Chapter 1

Exploratory Data Analysis and Design of Experiments

Section 1.1 Exploratory Data Analysis

Discussion 1.1.1 Data exists in our everyday life. As we flip through our newspapers each day, we see evidence of data being used and many questions being asked about data that has been collected. In other words, we see that research is becoming data driven and it is fast becoming necessary for one to be proficient in reasoning quantitatively. The ability to investigate and make sense of a data set is a core 21st century skill that any undergraduate, regardless of discipline should acquire.

An online article in 2021 shows the following:

(Source: https://www.todayonline.com/singapore/fall-singapore-marriages-divorces-2020-amid-covid-19-restrictions-uncertainty)

After reading the article, it is natural for one to ask questions on how the conclusion was arrived at. What kind of data was collected that supported this conclusion? Is the conclusion made correctly?

Definition 1.1.2 A *population* is the entire group (of individuals or objects) that we wish to know something about.

Definition 1.1.3 A *research question* is usually one that seeks to investigate some characteristic of a population.

Example 1.1.4 The following are some examples of research questions.

1. What is the average number of hours that students study each week?

2. Does the majority of students qualify for student loans?

3. Are student athletes more likely than non-athletes to do final year projects?

Broadly speaking, we can classify research questions into the following categories.

1. To make an estimate about the population.

2. To test a claim about the population.

3. To compare two sub-populations / to investigate a relationship between two variables in the population.

Example 1.1.5 Having a well designed research question is a critical beginning to any data driven research problem. While an in-depth discussion on *how* research questions can be designed is beyond the scope of this course, the following table gives a few examples and provides some insights into what are some considerations and desirable features that good research questions should have.

Considerations	Example of a neutral research question	Example of a better research question	Explanation
Narrow vs. Less Narrow	Q1: Do Primary Six students have an average sleep time of 7 hours a day?	Q2: Do Primary Six students have an average sleep time of 7 hours a day? What are some variables that may play a part in affecting the number of hours they sleep?	Q1 is too narrow as it can be answered with a simple statistic. It does not look at any other context surrounding the issue. Q2 is less narrow and attempts to go beyond simply finding some data or numbers. It seeks to understand the bigger picture too.
Unfocussed vs. Focussed	Q1: What are the effects of eating more than 2 meals of fast food per week?	Q2: How does eating more than 2 meals of fast food per week affect the BMI (Body Mass Index) of children between 10 to 12 years old in Singapore?	Q1 is too broad which makes it difficult to identify a research methodology. Q2 is focussed and clear on what data to be collected and analysed.
Simple vs. Complex	Q1: How are schools in Singapore addressing the issue of mental health among school children?	Q2: What are the effects of intervention programs implemented at schools in Singapore on the mental health among school children aged 13 to 16?	Q1 is simple and such information can be obtained with a search online with no analysis required. Q2 is more complex and requires both investigation and evaluation which may lead the research to form an argument.

We will now proceed to describe the process of Exploratory Data Analysis (EDA).

Definition 1.1.6 *Exploratory Data Analysis* (EDA) is a systematic process where we explore a data set and its variables and come up with summary statistics as well as plots. EDA is usually done iteratively until we find useful information that helps us answer the questions we have about the data set.

In general, the steps involved in EDA are

1. Generate research questions about the data.
2. Search for answers to the research questions using data visualisation tools. In the process of exploration, we could also perform data modelling (e.g. regression analysis).
3. We ask ourselves the following question: To what extent does the data we have, answer the questions we are interested in?
4. We refine our existing questions or generate new questions about the data before going back to the data for further exploration.

Section 1.2 Sampling

Definition 1.2.1 A *population* of interest refers to a group in which we have interest in drawing conclusions on in a study.

Definition 1.2.2 A *population parameter* is a numerical fact about a population.

Example 1.2.3 The following are some examples of a population and an associated population parameter.

1. The average height (population parameter) of all primary six students in a particular primary school (population).
2. The median number of modules taken (population parameter) by all first year undergraduates in a University (population).
3. The standard deviation of the number of hours spent on mobile games (population parameter) by pre-schoolers aged 4 to 6 in Singapore (population).

Definition 1.2.4

1. It is usually not feasible to gather information from every member of the population, so we look at a *sample*, which is a proportion of the population selected in the study.

2. Without the information from every member of the population, we will not be able to know exactly what is the population parameter. The hope is that the sample will be able to give us a reasonably good *estimate* about the population parameter. An *Estimate* is an inference about the population's parameter based on the information obtained from a sample.

3. A *sampling frame* is the list from which the sample was obtained.

Remark 1.2.5

1. Suppose the population of interest are people who drink coffee in Singapore. How should we design a sampling frame for this population? The sampling frame may or may not cover the entire population or it may contain units not in the population of interest. The all important question is whether the sample obtained from such a sampling frame is still able to tell us something about the population parameter. The following are some of the characteristics of the sampling frame that we should pay attention to:

    - Does the sampling frame include all available sampling units from the population?
    - Does the sampling frame contain irrelevant or extraneous sampling units from another population?
    - Does the sampling frame contain duplicated sampling units?
    - Does the sampling frame contain sampling units in clusters?

