Quantitative Reasoning with Data

Overview

- Introduction
- Population
- Research Question

Quantitative Reasoning with Data

- Population & Research Question

Introduction

Quantitative Reasoning Data

Introduction

Fall in Singapore marriages, divorces in 2020 amid Covid-19 restrictions, uncertainty

The number of marriages in Singapore in 2020 fell 10.9 per cent from the year before, the Singapore Department of Statistics said.

("Fall in Singapore marriages, divorces in 2020 amid COVID-19 restrictions, uncertainty," 2021)

Introduction

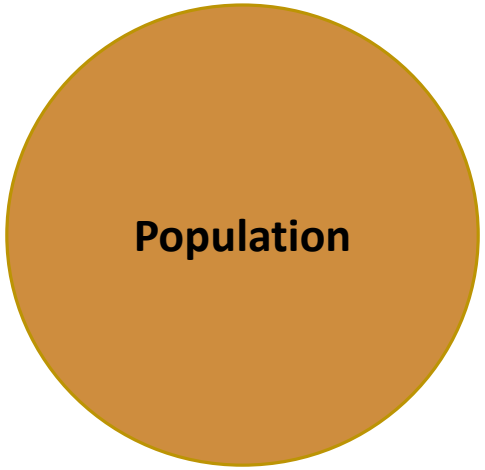
"Covid-19 is to blame for the drop in marriages?"

Singapore's population size for 2019 and 2020?

How did the study obtain the number 22,651?

Population

The population is the entire group (of individuals or objects) that we want to know something about.



Research Question

The research question is usually one that seeks to investigate some characteristic of a population

- *What percentage of Singapore adults owns a car?*

  (Population: Singapore adults)

- *Does Brand X pesticide work against mosquitoes?*

  (Population: Mosquitoes)

- *Do NUS students that take notes using pen and paper score better than those using laptops?*

  (Population: NUS students)

Research Question

Type of Research Questions	Examples
Make an estimate about the population	What is the average number of hours that students study each week?
	What proportion of all Singapore students is enrolled in a university?

Research Question

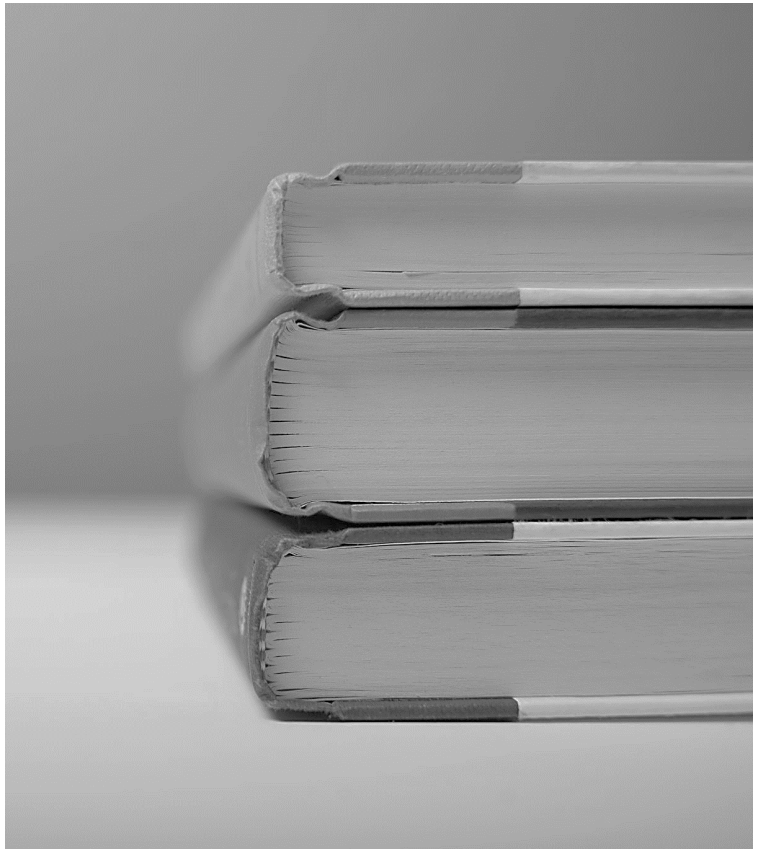
Type of Research Questions	Examples
Make an estimate about the population	What is the average number of hours that students study each week?
	What proportion of all Singapore students is enrolled in a university?
Test a claim about the population	Is the average course load for a university student greater than 20 units?
	Does the majority of students qualify for student loans?

Research Question

Type of Research Questions	Examples
Make an estimate about the population	What is the average number of hours that students study each week?
	What proportion of all Singapore students is enrolled in a university?
Test a claim about the population	Is the average course load for a university student greater than 20 units?
	Does the majority of students qualify for student loans?
Compare two sub-populations	In university X, do female students have a higher GPA score than male students?
	Are student athletes more likely than non-athletes to do final year projects?
Investigate a relationship between two variables in the population	Is there a relationship between the average number of hours students spend each week on Facebook and their GPA?
	Does drinking coffee help students pass the math exam?

Summary

- Introduction
- Population
- Research Question



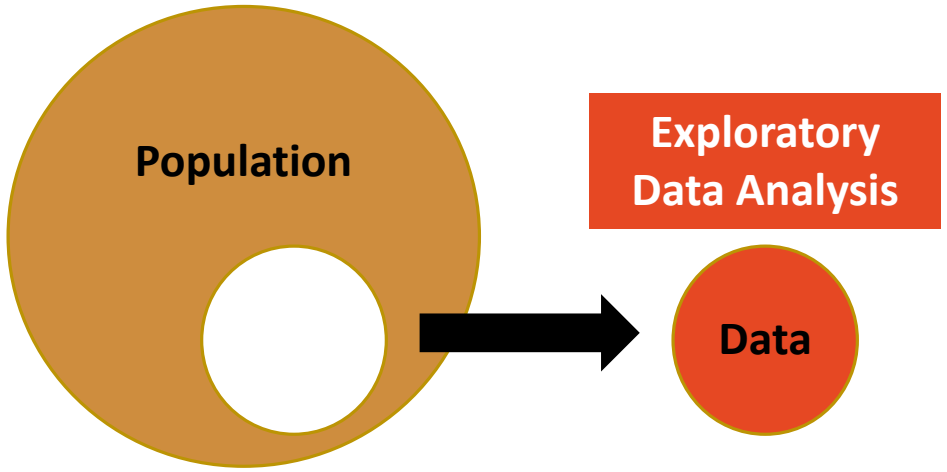
Sampling

Overview

- Exploratory Data Analysis (EDA) – An Introduction
- Understanding Population, Sample and Sampling Frame
- Understanding Bias and Error
- Probability Sampling
- Non-Probability Sampling
- Generalisability

Sampling
- Introduction to Sampling

Exploratory Data Analysis (EDA)

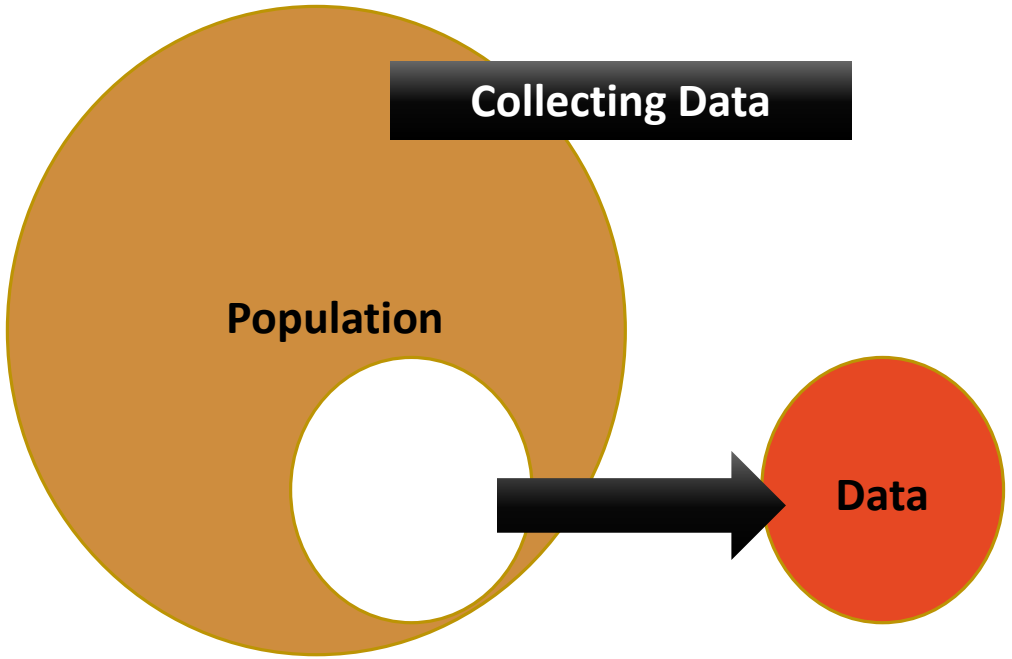


- Generate questions about data
- Search for answers by visualising, transforming and modelling data
- Use what is derived from the data to either
  - Refine our existing questions or
  - Generate new questions

Different Aspects of EDA in this module

Aspects of EDA
• Types of Variable (Categorical, Numerical)
• Correlation
• Visualisation Tools such as ✓ Bar Charts ✓ Scatter Plots ✓ Histogram ✓ Box Plots

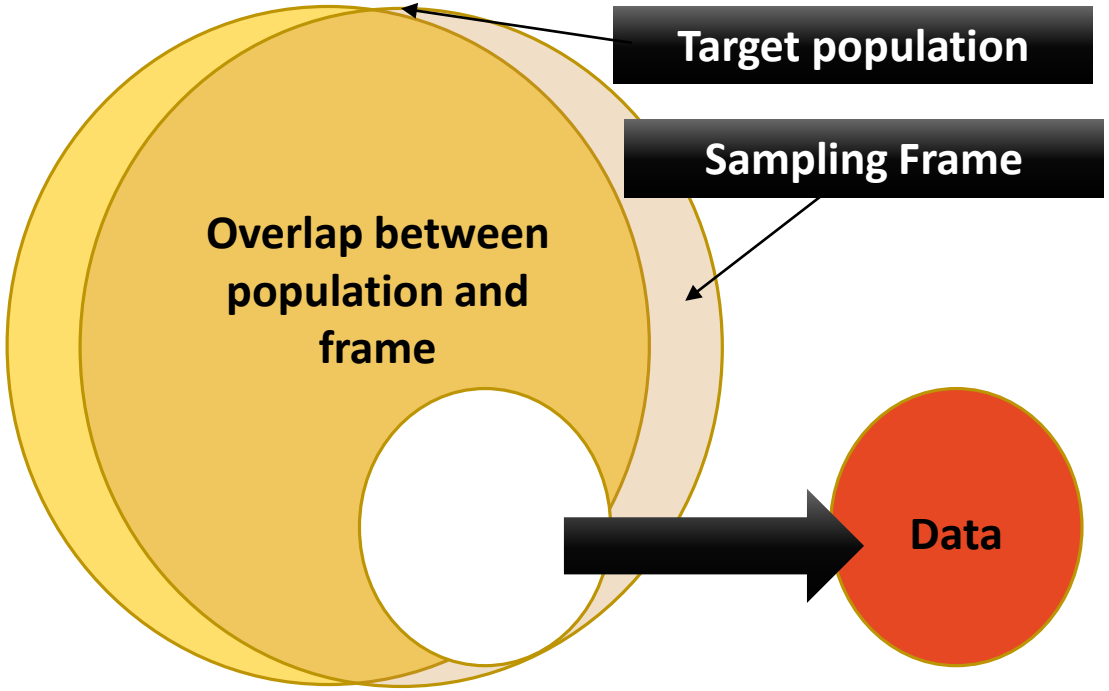
Population vs Sample



- Population of Interest – A group in which researcher has interest in drawing conclusions of the study.
  - Population Parameter – a numerical fact about a population.
- Sample – A proportion of the population selected in the study.
- Estimate – An inference about the population's parameter, based on information obtained from a sample

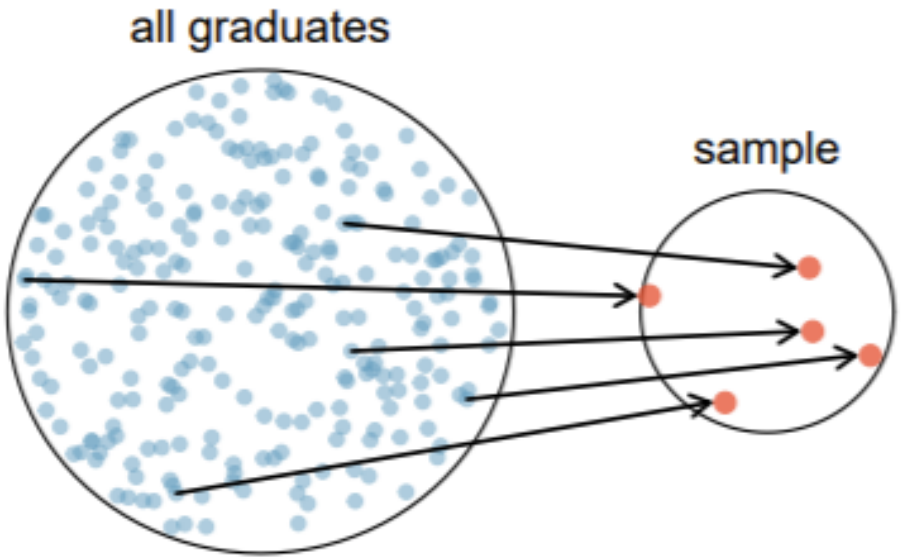
Sampling Frame

- "Source Material" from which sample is drawn.
  - Sample is drawn from the sampling frame.
- May not cover the population of interest, or may contain units that are not in the population of interest
- One of the conditions for generalisability – Sampling Frame equal to or greater than population of interest.



Census VS Sample

- Census – An attempt to reach out to the entire population of interest
- Sample – A proportion of the population selected
- Why sampling over population data?
  - Cost
  - Speed



To Watch: Bias

Selection Bias	Non-response Bias
Associated with the researcher's biased selection of units	Associated with the participants' non-disclosure of information related to the study
• Imperfect Sampling Frame • Non-Probability Sampling	• Disinterested • Inconvenient • Unwilling to disclose sensitive information

Sampling
- Sampling Methods

What is Probability Sampling?

- Sampling scheme such that the selection process is via a **known** randomised mechanism.
- The probability of selection may not be the same throughout all units of the population.
- General idea is to have the element of chance in the selection process, so as to eliminate biases associated with selection.

Picture Credits: https://www.scribbr.com/

Types of Probability Sampling

Simple Random Sampling (SRS)

- Units are selected randomly **without replacement** from the sampling frame
- Mechanism: Random Number Generator (Example: Random-Digit Dialling)
- An SRS of size n consists of n units from the population chosen in such a way that every set of units has equal chance to be the sample actually selected.
- Sample results do not change haphazardly from sample to sample. Variability is due to chance.

Types of Probability Sampling

Simple Random Sampling

- Advantage: Sample tends to be a good representation of population
- Disadvantage: Subject to non-response; accessibility of information

Types of Probability Sampling

Systematic Sampling

- A method of selecting units from a list by applying a selection interval *K, and random starting point from the first interval*.

