MATH9102: Fundamentals of Data Analysis

W1 – FUNDAMENTALS

LECTURER: DR. DEIRDRE LAWLESS

"Data don't make any sense,
we will have to resort to statistics."

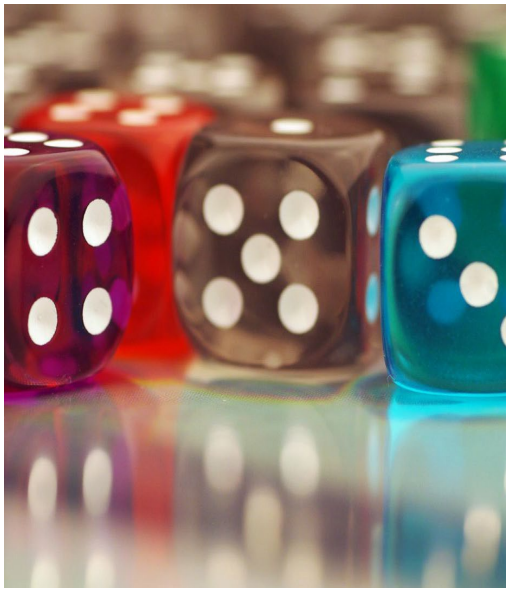
Probability v Statistics v Data Science

Probability v Statistics V Data Science

A person goes to a casino and encounters their first dice game. A six-sided dice is used.

If the person is a **probabilist**:

- They would think the dice is six sided
- If each side is equally likely to land face up, then the probability of getting a particular face of the dice is 1/6
- Using probability theory, they will be confident they can predict the likelihood that they will guess right (or wrong)



Probability v Statistics V Data Science

A person goes to a casino and encounters their first dice game. A six-sided dice is used.

If the person is a **statistician**:

- They would see the dice and think:
- The dice is six-sided. It could be that each side is equally likely to land face up.
- BUT – the dice might be loaded (some sides more likely than others).

They would:

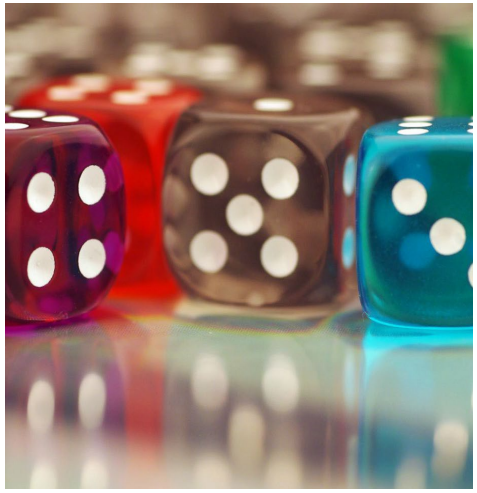
- Observe rolls for a while and record outcomes.
- Compare observed frequencies with what would be expected if the dice were fair.
- Use statistical tests (e.g., chi-squared) to check whether results are consistent with fairness.
- Once they had enough evidence,they could confidently decide whether to accept or question the fairness assumption.
- If consistent, they rely on probability theory to make predictions.

Probability v Statistics v Data Science

A person goes to a casino and encounters their first dice game. A six-sided dice is used.

If the person is a **data scientist :**

- Instead of immediately assuming each face has a probability of 1/61/61/6, they will want to collect data.
- They might roll the dice 100 times, record the outcomes, and build a dataset.
- OR
- They may source a dataset recording the outcomes of other people rolling a dice 100 times.

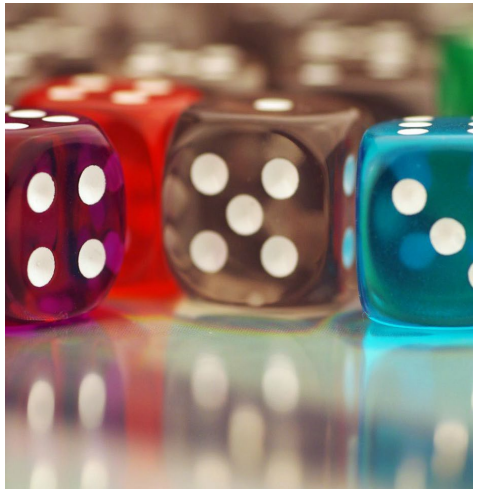


Probability v Statistics v Data Science

A person goes to a casino and encounters their first dice game. A six-sided dice is used.

If the person is a **data scientist :**

- Using the dataset, the data scientist will undertake some exploratory data analysis.
- They would create frequency tables and bar plots to check if each face appears roughly equally often.
- Example: after 100 rolls, the data might look like this:
- Face 1: 14 times; Face 2: 17 times; Face 3: 18 times; Face 4: 15 times; Face 5: 19 times; Face 6: 17 times
- This suggests relative frequencies close to 1/6, but the data scientist knows more evidence is needed.



Probability v Statistics v Data Science

The data scientist will then apply some statistical testing.

They may apply a statistical test (e.g. chi-squared test) to check whether the observed frequencies significantly deviate from the expected uniform distribution.

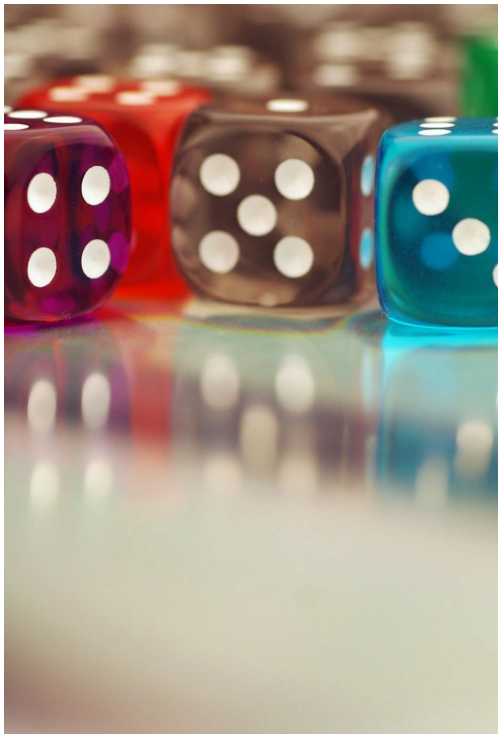
- If the p-value is large, they conclude the dice is fair (at least no evidence against fairness).
- If small, they suspect the dice is biased and might explore which face(s) occur more often.

The data scientist will then try to build a model

The probabilist assumes fairness, but the data scientist may model probabilities empirically:

$$P(\text{rolling face } i) \approx \frac{\text{observed count of } i}{\text{total rolls}}$$

Over many rolls, this estimate converges toward the true underlying probability distribution.



Probability v Statistics v Data Science

The data scientist will address prediction and uncertainty in their communication of their findings:

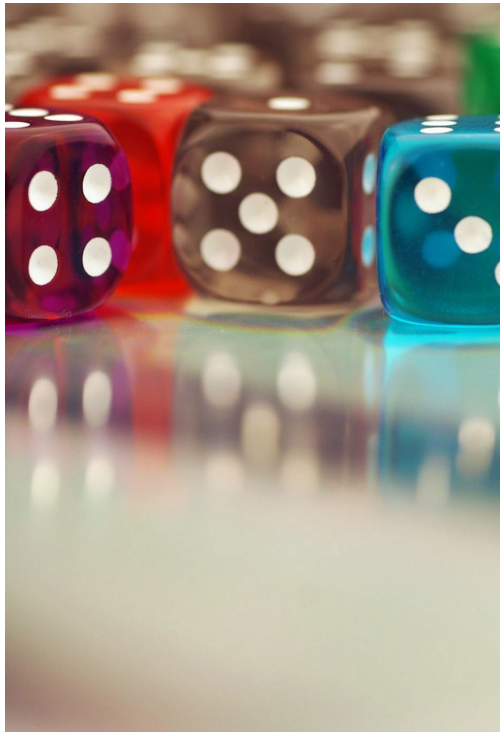
Instead of saying, “The chance is exactly 1/6,” the data scientist might say:

“Based on the observed data, the probability of rolling a 6 is estimated at 0.19 with a 95% confidence interval of [0.12, 0.26].”

This acknowledges both the data-driven estimate and the uncertainty around it.

The data scientist may also consider:

- Long-term expectation: Does the game favour the casino (house edge)?
- Simulation: Running Monte Carlo simulations of thousands of dice games to estimate likely profit/loss over time.
- Decision making: If the dice looks biased toward certain numbers, a betting strategy could exploit it.



Probability v Statistics V Data Science

Probabilist	**Theory-first.** Works with mathematical models and assumptions about randomness.	Assumes the die is fair → each face has probability 1/61/61/6. Uses probability theory to compute chances of winning/losing.
Statistician	**Inference and validation.** Designs experiments, collects data, and applies formal methods to test if the probabilist's assumptions hold.	Rolls the die many times, uses a chi-squared test to check if outcomes deviate from fairness. Reports whether data provides evidence for or against bias.
Data Scientist / Data Analyst	**Data-first, applied, and exploratory.** Gathers large datasets, explores patterns, builds predictive models, and communicates insights.	Collects thousands of dice rolls, visualises frequency distributions, estimates probabilities empirically, simulates casino outcomes, and advises on betting strategies if bias is found.

The Relationship

Probability provides the framework → what the model of fairness looks like.

Statistics provides the bridge → testing whether reality (data) aligns with the framework.

Data Science/Analysis provides the application → exploring data at scale, finding patterns, and generating actionable insights.

Together, they form a **theory–inference–application pipeline**:

Theory (probability)

Inference (statistics)

Application (data science/analysis)

What is Data Science?	The science of extracting knowledge and insights from data by combining statistics with computing, machine learning, and domain expertise
What is Data Analysis?	The process of inspecting, cleaning, and modelling data, using statistical methods to discover useful information and support decision-making.

What can statistics do?

Make data more manageable

6, 1, 8, 3, 5, 4, 9

- We can calculate an average (5.147…)
- We can state the range (1 to 9)
- We can create a visualization of it

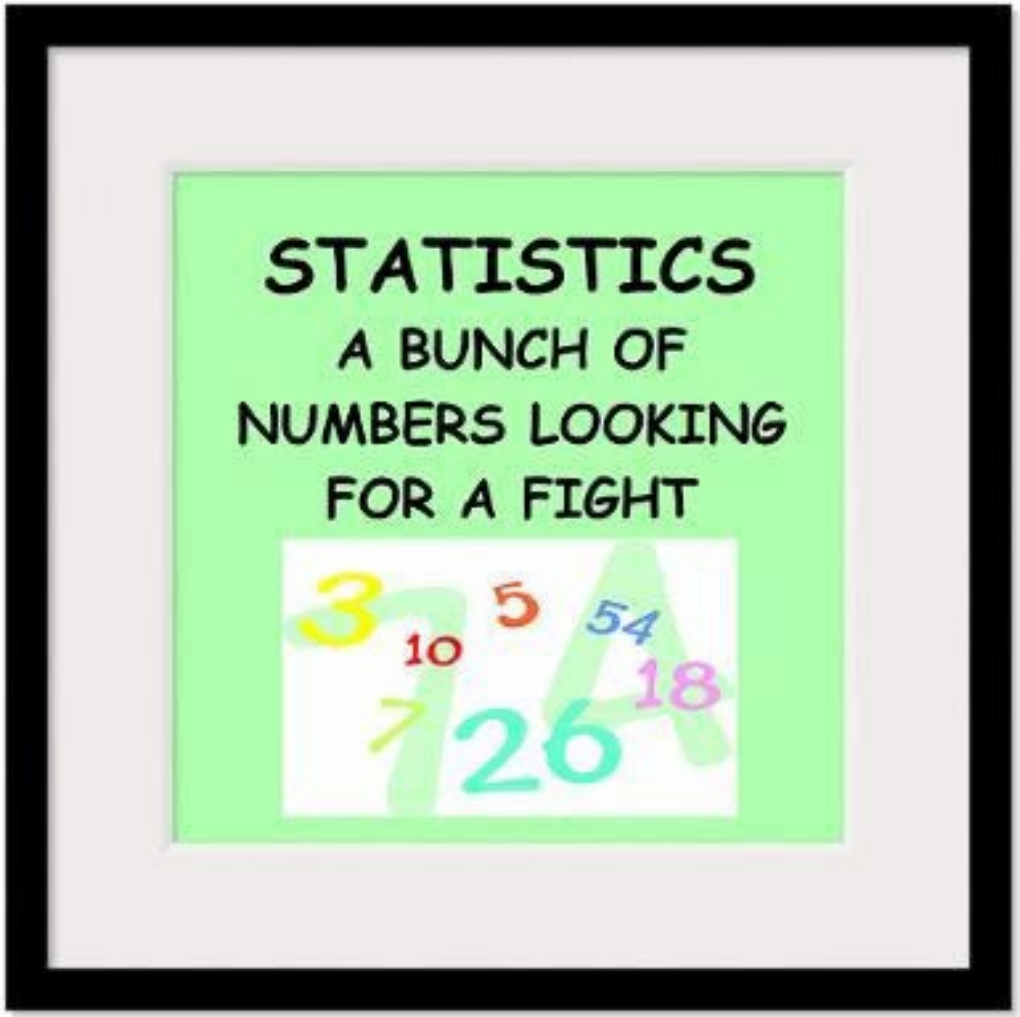
What can statistics do?

