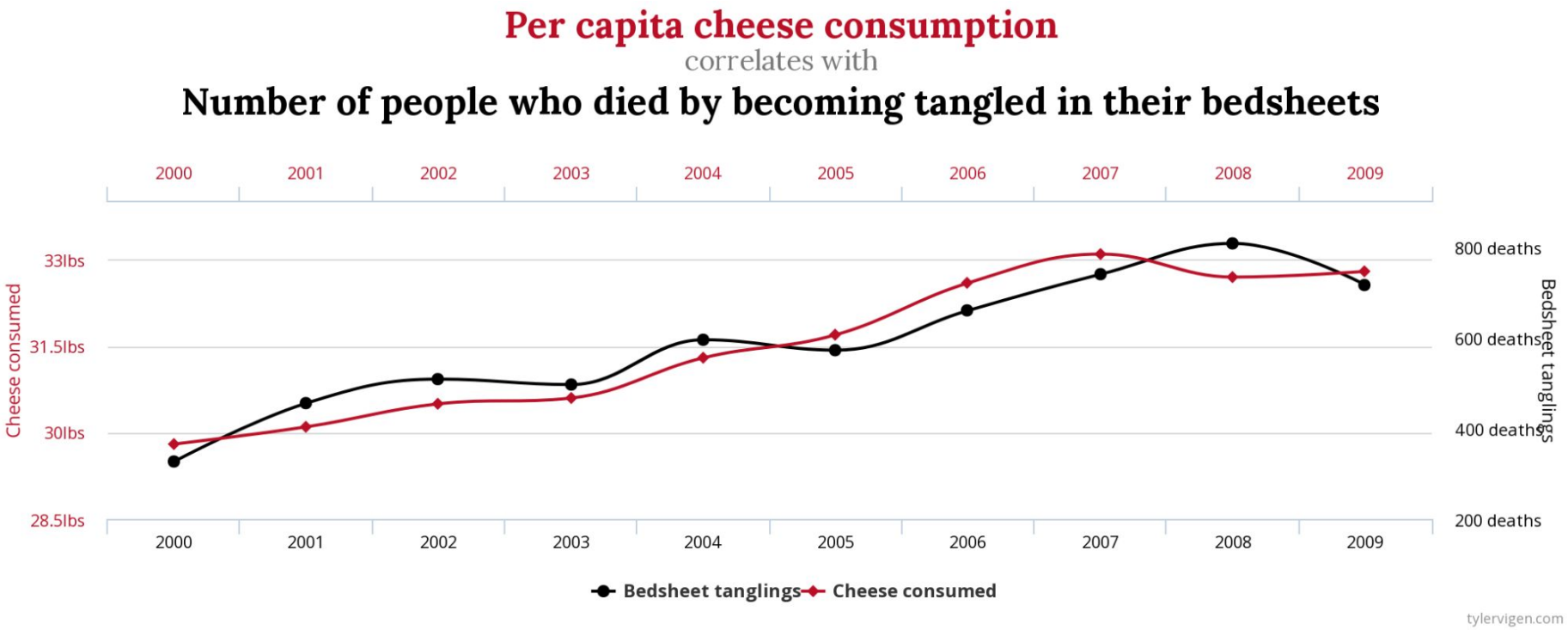
Does *x* cause *y*?

Correlation





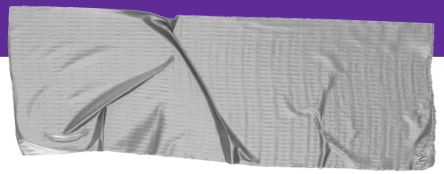
More here:
https://www.tylervigen.com/spurious-correlations



More here:
https://www.tylervigen.com/spurious-correlations

Correlation does not equal causation.

More here:
https://towardsdatascience.com/4-reasons-why-correlation-does-not-imply-causation-f202f69fe979



Establishing causation

- **1. Temporal precedence**
  The cause must happen before the effect
- **2. Association**
  The cause must logically have a connection to the effect
- **3. Ruling out alternative explanations**
  Did we miss something?

The Rubin causal model

Based on the idea of ***potential outcomes*** and ***counterfactual conditions***

Lets pose a question: does choosing a Business degree at Exeter **cause** higher wages?

So if we want to find the cause of the ***outcome*** wage, we are interested in the ***counterfactual condition*** of whether you chose an Exeter Business degree or not

---

A simple example:

James' ***potential outcomes*** are £28,000 a year and £42,000 a year

James' ***counterfactual conditions*** are he chooses an Exeter Business degree or he doesn't choose an Exeter Business degree

The Rubin causal model

But! The fundamental problem of causal inference is:

We can only observe *one* of these conditions and outcomes

One person cannot have both outcomes and/or both conditions (hence why they are called counterfactual)

Was it the Exeter Business degree that caused James to have a £42,000 wage? Or was it his gender, social class, work experience... etc.?

So, still in search of **ruling out alternative explanations**...

The Rubin causal model

We seek to find the *average* effects between a **treatment group** and a **control group**

Get a number of people together, some who have received the treatment of an Exeter Business degree and others who haven't, and we measure their wages

With the right balance of people, or better yet a *randomised controlled trial (RTC)*, we can rule out many of the alternative explanations and get at causality

---

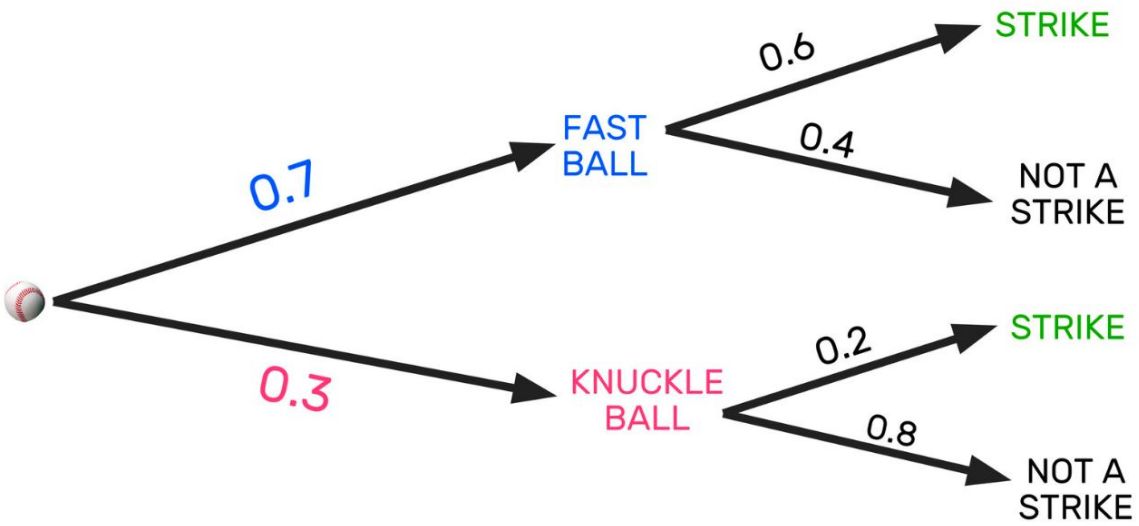
RTCs randomise selection into either the treatment or control group

There will be no systematic differences (random ones remain) so if there is an effect, it's as close to causal as we can get

Something all of this relies on: probability.