2. One of the conditions of *generalisability*, which is the ability to generalise the findings from a sample to the population is that the sampling frame must be equal to or greater than the population of interest. Note that this does not mean that when our sampling frame covers the entire population of interest, our findings from the sample will always be generalisable to the population. It is still an important question to know **how** the sample was collected. (See Remark 1.2.17 for more information on the criteria for generalisability.)

Definition 1.2.6 A *census* is an attempt to reach out to the entire population of interest while a sample is a proportion of the population.

While it is obviously nice to have a census, this is often not possible due to the high cost of conducting a census. In addition, some studies are time sensitive and a census typically takes a long time to complete, even when it is possible to do so. Furthermore, in a census attempt, one may not be able to achieve 100% response rate.

Definition 1.2.7 When we sample from a population, we must try to avoid introducing *bias* into our sample. A biased sample will almost surely mean that our conclusion from

the sample cannot be generalised to the population of interest. There are two major kinds of biases.

1. *Selection bias* is associated with the researcher's biased selection of units into the sample. This can be caused by imperfect sampling frame, which excluded units from being selected. Selection bias can also be caused by *non-probability sampling* (see Definition 1.2.15 and Example 1.2.16).

2. *Non-response bias* is associated with the participants' non-disclosure or non-participation in the research study. This results in the exclusion of information from this group. There can be various reasons for non-response, for example, inconvenience or unwillingness to disclose sensitive information. Note that non-response bias may occur regardless of whether the sampling method is probabilistic or non-probabilistic in nature.

Example 1.2.8

1. Suppose we would like to study the number of modules taken by all first year undergraduates in a University. To collect a sample, the researcher went to two different lecture theatres to survey undergraduates who were taking two different first year Engineering foundation (compulsory) modules. The sampling frame in this case consists of all undergraduates who were registered in the two modules in the semester. Undergraduates who are not taking either of the two modules will not have a chance to be sampled and thus the sampling frame is imperfect, leading to selection bias.

2. Suppose we would like to find out the proportion of students living at a boarding school who have received some form of financial assistance in the past and if they had received financial assistance, what was the quantum they received. A questionnaire was distributed to all students via a survey form slipped under their room doors and instructions were given to them to complete the form and drop it in a collection box if they had received financial assistance before. Students do not need to return the form if they had not received any form of financial assistance previously. The data collected from this is likely to be biased due to non-response as students who actually had received financial assistance in the past may be reluctant to share this information or be seen by their friends when they have to drop the form at the collection box. This will likely result in an underestimate of the proportion of students who had received financial assistance.

Definition 1.2.9 *Probability sampling* is a sampling scheme such that the selection process is done via a known randomised mechanism. It is important that every unit in the sampling frame has a **known** non-zero probability of being selected but the probability of being selected does not have to be same for all the units. The randomised mechanism is important as it introduces an element of chance in the selection process so as to eliminate biases.

We will introduce four main types of probability sampling methods.

1. *Simple random sampling* (SRS) - this happens when units are selected randomly from the sampling frame. More specifically, a simple random sample of size n consists of n units from the population chosen in such a way that every set of n units has an equal chance to be the sample actually selected. We are referring to *sampling without replacement* here, where a unit chosen in the sample is removed and has no chance of being chosen again into the same sample. A useful way to perform simple random sampling is to use a random number generator. While it is expected that different samples sampled from the same sampling frame using SRS would be different, the variability between the samples is entirely due to chance.

   Example 1.2.10 The classic lucky draw that is carried out during dinners is the best example of simple random sampling. In this case, every attendee has his/her lucky draw ticket placed inside a box and a simple random sample of these tickets are drawn out of the box, one at a time, without replacement. If we assume that before each draw, the remaining tickets in the box are mixed properly such that every ticket has a equally likely chance of being drawn out, then the probability of each ticket being drawn at any instance is $\frac{1}{n}$ where n is the number of tickets remaining inside the box.

   Example 1.2.11 Suppose we would like to sample 500 households in Singapore and find out how many household members there are in each household. Let us assume that every household has a unique home phone number. If we have a listing of all such phone numbers and list them from 1 to n, we can use a random number generator to select 500 phone numbers from the list to form our sample. Unique phone calls (i.e. sampling without replacement) can then be made to these households to survey the number of household members. This is another example of simple random sampling. Notice that this example also illustrates a common shortcoming of SRS, in that it can possibly be subjected to non-response from the units that are sampled.

2. *Systematic sampling* is a method of selecting units from a list by applying a selection interval k and a random starting point from the first interval. To carry out systematic sampling:

   (a) Suppose we know how many sampling units there are in the population (denoted by n);

   (b) We decide how big we want our sample to be (denoted by k). This means that we will select one unit from every $\frac{n}{k}$ units;

   (c) from 1 to $\frac{n}{k}$, select a number **at random**, say r;

   With this, the sample will consist of the following units from the list:

$$r, \quad r + \frac{n}{k}, \quad r + \frac{2n}{k}, \quad \cdots, r + \frac{(k-1)n}{k}.$$

However, it is often that we do not know the number of sampling units n in the population. In such a situation, systematic sampling can still be done by deciding on the selection interval k and **randomly selecting** a unit from the first k units and then subsequently every kth unit will be sampled. For example, if $k = 10$, we can sample the 5^{th}, 15^{th}, 25^{th} units and so on.