Provide us with evidence from the data that allows us to draw conclusions from the data

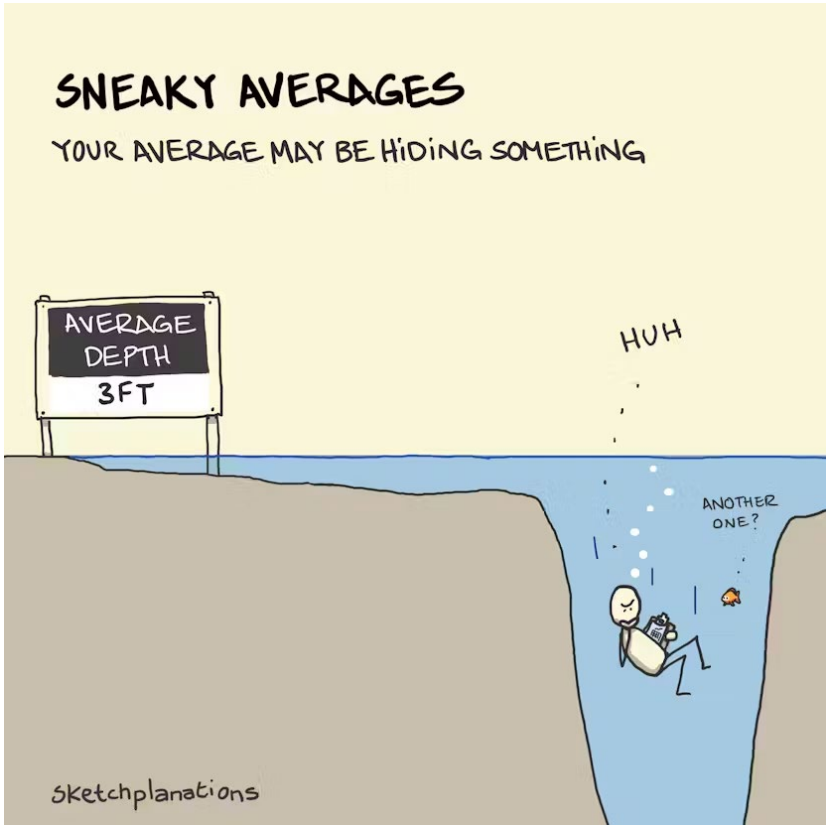
- Group of numbers #1: 6, 1, 8, 3, 5, 4, 9
  - Average is 5.14
- Group of numbers #2: 8, 3, 4, 2, 7, 1, 4
  - Average is 4.25

Allows us to compare groups of data

Allows us to do this **objectively** and **quantitatively**


STATISTICS
A BUNCH OF
NUMBERS LOOKING
FOR A FIGHT
3
10
5
54
18
7
26

Example

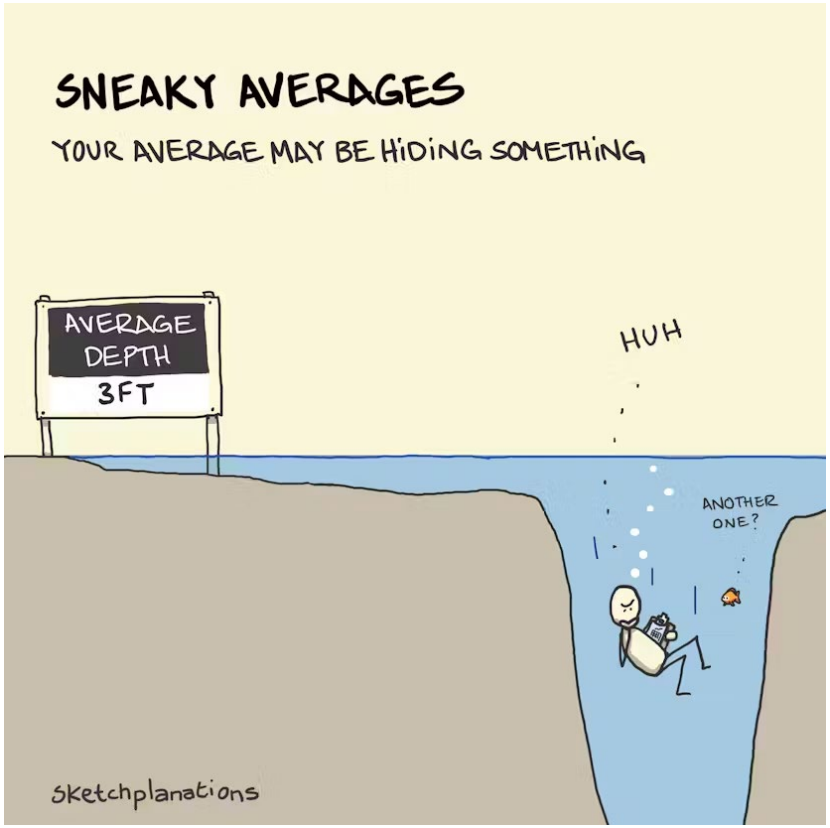


Averages summarise data into a single value, making complex information easier to compare.

But while they simplify, they also hide variation — and sometimes what's really going on.

https://sketchplanations.com/sneaky-averages

Example



Suppose you run a delivery service and have a mean delivery time of one day;

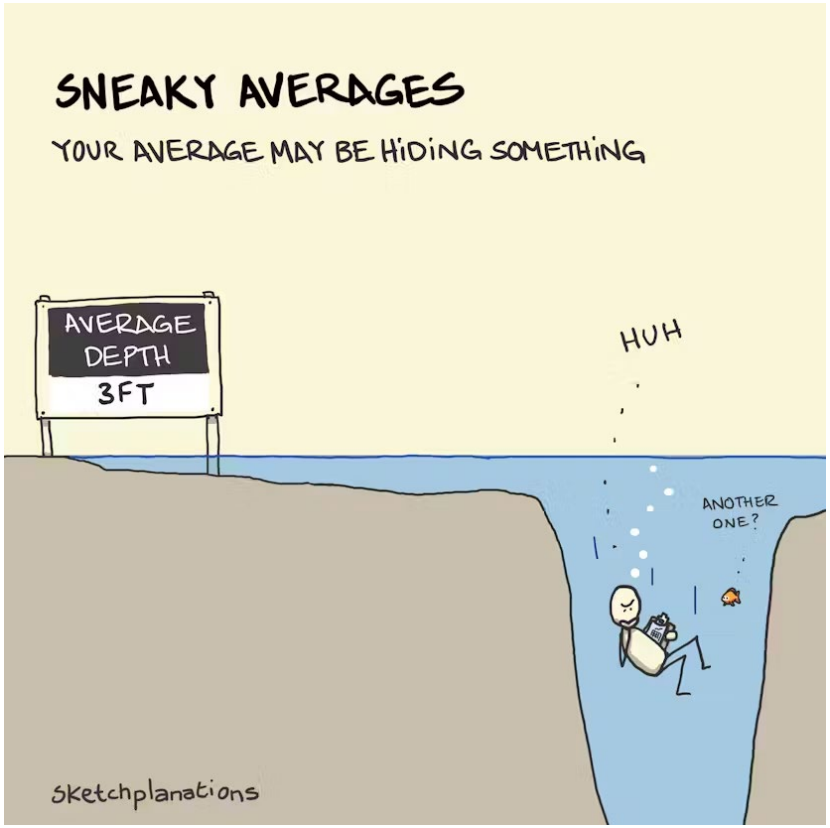
- it could be that most deliveries are actually a few hours, while just a few have people waiting for a week.

A figure of mean incomes might mask the fact that most people have low salaries, while a few are millionaires.

Sometimes a different measure of central tendency, like the median, can provide a clearer picture.

https://sketchplanations.com/sneaky-averages

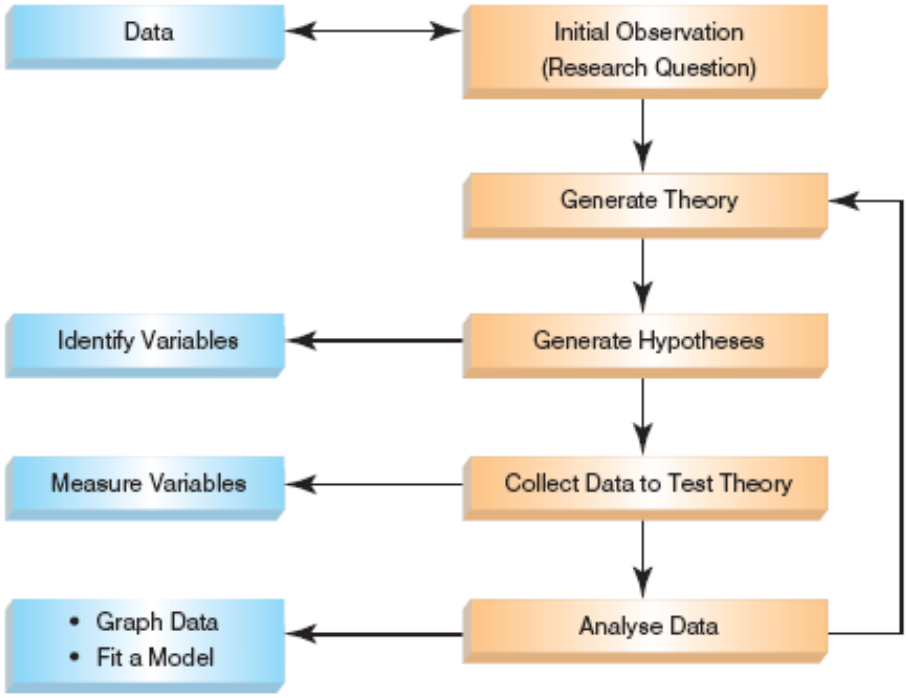
Example



But you also need other statistics to give you the full picture – variance, std deviation, range etc.

You need to "Spend time with your data." Sometimes it's the only way to know what's happening for sure.

https://sketchplanations.com/sneaky-averages



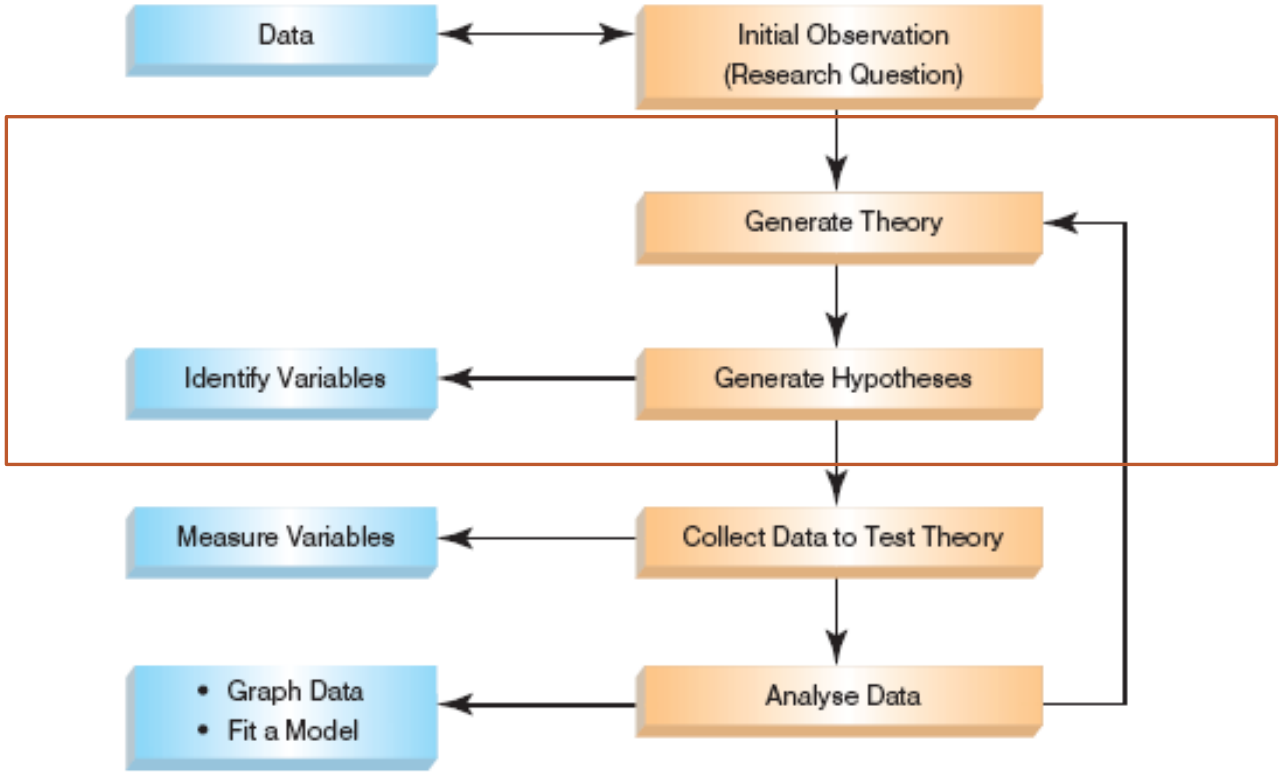
The Process (Source Andy Field)

Initial Observation

Find something that needs explaining

- Observe the real world
- Experience the real world
- Read related research
- ……

The Process


Data
Initial Observation
(Research Question)
Generate Theory
Identify Variables
Generate Hypotheses
Measure Variables
Collect Data to Test Theory
Graph Data
Fit a Model
Analyse Data

Conceptual Framework

System of

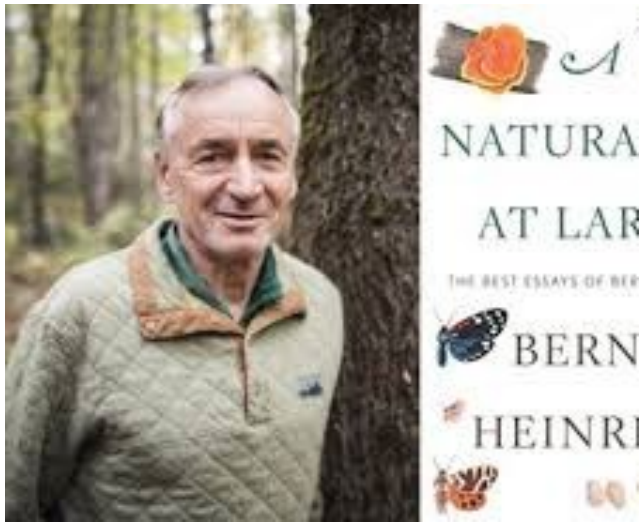
- concepts,
- assumptions,
- expectations,
- beliefs, and
- theories that supports and informs your research

A key part of your design

A conception or model of what is out there that you plan to study, and of what is going on with these things and why

- a tentative *theory* of the phenomena that you are investigating

Influences your design and in particular the data you need to collect/use



Conceptual Framework – Why is it important?