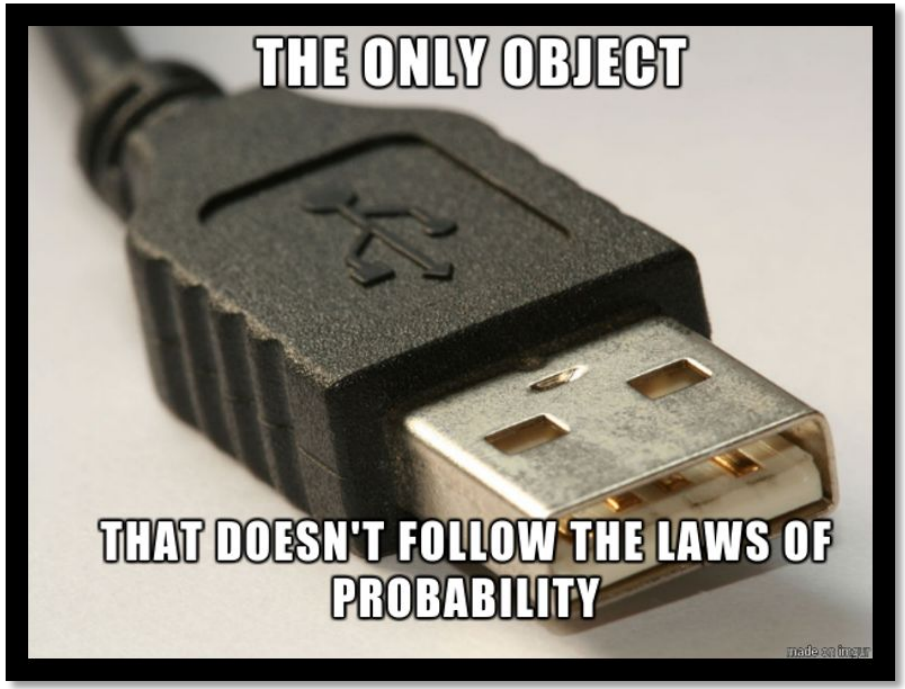
Remember

Every event has a probability of occurrence from 0 to 1.



Calculations

Fastball - Strike:	0.7 x 0.6 = 0.42
Fastball - No Strike:	0.7 x 0.4 = 0.28
Knuckler - Strike:	0.3 x 0.2 = 0.06
Knuckler - No Strike:	0.3 x 0.8 = 0.24
	+ ———
	1.00



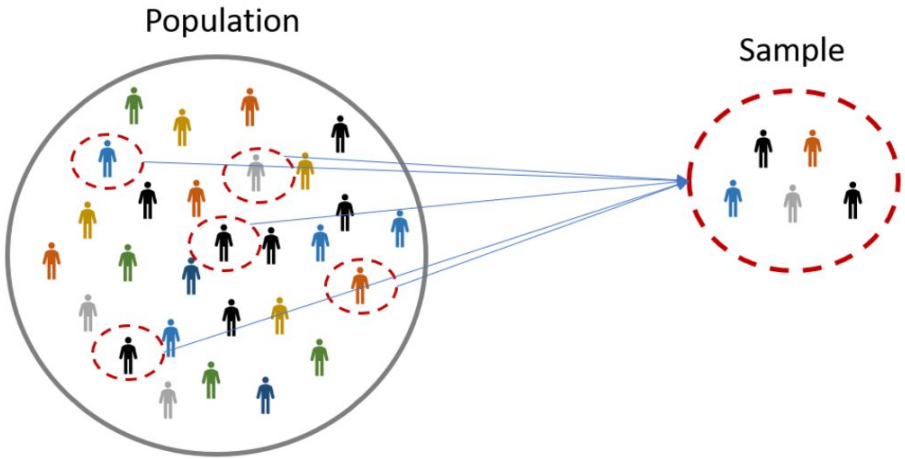
Probability isn't always intuitive

But it is important for statistics

https://www.youtube.com/watch?v=KtT_cgMzHx8

https://www.youtube.com/watch?v=4Lb-6rxZxx0

Uncertainty.

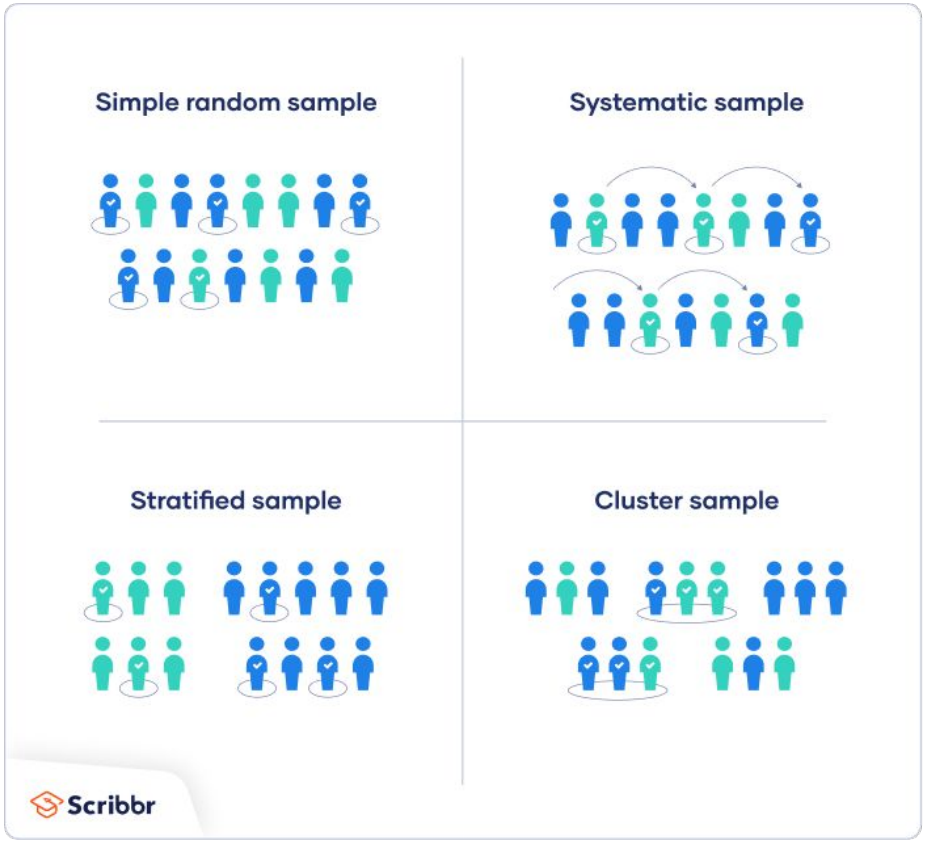


Dealing with uncertainty is at the heart of inferential statistics

We use things that we know - sample estimates - to make an educated guess about the things that we really want to know - population parameters

Estimate (statistic) – a numerical summary of **sample** e.g a mean or proportion (usually known)

Parameter – a numerical summary of **population**, e.g. a mean or proportion (usually unknown)



What kind of sample we need?

It should be a probability sample

The main principle is to avoid selection bias i.e. provide each person with an equal chance of being selected into the sample

So how it all works?