Compared to simple random sampling, systematic sampling is a simpler sampling process as we do not need to know how many sampling units there are exactly. On the other hand, if the listing is not random, but instead contains some inherent grouping or ordering of the units, then it is possible that a sample produced by systematic sampling may not be representative of the population.

Example 1.2.12 Suppose we know there are 110 sampling units in the population (so $n = 110$) and we would like to select a sample with 10 units (so $k = 10$). Imagine the sampling units are numbered 1 to 110 in a list and arranged according to the table below.

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

Since $n = 110$ and $k = 10$, we select one unit from every $\frac{110}{10} = 11$ units. So we randomly select a number from 1 to 11 which will start off the sampling process. For example, if the number selected was 5, then our sample will comprise of the elements

$$5, 16, 27, 38, 49, 60, 71, 82, 93, 104.$$

Similarly, if the number selected was 9, then our sample will comprise of the elements

$$9, 20, 31, 42, 53, 64, 75, 86, 97, 108.$$

From this example, it should be clear that if the sampling units are listed with some inherent pattern, then it is possible that the sample obtained could have selection bias.

3. *Stratified sampling* is a method where the population is divided into groups called strata. Each stratum is similar in that they share similar characteristics but the size of each stratum does not necessarily have to be the same. We then apply simple random sampling to each stratum to generate the overall sample. While stratified sampling is a commonly used probability sampling method, there are some situations where it may not be possible to have information on the sampling frame of each stratum in order to perform simple random sampling properly. Furthermore, depending on how the strata are defined, we may face ambiguity in determining which stratum a particular unit belongs to. This can complicate the sampling process.

   Example 1.2.13 An example of stratified sampling can be seen during elections, for example, a Presidential Election. Voters visit their designated polling stations to cast their votes for the candidate that they wish to support. In countries where the number of voters is very large, it may take a long time before all the votes are counted. Stratified sampling can be employed if we wish to make a reasonably good prediction of the outcome. This is done by taking a simple random sample of the voters at each polling station (stratum) and then computing the *weighted average* of the overall vote count, based on the size of each stratum, for each candidate. This way, we would be able to have a reasonably good estimate of the total votes each candidate would receive.

4. *Cluster sampling* - is a method where the population is divided into clusters. A fixed number of clusters are then selected using simple random sampling. All the units from the selected clusters are then included in the overall sample. One advantage of this sampling method is that it is usually simpler, less costly and not as resource intensive than other probability sampling methods. The clusters are usually naturally defined which makes it easy to determine which cluster a unit belongs to. The main disadvantage of this sampling method is that depending on which clusters are selected, we may see high variability in the overall sample if there are largely dissimilar clusters with distinct characteristics. In addition, if the number of clusters sampled is small, there is also a risk that the clusters selected will not be representative of the population.

   Example 1.2.14 Suppose a study wants to survey the mental wellness of Primary school students in Singapore. Cluster sampling can be done by treating each Primary school as a cluster and this way of clustering the population of interest is natural and unambiguous since all students in the population belongs to exactly one Primary school. A number of schools are then selected using simple random sampling for this survey and all the students in the selected schools will be part of the sample while those not in the selected schools will not be included. Another approach is of course to apply simple random sampling with the list of all students (from all Primary schools) as the sampling units. If this was done, then there is a possibility that all schools will

have students forming part of the sample. Cluster sampling would not provide such a characteristic.

We have presented four different probability sampling methods, below is a summary table of the advantages and disadvantages of the methods.

Sampling Plan	Advantages	Disadvantages
Simple Random Sampling	Good representation of the population	Time-consuming; accessibility of information and sampling frame
Systematic Sampling	Simple selection process as opposed to simple random sampling	Potentially under-representing the population
Stratified Sampling	Good representation of the sample by stratum	Require sampling frame and criteria for classification of the population into stratum
Cluster Sampling	Less time-consuming and less costly	Require clusters to be reasonably heterogeneous and not have cluster-specific characteristics

Remember:

There is no single universally best probability sampling method as each has its advantages and disadvantages. All probability sampling methods can produce samples that are representative of the population (that is, sample is unbiased). However, depending on the situation, some methods would further reduce the variability, resulting in a more precise sample.

Definition 1.2.15 A *non-probabilty sampling* method is when the selection of units is not done by randomisation. There is no element of chance in determining which units are selected, instead it is usually down to human discretion.

Example 1.2.16

1. *Convenience sampling* is a non-probability sampling method where a researcher chooses subjects to form a sample among those that are most easily available to participate in the study. A common occurrence of convenience sampling is at shopping malls where surveyors approach shoppers at a location convenient to them. Such a sampling method introduces selection bias since malls are frequently visited by those who are more affluent. Other demographics of the population may be left out. Another issue that may arise from convenience sampling done at shopping malls is non-response bias

as shoppers may not want to be stopped for questionnaires as they feel it is time consuming and not what they are meant to be doing in a mall.