Bernd Heinrich

Spent a summer conducting detailed, systematic research on ant lions

- Small insects that trap ants in pits they have dug.

When he conducted his analysis, he found his results were very different to other researchers…

- A good thing ? Was his research revolutionary?

Conceptual Framework

Repeated experiments following summer

Found that he and his team had misunderstood ant lion behaviour

- In particular the time frame involved
- Missed specific behaviour that impacted the results

“Even carefully collected results can be misleading if the underlying context of assumptions is wrong”

Source: Maxwell, J. A. (2005). Conceptual framework: What do you think is going on? Qualitative research design: An interactive approach (3rd. ed., pp. 39-72

Generating and Testing Theories

Theory

- A hypothesized general principle or set of principles that explains known findings about a topic and from which new hypotheses can be generated.
  - e.g. Computer Science attracts students with strong mathematical ability

Hypothesis

- A prediction from a theory.
- States an assumed relationship between concepts of interest.
  - E.g. the number of people applying for an MSc in Computer Science will have basic mathematical ability greater than the general level in the population.
  - We want to test our hypothesis/hypotheses to see if they hold in the world that we are modelling based on our observations/data.

Identify the variables

From your theory and hypotheses, you will be able to identify the items that can be observed or changed or manipulated or calculated or simulated.

Example:

If we are interested in researching the impact part-time work has on students' study time, we might state a hypothesis:

The number of hours a student spends working for their employer impacts the number of hours spent on working independent study

What are the variables?

- Number of hours spent working
- Number of hours spent independently studying

A **variable** is anything that can take on a different quantity or quality

- It is anything that can vary.

Example Theory

- The use of stimulants positively impact PSI students' level of alertness, mental and physical activation, or readiness to respond.
- We have a theory that the combination of caffeine and sugar as stimulants can increase these levels.
  - Caffeine, in particular, is known to block adenosine receptors in the brain, reducing feelings of fatigue and increasing alertness and concentration.
  - We choose Coca Cola as a suitable stimulant.
  - We decide on the dosage levels, timings etc.

Example Hypothesis:
Consumption of Coca-Cola improves a PSI student's ability to concentrate during lectures.

VARIABLE ONE

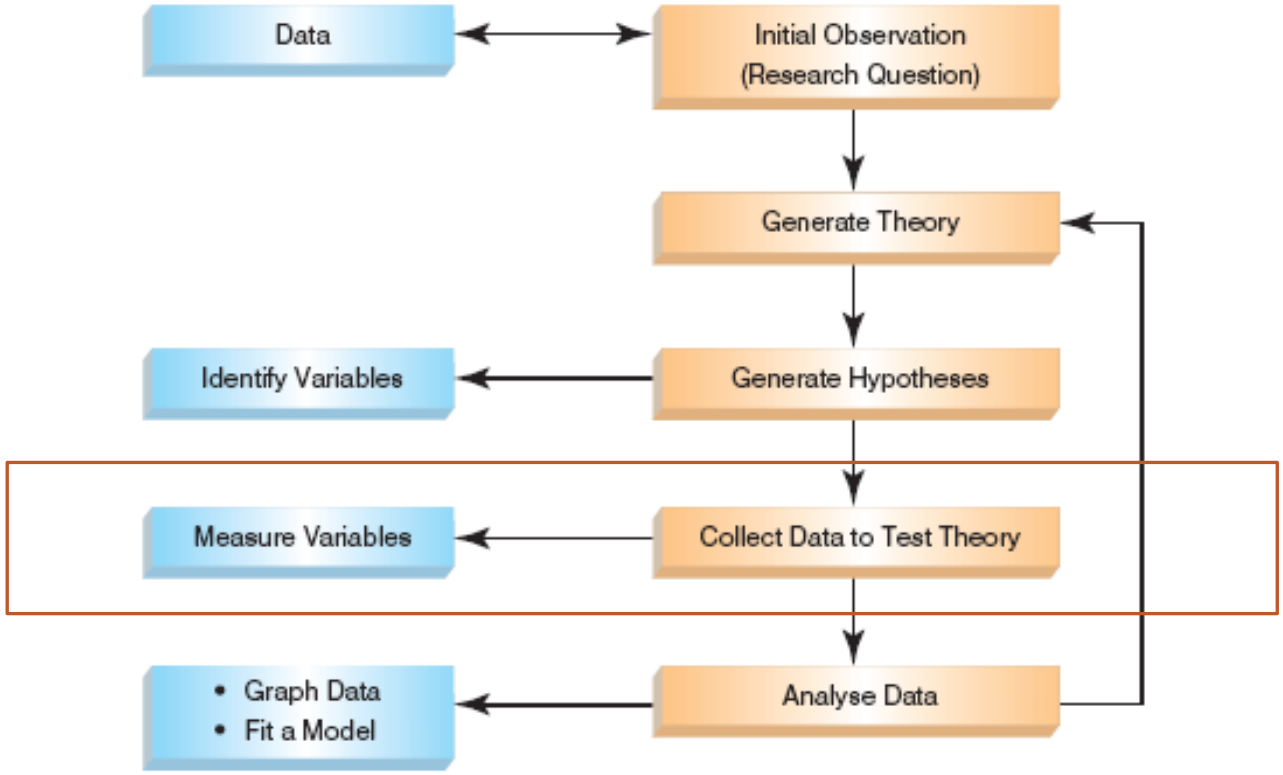
- Coca-Cola consumption
  - The proposed cause
  - A predictor variable
  - Could be a manipulated variable (in experiments)
  - **Independent Variable**

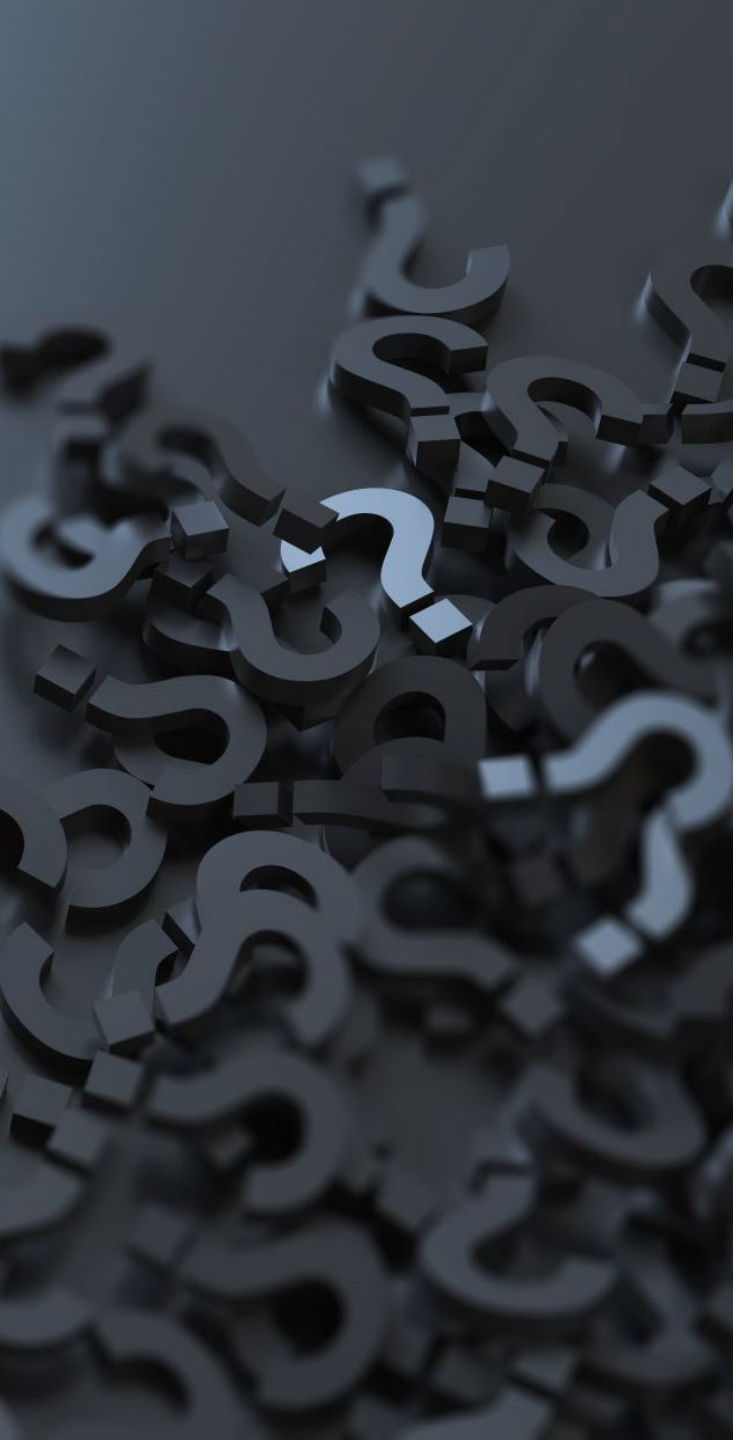
VARIABLE TWO

- A student's ability to concentrate in the hypothesis above
  - The proposed effect
  - An outcome variable
  - Measured not manipulated (in experiments)
  - **Dependent Variable**

"Now, keep in mind that these numbers are only as accurate as the fictitious data, ludicrous assumptions and wishful thinking they're based upon!"

The Process


Data
Initial Observation
(Research Question)
Generate Theory
Identify Variables
Generate Hypotheses
Measure Variables
Collect Data to Test Theory
Graph Data
Fit a Model
Analyse Data



Collect the data

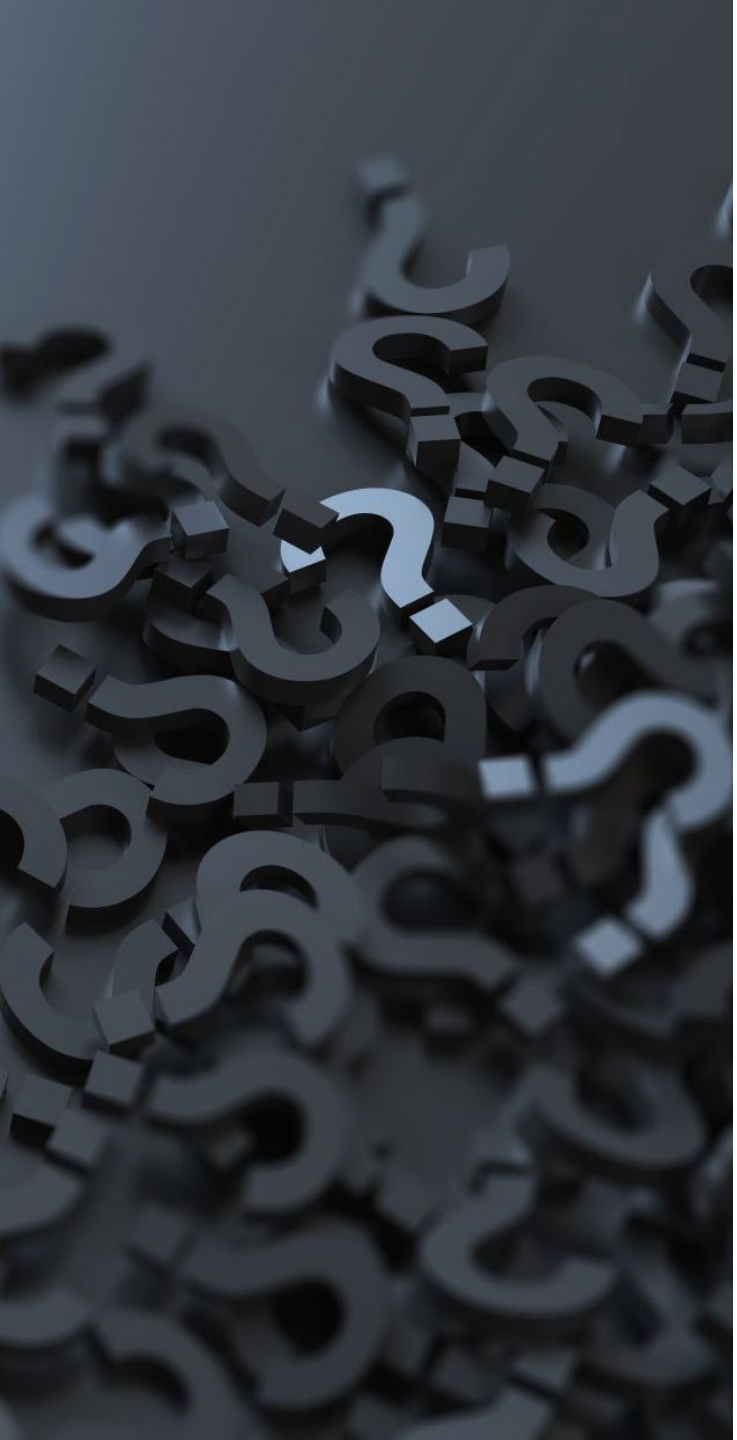
Once we have identified the concepts, we need to collect information about them we can use in statistical analysis to test our hypothesis/hypotheses to see whether our theory holds.

The information we collect about each concept is going to be stored in **variables**

- E.g. level of income, level of education, whether you smoke or not, number of children in a household, rate of water flow through a pipe, percentage of chlorine in a water supply etc.

A **variable** is anything that can take on a different quantity or quality

- It is anything that can vary.