Types of Probability Sampling

Systematic Sampling

Example: Suppose there are 110 sampling units in the population. A study requires us to select a sample of 10 units. Then a random number is selected from 1 to 110/10=11. If the selected number is 6, then units 6, 17, 28, ……105 are selected to form the sample.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110

Types of Probability Sampling

Systematic Sampling

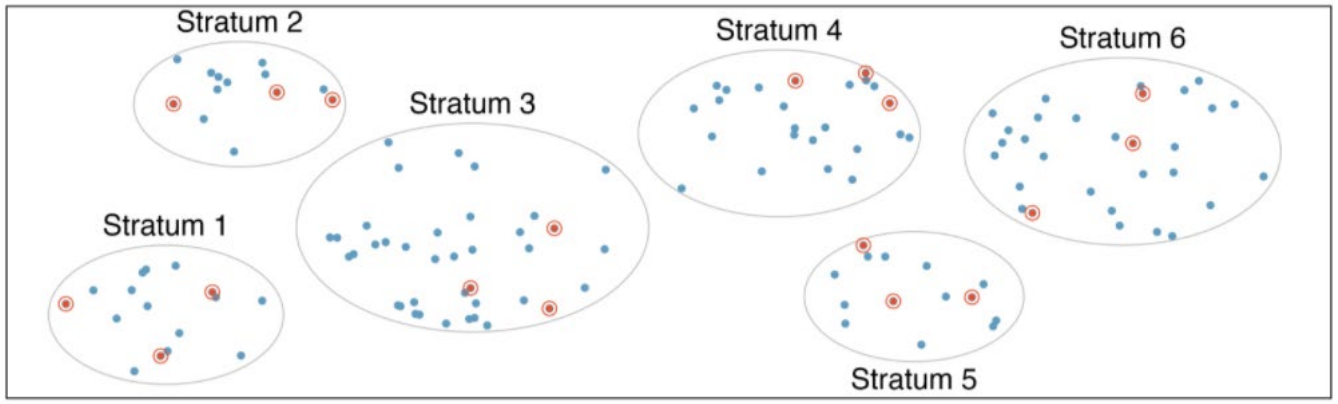
- Advantage: Simpler selection process than simple random sampling
- Disadvantage: May not be representative of population if list is non-random.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110

Types of Probability Sampling

Stratified Sampling

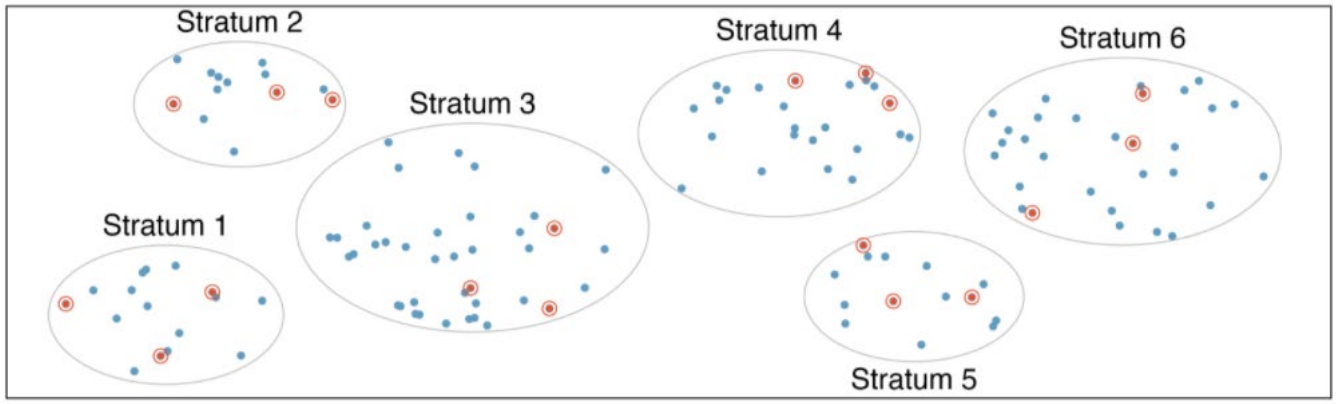
- Breaking down the population into strata.
- Each stratum are similar in nature, but size may be vary across strata
- Simple random sample from every strata
- Example: Sample Count (General Election)



Types of Probability Sampling

Stratified Sampling

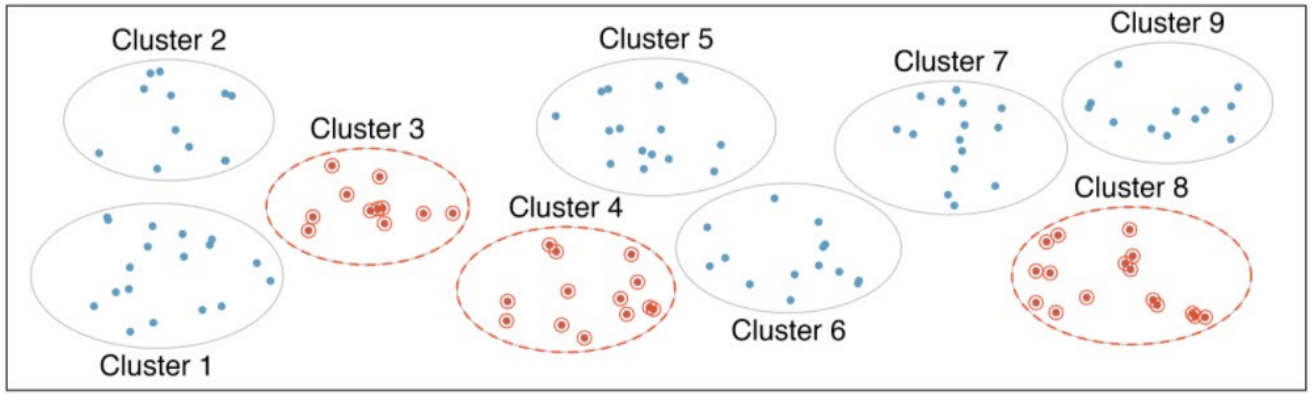
- Advantage: able to get a representative sample from every stratum
- Disadvantage: Need information about sampling frame and stratum



Types of Probability Sampling

Cluster Sampling

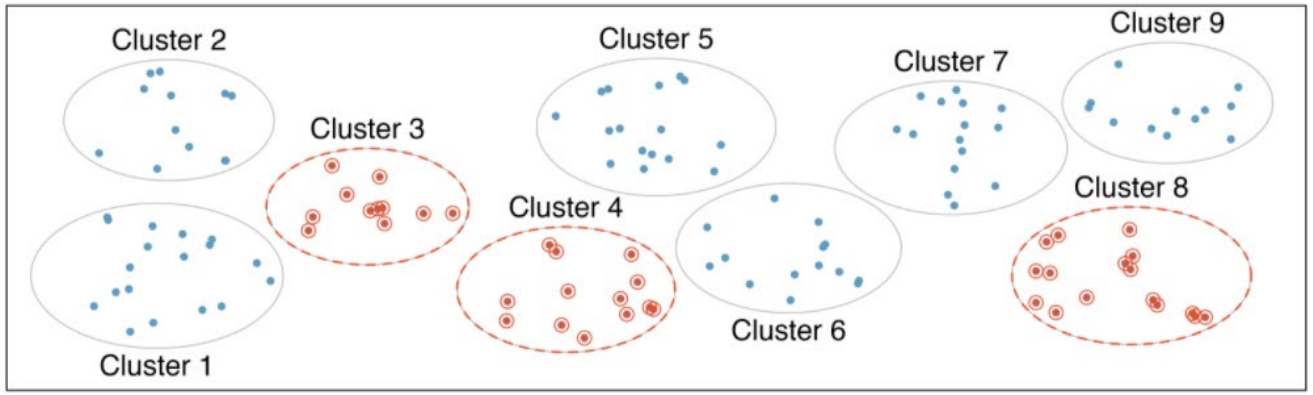
- Breaking down the population into clusters.
- Randomly sample a fixed number of clusters
- Include all observations from selected clusters
- Example: Mental Wellness Survey in schools



Types of Probability Sampling

Cluster Sampling

- Advantage: Less tedious, time-consuming and costly
- Disadvantage: High variability due dissimilar clusters or small number of clusters



Probability Sampling – Pros and Cons

Sampling Plan	Advantages	Disadvantages
Simple Random Sample	Good Representation of the Population	Time-consuming; accessibility of information
Systematic Sample	Simpler selection process as opposed to Simple Random Sampling	Potentially under-representing the population
Stratified Random Sample	Good Representation of Sample by Stratum	Require Sampling Frame and criteria for classification of population into stratum
Cluster Random Sample	Less time-consuming and less costly	Require larger sample size in order to achieve low margin of error

What is Non-Probability Sampling?

- The selection of individuals/unit were not done by randomisation, but by human discretion.

Example 1: Convenience Sampling

It is a non-probability sampling method in which the researcher uses the subjects that are most easily available to participate in the research study.

- An example: Mall surveys.
  - Issue 1: Demographics of mall goers – teenagers, retired people, people who are more affluent. Other groups (non-teenagers and retirees, and the not so affluent) are left out. This is a good example of **selection bias**.
  - Issue 2: Individuals asked to do the survey may not respond. This could lead to **non-response bias.**

Example 2: Volunteer Sampling

- It is a non-probability sampling method in which the researcher *actively* seek volunteers to participate in the study.
  - An example: Online Polls (American Family Association 2004)
    - Those who did not respond are left out of the study. This presents to us the clear problem on non-response bias, in a volunteer sample.

Selection Bias? | **Non-response Bias?**

General Approach to Sampling

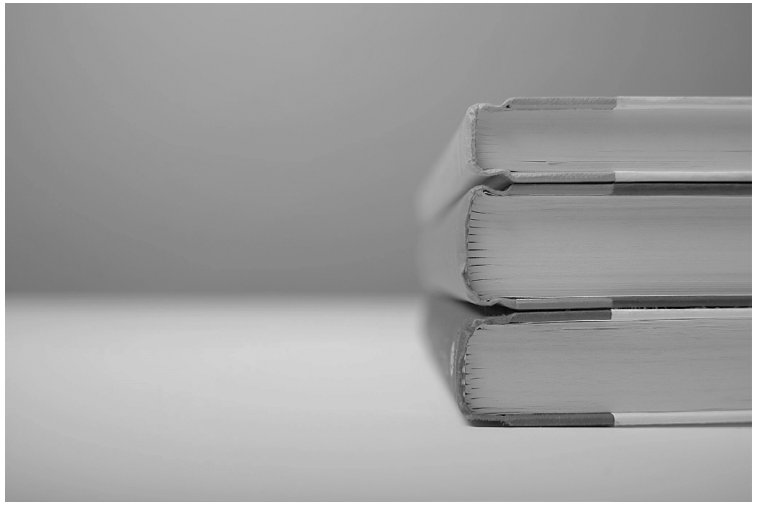
- Choose Sampling Frame
- Sample from Sampling Frame
- Remove unwanted units

Generalisability Criteria

- Good Sampling Frame
- Probability-based Sampling
- Large Sample Size
- Minimum Non-response

Summary

- Exploratory Data Analysis (EDA) – An Introduction
- Understanding Population, Sample and Sampling Frame
- Understanding Bias and Error
- Probability Sampling
- Non-Probability Sampling
- Generalisability



Understanding Variables

Overview

- Types of variables
  - Definition of variables
  - Independent and dependent variables
  - Categorical and Numerical variables
- Summary statistics
  - Mean and standard deviation
  - Median, quartiles and interquartile range
  - Mode

Understanding Variables

- Types of Variables

What is a variable?

- A **variable** is an attribute that can be measured or labelled.

- A data set consists of individuals and variables pertaining to the individuals.

Independent and dependent variables

- In research questions involving examining relationships between variables there are typically 2 sets of variables, namely independent variables and dependent variables.
  - An **independent variable** is a variable that maybe subject to manipulation (either deliberately or spontaneously) in a study
  - A **dependent variable** is a variable which is *hypothesised to change* depending on how the independent variable is manipulated in a study.

Examples

Research question	Dependent variable/Independent variable
Do NUS students who make notes using pen and paper score better in GEA1000 than those who use laptops?	Independent variable : Method of note taking for GEA1000
	Dependent variable : GEA1000 grade.
Does amount of caffeine consumed per day affect the quality of sleep amongst Singaporean adults?	Independent variable : Amount of caffeine consumed per day
	Dependent variable : Quality of sleep

Types of Variables

There are two main types of variables: numerical and categorical.

Categorical variables take category or label values. Each observation can be placed in only one label, and the labels are mutually exclusive (i.e no 2 labels overlap with each other).

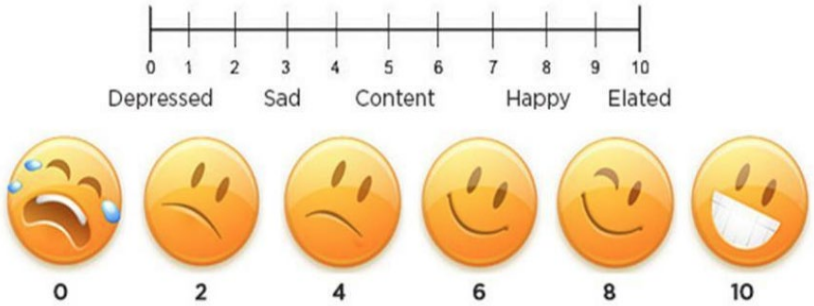
- Example : Smoking status can be a categorical variable, with two groups (smoker or non-smoker). Education Level is another example with multiple labels.

Numerical variables take numerical values for which **arithmetic operations such as adding and averaging make sense.**

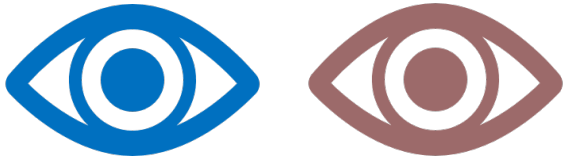
- Age, measured in years for example, is a numerical variable. Mass (kg) and height (m) are also numerical variables.

Categorical variables : Ordinal vs Nominal

- Sometimes categories come with some natural ordering and numbers are often used to represent the ordering. We call such variables ordinal.
- For example a happiness index can be rated from 0-10 in order of increasing happiness.
- Does this make happiness a numerical variable?



- In other cases where there is no intrinsic ordering for the variables, we will refer to these variables as nominal.
- For example if one were trying to collect basic information on a sample of birds, the eye colour (Blue / Brown) can be considered a nominal variable since there is no intrinsic ordering.



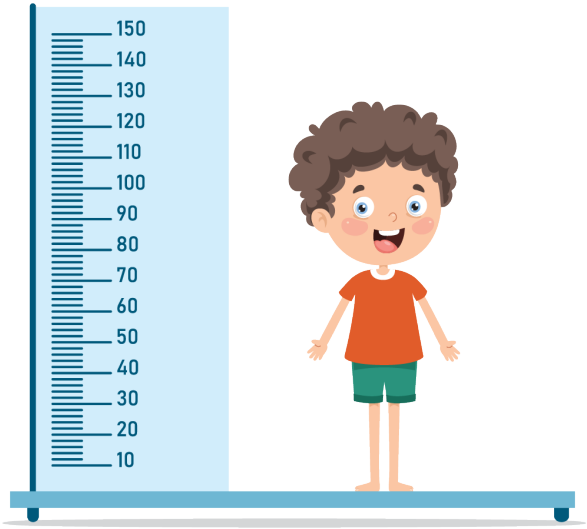
Numerical variables : Discrete vs continuous

Discrete variable : Is one where possible values of the variable form a set of numbers with "gaps" .

Example : Population count

Continuous variable : Is one that can take on all possible numerical values in a given range or interval.

Examples : Time, length.