2. *Volunteer sampling* happens when subjects volunteer themselves into a sample. Such a sample is also known as a *self-selected* sample and very often, the sample contains subjects who have a strong opinion (either positive or negative) on the research question than the rest of the population. Such a sample is unlikely to be representative of the population of interest. For example, the host of a "popular" radio talkshow may wish to find out how well received is his show. To do this, he asked his listeners to go online and submit a rating of this show, out of a score of 10. Each listener can voluntarily decide if they wish to be part of this rating exercise or not. By collecting a sample of opinions this way, it is likely that the sample will be skewed towards a high rating because listeners who did not like the talkshow would not even be aware of such a survey and therefore their opinion would have been left out. On the other hand, listeners who are strong supporters of this show would be more enthusiastic to go online to support their favourite radio show.

Let us summarise our discussion on sampling. In most instances where a census is not possible, obtaining a sample of the population of interest is necessary. The following outlines the general approach to sampling:

1. To design a sampling frame. Recall that a sampling frame should ideally contain the population of interest so that every unit in the population has a chance to be sampled.
2. Decide on the most appropriate sampling method to generate a sample from the sampling frame. Probability sampling methods are generally preferred over non-probability sampling methods as non-probability sampling methods have a tendency to generate a biased sample.
3. Remove unwanted units (those that are not from the population) from the generated sample.

Remark 1.2.17 If the following *generalisability criteria* can be met, we will be more confident in generalising the conclusion from the sample to the population.

1. Have a good sampling frame that is equal to or larger than the population;
2. Adopt a probability-based sampling method to minimise selection bias;
3. Have a large sample size to reduce the variability or random errors in the sample;
4. Minimise the non-response rate.

Section 1.3 Variables and Summary Statistics

Definition 1.3.1

1. A *variable* is an attribute that can be measured or labelled.
2. A *data set* is a collection of individuals and variables pertaining to the individuals. Individuals can refer to either objects or people.

In a research question where we are examining relationships between variables, there is usually a distinction between which are *independent* and which are *dependent* variables.

Definition 1.3.2

1. Independent variables are those that may be subjected to adjustments, either deliberately or spontaneously, in a study.
2. Dependent variables are those that are hypothesised to change depending on how the independent variable is adjusted in the study.

It is important to note that the dependent variable is **hypothesised to change** when the independent variable is adjusted. It does not mean that the dependent variable **must** change. It is perfectly possible that any changes to the independent variable does not result in any change in the dependent variable.

Example 1.3.3

1. In a study, if we wish to investigate the relationship between time spent on computer gaming and examination scores, the independent variable is the amount of time one spends on computer gaming while the dependent variable is the examination score.
2. In a study where we investigate which brand of tissue paper is able to absorb the most water, the independent variable is the brand of the tissue paper and the dependent variable is the amount of water a piece of tissue paper (from a particular brand) can absorb. In this study, we will vary the different brands of tissue paper used and record the different amounts of water absorbed.
3. We would like to study whether drinking at least 2 glasses of orange juice per day for a year is associated [1] with having lower cholesterol levels in a year's time. In this case, the independent variable is whether (or not) a person drinks at least 2 glasses of orange juice a day. Each individual will have an attribute labelled either as "YES" or "NO" with regards to this variable. The dependent variable would be whether

[1] The notion of association between variables will be discussed extensively in Chapter 2.

an individual's cholesterol level next year is lower than this year's level. Again, each individual will have an attribute labelled either as "YES" or "NO" with regards to this variable.

Definition 1.3.4

1. *Categorial variables* are those variables that take on categories or label values. The categories or labels are mutually exclusive, meaning that an observation cannot be placed in the same category or given two different labels at the same time.

2. *Numerical variables* are those variables that take on numerical values and we are able to meaningfully perform arithmetic operations like adding and taking average.

3. Among categorical variables, there are generally two sub-types. An *ordinal* variable is a categorical variable where there is some natural ordering and numbers can be used to represent the ordering. A *nominal* variable is a categorial variable where there is no intrinsic ordering.

4. Among numerical variables, there are also generally two sub-types. A *discrete* numerical variable is one where there are gaps in the set of possible numbers taken on by the variable.

5. A *continuous* numerical variable is one that can take on all possible numerical values in a given range or interval.

Example 1.3.5

1. The happiness index used to measure how happy a group of Secondary school students are, is an ordinal variable. For instance, we can specify "1" as "not happy", "2" as 'somewhat not happy', "3" as neutral, "4" as "somewhat happy" and "5" as "happy". Whether a subject drinks at least 2 glasses of orange juice or not is an example of a nominal variable.

2. The number of children in the school who scored an A grade in Mathematics for PSLE is a discrete numerical variable. In this case, the gaps are the non integer values that lie between every two integer values. It is clear that we cannot have, for example, 134.5 children scoring A in the school, so this is a gap between 134 and 135.

3. The height or the weight of a person is a continuous numerical variable, as the weight can take on all numerical values, not necessarily only the integer values.