Say you take a random sample of 225 customers and ask them about their satisfaction with your product on a scale from 0 to 10..

..and you calculate that the sample estimate of average product satisfaction is 7.3

Could you then say that among all customers the average product satisfaction is exactly 7.3?

Unfortunately, it's not that easy. If you were to retake a sample of the same size, it's very likely that the average satisfaction level would be somewhat different:

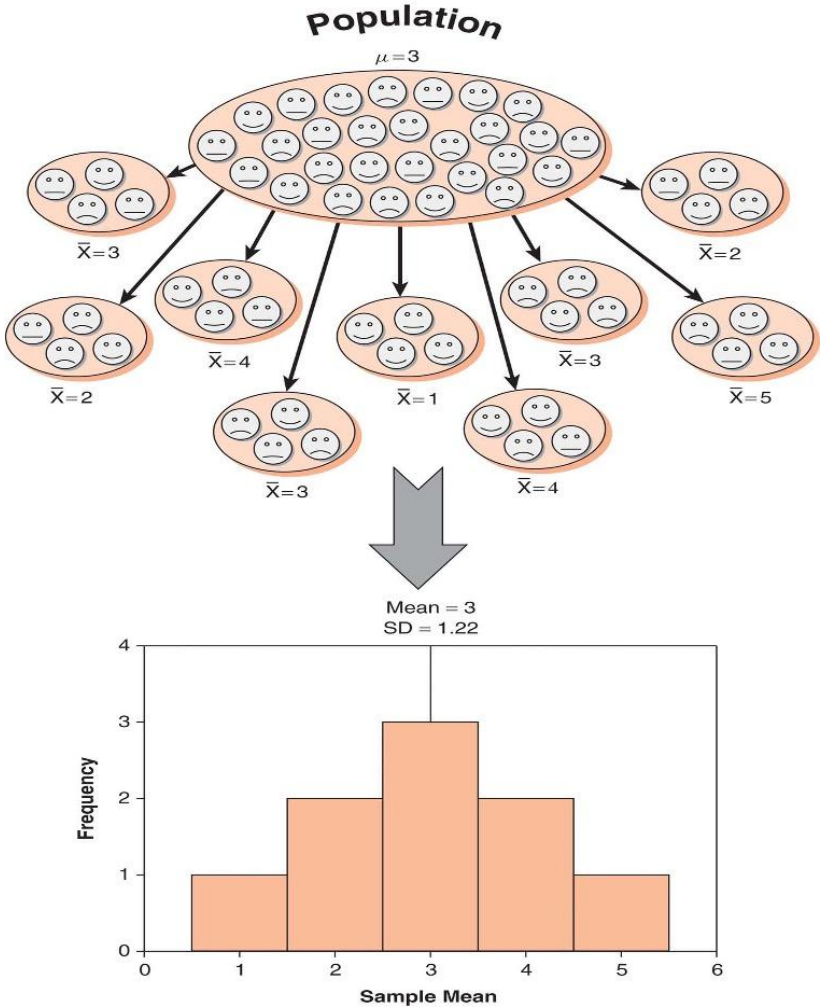
1st sample: 7.3 out of 10

2nd sample: 6.5 out of 10

3rd sample: 8.8 out of 10

...

nth sample: 7.1 out of 10



In this hypothetical scenario of drawing n number of samples of the same size, you would end up with a distribution of average values

This distribution is called a sampling distribution

It is a theoretical construct

Sampling distribution shows **all the possible values** that the statistic (mean, proportion, etc.) can take, and **how often they would occur**, if you took **all possible random samples** of the same size from the population

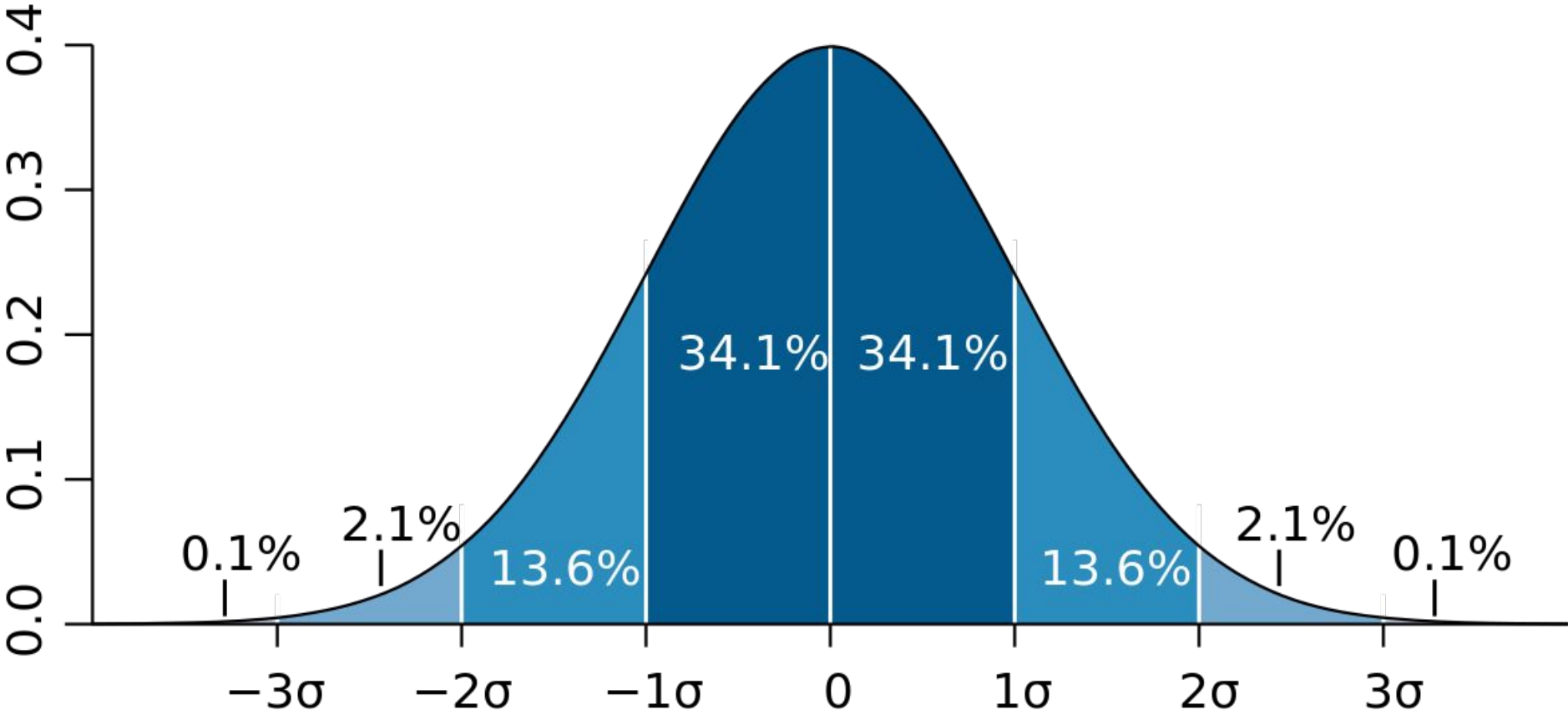
Turns out that sampling distribution has some useful statistical properties:

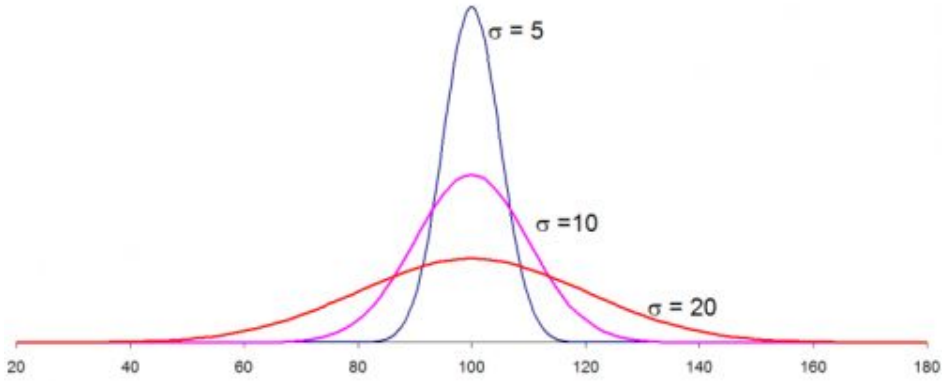
1. It follows a "bell curve" (normal) distribution, regardless of how skewed the true population distribution is (e.g. income)

2. The mean of a sampling distribution is in fact a population mean - the parameter that we are looking for!

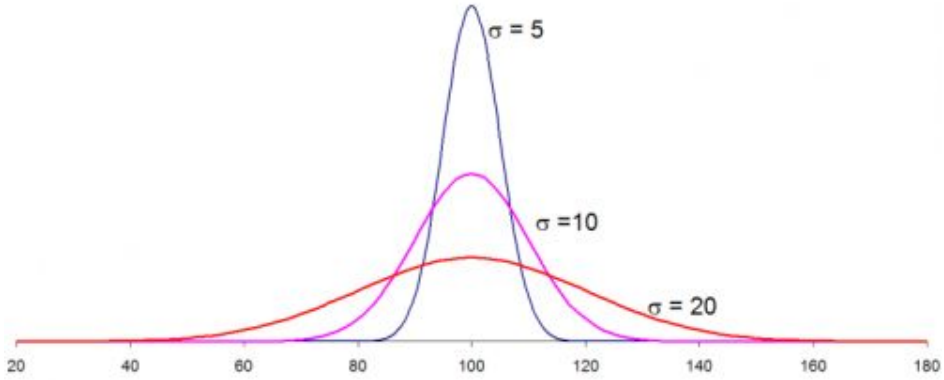
Let’s look at the example

The very fact that sampling distribution has a “bell curve” (normal) shape allows us to say quite a lot about the accuracy of our single point sample estimate..


0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ



..however, even normal sampling distribution can be narrow or wide, depending on its standard deviation, which is called standard error



$$SE = \frac{\sigma}{\sqrt{n}}$$

Standard error can be obtained with a simple formula:

Sample Standard Deviation / Square Root of Sample Size

Larger n -> smaller SE -> narrower sampling distribution -> more accurate point estimate

Let's now get back to the example with product satisfaction

We know that the sample estimate of average product satisfaction is 7.3

Even if the sample size was large enough, you still cannot be entirely sure that your sample estimate adequately represents the population average

What can we do about that?

Report an interval estimate instead of a point estimate

Point estimate - your single best “guess” about the population parameter - what you’ve got in your sample (e.g. 7.3)

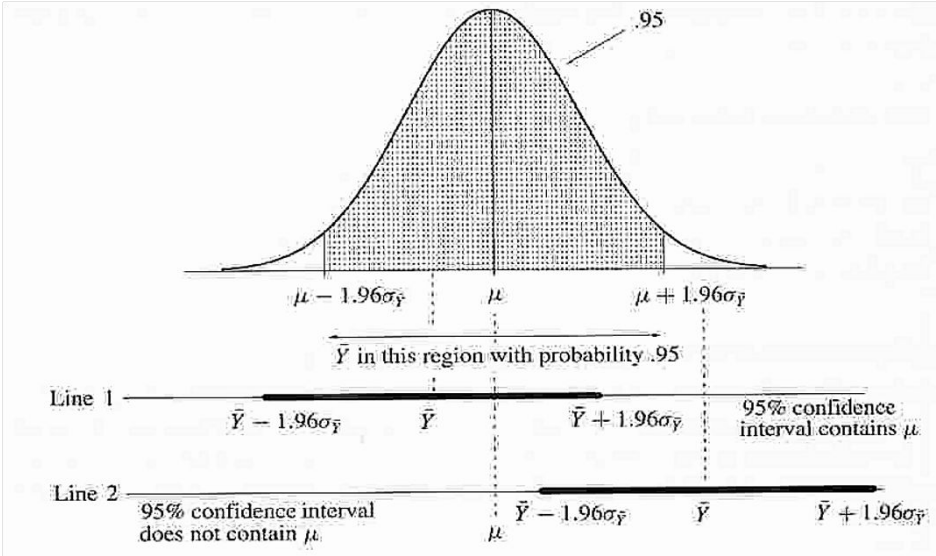
Interval estimate - a range of values around the point estimate that the population parameter is likely to fall into (e.g. from 6.5 to 7.8)

This is also called a confidence interval

Confidence interval represents a probability that a population parameter will lie within its range

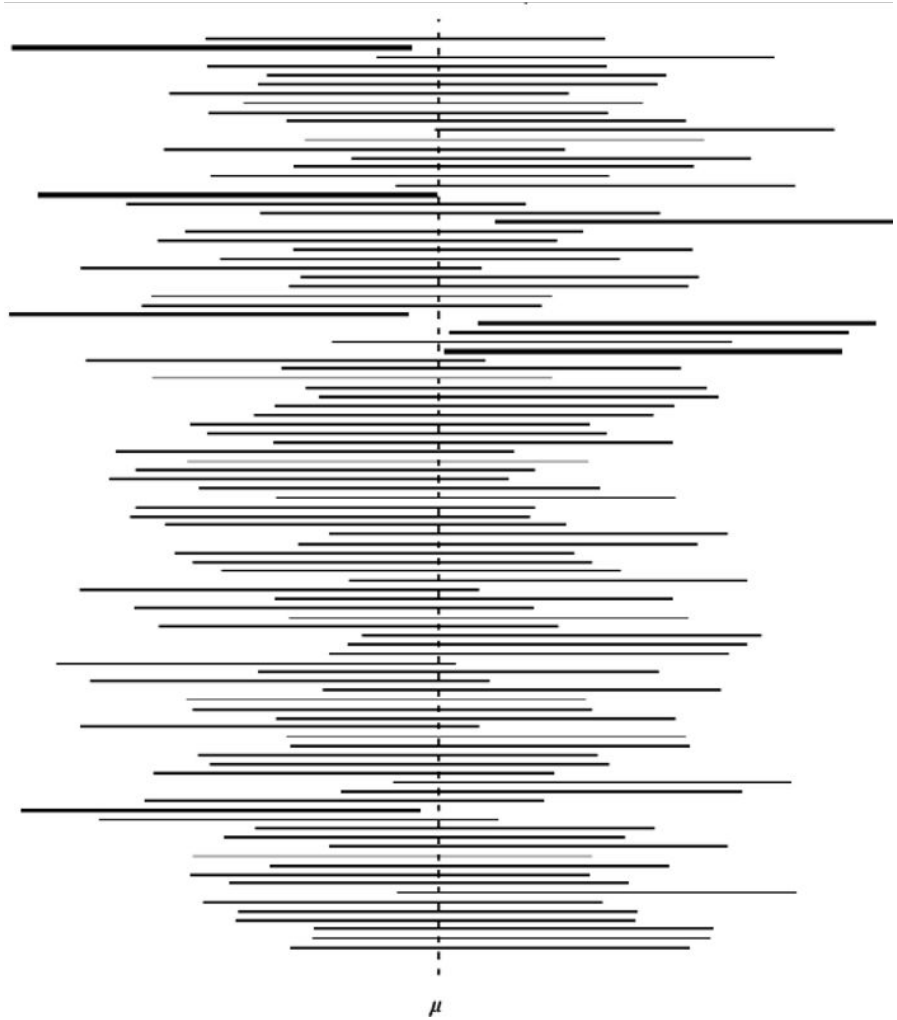
Its width depends on how certain you want to be that the drawn confidence interval indeed contains a population parameter