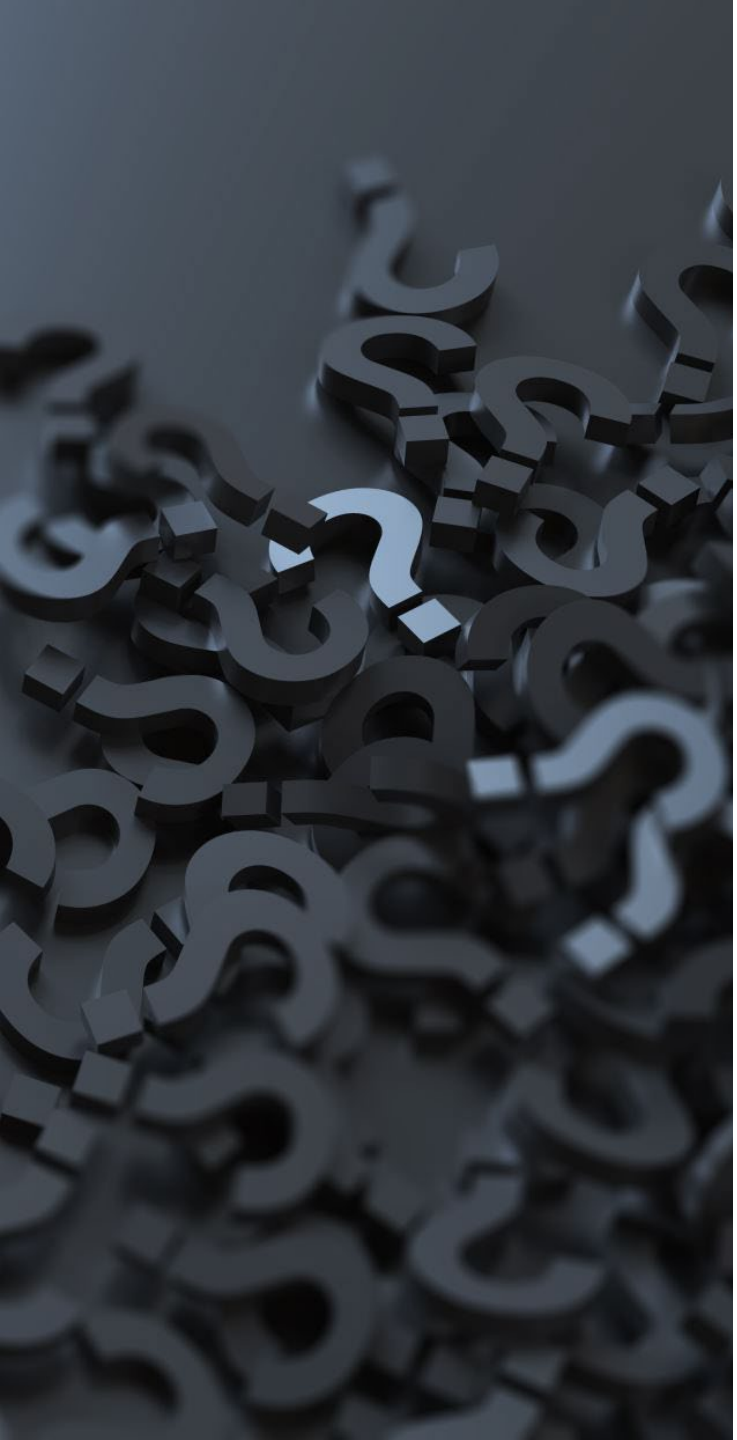
Collect the data

The information about different variables is referred to as **data**

When the data relative to some specific variables is collected and organised, we refer to the collection as a **dataset**.

Datasets are collected and organised around specific observations or instances in the world we are interested in.

- E.g. we are interested in 40 students who are studying PSI in Dublin at MSc level, we collect their age, previous education level and their confidence level studying mathematics
- The dataset would be made up of 40 **observation or cases**.
- Each case should include a measurement for each of our values of interest.



Measure the variables

Once we have identified the variables we need, we need to decide their **statistical type**

Types of Variable (In Statistical Analysis)

QUANTITATIVE

Involves measurement

- Data in numerical form
- Objective and results in unambiguous conclusions
- 5.14 versus 4.25
- 25% versus 50%
- 1 hour versus 24 hours

QUALITATIVE

Describes the nature of something

- Often evaluative and ambiguous
- “Good” versus “Bad”
- “Right” versus “Wrong”
- “A Lot” versus “A Little”
- May be numerical
  - e.g. 1 = Group 1, 2 = Group 2

Qualitative, or Attribute, or Categorical, Variable

A variable that **categorizes** or **describes** an element of a population.

Identifies basic differentiating characteristics of the population

Note:

- Category values can be numerical e.g., 1 for Urban and 2 for Rural, 1 for Blue Eyes, 2 for Brown Eyes

BUT

- Arithmetic operations, such as addition and averaging, **are not meaningful** for data resulting from a qualitative variable.

Qualitative Datatype Nominal

E.g. Country of Origin: 1=Ireland, 2=India, 3=China, 4=……..; Type of residence: A=House, B=Apartment, C=Other

Categories should be

- mutually exclusive
  - You must be able to assign every case in your dataset into one of the categories but only one category

AND

- collectively exhaustive
  - You have to have a category for every case in your dataset.

Example

- Suppose we are collecting data using a survey
- We want to collect information about respondents Religious Belief. We decide the categories.
  - We decide to have categories for Catholic, Protestant, Jewish, Muslim, Hindu and Other
  - What would we do if a respondent was an atheist?

Qualitative Datatype Ordinal

Categorises but also introduces some order to the categories

E.g.

- Ranking: 1st, 2nd, 3rd
- Answer: Strongly Disagree, Disagree, Agree, Strongly Agree

Suppose in our survey we want to collect information about respondents' highest level of educational achievement.

- We decide to include the categories:
  - completed some primary level, completed primary level, completed some second level, completed second level, completed a primary degree, completed a post-graduate degree
- Here there is an inherent notion of order
  - Those who respond that they completed primary have less education that those that have completed a primary degree

Quantitative variable

A variable that quantifies an element of a population.

Can be **discrete**

- have values that are counted

OR

Can be **continuous**

- takes on any value within a range, and the number of possible values within that range is infinite

Observations or measurements take on numerical values

Note:

- Arithmetic operations such as addition and averaging, **are meaningful** for data resulting from a quantitative variable.

Quantitative variable – General Types

Discrete

- Isolated points along a number line
- Usually counted
- Can only take certain values
- E.g. rolling a dice, #students attending class

Continuous

- Variable that can be any value in a given range
- Usually measured
- E.g. heights of students attending class, time spent concentrating in class

Levels of measurement – Interval Datatype

Interval

- Data can be ordered but there is meaning between the values of order
- Values are measure on a scale and the difference between the points on the scale are the **interval**
- Key points
  - There are **equal intervals** between the values
    - The intervals are fixed and constant
  - There is no true zero point

- Allows comparison between the data values
- While you can add and subtract interval data, you cannot meaningfully multiply or divide it.
- Example:
  - Temperature in Celsius or Fahrenheit.
  - Zero degrees doesn't mean the absence of temperature; it's just another point on the scale.

Interval Datatype

Interval

- Example:
- Suppose I ask all the students in the class to tell me what time you usually wake in the morning (hours and minutes 24-hour clock format, nearest quarter hour)
- The measurement type is interval:
  - The interval is 15 mins
  - The difference between 9:00 and 9:15 is the same as the difference between 18:00 and 18:15
  - But 0:00 is not a true zero - it is a valid time, it is not the absence of time.

Levels of measurement – Ratio Datatype

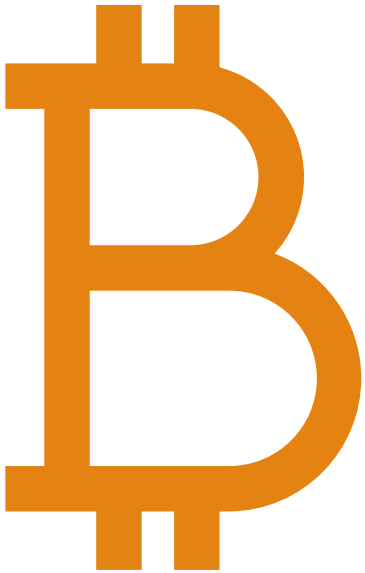
Ratio

- All the properties of interval
- Plus
  - Can take any value within a given range
  - But it is not restricted to fixed increments
  - It can be measured with increasing precision
  - Has a **true zero**
    - **Represents the absence of the concept**
- You can find the ratio - it makes sense to say for example one value is twice as large as another

Levels of measurement – Ratio Datatype

Example:

- Weight is an example of ratio data.
  - A weight of 0 kilograms means the absence of weight
  - And you can meaningfully say that 10 kg is twice as heavy as 5 kg.
- In a survey of students on how much they spend on buying lunch in the university café:
  - This could be any value 7.01, 14.02, 10.56, 63.09
    - We can say 14.02 is twice as much as 7.01, 63.09 is 9 times as much as 7.01
  - But also, a student may spend 0 which is this case represents the absence of spending



Key Note

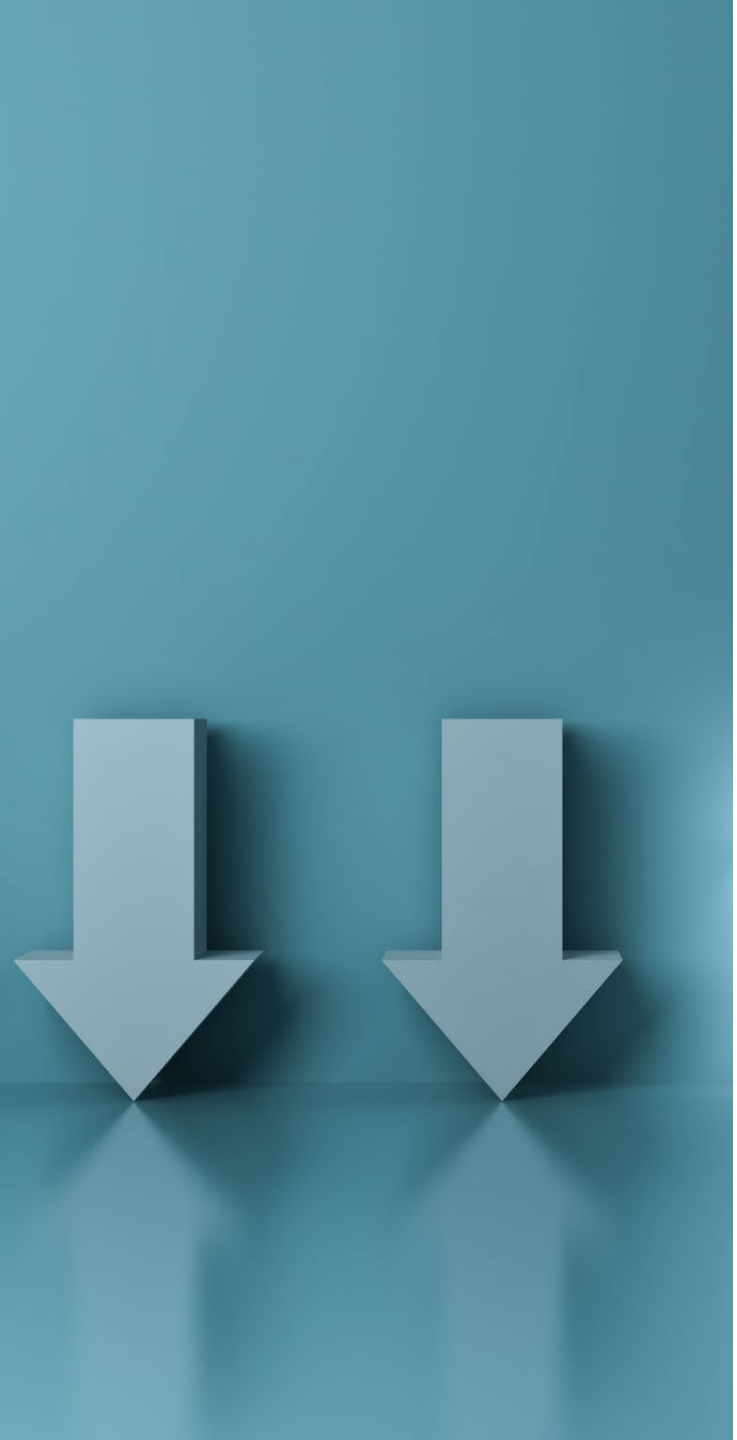
You don’t need to have every possible value of a variable represented in your dataset

BUT

It should be possible for that value to be included if it exists in the world you are modelling