Putting it altogether

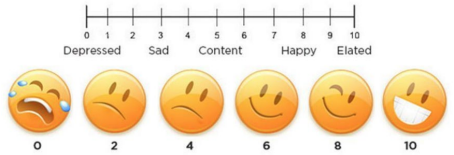
Types of Variables

Categorical
- Measured in categories

Nominal
- Named category with no definite order

Eye Colour

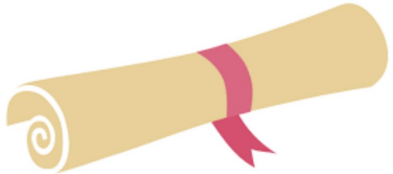
Ordinal
- Values can be ordered or ranked where distances between values are incomparable



Happiness Level

Numerical
- Measured in numerical values

Discrete
- Is one where possible values of the variable form a set of numbers with “gaps”



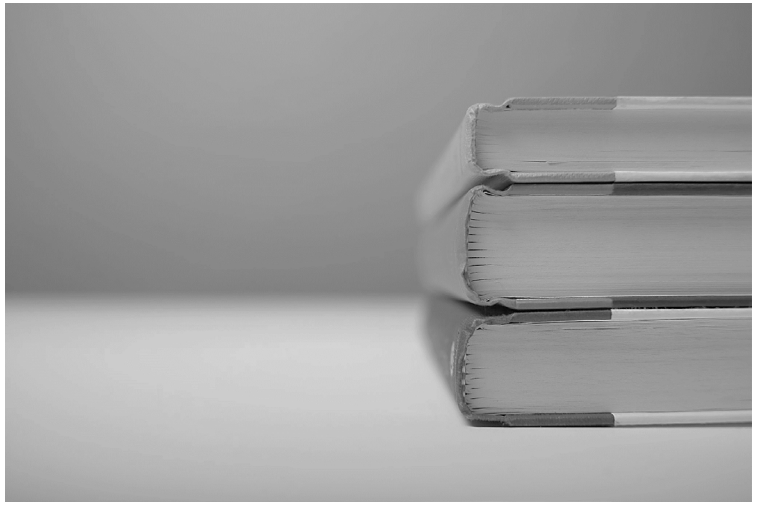
Module Credits (MCs)

Continuous
- Can take any numeric value on a given range

Time

Summary

- Types of variables
  - Definition of variables
  - Independent and dependent variables
  - Categorical
    - Ordinal and Nominal variables
  - Numerical variables
    - Discrete and Continuous variables



Understanding Variables
- Summary Statistics Part 1

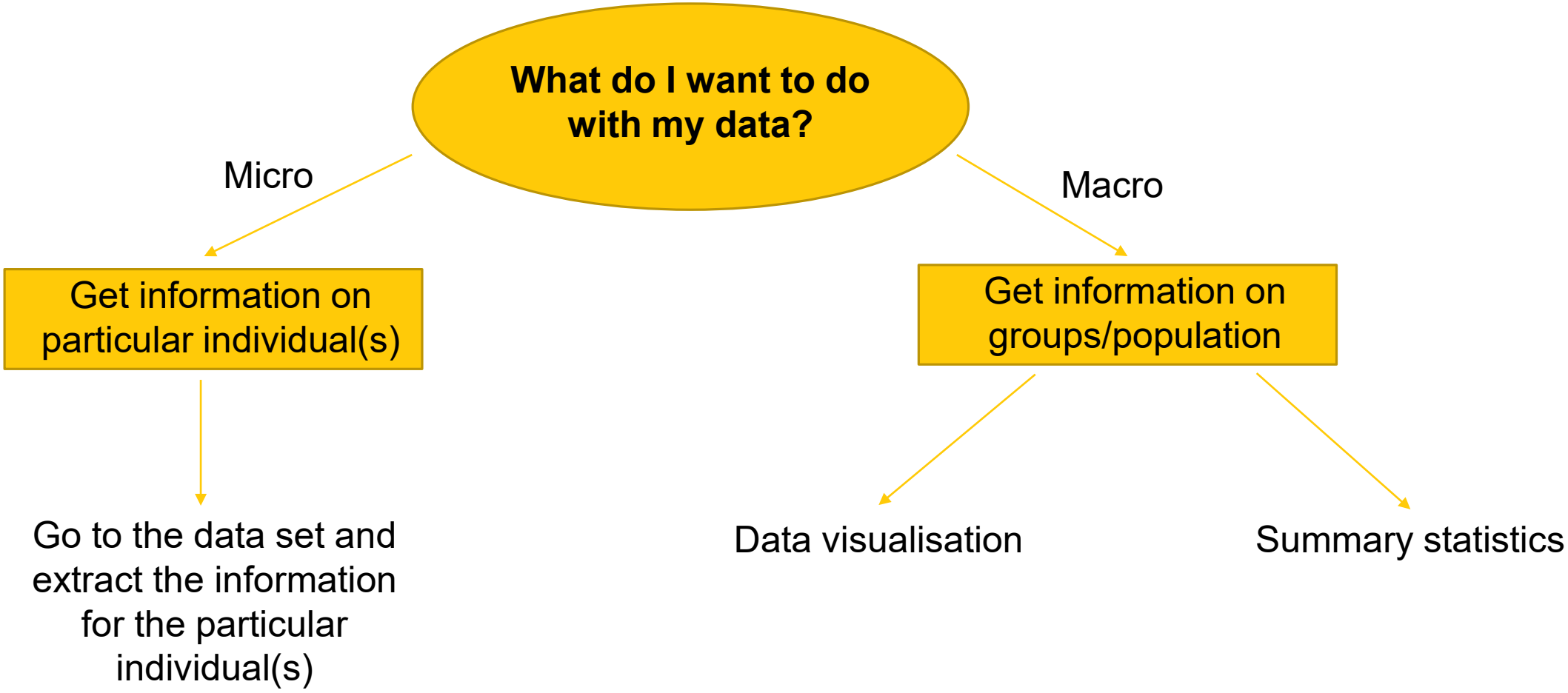
Presentation of data

Case	Age	Gender	Nationality	Days to Recover	Education Level	Confirmed At	Recovered At
1	66	Male	Chinese	26	Diploma	23rd, Jan 2020	19th, Feb 2020
2	53	Female	Chinese	14	University	24th, Jan 2020	7th, Feb 2020
3	37	Male	Chinese	27	High School	24th, Jan 2020	21st, Feb 2020
4	36	Male	Chinese	17	University	25th, Jan 2020	12th, Feb 2020
5	56	Female	Chinese	21	Diploma	27th, Jan 2020	18th, Feb 2020
6	56	Male	Chinese	23	Diploma	27th, Jan 2020	20th, Feb 2020
7	35	Male	Chinese	7	High School	27th, Jan 2020	4th, Feb 2020
8	56	Female	Chinese	20	Diploma	28th, Jan 2020	18th, Feb 2020
9	56	Male	Chinese	25	University	29th, Jan 2020	23rd, Feb 2020
10	56	Male	Chinese	10	High School	29th, Jan 2020	9th, Feb 2020
11	31	Female	Chinese	11	University	29th, Jan 2020	10th, Feb 2020
12	37	Female	Chinese	13	University	29th, Jan 2020	12th, Feb 2020

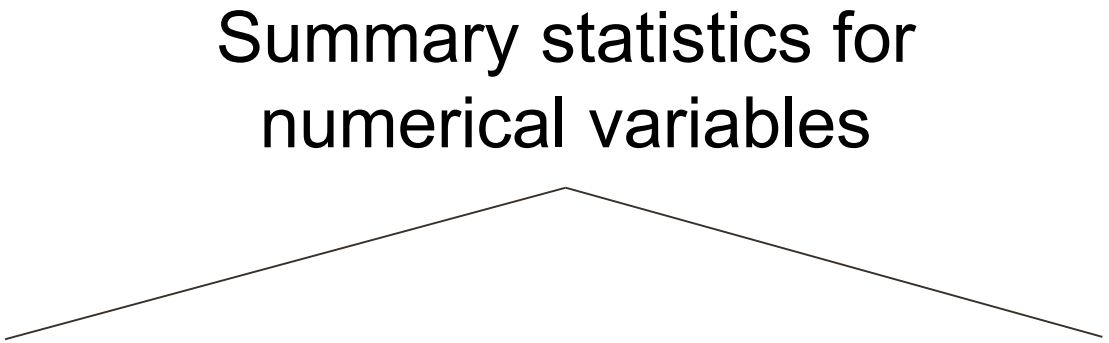
Row

Column

The macro and the micro



Summary statistics



Measures of central tendencies

- Mean
- Median
- Mode

Measures of dispersion

- Standard deviation
- Interquartile range

Mean and standard deviation

Mean

Case	Age	Gender	Nationality	Days to Recover
1	66	Male	Chinese	26
2	53	Female	Chinese	14
3	37	Male	Chinese	27
4	36	Male	Chinese	17
5	56	Female	Chinese	21
6	56	Male	Chinese	23
7	35	Male	Chinese	7
8	56	Female	Chinese	20
9	56	Male	Chinese	25
10	56	Male	Chinese	10
11	31	Female	Chinese	11
12	37	Female	Chinese	13

The mean of a numerical variable x, denoted by $\bar{x}$ (read as "x bar") is given by the formula

$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$ which is also written as $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Sigma notation

where n denotes the number of data-points and x_1 to x_n denotes the values of the numerical variable x in the data set.

EXCEL COMMAND : "=AVERAGE"

Properties of Mean

- $x_1+x_2+\ \ldots+x_n = n\bar{x}$. (This follows from the formula for mean in the previous slide)

- Adding a constant value to all the data points (be it positive or negative) changes the mean by that constant value.

- Multiplying all the values to all the data points by a constant number c will result in the mean also being multiplied by c.

Means in real-life scenarios

Whether the weather be fine

	A	B	C	D	E
1	Station	Year	Month	Day	Daily Rainfall Total (mm)
2	Changi	2020	January	1	0
3	Changi	2020	January	2	0
4	Changi	2020	January	3	0
5	Changi	2020	January	4	0
6	Changi	2020	January	5	0
7	Changi	2020	January	6	11
8	Changi	2020	January	7	4.8
9	Changi	2020	January	8	0
10	Changi	2020	January	9	0
11	Changi	2020	January	10	2.4
12	Changi	2020	January	11	0
13	Changi	2020	January	12	0
14	Changi	2020	January	13	0
15	Changi	2020	January	14	0
16	Changi	2020	January	15	0
17	Changi	2020	January	16	0
18	Changi	2020	January	17	0
19	Changi	2020	January	18	0
20	Changi	2020	January	19	0
21	Changi	2020	January	20	0

- Which were the hottest and coolest months?
- Which were the wettest months? How much rain do we typically get in a month?
- Is there any relationship between wind-speed, temperature and rainfall?
- Does the weather pattern for 2020, serve as a good prediction for how the weather will be in 2021?

Aggregating the data

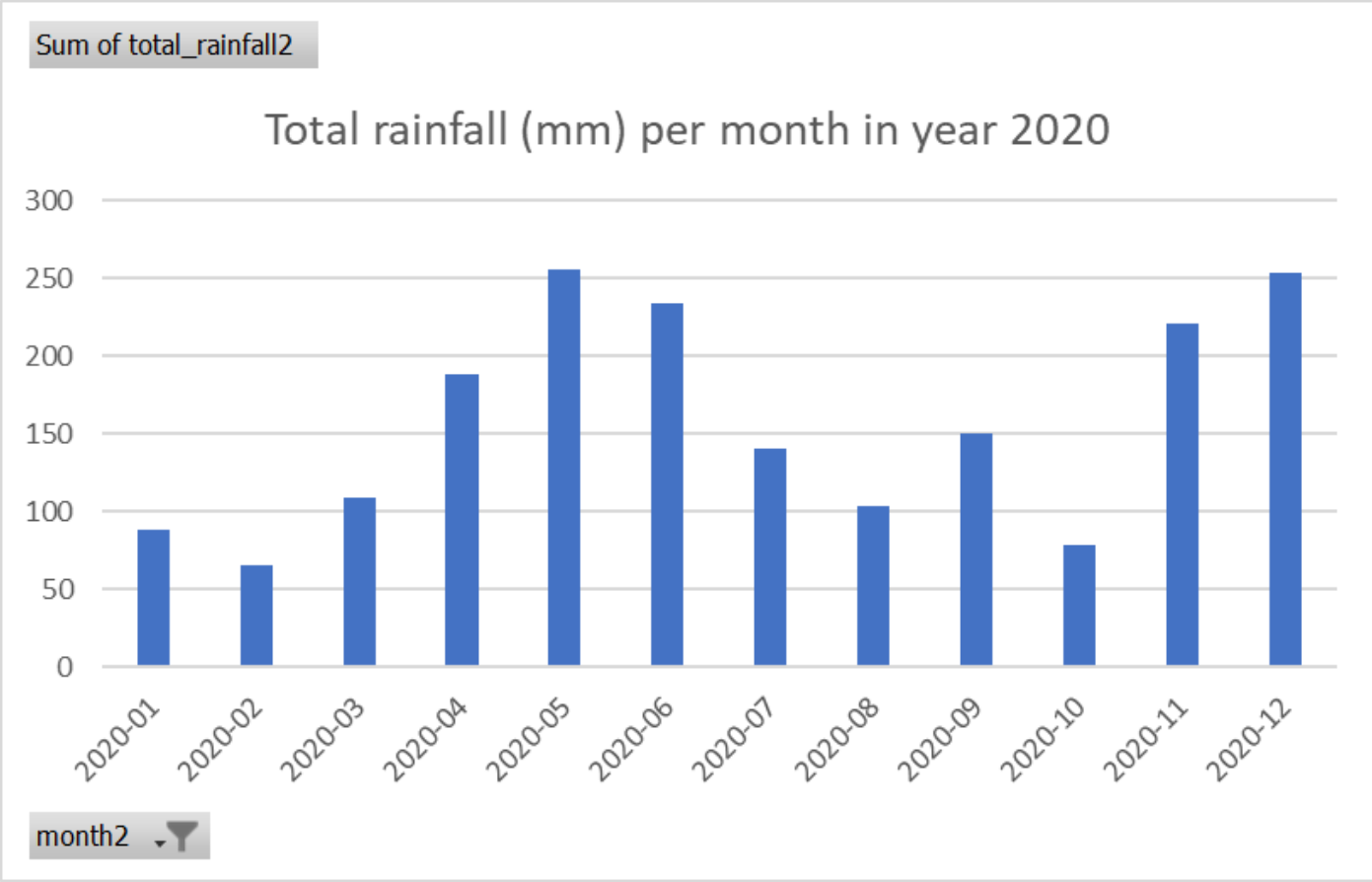
	A	B	C	D	E
1	Station	Year	Month	Day	Daily Rainfall Total (mm)
2	Changi	2020	January	1	0
3	Changi	2020	January	2	0
4	Changi	2020	January	3	0
5	Changi	2020	January	4	0
6	Changi	2020	January	5	0
7	Changi	2020	January	6	11
8	Changi	2020	January	7	4.8
9	Changi	2020	January	8	0
10	Changi	2020	January	9	0
11	Changi	2020	January	10	2.4
12	Changi	2020	January	11	0
13	Changi	2020	January	12	0
14	Changi	2020	January	13	0
15	Changi	2020	January	14	0
16	Changi	2020	January	15	0
17	Changi	2020	January	16	0
18	Changi	2020	January	17	0
19	Changi	2020	January	18	0
20	Changi	2020	January	19	0
21	Changi	2020	January	20	0

After aggregating the data, we can see that May and December were the wettest months

We can calculate the average rainfall per month which works out to be 157.22mm. But what does this average tell us?

month	total_rainfall
2020-01	88.4
2020-02	65
2020-03	108.8
2020-04	188
2020-05	255.6
2020-06	233.8
2020-07	140.8
2020-08	103.4
2020-09	150.2
2020-10	78.8
2020-11	220.6
2020-12	253.2

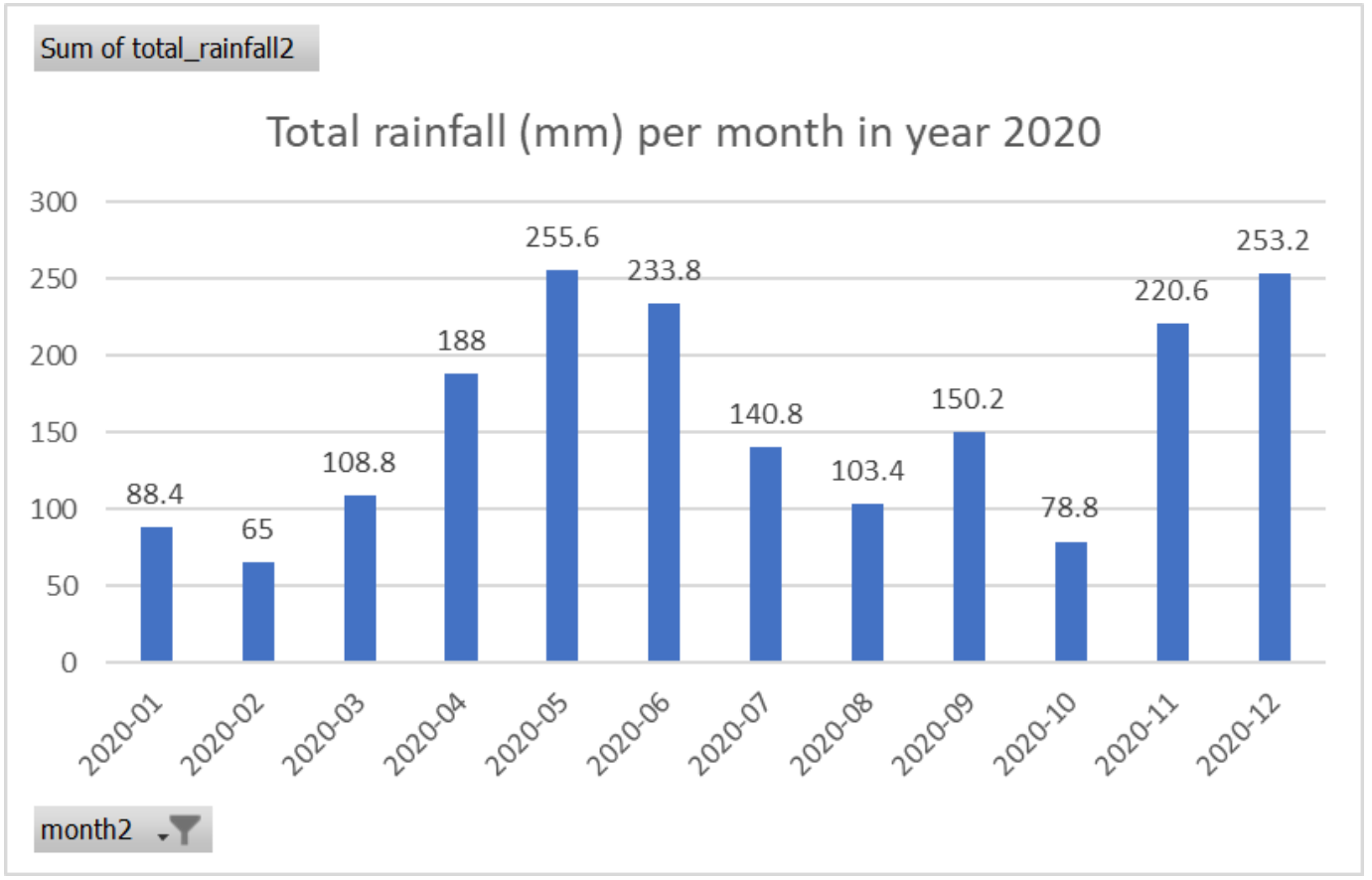
What the mean can tell us about rainfall



The bar graph shows the total rainfall for each month across the year 2020. The average rainfall per month in 2020 is 157.22mm.