Usually, social sciences accept 95% confidence level



So if you know from the properties of a normal distribution that 95% of all its values lie within 1.96 standard errors from the mean..

..you can extend the value of your sample estimate by 1.96 of standard errors, so you can be 95% sure that the mean of sampling distribution (population parameter) will fall within this range



It’s also true that among 100 drawn samples, for 5 of them the confidence interval won’t capture the true population mean

This is a trade-off between the confidence and precision

Higher confidence - lower precision

Lower confidence - higher precision

You can only be 100% sure that the population mean lies somewhere from 0 to 10 (if this is the range of your scale)

This, however, doesn’t tell us anything

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

Below is an example of confidence interval calculation:

Sample mean: 7.3

Sample SD: 2.5

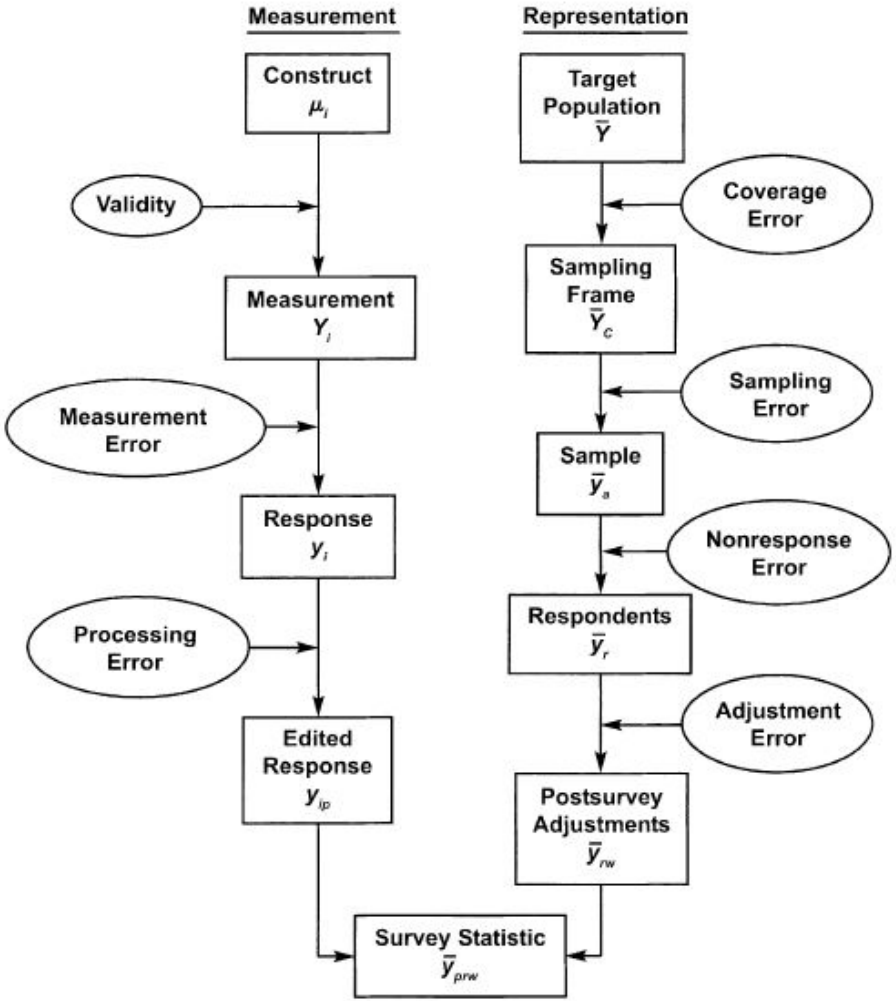
Sample size: 225

Confidence level value: 1.96 for 95%

CI_{lower} = 7.3 - 1.96 * (2.5 / 15) = 7.3 - 0.33 = 6.97

CI_{upper} = 7.3 + 1.96 * (2.5 / 15) = 7.3 + 0.33 = 7.63

Hence, you can claim with 95% confidence that the true average value of population satisfaction with your product lies somewhere between 6.97 and 7.63



P.S. Total Survey Error Framework

Hypothesis testing.

Inferential statistics is all about testing the **statistical hypotheses**

Statistical hypothesis is a statement about **population parameter** which one can test with **sample data**

There are two main types of statistical hypotheses:

Null hypothesis (H0)
Alternative hypothesis (Ha)

While the exact wording of H0 and Ha depends on a statistical test (e.g. correlation, T-test, etc.), they all follow the same principle:

Null hypothesis (H0)	Alternative hypothesis (Ha)
Represents the “default” state of the world	Represents a statement about the world that the researcher would like to test
Assumed to be true	Rejects the assumption that H0 is true
Contains equality (=)	Does not contain equality (!=, >, <)

Both statistical hypotheses refer to the general population, not sample data!

A few examples

Correlation analysis:

H0: The value of the correlation coefficient between salary and years of work experience is equal to 0

Ha: The value of the correlation coefficient between salary and years of work experience is not equal to 0

T-test:

H0: The average number of hours students spend on studying stats is equal to 32

Ha: The average number of hours students spend on studying stats is not equal to/more than/less than 32

Chi-square test:

H0: There is no association between smoking (Yes/No) and self-perceived life satisfaction (Low/Middle/High)

Ha: There is an association between smoking (Yes/No) and self-perceived life satisfaction (Low/Middle/High)

Knowing that we assume H0 to be a true representation of the world, the question that we implicitly ask running any statistical test is:

Does our sample data provide us with enough evidence to reject the Null Hypothesis (H0) and accept the Alternative Hypothesis (Ha)?

In other words: how likely it is to observe what we actually see in the sample data under the assumption that H0 holds true in the general population?