Quiz

https://www.med.soton.ac.uk/stats_eLearning/typesofdataquiz/index.html

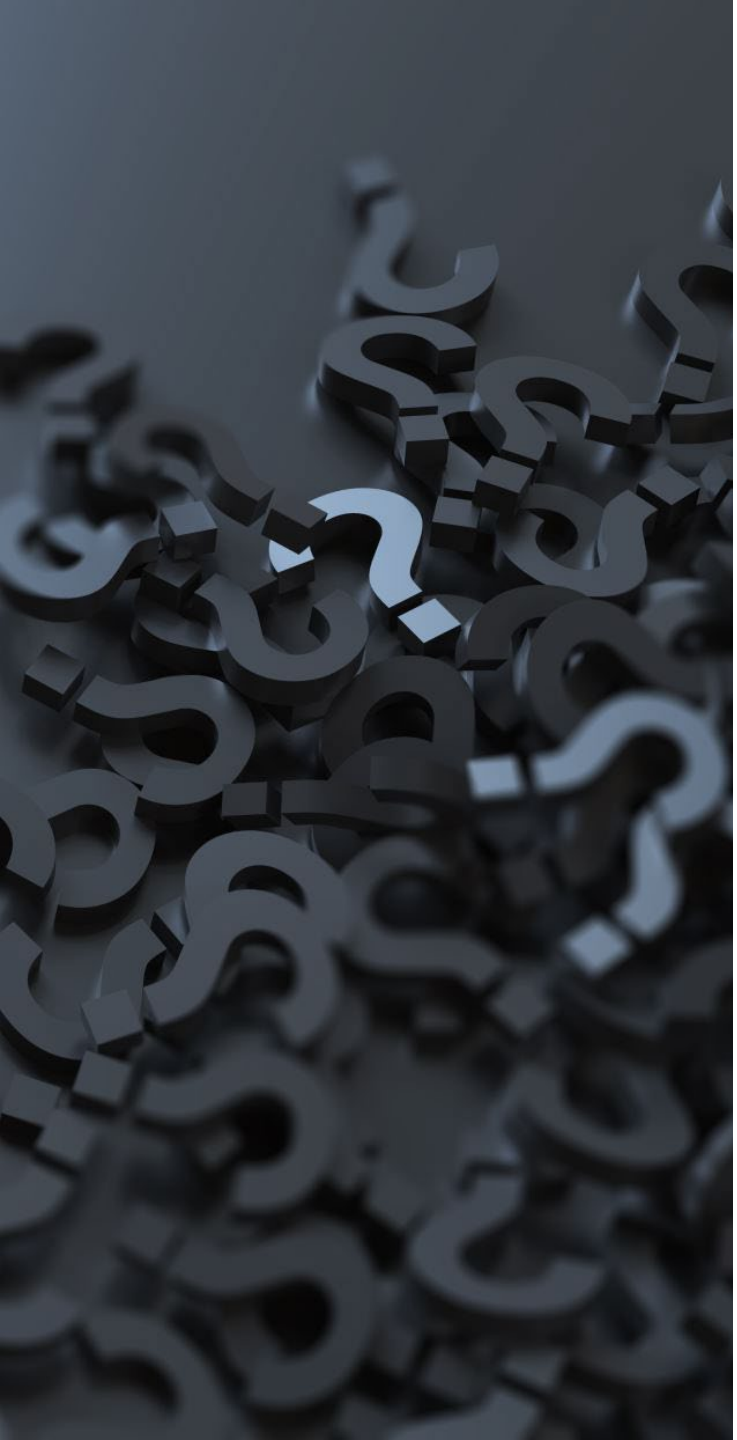


Collect the data

Collect data that will provide evidence you can test to see if your idea is valid

You can use data you have already collected – but you should not fish or dredge it to generate your hypothesis

- Your theory and hypotheses should make sense in the real world



What data am I interested in/do I need?

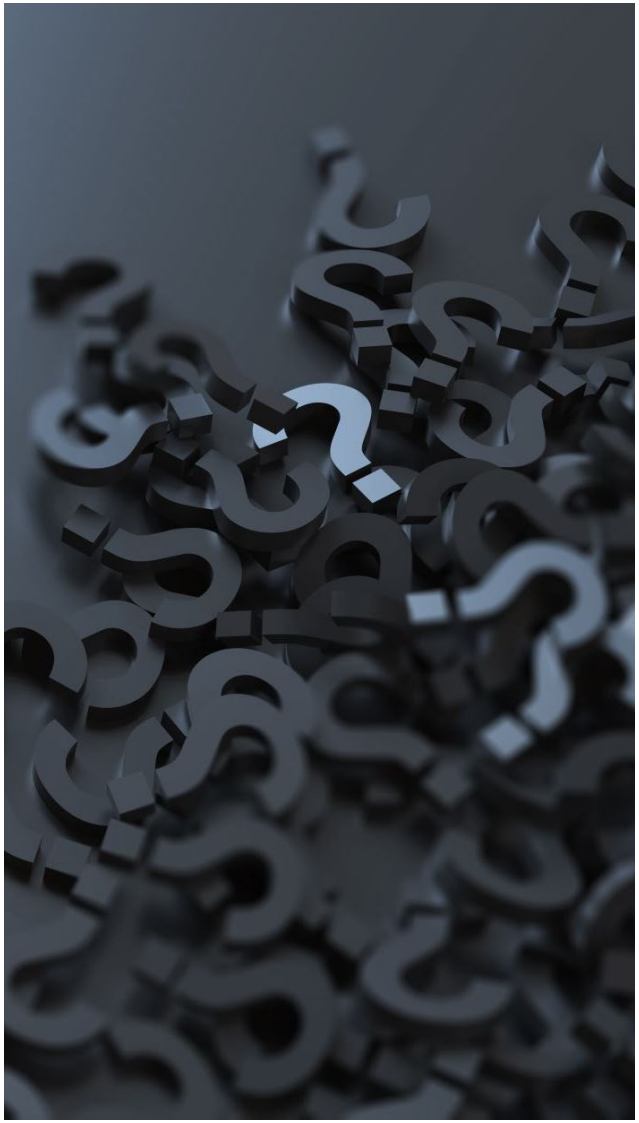
Data Analysis is usually carried out with a SAMPLE

Population

- The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
- The world you are interested in and trying to build a model of.
- Two kinds of populations:
  - finite or infinite.

Sample

- A portion, or part, of the population of interest



Population

The entire set of individuals or objects of interest or the measurements obtained from all possible individuals or objects of interest

The world you are interested in and trying to build a model of.

- E.g., All registered voters in Ireland.

But the population is constantly changing.

- E.g. All registered voters in Ireland
  - People are dying and being removed, turning 18 and being added

It is also usually incredibly large

- E.g. as of 2024 there are 3.4 million registered voters in Ireland.
- If you want to collect information about every single one of them, how do you do it, how long would it take?

So it is difficult to work with the full population

Sample

A portion of your population.

E.g., we may choose to work with a sample of 2000 registered voters in Ireland.

But we need to decide which voters?

What do we need the sample to achieve?

- It will be a proxy for the population so it should be **representative** of that population
- If Sample is not representative, it is **biased**

Sample

Need to understand the nature of the population before collection/creating/choosing the sample

Need the collection of the sample to provide the same opportunities for all possible values of a variables to appear as they would in the population.

Population v Sample

- Example
  - You are working for a company
  - You have been asked to analyse employee salaries.
  - You have access to all employee records.
  - If your task is to analyse the salaries on a particular day
    - You could say you have a population to work with
  - If your task is to analyse the salaries over a year, some people quit, retire, be fired, get hired, be promoted etc.
    - You can't really work with the population, you would work with a sample

Population v Sample

- Example
  - You are working for a retail company, and you want to conduct a customer satisfaction survey
  - You want to survey every customer who has purchased something from the company in the past calendar year
    - That's the population
  - But could you logistically contact all of these customers? (Unlikely)
    - And if you could, would they all respond? (Unlikely)
  - In this case you would work with a sample.
  - You need to make it **representative**.

Sample

Factors influencing the accuracy of a sample's ability to represent a population:

- Size
- Randomness

Biased Sampling Method:

- A sampling method that produces data which systematically differs from the population from which it is taken.

Aim for a Simple Random Sample

- A sample of *n* measurements from a population is a subset of the population selected in such a manner that every sample of size *n* from the population has an equal chance of being selected
- Some things to note:
  - Researcher bias should not occur in the sample selection
  - May not always reflect the diversity of the population

Sample

In an ideal situation, the entire population should be studied but this is almost impossible.

Majority of studies are performed on limited subjects drawn from the concerned population known as “sample population”.

The data obtained is analysed and conclusions are drawn which are **extrapolated** to the population under study.

Sample size - The number of *cases* n

- E.g. n = 90

Too small a sample

- May not lead to conclusions that are valid for the population

Too large a sample

- Wasteful if a smaller sample would do
- May have ethical implications dependent on the type of experiment/observation being undertaken

How do decide on sample size is something we will talk about

Data Collection: What to Measure?

Hypothesis:

- *Consumption of Coca-Cola improves a student's ability to concentrate.*

Decide what variables you need

- Variable One
  - The proposed cause
  - A predictor variable
  - A manipulated variable (in experiments)
  - Coca-Cola consumption in the hypothesis above
- Variable Two
  - The proposed effect
  - An outcome variable
  - Measured not manipulated (in experiments)
  - A student's ability to concentrate in the hypothesis above

Data Collection: Measurement Error

Measurement error

- (also referred to as observational error)
- The discrepancy between the actual value we're trying to measure, and the number we use to represent that value.

There are always random errors

- We need to expect them and handle them in our reporting

Data Collection: Measurement Error

Example:

- Suppose you are measuring the weight of 100 athletes
  - The scale you use is 0.1 kg out
  - This is a **systematic measurement error** of 0.1kg
  - However,
    - If our scale is accurate but our athletes all have different levels of hydration or are wearing different types of clothes
    - Then we still have a measurement error – this is **random**

Data Collection: Validity

Whether an instrument measures what it set out to measure.

Content validity

- Evidence that the content of a test corresponds to the content of the construct it was designed to cover

Ecological validity

- Evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions.

Data Collection: Reliability

Degree to which an indicator is a consistent measuring device

The ability of the measure to produce the same results under the same conditions.

Test –Retest Reliability

- The ability of a measure to produce consistent results when the same entities are tested at two different points in time.

Confounding Variables

Sometimes we are interested in establishing relationships/differences and then inferring cause and effect[1]

This can be complicated by *confounding variables*

- Variables which may or may not directly measured which has influence on other variables in the study

Two variables are *confounding* when one cannot be distinguished from the effects of another

- E.g. amount of petrol and time used to commute to work, it is likely that level of congestion will be a factor

[1]: We won't actually be able to 100% demonstrate cause and effect – we will look at this issue later.

Confounding Variables

What could the confounding variables be for our Coca-Cola hypothesis?

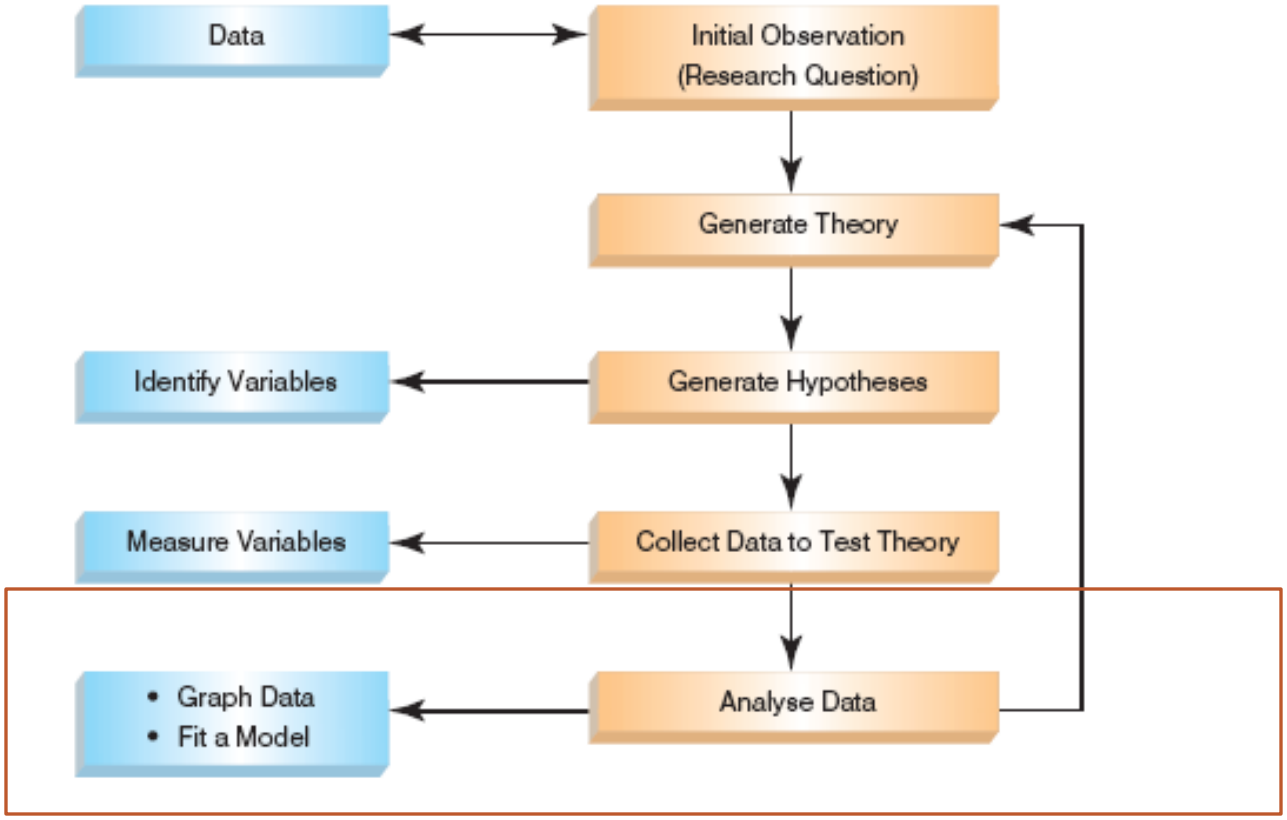
Hypothesis:

- *Consumption of Coca-Cola improves a student's ability to concentrate.*

Decide what variables you need

- Variable One
  - Coca-Cola consumption in
- Variable Two
  - A student's ability to concentrate

The Process


Data
Initial Observation
(Research Question)
Generate Theory
Identify Variables
Generate Hypotheses
Measure Variables
Collect Data to Test Theory
• Graph Data
• Fit a Model
Analyse Data

Important Step - Describing your data

Must describe before analyzing

Need to include sufficient detail that your consumer can

- See what you are basing your analysis/conclusions on
- Understand anything that may constrain those analysis/conclusions

You are trying to present information about a large body of data so that your consumer can understand it without having to view every individual case you have collected

- Usually start with a picture and
- Include some meaningful numbers to summarise and illustrate variability