We can add the total rainfall of each month to get the total rainfall for the entire year but knowing the mean gives us an easier way to get this quantity.

What the mean can’t tell us about rainfall



Overall mean vs means in subgroups

The table below shows the average performance of students when categorized by school. The maximum score attainable is 60.

	Number of students	Average Mark
School A	349	32.21
School B	46	30.72
Overall	395	?

Clinical trials : Means disguised as proportions

- Imagine we want to investigate the effectiveness of a new drug for treating asthma attacks compared to an existing drug. Is it fair to say that the new drug is performing better since there are only 200 asthma attacks for those taking the new drug as compared to 300, for those taking the existing drug?

	New drug	Existing drug
Number of patients	500	1000
Total asthma attacks	200	300

Proportion of asthma attacks with new drug = 0.4

Proportion of asthma attacks with existing drug = 0.3

How is proportion an example of mean?

	New drug	Existing drug
Number of patients	500	1000
Total asthma attacks	200	300

Proportion of asthma attacks with new drug = 0.4

No. of people receiving new drug = 500

Person who get an asthma attack : 1

Person who doesn't get an asthma attack : 0

Sample Variance and Standard deviation

- Recall that the mean doesn't tell us anything significant about how data is distributed which also includes the spread of the data points.
- The standard deviation is one way (yes there are other ways) of quantifying the "spread" of the data **about the mean**. The formula is derived via the variance.

$$Sample\ Variance = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \qquad s_x = \sqrt{Variance}$$

Notation for standard deviation of *x*

where n denotes the number of data-points and x_1 to x_n denotes the values of the numerical variable x in the data set. We assume this is sample data.

More intuition behind the formula

- **Question : Why can't we do the following to quantify spread?**
  - Take the difference between each value and the mean
  - Add up the differences to get the "total spread".
  - Divide by the number of points to get an "average spread"

- **Answer : Try applying that idea to a simple dataset {5, 10, 15, 20, 25}**
  - The mean is 15. The difference between each point and the mean gives us {-10, -5, 0, 5, 10}
  - What happens when you add up these differences to calculate the "total spread"?

An explicit computation of Standard deviation

Consider a simple sample data set of just 3 points. Let the points be {1, 4, 7} and suppose we wish to find the standard deviation of this data set.

Step 1: Find the average value of the data set.

In this case the average is $\frac{1+4+7}{3} = 4$.

Step 2 : Subtract the average value from each of your data points and square the answer.

We get $(1 - 4)^2 = 9, \quad (4 - 4)^2 = 0, \quad (7 - 4)^2 = 9.$

An explicit computation of Standard deviation

Step 3: Add up the results in Step 2 and divide by the number of points minus 1 to get the variance

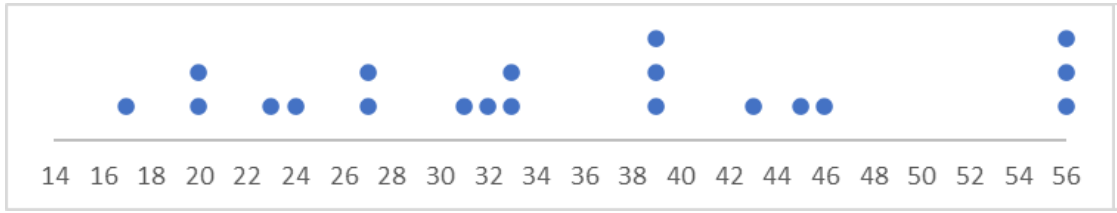
We get the variance to be $\frac{9+9}{2} = 9$

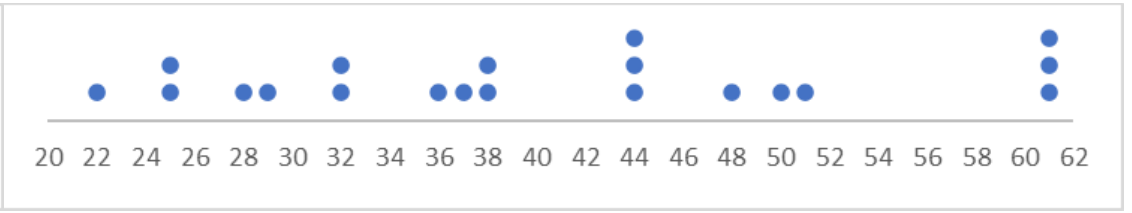
Step 4 : Take square root of the variance.

The standard deviation in this case is $\sqrt{9} = 3$

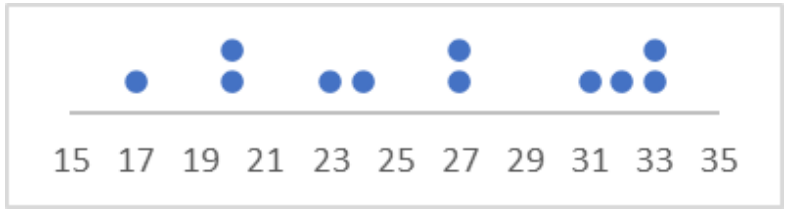
Properties of Standard deviation

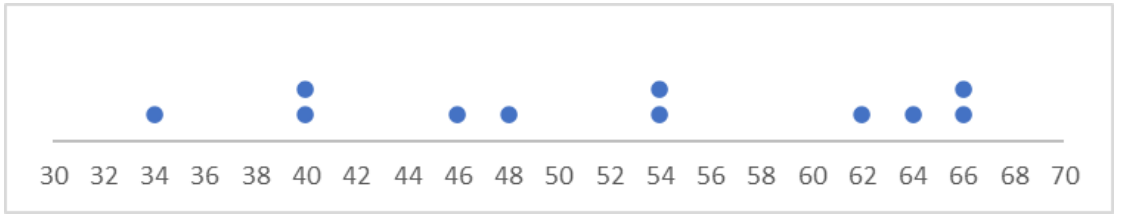
- The standard deviation is always non-negative. (i.e it is either 0 or a positive number) with same units (if any) as the numerical variable
- Adding a constant value, c (positive or negative) to all the data points **does not change the standard deviation.**





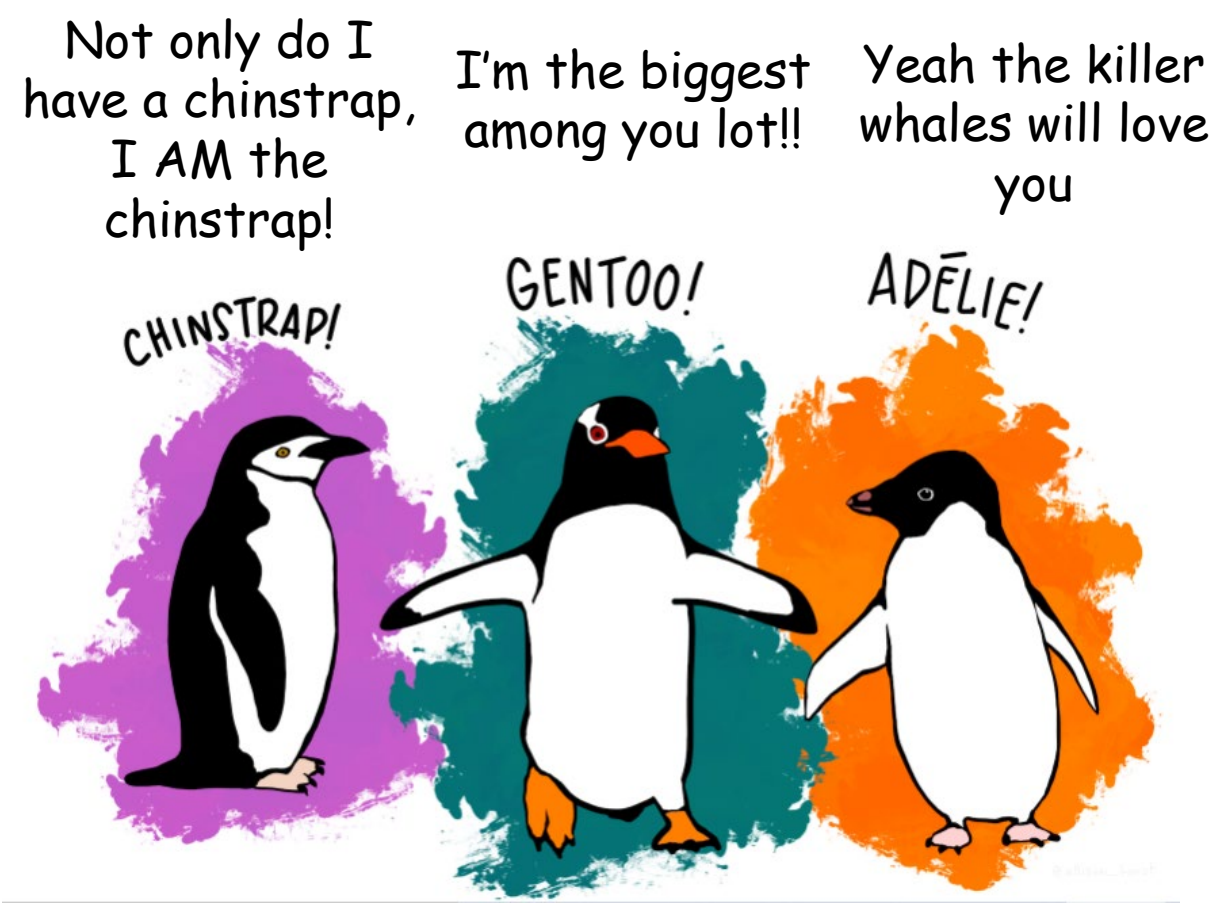
- Multiplying all the data-points by a constant value c results in the standard deviation being multiplied by $|c|$ where $|c|$ is the absolute value of c.





Standard deviations in real-life scenarios

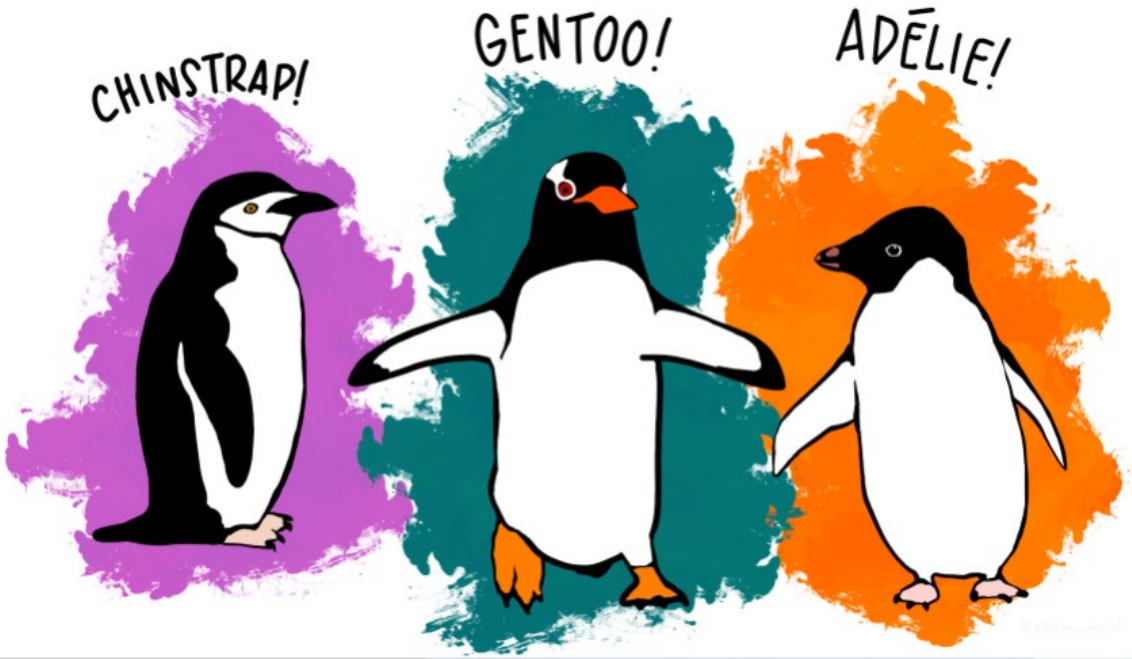
Meet the Palmer Penguins



species	island	bill_length_mm	bill_depth_mm
Chinstrap	Dream	46.9	16.6
Adelie	Biscoe	36.5	16.6
Adelie	Biscoe	36.4	17.1
Adelie	Biscoe	34.5	18.1
Adelie	Dream	33.1	16.1
Adelie	Torgersen	38.6	17
Chinstrap	Dream	43.2	16.6
Adelie	Biscoe	37.9	18.6
Adelie	Dream	37.5	18.9
Adelie	Dream	37	16.9
Adelie	Dream	37.3	16.8
Adelie	Torgersen	35.9	16.6
Adelie	Torgersen	35.2	15.9
Adelie	Torgersen	39	17.1
Adelie	Dream	32.1	15.5
Adelie	Biscoe	37.7	16
Adelie	Dream	36	18.5
Adelie	Biscoe	37.9	18.6
Adelie	Dream	36.5	18
Adelie	Biscoe	35.7	16.9

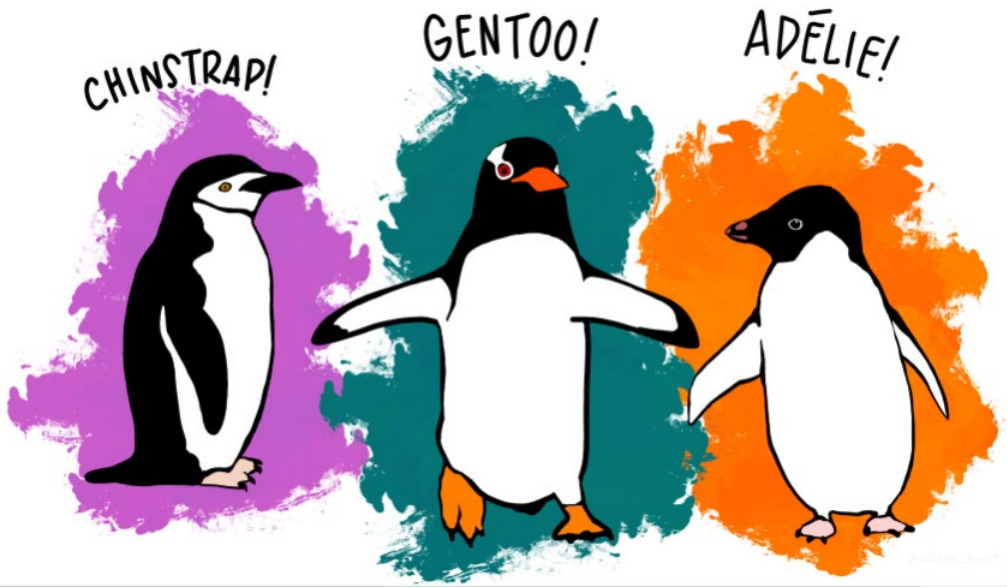
An overarching question

How different are these species of penguins?



species	island	bill_length_mm	bill_depth_mm
Chinstrap	Dream	46.9	16.6
Adelie	Biscoe	36.5	16.6
Adelie	Biscoe	36.4	17.1
Adelie	Biscoe	34.5	18.1
Adelie	Dream	33.1	16.1
Adelie	Torgersen	38.6	17
Chinstrap	Dream	43.2	16.6
Adelie	Biscoe	37.9	18.6
Adelie	Dream	37.5	18.9
Adelie	Dream	37	16.9
Adelie	Dream	37.3	16.8
Adelie	Torgersen	35.9	16.6
Adelie	Torgersen	35.2	15.9
Adelie	Torgersen	39	17.1
Adelie	Dream	32.1	15.5
Adelie	Biscoe	37.7	16
Adelie	Dream	36	18.5
Adelie	Biscoe	37.9	18.6
Adelie	Dream	36.5	18
Adelie	Biscoe	35.7	16.9

Comparing mass across species



Does this mean the heaviest penguin has a mass of 4201g + 802.0g = 5003g?