A common way of presenting data is to use a table with rows and columns. Each row of the table usually gives information pertaining to a particular individual while each column is a variable. So if we look across a row in the table, we will see the variables' information for that particular individual.

Penguins

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Chinstrap	Dream	46.9	16.6	192	2700	female	2008
Adelie	Biscoe	36.5	16.6	181	2850	female	2008
Adelie	Biscoe	36.4	17.1	184	2850	female	2008
Adelie	Biscoe	34.5	18.1	187	2900	female	2008
Adelie	Dream	33.1	16.1	178	2900	female	2008
Adelie	Torgersen	38.6	17	188	2900	female	2009
Chinstrap	Dream	43.2	16.6	187	2900	female	2007

The table above shows part of a data set involving different species of penguins and some of the physical attributes of the penguins. Each row represents a particular penguin and the columns are the variables pertaining to that particular penguin. Some of the variables are categorical variables while others are numerical. Can you figure out whether the categorial variables are ordinal or nominal? Can you figure out whether the numerical variables are discrete or continuous?

With a data set, we are able to zoom into a particular individual's information at a micro level. If we do this, we can extract all the information on that particular individual for our use. However, we may also be interested in looking at the entire data set at the macro level, obtaining information on groups of individuals or the entire population. Useful information like trends and patterns can be observed from the data through data visualisation, which is very useful. While calculations cannot be done through visualisations, we can use *summary statistics* to do numerical and quantitative comparisons between groups of data.

Summary statistics for numerical variables can be broadly classified into two types. Firstly, there are those that measure the central tendencies of the data, like *mean*, *median* and *mode*. Secondly, there are those that measure the level of dispersion (or spread) of the data, like *standard deviation* and *interquartile range*.

Section 1.4 Summary Statistics - Mean

Definition 1.4.1 The *mean* is simply the average value of a numerical variable x. We denote the mean of x by $\overline{x}$ and the formula to compute $\overline{x}$ is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Here, n is the number of data points and $x_1, x_2, \ldots, x_n$ are the numerical values of the numerical variable x in the data set.

Example 1.4.2 Suppose the bill length (in mm) of 7 penguins were

$$46.9, \quad 36.5, \quad 36.4, \quad 34.5, \quad 33.1, \quad 38.6, \quad 43.2.$$

Then the mean bill length is

$$\frac{46.9 + 36.5 + 36.4 + 34.5 + 33.1 + 38.6 + 43.2}{7}$$

which is approximately 38.46 (rounded to 2 decimal places).

Remark 1.4.3 These are some properties of the mean of a variable.

1. $x_1 + x_2 + \cdots + x_n = n\overline{x}$. This means that we may not know each of the individual values $x_1, x_2, \ldots, x_n$, but we can calculate their sum if we know their mean ($\overline{x}$) and the number of data points (n) that is used to compute the mean.

2. Adding a constant value c to all the data points changes the mean by that constant value. So if the mean of the values $x_1, x_2, \ldots, x_n$ is $\overline{x}$, then the mean of

   $$x_1 + c, \quad x_2 + c, \quad \ldots, \quad x_n + c$$

   will be $\overline{x} + c$. For example, the mean of $1, 6, 8$ is $\frac{1}{3}(1 + 6 + 8) = 5$ and the mean of $(1 + 3), (6 + 3), (8 + 3)$ (adding 3 to each of the 3 numbers $1, 6$ and 8) is

   $$\frac{(1 + 3) + (6 + 3) + (8 + 3)}{3} = \frac{4 + 9 + 11}{3} = 8 = 5 + 3.$$

3. Multiplying a constant value of c to all the data points will result in the mean being changed by the same factor of c. So if the mean of the values $x_1, x_2, \ldots, x_n$ is $\overline{x}$, then the mean of

   $$cx_1, cx_2, \ldots, cx_n$$

   will be $c\overline{x}$. For example, the mean of $2, 7, 12$ is $\frac{1}{3}(2 + 7 + 12) = 7$ and the mean of $(2 \times 2), (2 \times 7), (2 \times 12)$ (multiplying 2 to each of the 3 numbers $2, 7$ and 12) is

   $$\frac{(2 \times 2) + (2 \times 7) + (2 \times 12)}{3} = \frac{42}{3} = 14 = 2 \times 7.$$

We will now look at several examples of means in real life.

Example 1.4.4 Consider a data set where we have daily weather data, collected at various weather stations in Singapore. Part of the data set is shown below.

Station	Year	Month	Day	Daily Rainfall Total (mm)	Mean Temperature (degree C)	Mean Wind Speed (km per h)
Admiralty	2020	1	1	0	27.5	22
Admiralty	2020	1	2	0	27.4	20.2
Admiralty	2020	1	3	0.2	27.5	22.7
Admiralty	2020	1	4	7	26.7	20.9
Admiralty	2020	1	5	0	27.6	22.3

With this data set, some of the questions that we can ask are

1. Which month in 2020 had the most amount of rainfall?
2. If the mean monthly rainfall in 2020 was 157.22mm, what was the total amount of rainfall recorded in 2020?
3. Is there any relationship between wind speed and temperature? What about between the amount of rainfall and wind speed?
4. Does the weather pattern for 2020 allow us to make a good prediction for how the weather will be like in 2021?