P-value is what quantifies the probability of obtaining your sample estimate (average value, correlation coefficient, T-test value, etc.) under the assumption that H0 holds true in the general population

Larger P-value - more likely that the statistic you observe in the sample data complies with the idea of H0 being true

Smaller P-value - less likely that the statistic you observe in the sample data complies with the idea of H0 being true

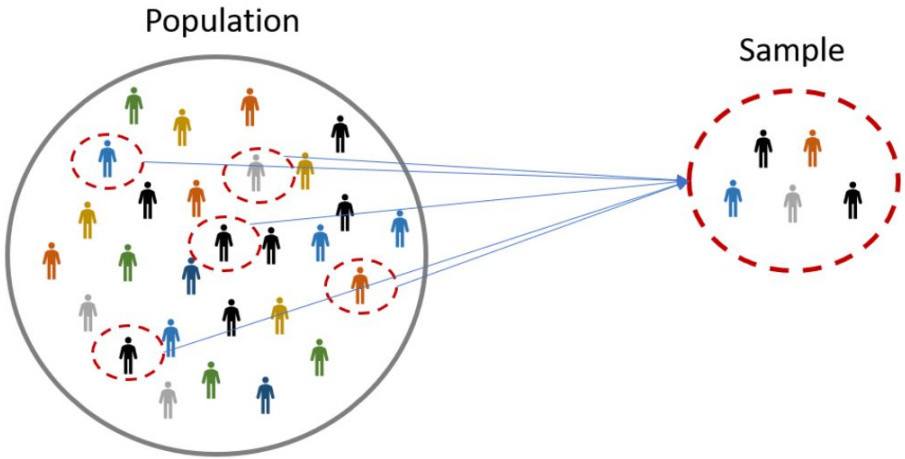
Example: product satisfaction among men and women

Say that the survey based on a random sample of customers has shown the following results:

Average product satisfaction among men: 5.5 out of 10

Average product satisfaction among women: 7.1 out of 10

Research question: do men and women have different levels of product satisfaction?



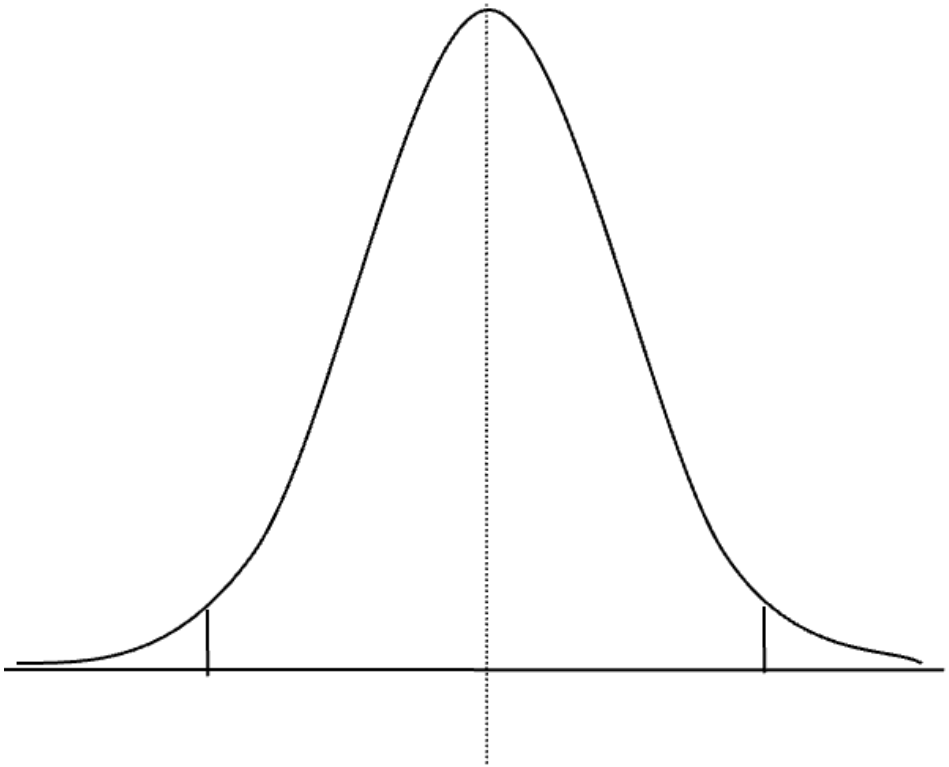
The levels of product satisfaction are clearly different in the sample data (5.5 vs 7.1)

But can we say the same about the general population? Do we have enough evidence?

Let's outline the statistical hypotheses for the test of equality of means in two groups (called Independent sample T-test)

H0: There is no difference in the average values of product satisfaction among men and women ($X_{men} = X_{women}$)

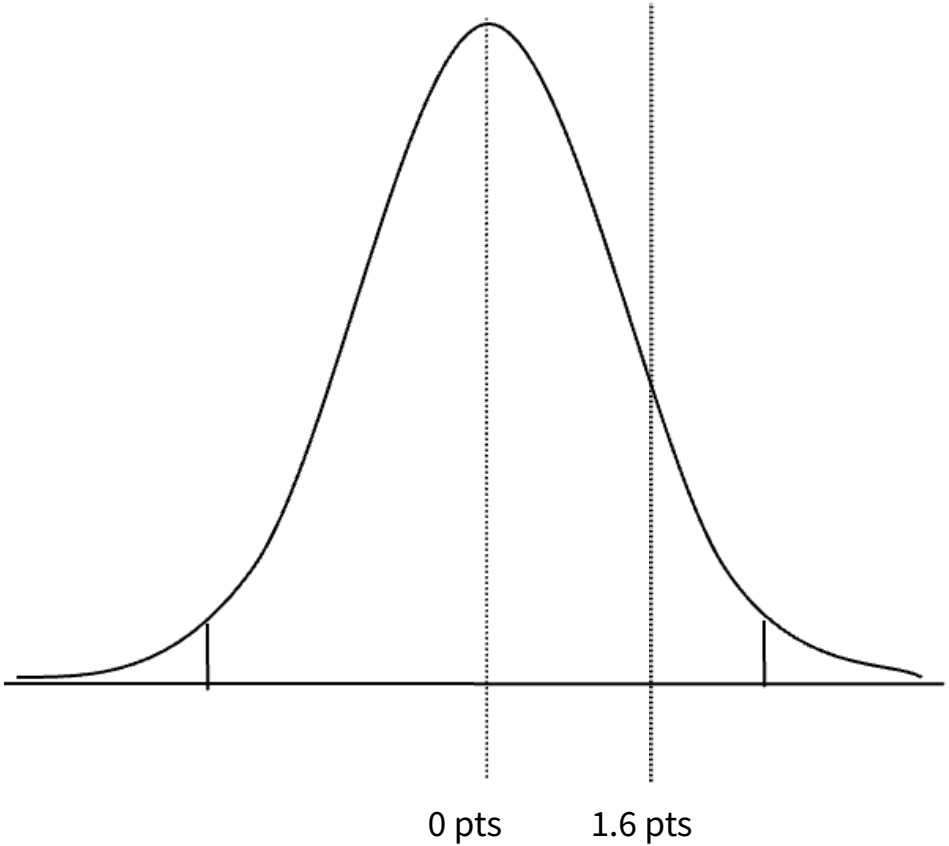
Ha: There is a difference in the average values of product satisfaction among men and women ($X_{men} != X_{women}$)