Summarizing v Analyzing

Descriptive Statistics

- Describing the population and the sample
- Summarizing

Inferential Statistics

- Inference from sample to population
- Inference from statistic to parameter

Getting Started With Statistical Study

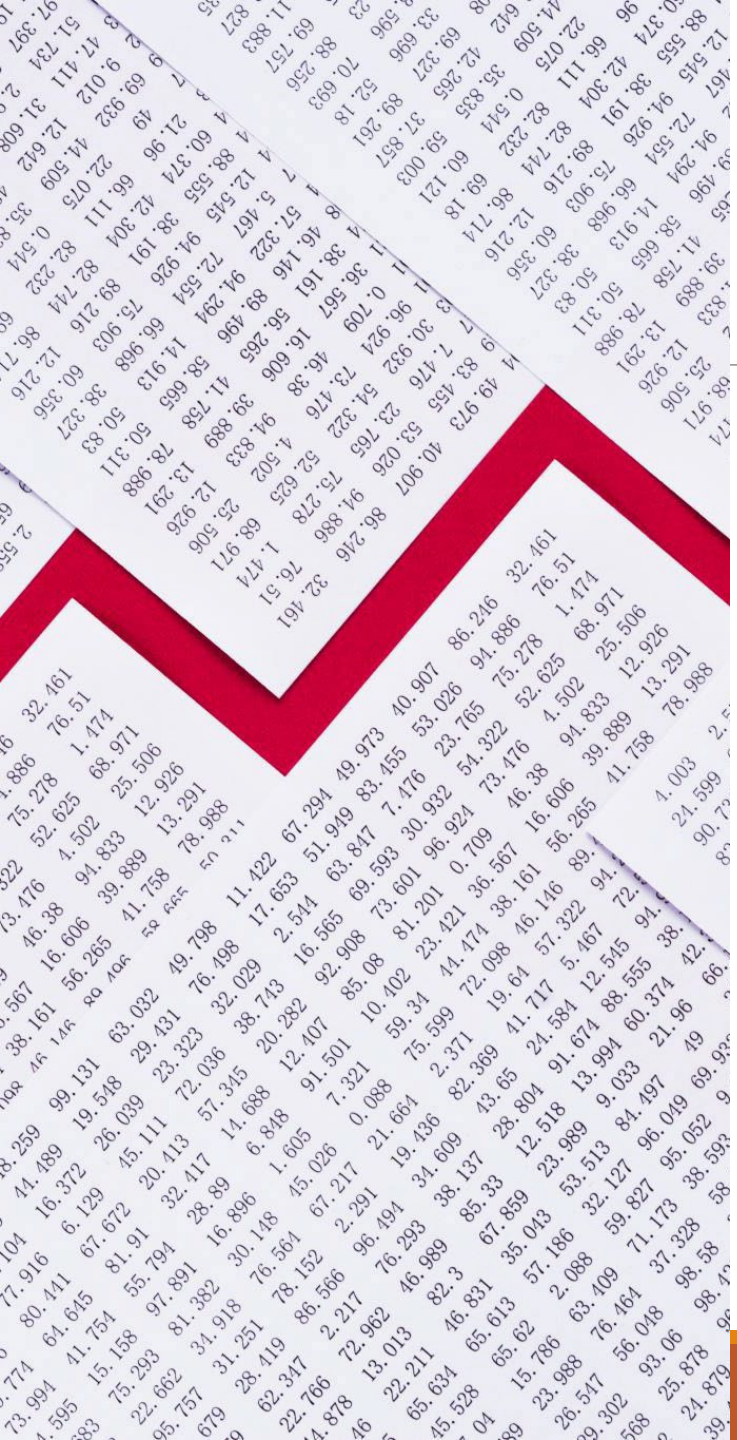
Need to decide the nature of the study and then decide what individuals or objects are of interest

We must understand our population

- The data we need to collect
  - Its structure and values, how it can be measured/represented
- Sample
  - Representativeness and size
- Be able to describe it simply
  - Using the appropriate statistics for the variables of interest
- Be able to represent it visually
  - Using the appropriate type of graph for variables of interest

The we will be able to analyse it appropriately

- Using the appropriate statistical tests to draw inference



Misleading Statistics

REFERS TO THE MISUSE OF NUMERICAL DATA EITHER INTENTIONALLY OR BY ERROR

What do you think when you see this?

IMAGE: MANCHESTER EVENING NEWS

What do you think when you see this?

- This does not disclose any information about the survey from which it was derived:
  - When dentists were surveyed, they could choose several brands — not just Colgate.
  - So other brands could be just as popular as Colgate.
- Use of this statistic for advertising was banned by Advertising standards in UK.

Always check who was surveyed, what they were asked and how

Robert Langkjæ r-Bain
First pub lished: 29 July 2020
https://doi.org/10.1111/1740-9713.01420

What's wrong with this?

Roll over image to zoom in

1.7 Fl Oz (Pack of 1)	1.7 Fl Oz (Pack of 2)
$39.98 ($23.52 / Fl Oz)	Currently unavailable.

Brand	L'Oréal Paris
Item Form	Cream
Unit Count	1.7 Fl Oz
Number of Items	1
Use for	Whole Body

About this item

- Luxurious, non-greasy cream visibly improves skin quality by working on uneven skin texture with visible pores and fine lines
- Formulated with Perline-P, an association of three perfecting actives that optimizes the quality of skin's surface by tightening pores, diminishing fine lines and revealing petal-soft skin texture
- Women self report skin instantly feels softer and texture is smoothed for a velvet touch
- In 1 month women self report pores appear smaller and fine lines are reduced, and 75% of women self report seeing an improvement in skin quality
- Suitable for all skin tones

https://www.amazon.com/LOreal-Paris-Youth-Texture-Perfector/dp/B00DG1B6V2?th=1 retrieved 13/09/2025

Roll over image to zoom in

1.7 Fl Oz (Pack of 1)	1.7 Fl Oz (Pack of 2)
$39.98 ($23.52 / Fl Oz)	Currently unavailable.

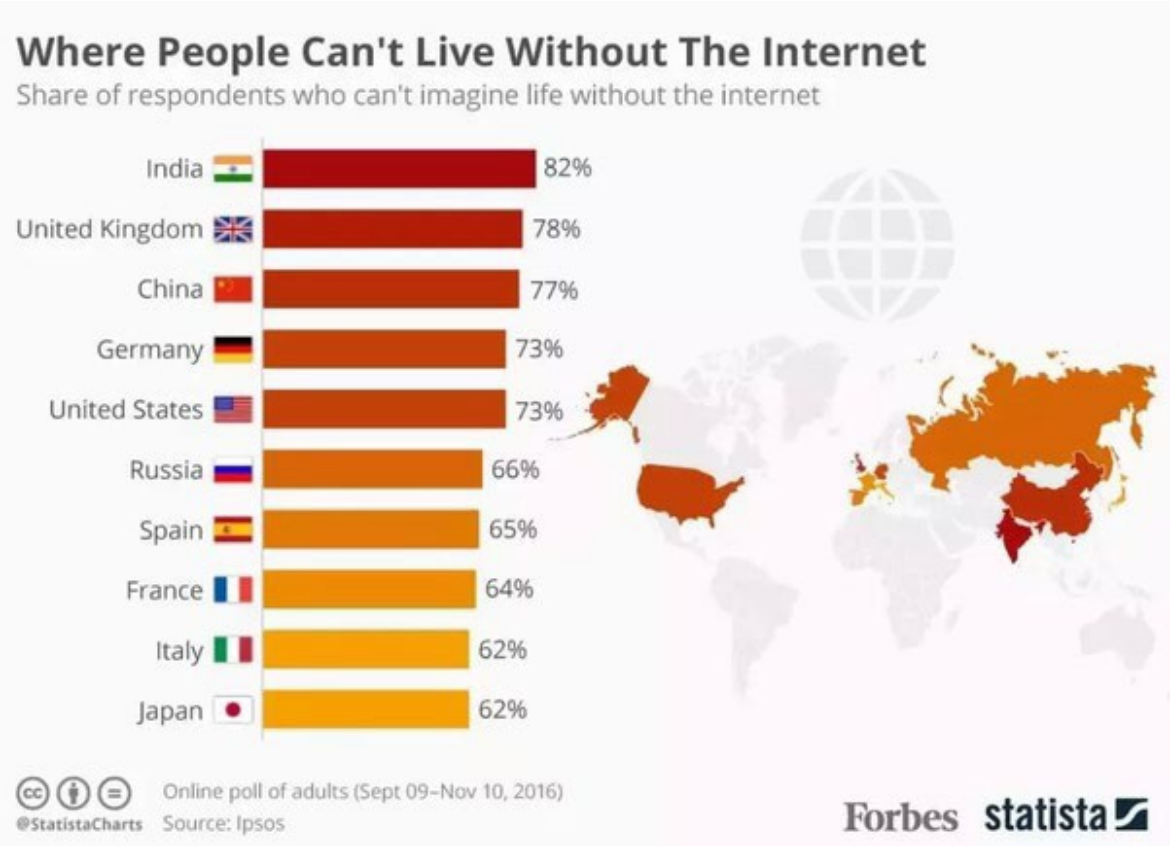
Brand	L'Oréal Paris
Item Form	Cream
Unit Count	1.7 Fl Oz
Number of Items	1
Use for	Whole Body

About this item

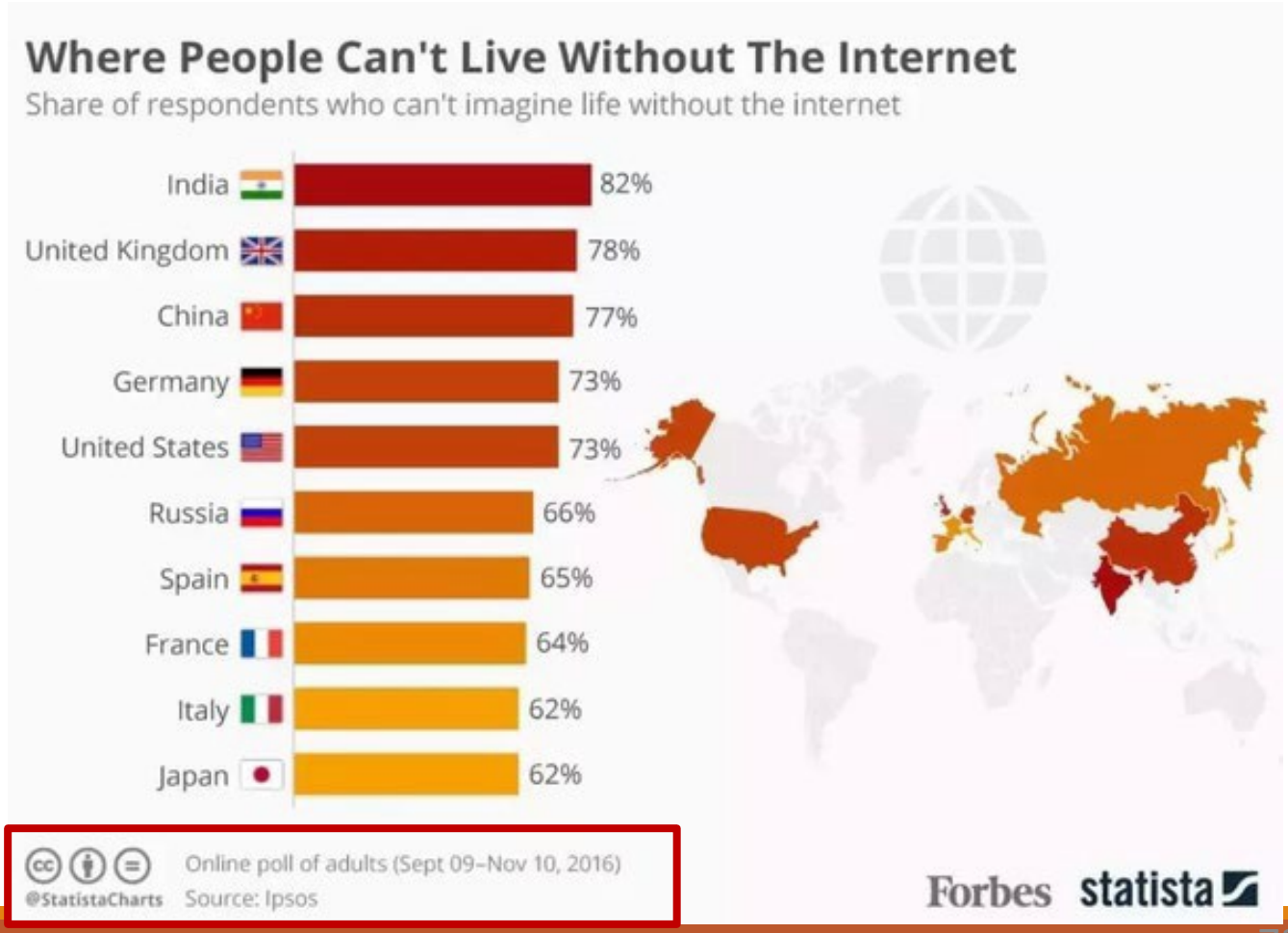
- Luxurious, non-greasy cream visibly improves skin quality by working on uneven skin texture with visible pores and fine lines
- Formulated with Perline-P, an association of three perfecting actives that optimizes the quality of skin's surface by tightening pores, diminishing fine lines and revealing petal-soft skin texture
- Women self report skin instantly feels softer and texture is smoothed for a velvet touch
- In 1 month women self report pores appear smaller and fine lines are reduced, and 75% of women self report seeing an improvement in skin quality
- Suitable for all skin tones