To answer the first question on the month with the most amount of rainfall, we need to add up the amount of rainfall recorded on each day of a month, for every month in the year in order to do a comparison. To answer the second question, using the information on the average rainfall ($\overline{x} = 157.22$), with the fact that

$$12\overline{x} = x_1 + x_2 + \cdots + x_{12},$$

we can find the total rainfall in 2020 to be $12 \times 157.22 = 1866.64$mm. This way, we can find the total rainfall in 2020 without having to add the total amount of rainfall for each of the twelve months. It is also useful to note that if the average rainfall in 2020 was 157.22mm, then

1. It is not possible for the amount of rainfall to be less than (or more than) 157.22mm **every** month in 2020.
2. It is not necessarily the case that the amount of rainfall is 157.22mm **every** month in 2020.
3. In fact, it may not even be the case that there were six (half of twelve) months where the monthly rainfall were higher than the mean and the other six months lower than the mean.

In conclusion, knowing the mean, while useful, does not tell us how the rainfall was distributed over the twelve months of 2020. We would not know which months had more than

the mean and which months had less. In order to have further information beyond the mean, we need to know a bit more about the *spread* of the data. This will be covered later in this chapter.

Example 1.4.5 Suppose students from two different schools (A and B) took a common examination and the table below shows the average performance of the students in both schools.

	No. of students	Average mark
School A	349	32.21
School B	46	30.72
Overall	395	?

The mean score of students in school A was 32.21 and the mean score of students in school B was 30.72. What would be the mean score of all the students in both schools if we consider them altogether? Would it be the simple average

$$\frac{(32.21 + 30.72)}{2} = 31.465?$$

The answer is no and the reason for this is because we do not know how many students in each school contributed to the mean sores recorded in their respective schools. Imagine the extreme case where school A had 500 students who took the examination while school B only had 5. In such a situation, you would expect that the overall average score of the 505 students in both schools to be very close to the mean score of school A. In order to know what is the overall mean for the students in both schools, we need to have the information on the number of students in each school, given below.

	Number of students
School A	349
School B	46

With this information, the overall mean can be computed using the *weighted average* of the two subgroup means. The overall mean for the $349 + 46 = 395$ students would be

$$\frac{349}{395} \times 32.21 + \frac{46}{395} \times 30.72 = 32.04.$$

The numbers $\frac{349}{395}$ and $\frac{46}{395}$ that were multiplied to their respective group means are called the *weights* of the subgroups. Observe that due to the much larger subgroup size of school A compared to that of school B, the overall mean as we expected, is much closer to the mean of school A.

Another useful observation is that the overall mean of 32.04 lies between the two subgroup means of 32.21 and 30.72 (although closer to 32.21). This is not a one-off coincidence. Generally, the overall mean will **always** be between the smallest and largest means among all the subgroups (when we have more than just two subgroups). This will be discussed in greater detail in the next chapter.

Example 1.4.6 In this final example on means, we introduce a related concept known as *proportions*. Suppose we would like to investigate the effectiveness of a new drug for treating asthma attacks compared to existing drugs. The table below shows the number of patients taking the new drug and the number taking the existing drug.

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

Since there are only 200 asthma attacks among those patients taking the new drug, compared to 300 asthma attacks among those taking the existing drug, can we conclude that the new drug is more effective? The answer is no. Notice that the number of patients taking the new drug and those taking the existing drug are vastly different. This means that we should not be simply looking at the *absolute* number of asthma attacks observed in the two groups of patients, but instead consider the *proportion* of patients in each group having asthma attacks. We see that the proportion is higher in the group taking the new drug compared to the group taking the existing drug and this makes us a lot less confident that the new drug is more effective than the existing one.

	New drug	Existing drug
Number of patients	500	1000
Total asthma attacks	200	300
Proportion of patients having asthma attacks	$\frac{200}{500} = 0.4$	$\frac{300}{1000} = 0.3$

The computation of proportion can actually be thought of as a mean in the following way. Imagine that among the 500 patients receiving the new drug, we assign a numerical value of 1 to those who had an asthma attack after the taking the new drug and a numerical value of 0 to those patients who did not have an asthma attack. If we do this, then the mean of these 500 observations of 0s and 1s would be

$$\frac{\overbrace{1+1+\cdots+1}^{200}+\overbrace{0+0+\cdots+0}^{300}}{500} = 0.4,$$

which coincides with what was computed as the proportion for this group of patients having asthma attack. Therefore, proportion can be thought of as a special case of mean.

Section 1.5 Summary Statistics - Variance and Standard Deviation

Definition 1.5.1 Recall that in Example 1.4.4, we saw that knowing the mean of a variable does not tell us about how the data is *distributed* and the *spread* of the data points. *Standard deviation* is one of the ways to measure the spread of the data about the mean. The computation of the standard deviation is done via the computation of the *sample variance* of the data as follows:

$$\text{Sample Variance, Var} = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1};$$

$$\text{Standard Deviation, } s_x = \sqrt{\text{Var}}.$$

Here, $x_1, x_2, \cdots, x_n$ are n observations of the variable x while $\overline{x}$ is the mean.