0 points

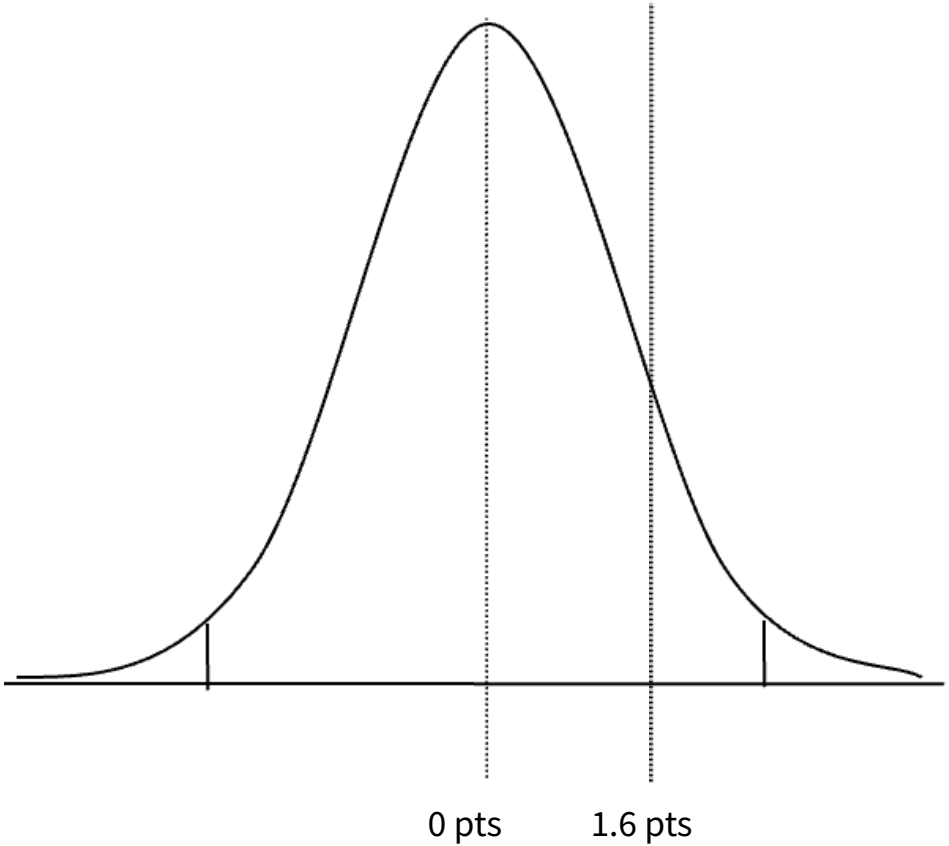
The sample mean difference between men and women is 1.6 (7.1 - 5.5)

However, the sampling distribution of this statistic that you would expect to see under the assumption that H0 is true, looks like this:



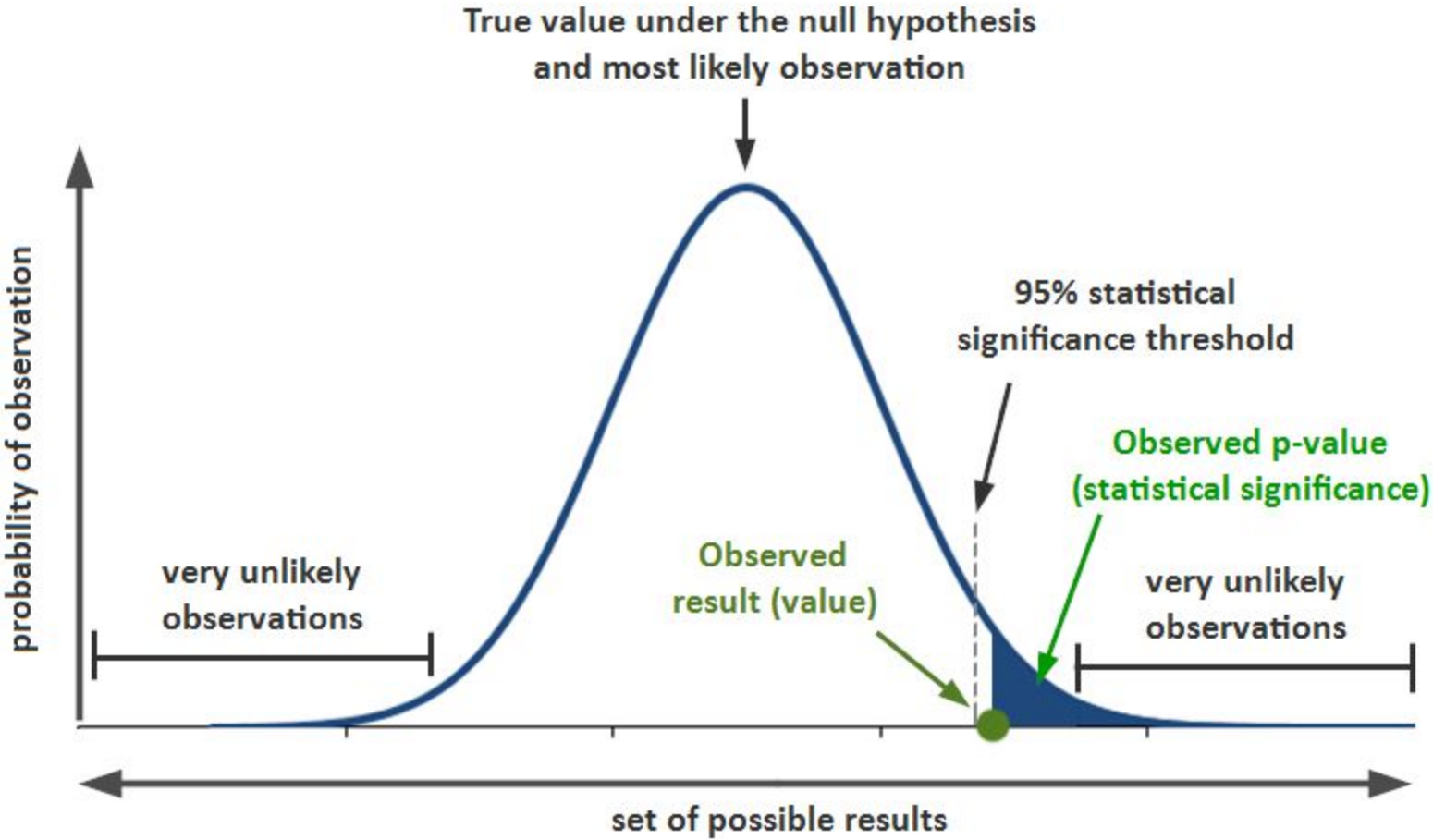
The sample mean difference between men and women is 1.6 (7.1 - 5.5)

However, the sampling distribution of this statistic that you would expect to see under the assumption that H0 is true, looks like this:



How likely it is to observe 1.6 point (or even more extreme) difference knowing that the sampling distribution looks like this?

This is where we come back to P-value!