*Based on consumer self-assessments

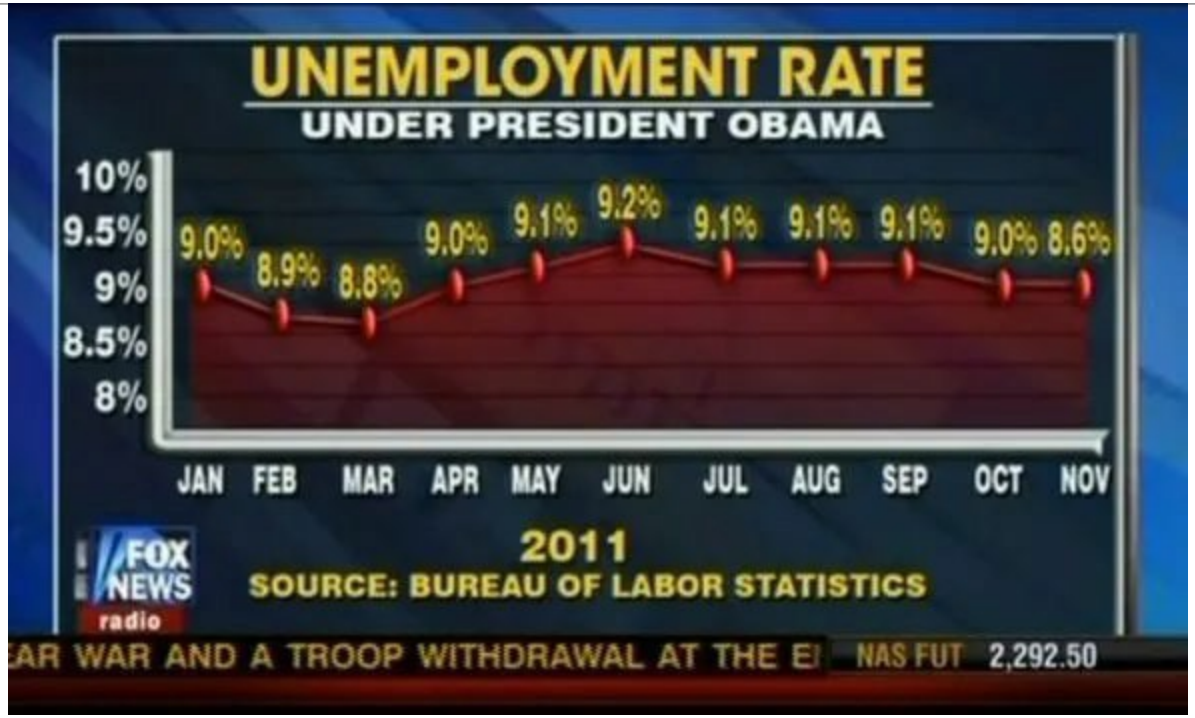
What's wrong with this?



What's wrong with this?

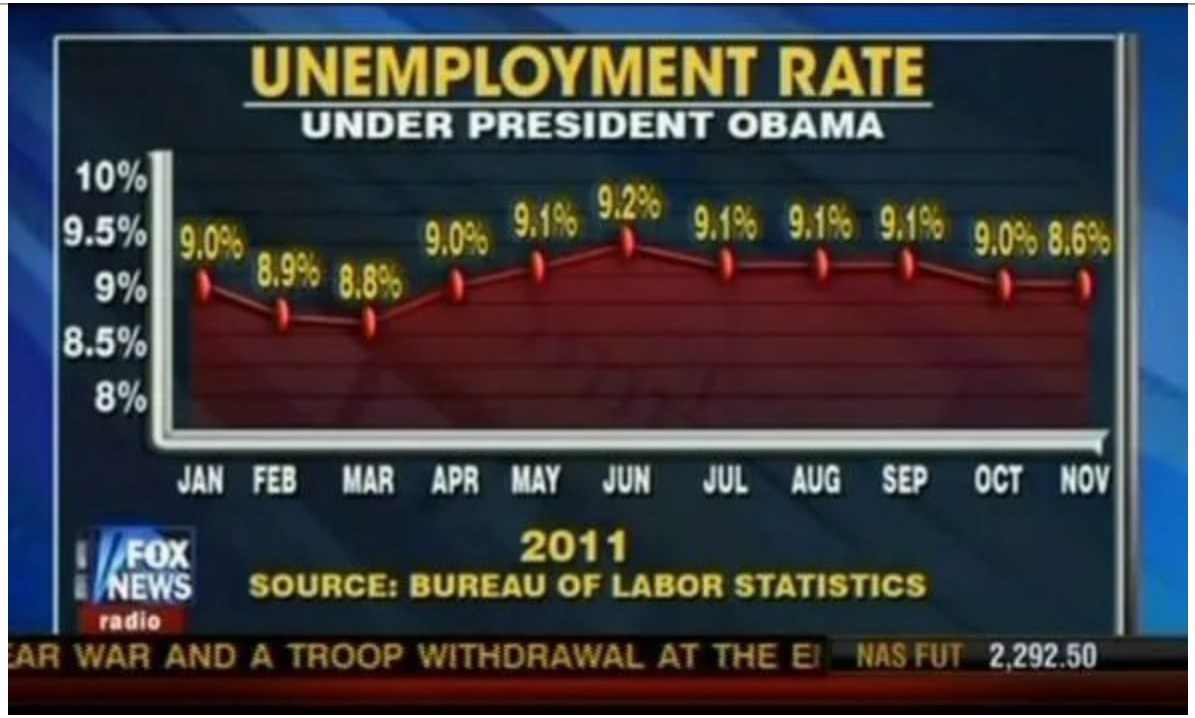


What's wrong with this ?



Source: Fox News

What's wrong with this ?

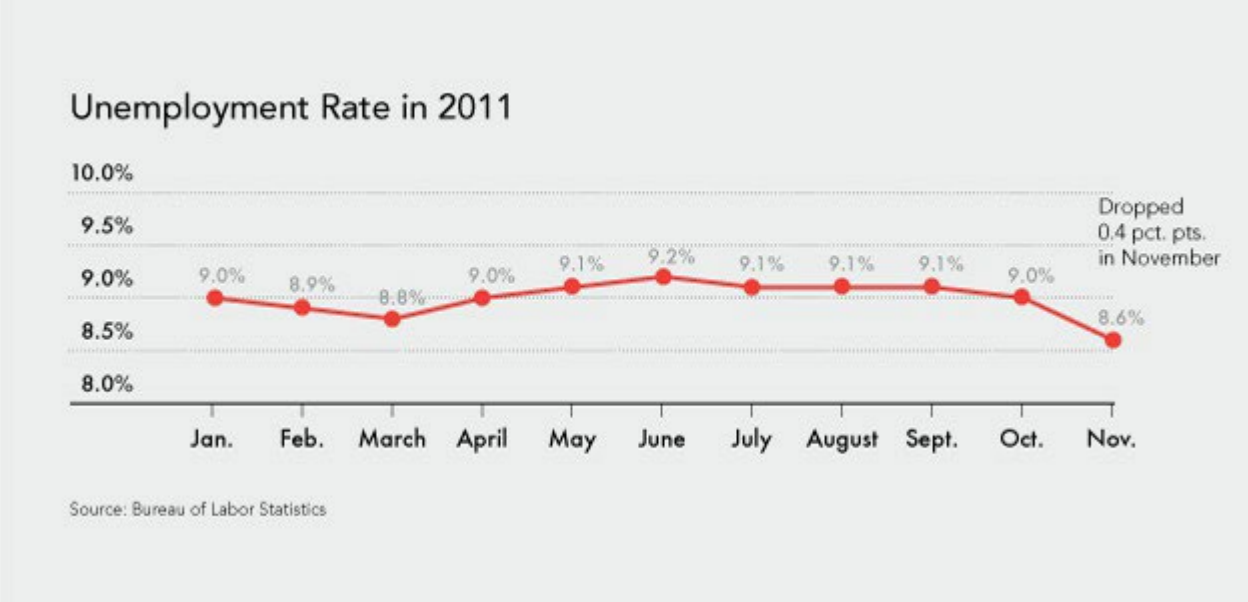


Source: Fox News

The scale is 0.5 on the y axis but the points are 0.1

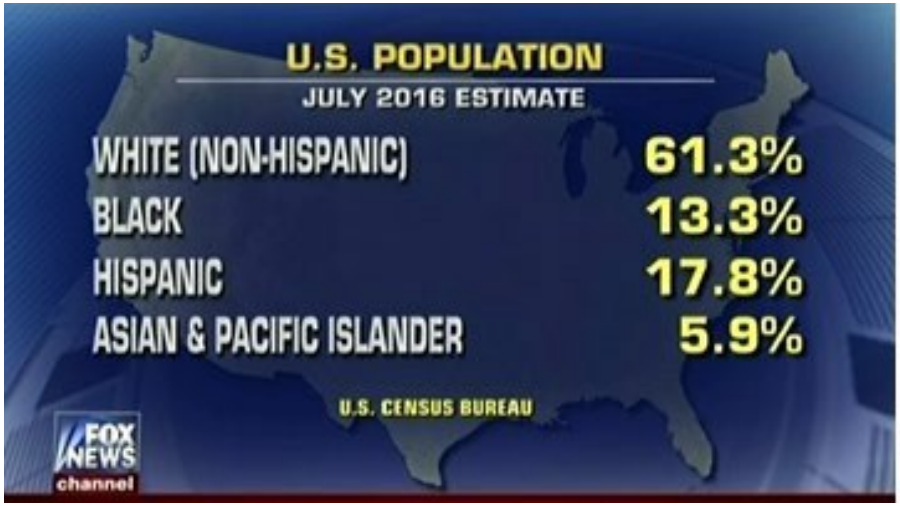
The 8.6% on the far right is higher than the 8.9% on the left

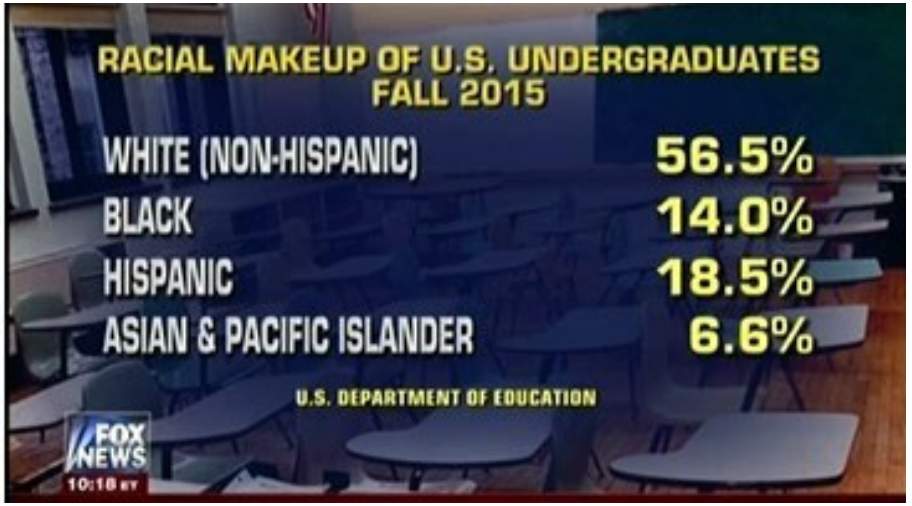
What's wrong with this ?



Corrected version

What is wrong with this?

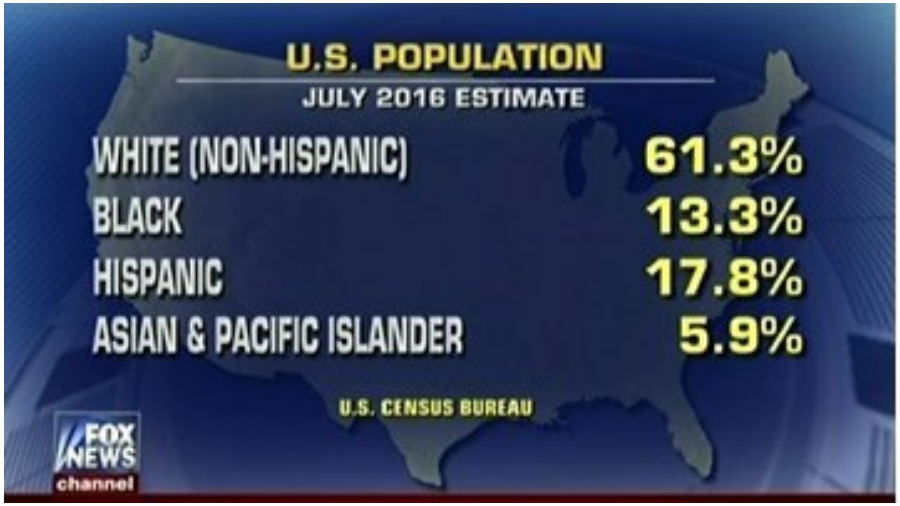


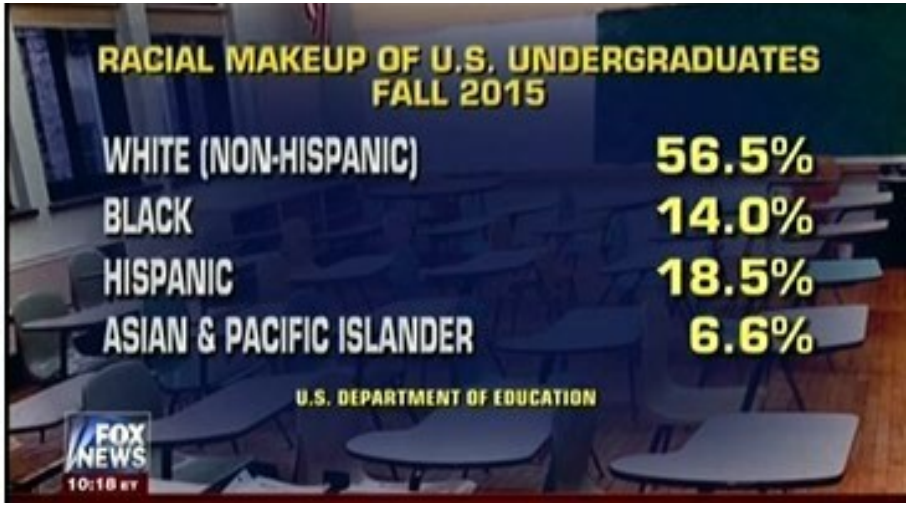


Source: Fox News

Attempt by Fox News to demonstrate under-representation of one section of population at undergraduate education using categorisations for ethnicity from US census.