You may wonder at this point why do we need to compute the **square** of the difference between each observation x_i and the mean $\overline{x}$ before proceeding to sum up these differences for all the n data points? Why can't we compute $(x_i-\overline{x})$ instead of $(x_i-\overline{x})^2$? Consider a set of 5 data points as follows: $x_1=1, x_2=3, x_3=5, x_4=7, x_5=9$. This would result in the mean being $\overline{x}=\frac{1}{5}(1+3+5+7+9)=5$. There is clearly a spread of the data points about the mean value of 5. However, if we were to consider

$$\begin{aligned} & (x_1-\overline{x})+(x_2-\overline{x})+(x_3-\overline{x})+(x_4-\overline{x})+(x_5-\overline{x}) \\ = \;& (1-5)+(3-5)+(5-5)+(7-5)+(9-5)=0; \end{aligned}$$

this would result in the wrong conclusion that there is no variance (and thus no spread) of the data points about the mean. The reason is simply because each data point could be *smaller* or *bigger* than the mean and if the differences $(x_i-\overline{x})$ are not squared, they will cancel out each other like in the example above, giving us the wrong impression that there is no variation or spread among the data points about the mean.

Remark 1.5.2 You may wonder why, in the computation of sample variance, we divide the sum of the squares $(x_i-\overline{x})^2$ by $n-1$ instead of n, since we have n data points and not $n-1$. The reason is because $x_1, x_2, \ldots, x_n$ are assumed to be a sample taken from a population. We are using the variance observed in such a sample to estimate the variance at the population level, which is usually unknown. You can think of dividing by $n-1$ instead of n as a 'correction' to make since our data is only a sample of the population. More detailed discussion on this is beyond the scope of this module.

Example 1.5.3 The highest temperature recorded on the 1st day of every month is shown below:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
30.1	31.1	31.8	32.1	31.9	32.6	33.0	32.4.	32.0	32.5	31.3	29.6

The mean is

$$\frac{30.1+31.1+31.8+32.1+31.9+32.6+33.0+32.4+32.0+32.5+31.3+29.6}{12} = 31.7.$$

The sample variance is

$$\text{Var} = \frac{1}{11}\left((30.1-31.7)^2+(31.1-31.7)^2+\cdots+(31.3-31.7)^2+(29.6-31.7)^2\right) \approx 1.038$$

The standard deviation is

$$s_x = \sqrt{\text{Var}} \approx 1.019.$$

Remark 1.5.4 The following are some properties of the standard deviation of a variable x.

1. The standard deviation s_x is always non negative. In fact, s_x is almost always positive and the only instance when $s_x = 0$ is when the data points are all identical, that is, $x_1 = x_2 = \cdots = x_n$. In this case, the variance is zero and so is the standard deviation.

2. The standard deviation shares the same unit as the numerical variable x. For example, if x measures the weight (in kilograms) of adult males in Singapore, then the unit for s_x is also kilograms.

3. Adding a constant c to all data points does not change the standard deviation. So the standard deviation for the set of data

   $$A = \{x_1, x_2, x_3, \ldots, x_n\}$$

   is the same as the standard deviation for the set of data

   $$B = \{x_1 + c, x_2 + c, x_3 + c, \ldots, x_n + c\}.$$

   Intuitively, since all the data points are adjusted by the same constant c, the spread of the data points about the new mean will be the same as the spread of the original data about the previous mean.

4. Multiplying all the data points by a constant c results in the standard deviation being multiplied by $|c|$, the absolute value of c. In other words, if s_x is the standard deviation for the set of data

   $$A = \{x_1, x_2, x_3, \ldots, x_n\},$$

   then the standard deviation for the set of data

   $$B = \{cx_1, cx_2, cx_3, \ldots, cx_n\}$$

   will be $|c|s_x$.

Example 1.5.5 Let us return to the data set involving three different species of penguins introduced earlier in the chapter. The three species were named Chinstrap, Adelie and Gentoo and the data set contained information on the physical attributes (e.g. mass, bill length, bill depth etc.) of various penguins in each of the three species. An overarching question that one may be interested to answer is - how different are these penguins? A common approach to answer this question is to compare those physical attributes across samples collected for the different species and see if they are significantly different. For example, we can compute the mean and standard deviation of the mass of the penguins, summarised as follows:

	Mean mass	Standard deviation of mass
Chinstrap	3733g	384.3g
Adelie	3710g	458.6g
Gentoo	5076g	504.1g
Overall	4201g	802.0g

1. Observe that the overall mean mass 4201g is indeed between the group with the highest mean mass (Gentoo) at 5076g and the group with the lowest mean mass (Adelie) at 3710g. This is consistent with our earlier discussion.

2. Even though the overall mean mass is 4201g with standard deviation 802g, it **does not** imply that the heaviest penguin weighs $4201 + 802 = 5003$g.