	Mean mass	Standard deviation of mass
Chinstrap	3733g	384.3g
Adelie	3701g	458.6g
Gentoo	5076g	504.1g
Overall	4201g	802.0g

Are the Adelie and Chinstrap similar?

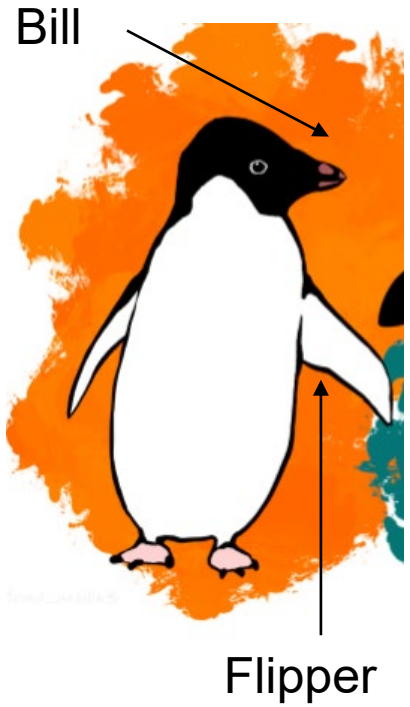
Why is it that the Adelie and Chinstrap have almost the same mass on average but yet the standard deviation for the Adelie species is more? Could it be due to

- **Gender?**
- **Age? (not found in the data set)**
- **Location?**

	Mean mass	Standard deviation of mass
Chinstrap	3733g	384.3g
Adelie	3701g	458.6g
Gentoo	5076g	504.1g
Overall	4201g	800.8g

Comparing the spread b/w variables

- Let's focus on the Adelie penguin species to define a notion that helps us quantify the degree of spread *relative to the mean*.



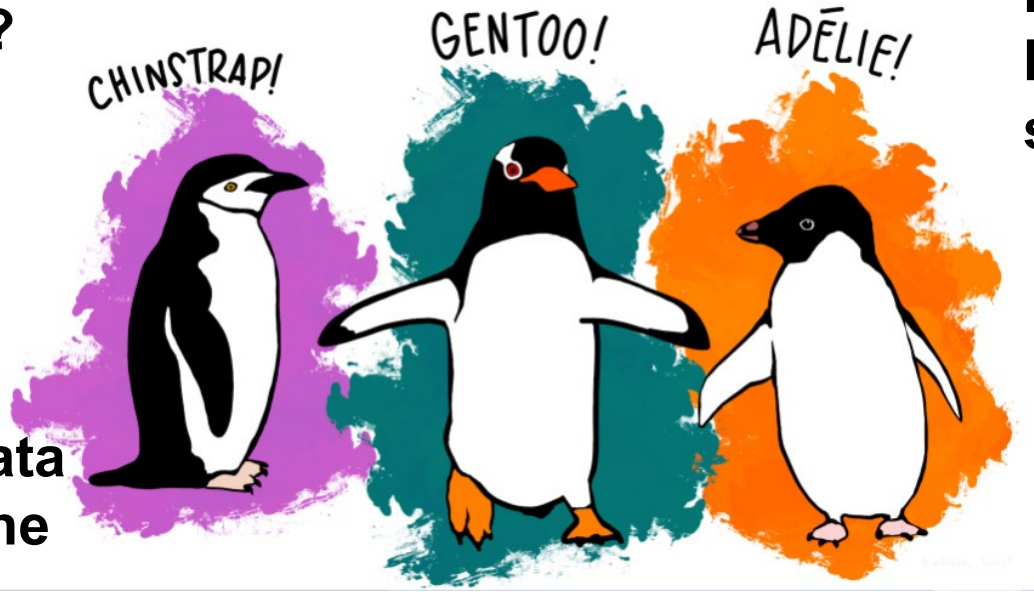
S.D of bill length	S.D of flipper length
2.65mm	6.52mm
Mean bill length	**Mean flipper length**
38.8mm	190.0mm
Coefficient of variation	**Coefficient of variation**
0.07	0.03

The coefficient of variation is a way of quantifying the degree of spread, *relative to the mean.*

$$coefficient\ of\ variation = \frac{s_x}{mean\ of\ x}$$

Questions, Questions and more questions

Are male penguins heavier than female penguins across all species?



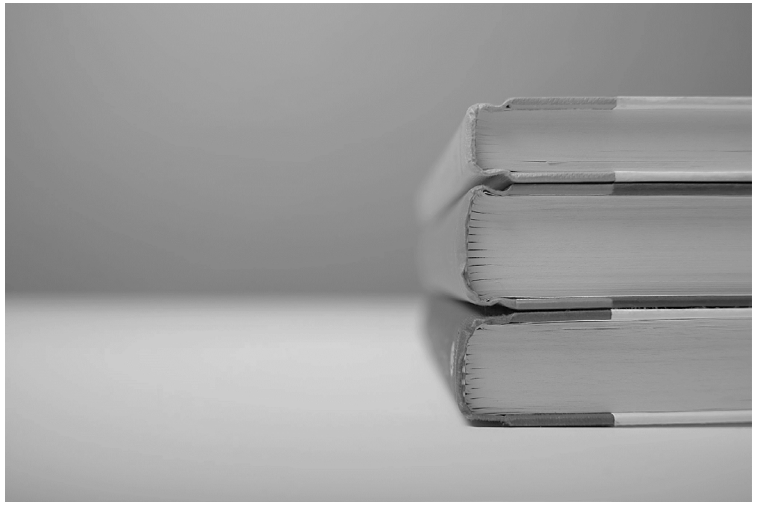
Is there a relationship between bill length and bill-depth across all species?

Can findings in this data be generalized to all the three species of penguins?

Do the heavier penguins come from colder locations?

Summary

- Summary statistics
  - Mean
    - Formula for mean
    - Basic properties of mean
    - Means in real-life scenarios
  - Standard deviation
    - Formula for standard deviation via variance
    - Properties of standard deviation
    - Standard deviations in real-life scenarios



Understanding Variables
- Summary Statistics Part 2

Median, quartiles and interquartile range

Median

The **median** of a numerical variable in a data-set is the middle value of the variable after arranging the values of the data-set in ascending/descending order.

Case	Age	Gender	Days to Recover
1	66	Male	26
2	53	Female	14
3	37	Male	27
4	36	Male	17
5	56	Female	21
6	56	Male	23
7	35	Male	7
8	56	Female	20
9	56	Male	25
10	56	Male	10
11	31	Female	11
12	37	Female	13

Case	Age	Gender	Days to Recover
7	35	Male	7
10	56	Male	10
11	31	Female	11
12	37	Female	13
2	53	Female	14
4	36	Male	17
8	56	Female	20
5	56	Female	21
6	56	Male	23
9	56	Male	25
1	66	Male	26
3	37	Male	27

→ Arranged in ascending order based on the number of days to recover

Median of the number of Days To Recover?

} (17+20)/2 = 18.5

EXCEL COMMAND : "=MEDIAN"

Properties of median

- Adding a constant value (positive or negative) to all the data points changes the median by that constant value.

- Multiplying all the data points by a constant value c results in the median being multiplied by c.

Medians in real-life scenarios

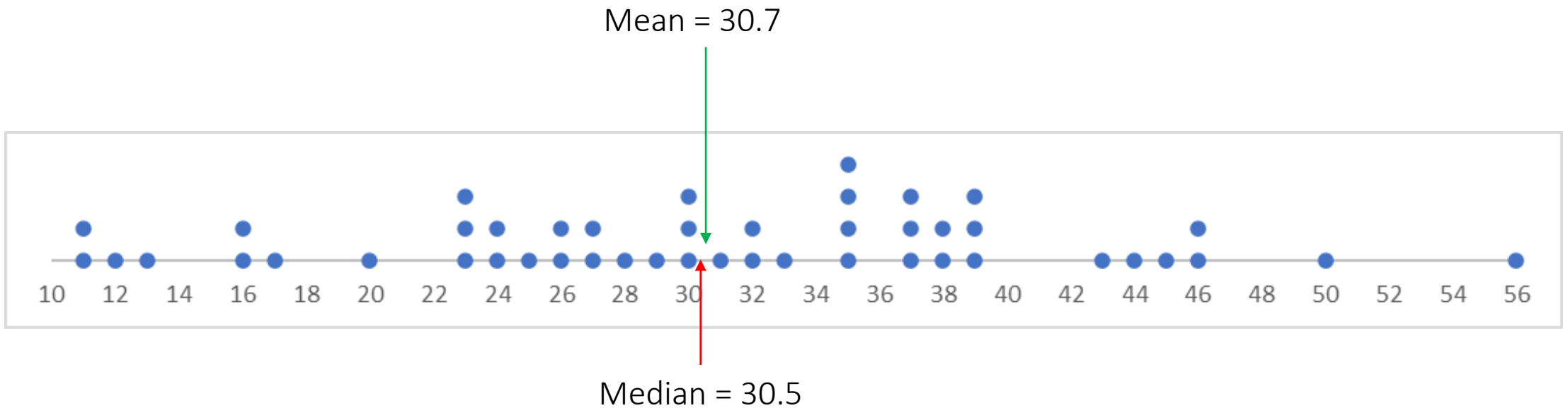
Performance in a test

- Recall that when we were discussing means, we showed this table regarding average performance in a test of 2 schools. Let's focus on School B which consists of 46 students.

	Number of students	Average Mark
School A	349	32.21
School B	46	30.72
Overall	395	?

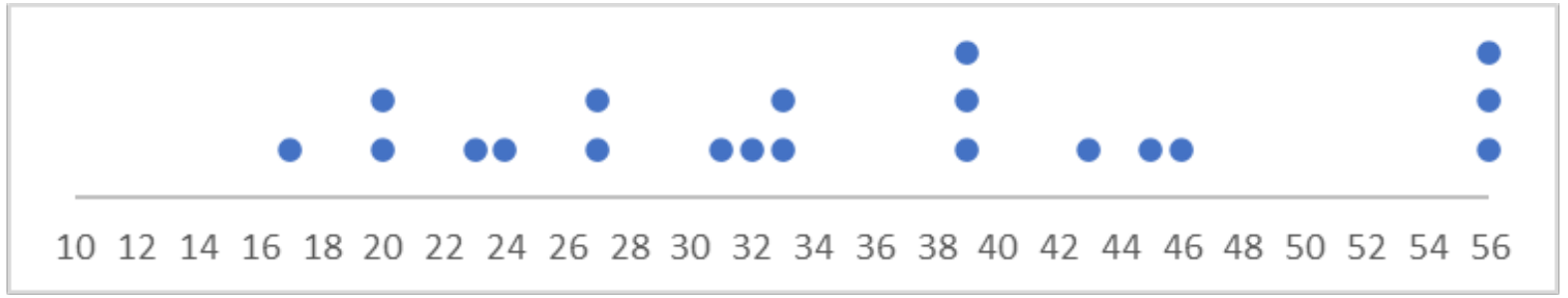
What the median can/can't tell us

- Since School B only has 46 students, a dot plot is quite handy in viewing how the scores are distributed. The median score is 30.5.
- **Is there any reason why the mean is so close to the median? Does this happen all the time?**



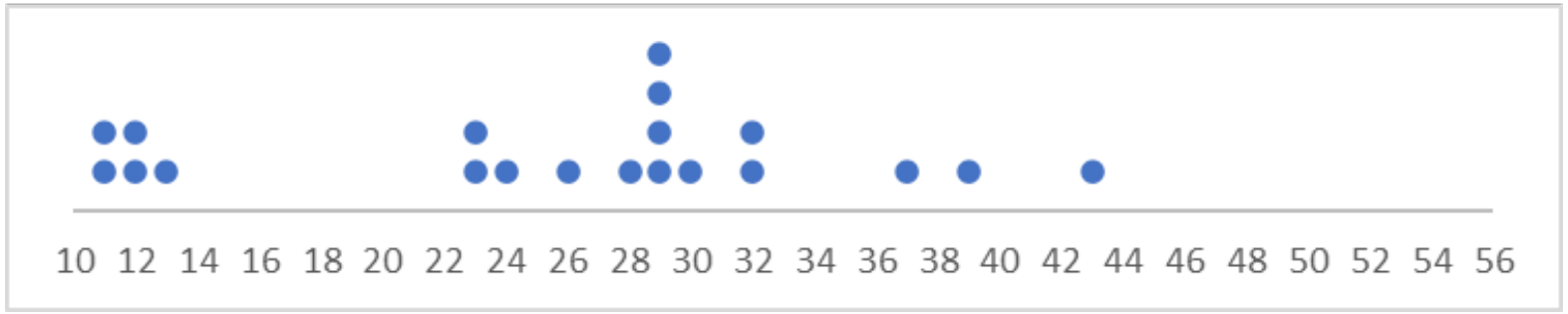
Overall median vs median in subgroups

Students in Class A



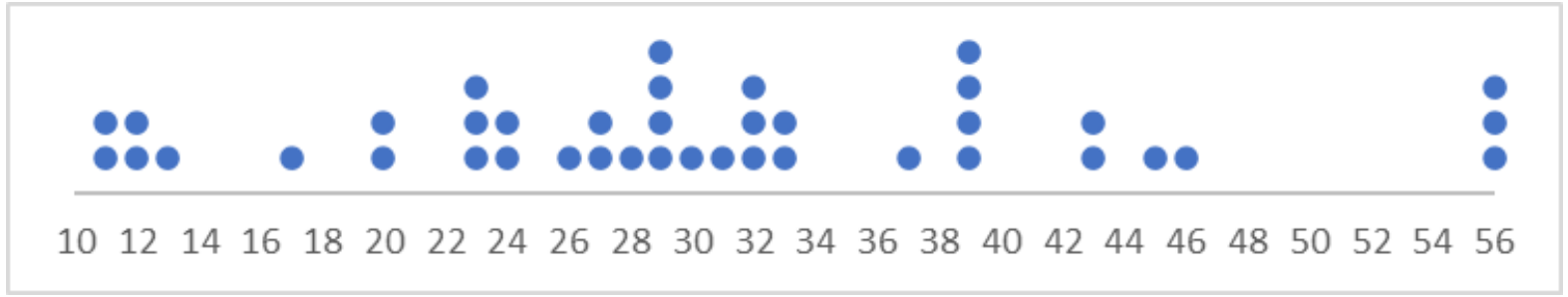
Median = 33

Students in Class B



Median = 28.5

Class A + Class B



Median = 29

Quartiles and InterQuartileRange

- The **first quartile** usually denoted by Q_1 is the 25^{th} percentile of the data-values and the **third quartile**, usually denoted by Q_3 is the 75^{th} percentile of the data-values.