True value under the null hypothesis
and most likely observation
95% statistical
significance threshold
Observed p-value
(statistical significance)
Observed
result (value)
very unlikely
observations
very unlikely
observations
probability of observation
set of possible results

P-value is essentially an area under the curve, and you have to set a significance level that the obtained P-value will be compared to prior running a statistical test

Say the significance level is 5%, in that case:

If your P-value is < 0.05, H0 should be rejected and Ha - accepted. The results are statistically significant

If your P-value is > 0.05, H0 should be retained and Ha - rejected. The results are statistically insignificant

More to come in the lab sessions!

Seeing the wood for the trees: communicating data analysis.

What does is it mean to be a data analyst?

Take observations

Gather our data.

Understand them

Perform our analyses.

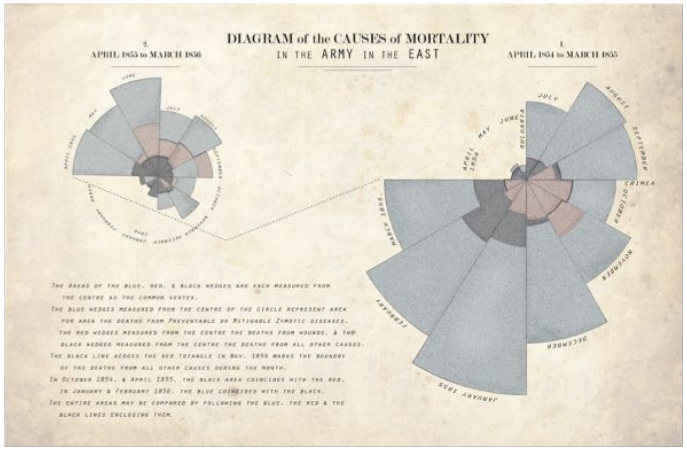
Communicate them

Share the results.

Tell a story with your data

Florence Nightingale

- Visualised her data in a compelling way
- Used it to enact change

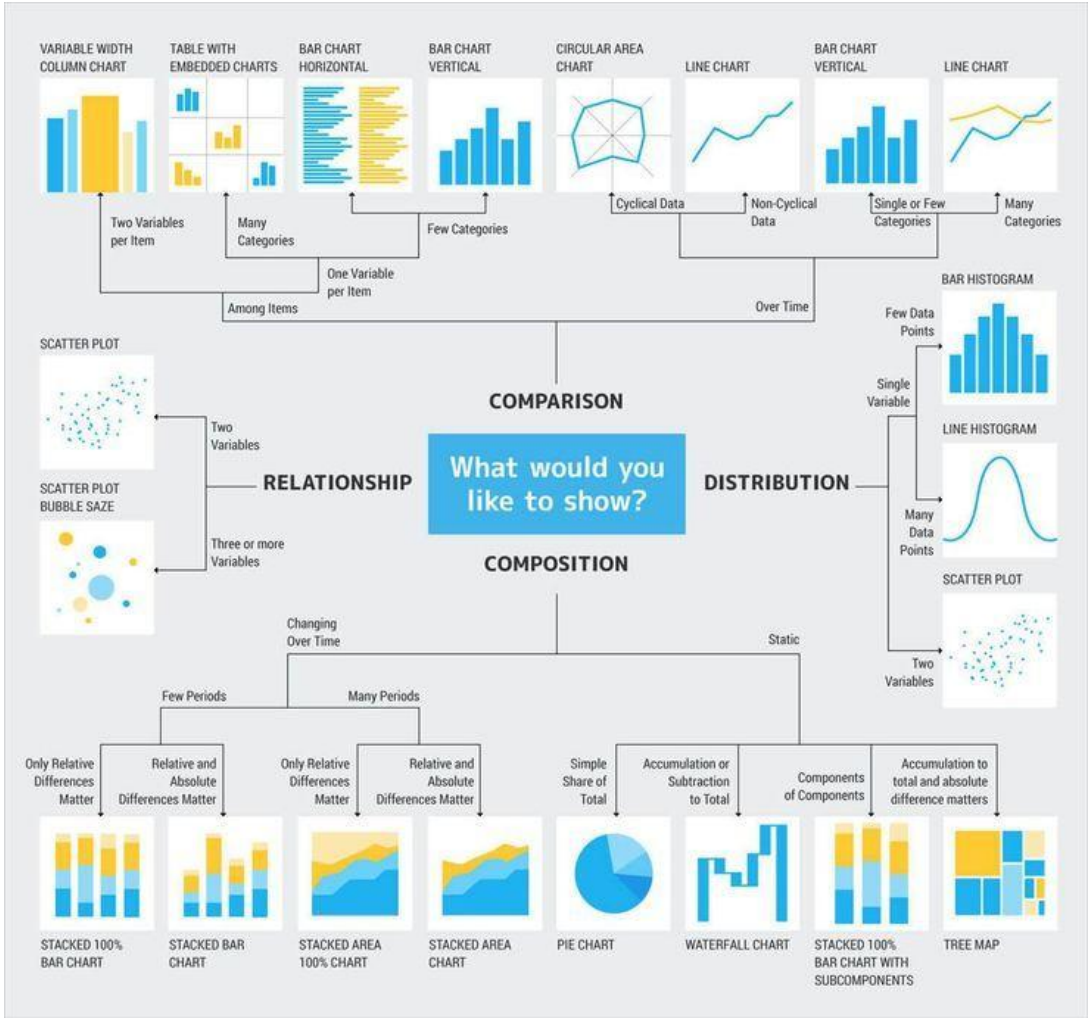


Gregor Mendel

- Thought his data would ‘speak for itself’
- The presentation was convoluted
- 30 years until someone made the same discovery

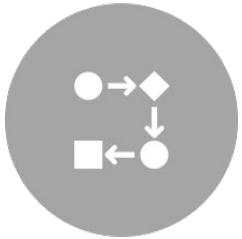
Principles for data vis

- Show the data, but don’t distort it.
- Induce the reader to think about the data, not the graph itself.
- Present many numbers with minimal ink.
- Encourage the reader to compare different pieces of data.
- Tell the story.


What would you like to show?
COMPARISON
RELATIONSHIP
DISTRIBUTION
COMPOSITION
VARIABLE WIDTH COLUMN CHART
TABLE WITH EMBEDDED CHARTS
BAR CHART HORIZONTAL
BAR CHART VERTICAL
CIRCULAR AREA CHART
LINE CHART
BAR CHART VERTICAL
LINE CHART
Two Variables per Item
Many Categories
Few Categories
One Variable per Item
Among Items
Cyclical Data
Non-Cyclical Data
Single or Few Categories
Many Categories
Over Time
SCATTER PLOT
Two Variables
SCATTER PLOT BUBBLE SAZE
Three or more Variables
BAR HISTOGRAM
Few Data Points
Single Variable
LINE HISTOGRAM
Many Data Points
SCATTER PLOT
Two Variables
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Accumulation to total and absolute difference matters
STACKED 100% BAR CHART
STACKED BAR CHART
STACKED AREA 100% CHART
STACKED AREA CHART
PIE CHART
WATERFALL CHART
STACKED 100% BAR CHART WITH SUBCOMPONENTS
TREE MAP

A parting note

Remember that some of statistics is convention: e.g. Why the 0.05 statistical significance cut-off?



Not everything is exact, there's often more than one way to do things.

The statistical tests we favour might have been different if we'd had a different history of statistical development.

There's a pragmatic rather than a pure spirit about statistical thinking.