What is wrong with this?

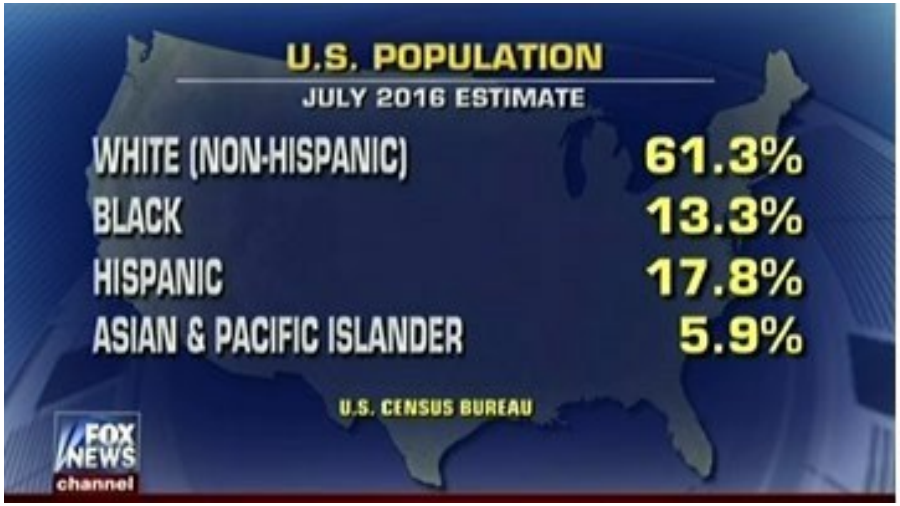


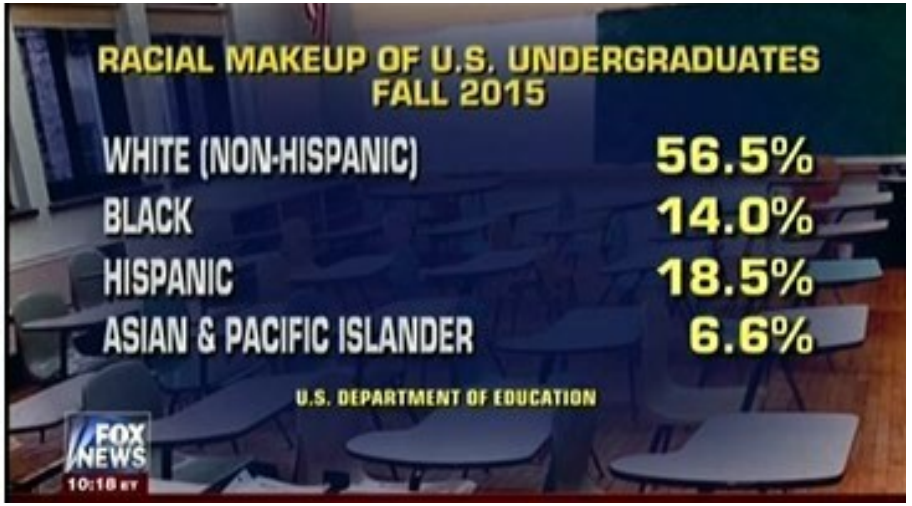


What is the breakdown for 18-24 year olds in the US? Sector of population that makes up undergraduate education

In 2015 this was estimated to be 55.7% white, 14.2% black, 22.3% Hispanic, and 4.5% Asian or Pacific Islander.

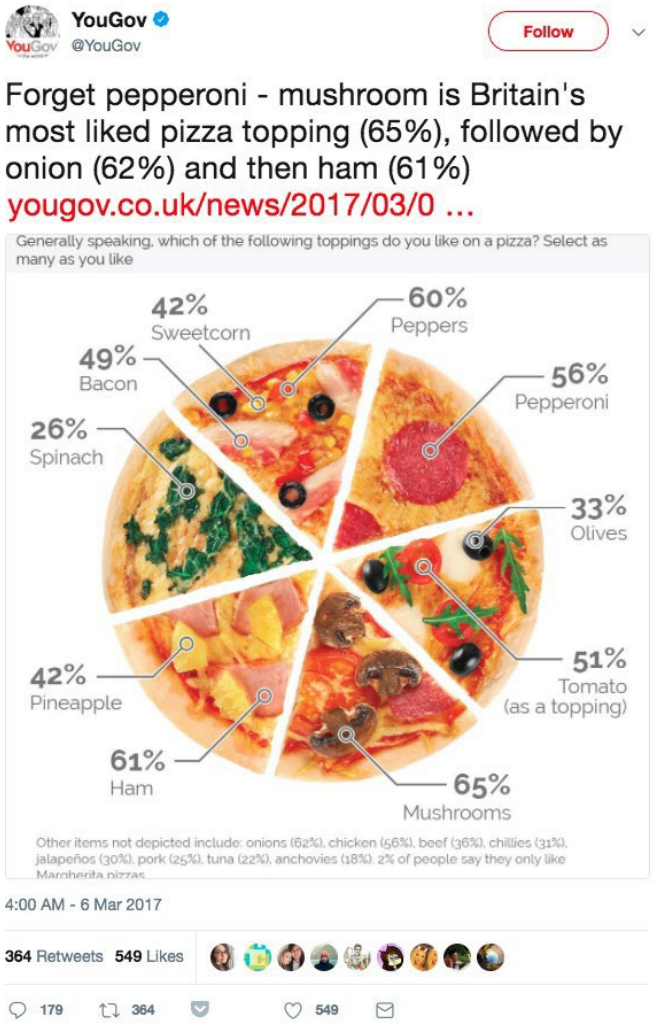
What is wrong with this?



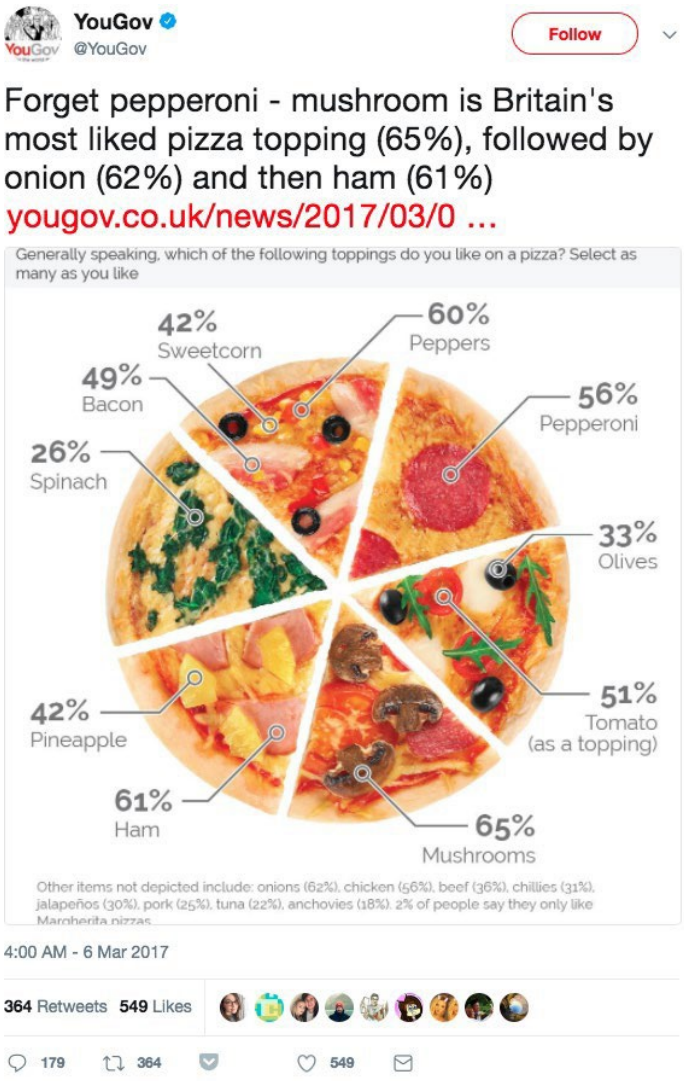


In 2015 18-24 yr old population was estimated to be 55.7% white, 14.2% black, 22.3% Hispanic, and 4.5% Asian or Pacific Islander.

So now what can we conclude from the numbers presented?



What about this?



Adds up to 485%

Respondents could choose more than one option ….

Is this the most appropriate visualization to use?

Does the amount of the ingredient shown reflect the percentage?

Let's look at some correlations

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

Findings

Look for source of data and context

Be careful about using statistics to justify statements that may not hold in the real world (correlation does not mean causation)

Examine the visuals (x, y axes, proportions, colours)

Sample used

Selective statistics - bias

Where can errors (or deceit) occur?

Collection

- While gathering the data

Processing

- When analyzing the data and its implications

Presentation

- When sharing your findings with others

Common Mistakes

Faulty Survey Design/Polling

- Who, what, where, when and how?

Insufficient Sample

Flawed Correlation

- Statistical correlation may be meaningless in the real world

Misleading presentation of results

- Particularly visualisation

Selective Reporting

- Focusing on only one particular time period/segment of population

Data Fishing/Dredging

- Trying to find relationships in data (not bad in itself) but without a hypothesis (leads to bias)

A really useful paper

Christopher Tong (2019) Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science, The American Statistician, 73:sup 1, 246-261, DOI: 10.1080/00031305.2018.1518264

Should we be worried?

Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*. 2009;4(5):e5738. Published 2009 May 29. doi:10.1371/journal.pone.0005738

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2685008/
- 2009 systematic review and meta-analysis found that 33.7% of scientists considered admitted to some form of questionable practice e.g.
  - modifying results to improve outcomes
  - subjective data interpretation,
  - withholding analytical details
  - dropping observations because of **gut feelings**….

If you look up this paper in Google Scholar and follow the citations of this paper (cited by) you will see more recent papers which investigate the issue further

How do we avoid falling in these traps?

Most important

Make sure you are asking the right question of the data

And that you know what it is before you start

No fishing!

Most important

Make sure that you understand the *conceptual framework of any research* you use

- to *justify* the *question(s) you are investigating*
- to *justify* the *approach you use*
- Or *other work* with which you *compare your findings*

At the end of the module, you will be able to

Present the question you are interested in
- In a way that makes sense to conduct a statistical analysis

Inspect and prepare the data you have
- To support a statistical analysis

Describe the data you have
- In a way that your consumer can understand any constraints the data may put on your analysis

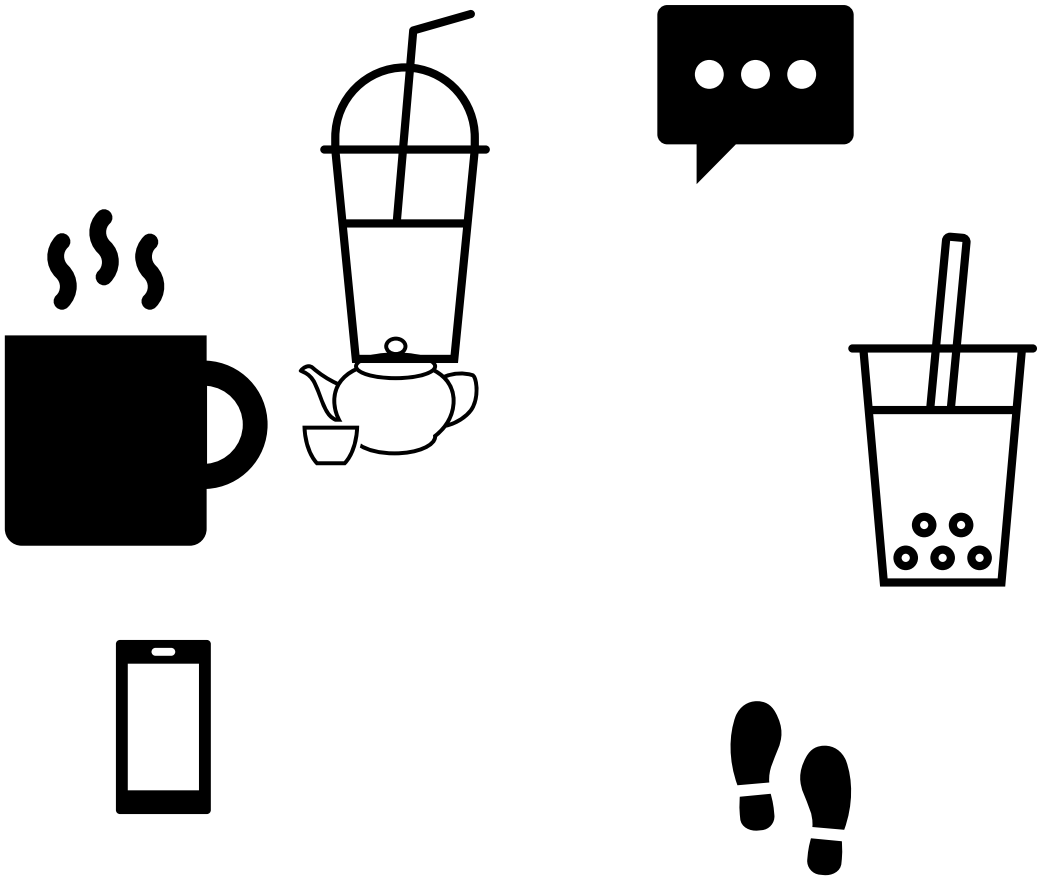
Conduct a statistical analysis
- Using appropriate statistical tests (using R to execute the analysis)

Report on your statistical analysis
- In a way that makes sense for your consumer

Interpret the outcomes of your statistical analysis
- Drawing appropriate conclusions

Report on the findings of your statistical analysis
- In a way that makes sense for your consumer

Time for a break!



Next Step

Get started with using R to help us to describe our data