  (The 25^{th} percentile, is a value such that 25% of the data is either equal or less than this value. Same idea for the 75^{th} percentile)

  EXCEL COMMAND : "=QUARTILE"

- The **interquartile range** is the difference between the third quartile and the first quartile.

$$IQR = Q_3 - Q_1$$

It gives us another way of quantifying the spread of the data.

Similarities between IQR and S.D

- The IQR is always non-negative and this follows from the fact that Q_3 is at least as large as Q_1.

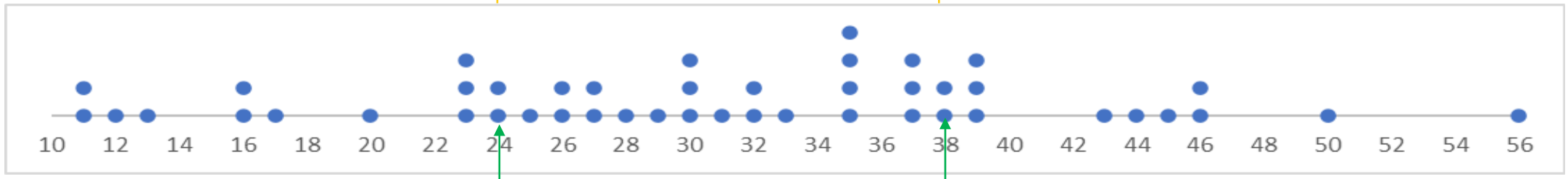
- Adding a constant value, c (positive or negative) to all the data points **does not change the IQR.**

- Multiplying all the data points by a constant value c results in the IQR being multiplied by $|c|$.

Explicit computation of Quartiles and IQR

Let's go back to the same 46 students from School B and explicitly work out the IQR

$$IQR = 38 - 24 = 14$$



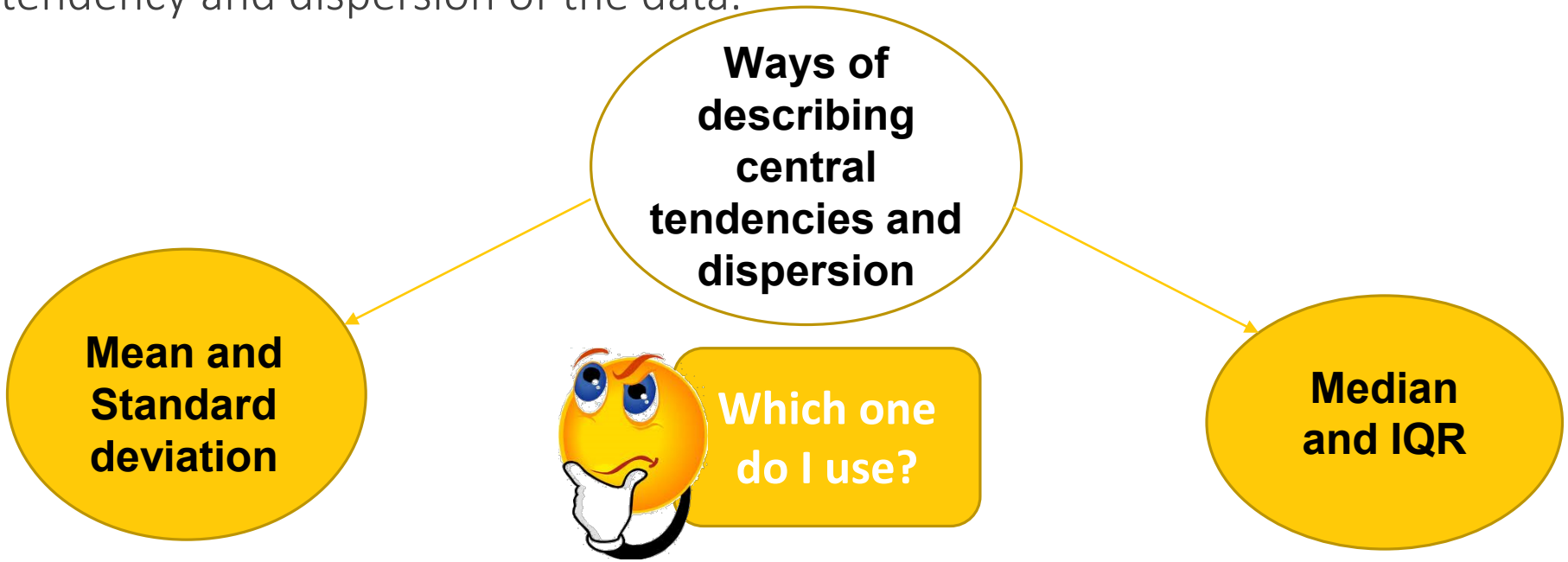
$Q_1 = 24$ $Q_3 = 38$

11, 11, 12, 13, 16, 16, 17, 20, 23, 23, 23, 24, 24, 25, 26, 26, 27, 27, 28, 29, 30, 30, 30.

- For this module, when we have an odd number of data points, we will **not include** the median in both halves if we are computing the quartiles manually.

Using summary statistics appropriately

- The mean and standard deviation are a pair of summary statistics that attempt to describe the central tendency and dispersion of the data
- The median and IQR are another pair of summary statistics that attempt to describe the central tendency and dispersion of the data.



Mode

Mode

Mode of a variable is the value of the variable that appears the most frequently.

Case	Age	Gender	Days to Recover
1	66	Male	26
2	53	Female	14
3	37	Male	27
4	36	Male	17
5	56	Female	21
6	56	Male	23
7	35	Male	7
8	56	Female	20
9	56	Male	25
10	56	Male	10
11	31	Female	11
12	37	Female	13

Mode for Gender?

Male

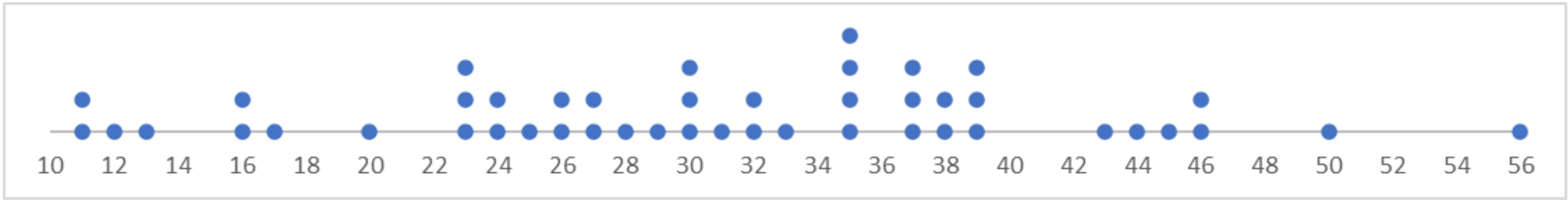
Mode for Age?

56

EXCEL COMMAND : “=MODE”

Interpretation of mode as “peaks”

- For the final time, lets go back to that dot-plot of the 46 students from School B and see what the mode is telling us.



- When we are describing the distribution of points of a discrete variable, the mode can be interpreted as a “peak” of the distribution. In the context of probability, a peak of the distribution, refers to the value that has the highest probability of occurring.
- We will touch upon this idea more in later chapters when we define a Discrete Random Variable (DRV).

Applications of mode

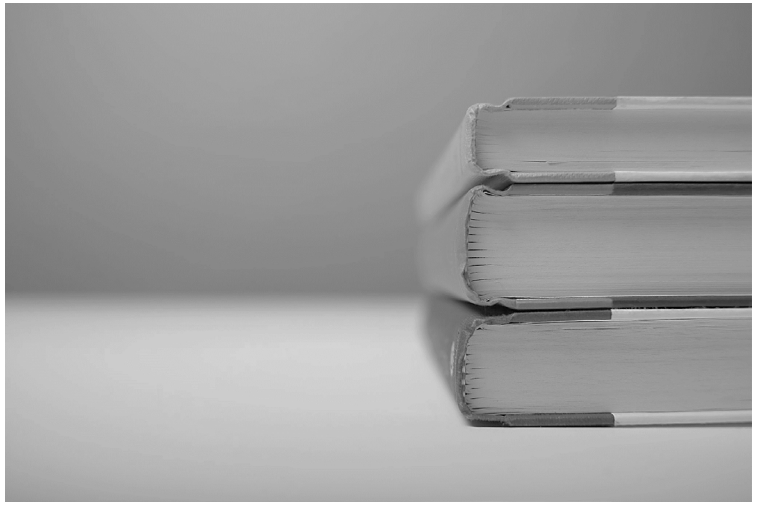
Real estate : Real estate agents need the mode of the number of bedrooms per house so they can inform their clients on how many bedrooms they can expect to have in houses located in a particular area.

HR : Human Resource managers also use the mode of different positions in the company so that they can be aware of the most common position of employees at their company.

Healthcare and Insurance : Actuaries also calculate the mode of the age of their customers (the most commonly occurring age) to find out which age group uses their insurance the most.

Summary

- Summary statistics
  - Median
    - Definition of median
    - Medians in real-life scenarios
  - Quartiles and interquartile range
    - Definition of 1st and 3rd Quartile.
    - Formula for Interquartile range
    - Explicit computation of interquartile range
  - Mode
    - Interpretation of mode
    - Applications of mode.



Study Designs

Overview

- Experimental Studies
  - Treatment and Control Groups
  - Random Assignment
  - Blinding
- Observational Studies
- Experimental vs Observational Studies

Study Designs
- Experimental Studies

Recall

Type of Research Questions
Make an estimate about the population
Test a claim about the population
Compare two sub-populations
Investigate a relationship between two variables in the population

Recall

Type of Research Questions
Make an estimate about the population
Test a claim about the population
Compare two sub-populations
Investigate a relationship between two variables in the population

Does
drinking coffee
help students
pass the math exam
?

Types of Study Designs

Experimental

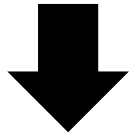
Observational

Experiments

An experiment intentionally manipulates one variable in an attempt to cause an effect on another variable.

The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.

Independent Variable



Dependent Variable

Experiments

An experiment intentionally manipulates one variable in an attempt to cause an effect on another variable.

The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.

Drinking coffee

Help students ⬇

Pass the math exam

Experiments – Treatment and Control

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*

Experiments – Treatment and Control

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*

Treatment Group

Control Group

Experiments – Treatment and Control

Coffee

Drink exactly one cup of coffee every day for one month

Treatment Group

Experiments – Treatment and Control

Coffee

Drink exactly one cup of coffee every day for one month

Treatment Group

90% passed

Coffee helps?

Experiments – Treatment and Control

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*

Treatment Group

90% passed

Maybe???

90% passed

Experiments – Treatment and Control

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*

Treatment Group

Control Group

The control group provides a baseline for comparison with the treatment group.

Experiments – Treatment and Control

Antidepressant	Nicotine patch

Treatment Group

Control Group

The control group provides a baseline for comparison with the treatment group.

Study Designs
- Experimental Studies (Random Assignment)

Experiments – Study results

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*
Treatment Group	**Control Group**

Experiments – Study results

	Coffee	No coffee
	Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*
Pass		
Fail		

Experiments – Study results

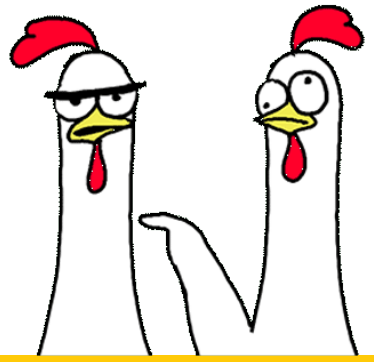
	Coffee	No coffee
	Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*
Pass	900	450
Fail	100	550

90% passed

45% passed

Experiments

	Coffee	No coffee
	Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*
Pass	900	450
Fail	100	550



Coffee causes an improvement in passing the math exam!

Not wise?

Experiments

Revision time

Pass the math exam

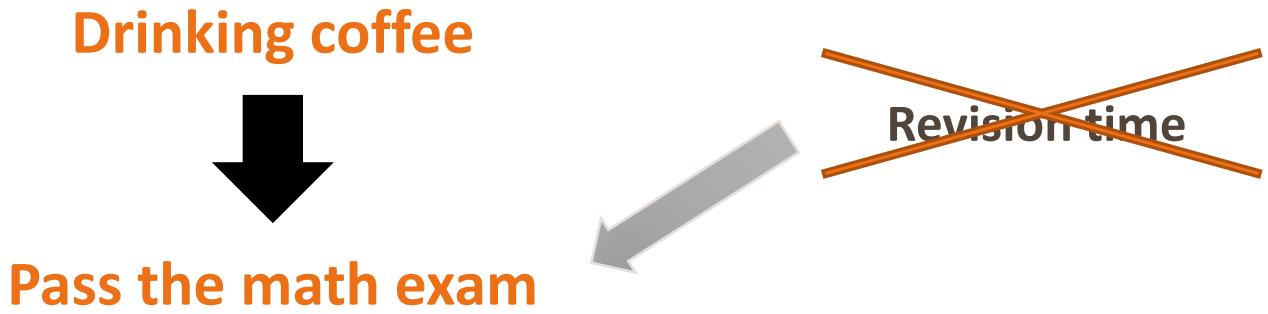
Experiments

	Long Revision Time	Short Revision Time
Pass	900	450
Fail	100	550

Possible?

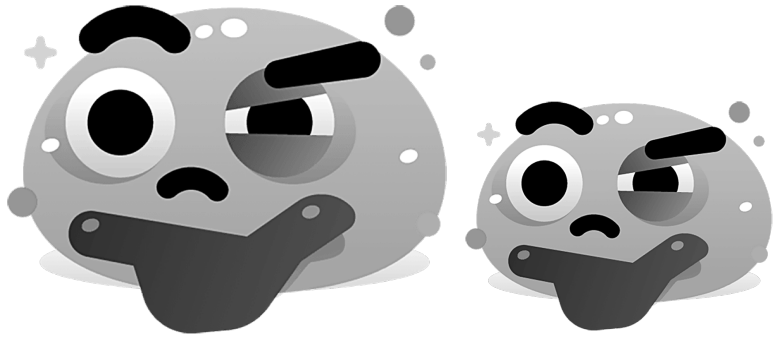
Experiments

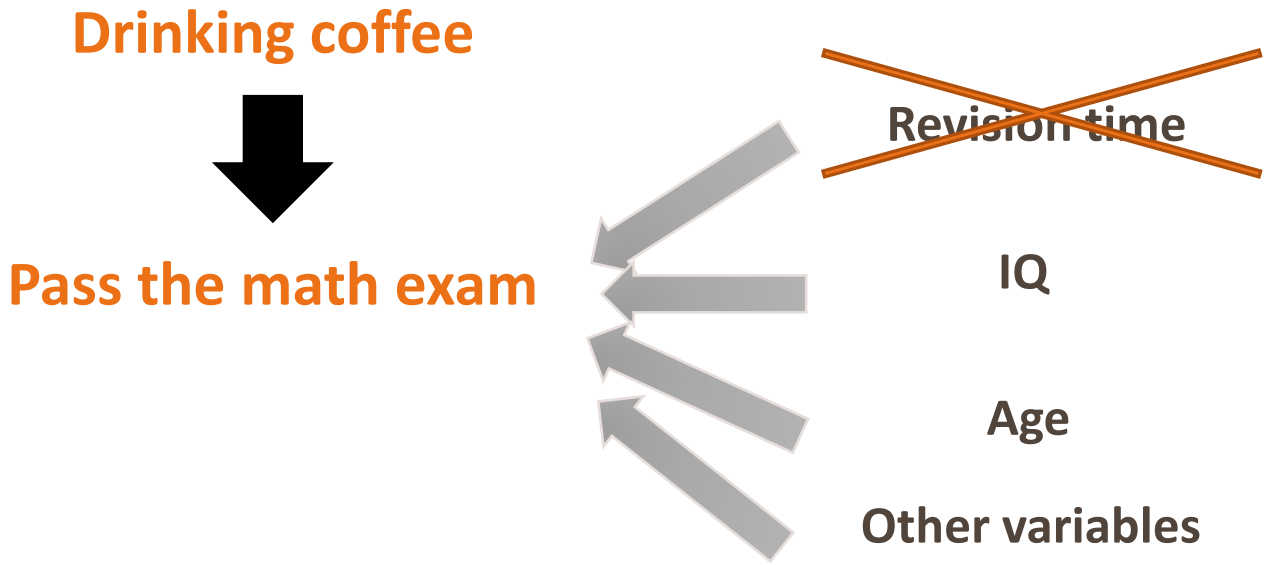
To establish a cause-and-effect relationship, we want to make sure that the independent variable is the only factor that impacts the dependent variable.



Experiments

To establish a cause-and-effect relationship, we want to make sure that the independent variable is the only factor that impacts the dependent variable.





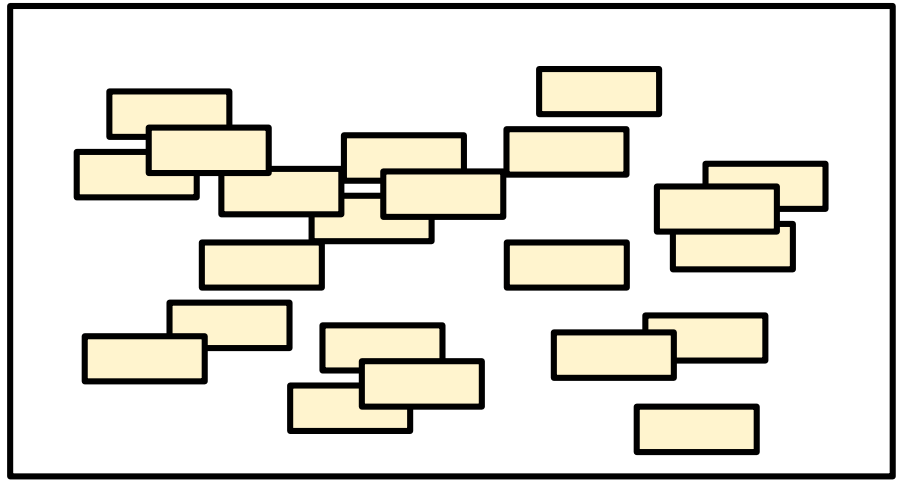
Experiments

How do we account for the effects from all these other variables?

Random Assignment

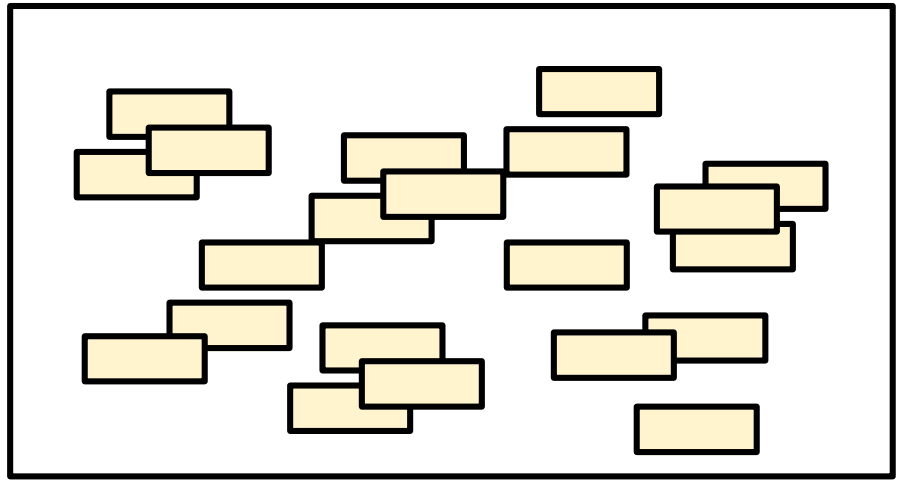
Random Assignment

Random assignment is an impartial procedure that uses chance.



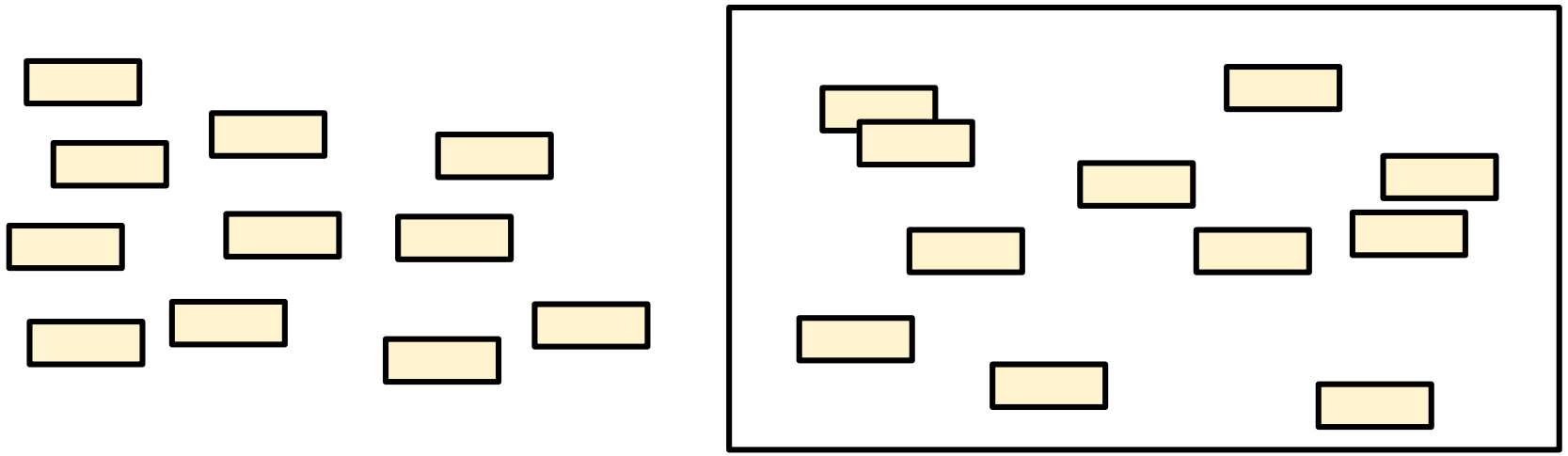
Random Assignment

Random assignment is an impartial procedure that uses chance.



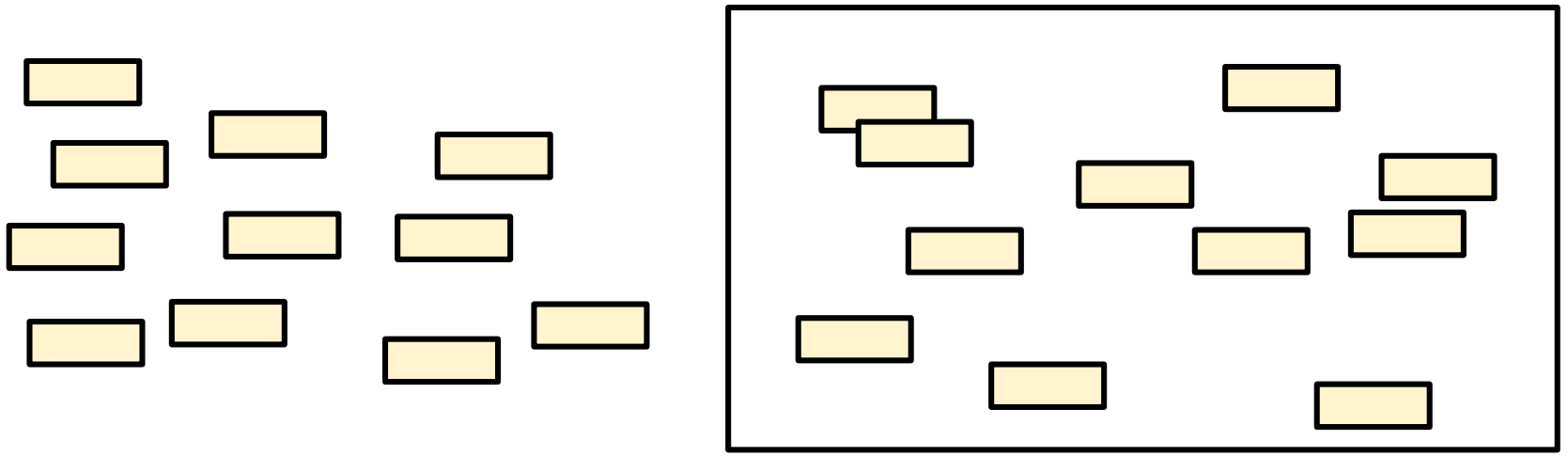
Random Assignment

Random assignment is an impartial procedure that uses chance.



Random Assignment

Random assignment is an impartial procedure that uses chance.

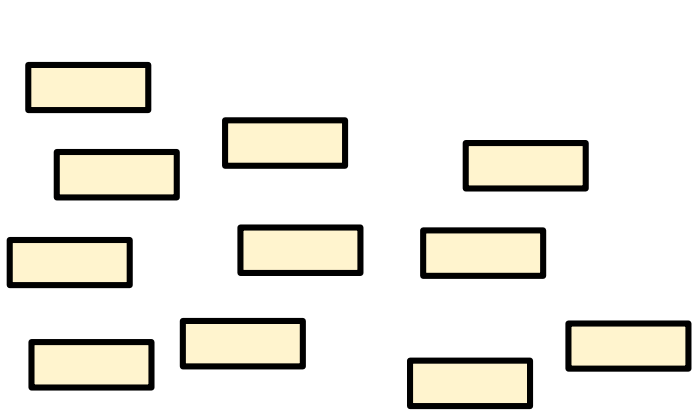


Treatment Group **Control Group**

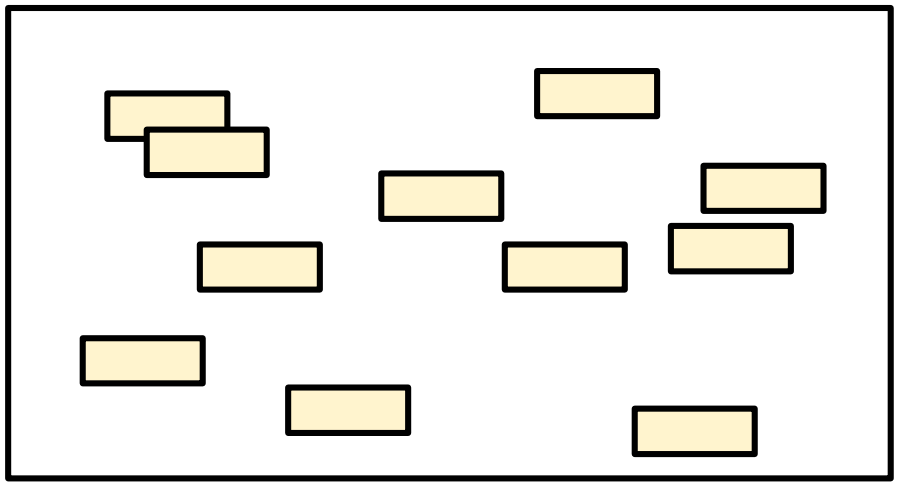
Random Assignment

Random assignment is an impartial procedure that uses chance.

If the number of subjects is large, by the laws of probability, the treatment and control groups will tend to be similar in all aspects.



Treatment Group



Control Group

Random Assignment

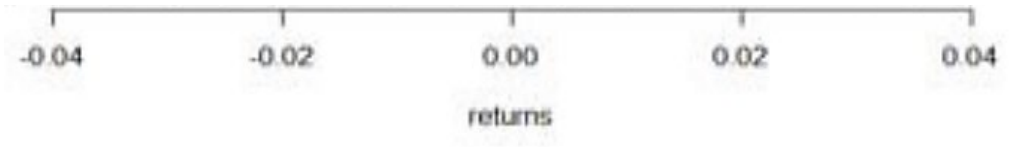
The goal of random assignment is to create similar treatment and control groups with respect to

- Revision time
- IQ
- Age
- Other variables

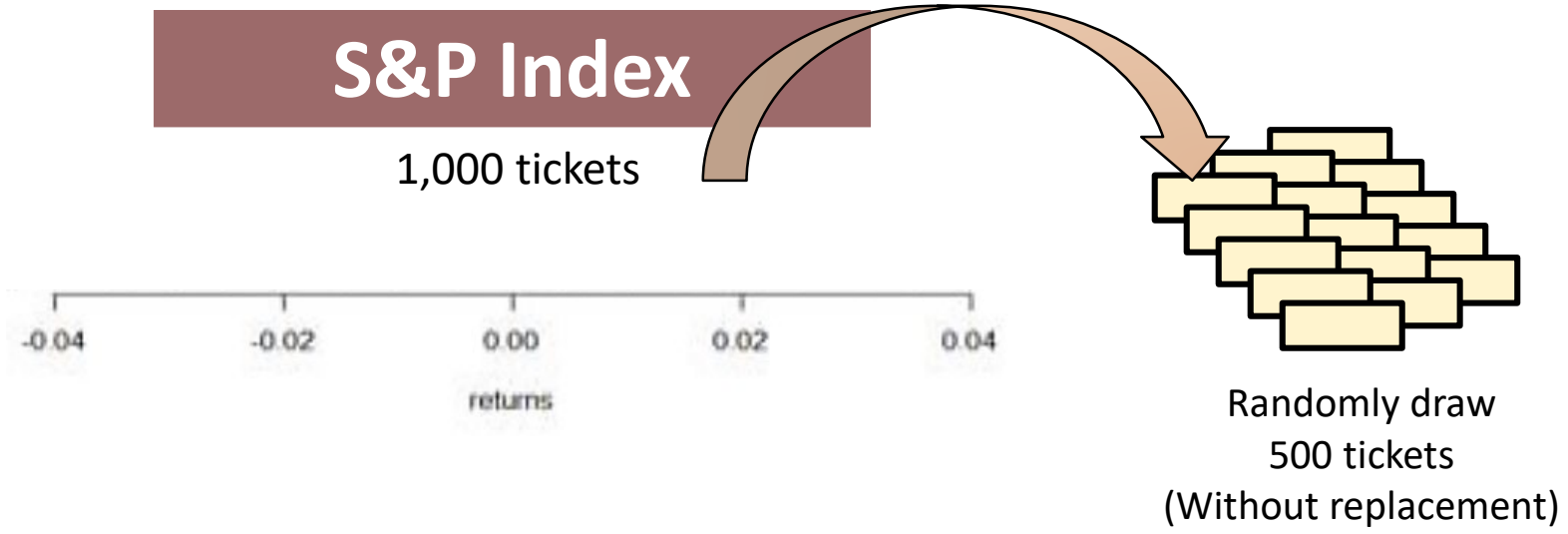
Random Assignment – S&P Example

S&P Index

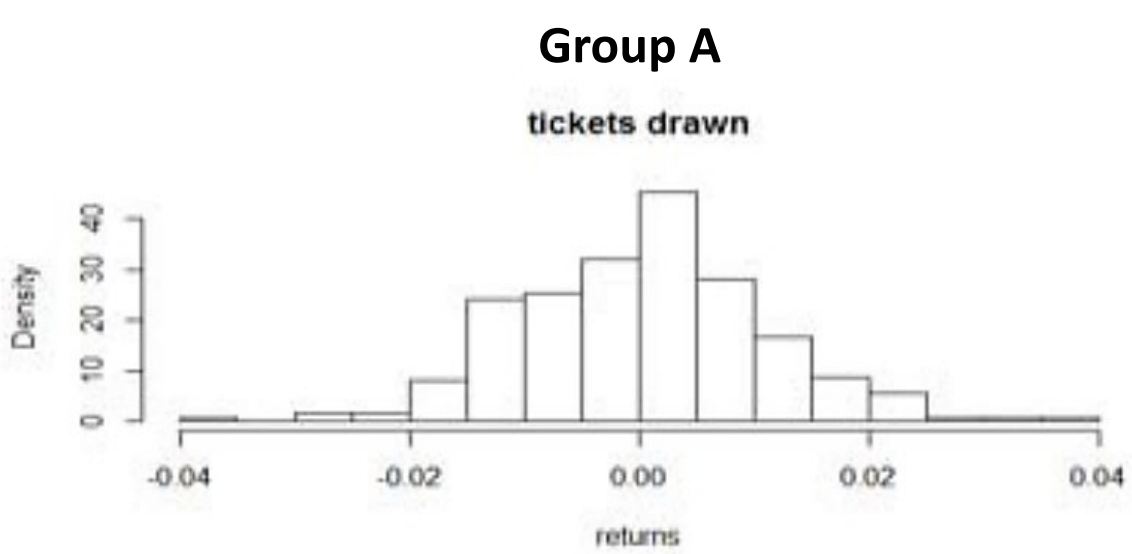
1,000 tickets

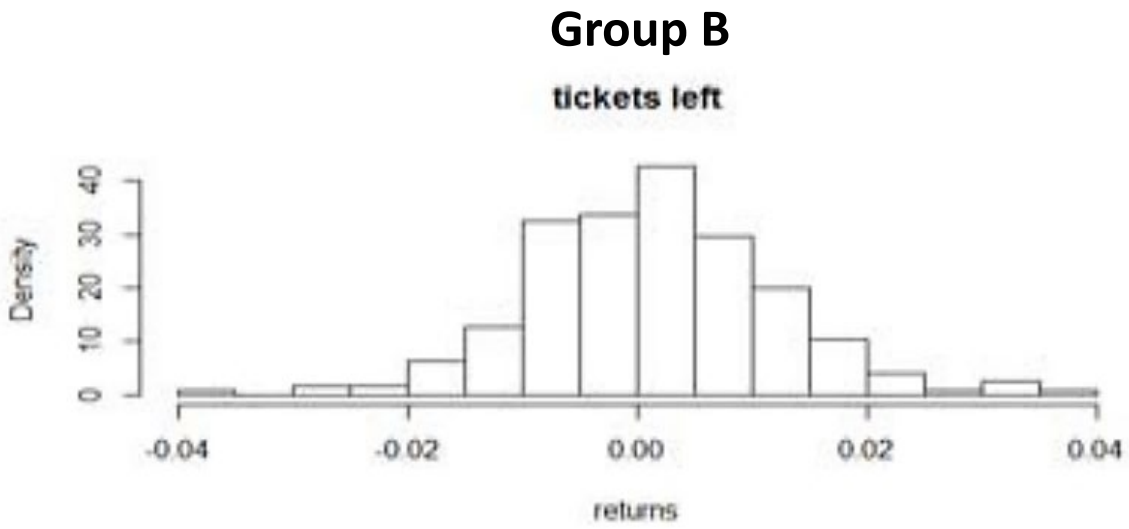


Random Assignment – S&P Example



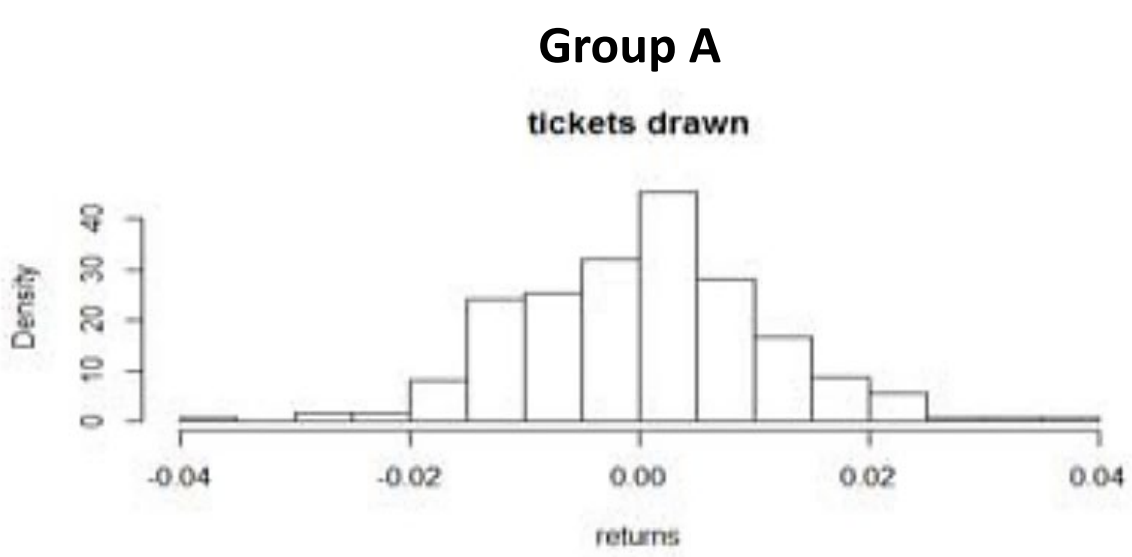
Random Assignment – S&P Example

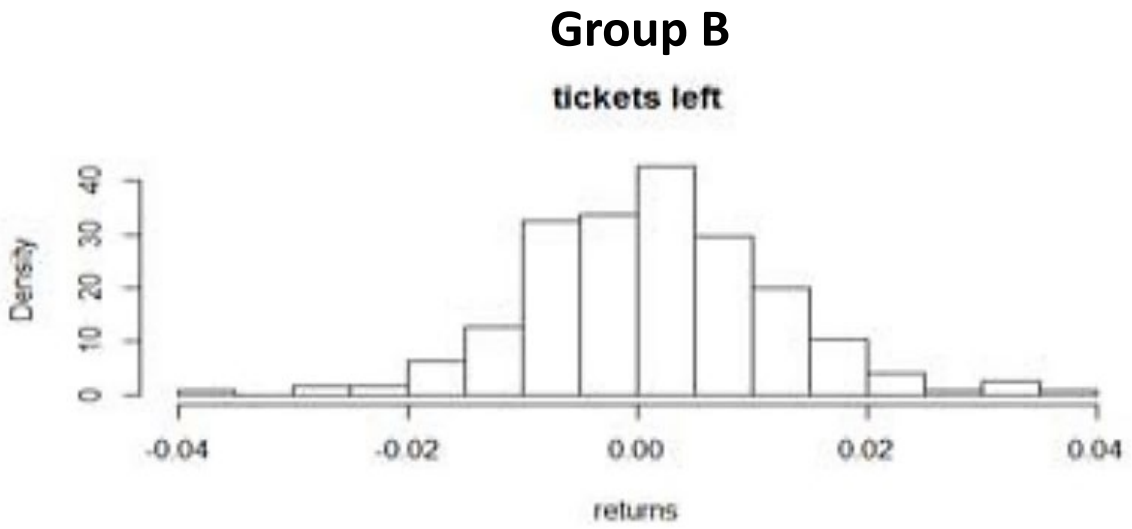




Similar distributions

Random Assignment – S&P Example





The treatment and control groups can have different sizes.
As long as the size of the groups are quite large, then a randomised assignment tends to produce two very similar groups.

Note on the term “Random”

“Random” has a much more strict meaning related to an impartial chance mechanism.

- “Random” does not mean “Haphazard”.

Study Designs
- Experimental Studies (Blinding)

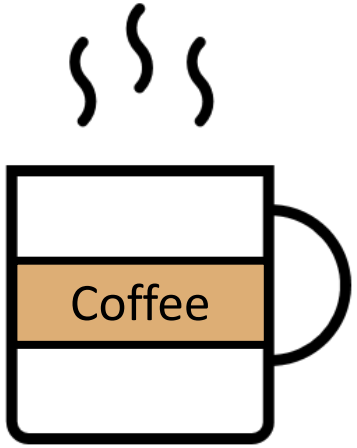
Experiments

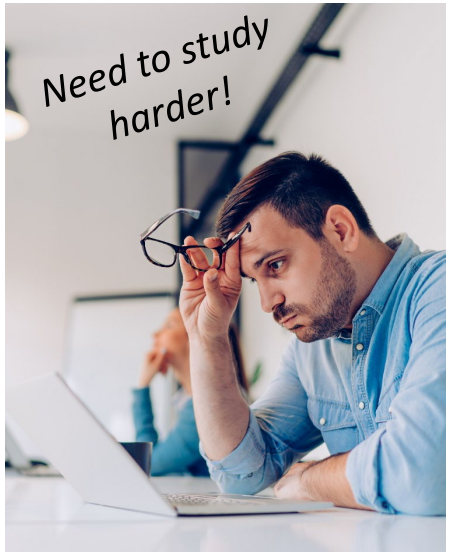
Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*
Treatment Group	**Control Group**

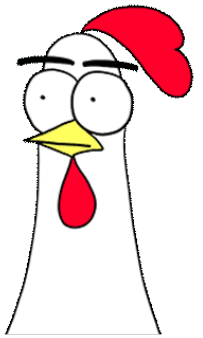
In some cases, leaving the control group alone may cause bias!

Experiments

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*

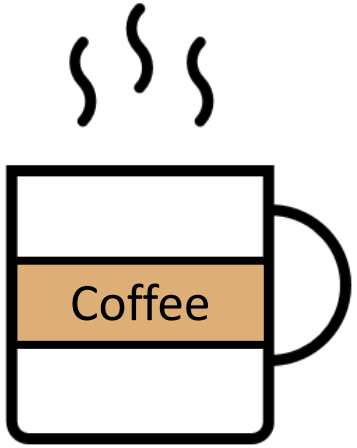


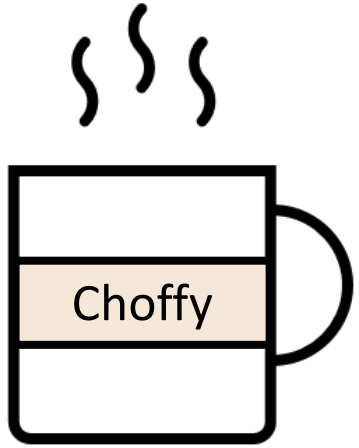




Experiments

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*

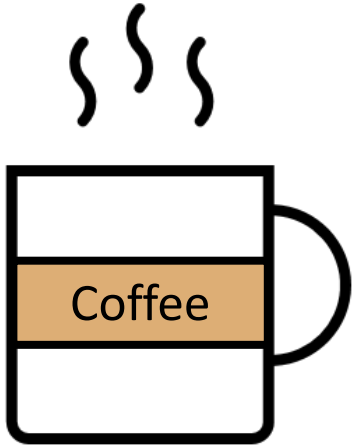


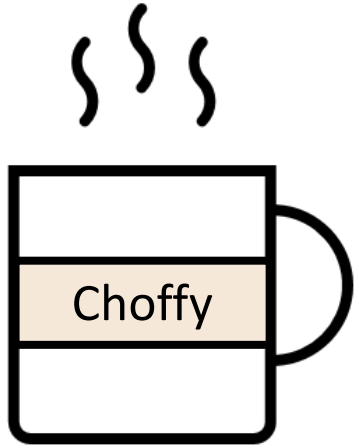


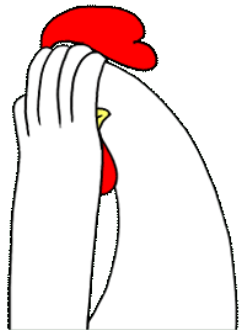
Experiments - Placebo

Placebo: Treatment with no active ingredients, and no effect.

Placebo Effect: The response observed when subjects receive a placebo treatment, but still show some positive effects.







Blinding

Blinded subjects do not know whether they are in the treatment or control group.

- A placebo that is very similar to the treatment can be chosen to help make the blinding effective.
- The subjects are blind to the treatment to prevent their own beliefs about the treatment from affecting the results.

Blinding

Blinded assessors do not know whether they are assessing the treatment or control group.

Double-Blinding

An experiment is called double-blind if both **subjects** and **assessors** are blinded about the assignment.

Study Designs
- Experimental vs Observational Studies

Experiments

Do vaccinations help reduce the effects of the coronavirus?

NEWS | 20 October 2020

Dozens to be deliberately infected with coronavirus in UK 'human challenge' trials

Proponents of the trials say they can be run safely and help to identify effective vaccines, but others have questioned their value.

("Dozens to be deliberately infected with coronavirus in UK 'human challenge' trials," 2020)

Ethical issues

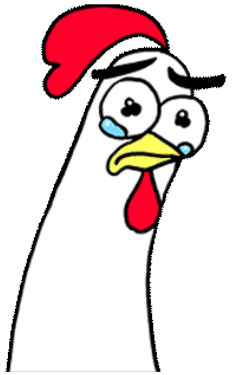
Experiments are useful in providing evidence for a cause-and-effect relationship.

However, an experiment has its issues too.

Injecting low dosages of the virus strain into humans

Deciding which subject to assign to the vaccine treatment and control group

Consent given at the start of the study



Types of Study Designs

Experimental

Observational

Observational Studies

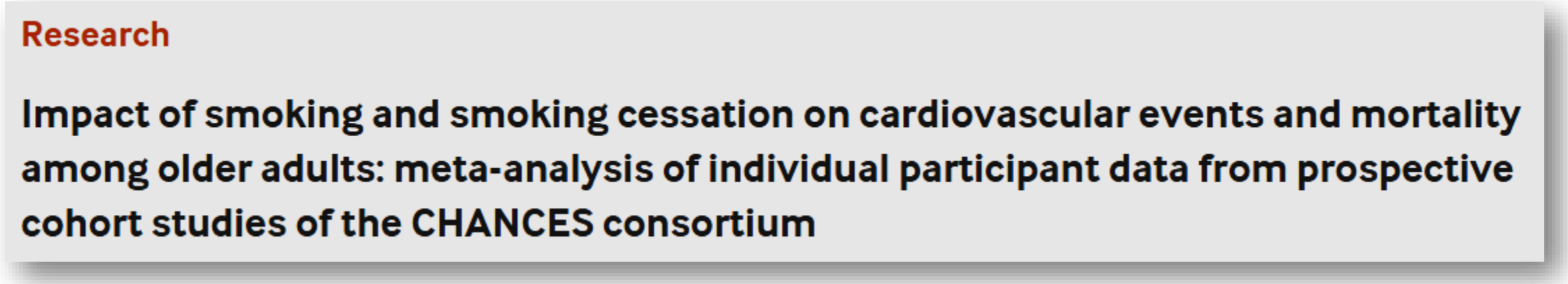
An observational study observes individuals and measures variables of interest.

However, researchers do not attempt to directly manipulate one variable to cause an effect in another variable.

- Does not provide convincing evidence of a cause-and-effect relationship.

Observational Studies

Is long term smoking linked to heart disease?



("Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: Meta-analysis of individual participant data from prospective cohort studies of the chances consortium," 2015)

Observational Studies

Is long term smoking linked to heart disease?

Note: For observational studies, while there is no actual treatment being assigned to the subjects, we still use the terms “treatment” and “control” groups in the same way as though we are dealing with an experiment.

Smokers	Non-smokers
Treatment Group (Exposure)	**Control Group** (Non-exposure)

Experiment vs Observational Study

Experimental Studies	Observational Studies
Assigned by researcher	Assigned by subjects themselves

Experiment vs Observational Study

Experimental Studies	Observational Studies
Assigned by researcher	Assigned by subjects themselves
Can provide evidence of a cause-and-effect relationship	Cannot provide evidence of a cause-and-effect relationship

Our study corroborates and expands evidence from previous studies in showing that smoking is a strong independent risk factor of cardiovascular events and mortality …

Experiment vs Observational Study

Experimental Studies	Observational Studies
Assigned by researcher	Assigned by subjects themselves
Can provide evidence of a cause-and-effect relationship	Can provide evidence of ‘Association’

Final note - Generalisability

If an experiment is well-designed, can we generalise the results?

Final note - Generalisability

If an experiment is well-designed, can we generalise the results?

Sampling frame?
Sampling method?
Sample size?

<u>The design of the experiment is not the only factor we look at.</u>

Summary

- Experimental Studies
  - Treatment and Control Groups
  - Random Assignment
  - Blinding
- Observational Studies
- Experimental vs Observational Studies