3. Suppose we wish to investigate whether the Adelie and Chinstrap species are similar in terms of their mass. First, we observe that the mean mass of these two groups are rather similar with the Adelie species having a mean mass of 3710g while the Chinstrap species has a mean mass of 3733g. However, the standard deviation of mass for these two species are rather different.

4. To examine further on the difference in physical attributes between the Adelie and the Chinstrap species, we need to delve into other factors or variables that we have information on from the data set, for example, variables like age, gender, location and so on. This is Exploratory Data Analysis in action, where we start off with a few questions about the data set and with exploration into the data, we ask new questions and go back to the data set to look more closely at the data in an attempt to answer the new questions. In data analysis, this process is often repeated several times. In relation to this penguin data set, here are some further questions that can be asked:

   - Are male penguins heavier than female penguins?
   - Is there a relationship between bill length and bill depth across all species?
   - Do heavier penguins come from colder locations?
   - Can findings in this data be generalised to all of the three species?

5. The concept of *coefficient of variation* is often used to quantify the degree of spread *relative* to the mean. The formula is

$$\text{coefficient of variation} = \frac{s_x}{\overline{x}}.$$

Observe that since s_x and $\overline{x}$ have the same units, the coefficient of variation has no units and is simply a number. The coefficient of variation is a useful statistic for comparing the degree of variation across different variables within a data set, even if the means are drastically different from one another.

Section 1.6 Summary Statistics - Median, quartiles, IQR and mode

Definition 1.6.1 In this section, we will introduce a few other summary statistics. We have already discussed the mean, which measures the central tendencies of a variable, as well as standard deviation which measures the spread of the data points about the mean. 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 or descending order. If there are two middle values (when there are an even number of data points), we will take the average of the two middle values as the median. The median is an alternative to the mean as a measure of central tendencies of a numerical variable.

Example 1.6.2 After arranging the following 12 numbers

$$6,\ \ 12,\ \ 5,\ \ 10,\ \ 11,\ \ 18,\ \ 9,\ \ 4,\ \ 12,\ \ 11,\ \ 3,\ \ 13$$

in increasing order, we have

$$3,\ \ 4,\ \ 5,\ \ 6,\ \ 9,\ \ 10,\ \ 11,\ \ 11,\ \ 12,\ \ 12,\ \ 13,\ \ 18.$$

The median is the average of the sixth (10) and seventh (11) numbers in the order, which is 10.5.

Remark 1.6.3

1. We have seen that when a constant c is added to every data point in a data set, the mean will also be increased by c. The median behaves in the same way, so if the median of the values $x_1, x_2, \ldots, x_n$ is r, then the median of

$$x_1 + c, x_2 + c, \ldots, x_n + c$$

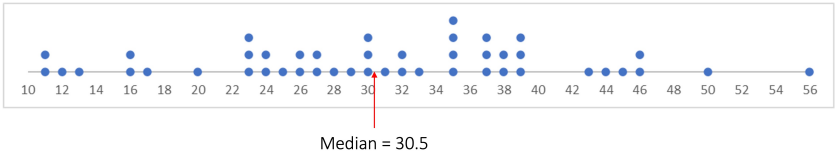
is $r + c$.

2. We have also seen that when a constant c is multiplied to all the data points, then mean is also multiplied by c. The effect on the median is similar, so if the median of the values $x_1, x_2, \ldots, x_n$ is r, then the median of

$$cx_1, cx_2, \ldots, cx_n$$

is cr.

Example 1.6.4 Returning to Example 1.4.5 we saw that school B had 46 students who took an examination and the mean of their scores was 30.72. The plot below, known as a *dot plot* shows the scores obtained by each of the 46 students.



Each dot placed on a particular number represents a student obtaining that score for the examination. Since there were 46 students, the median score would be the average of the 23$^{\text{rd}}$ and 24$^{\text{th}}$ ranked students' scores. The 23$^{\text{rd}}$ ranked student scored 30 marks while the 24$^{\text{th}}$ ranked student scored 31 marks. So the median score is 30.5. This also means that 50% of the students scored below 30.5 marks and the other 50% scored more than 30.5 marks.

It is interesting to note that the mean score for school B was 30.72, which is very close to the median score. The main reason for this is because the spread of the scores are quite symmetrical about the mean and the median. Can you construct a data set where the mean and median are far apart?

We can also compute the median score for students in school A, as well as the overall median score when we combine the students from both schools together. The median and mean (computed in Example 1.4.5) for each subgroup as well as the overall median and mean scores are shown in the table below.

	Median score	Mean score
School A	32	32.21
School B	30.5	30.72
Combine schools A and B	32	32.04

Similar to what we observed for means, the overall median score (32) lies between the subgroup with the higher median (32) and subgroup with the lowest median (30.5). This is by no means a coincidence. Even when there are more than 2 subgroups, the overall median will always be between the lowest median and the highest median among all the subgroups. However, if we know each of the subgroup medians, it is not possible to use this information to derive the overall median. This is unlike the case for mean where, if we know the mean of

each subgroup, together with the "weights" of each group (meaning the number of members in each subgroup) we can take a weighted average to compute the overall mean exactly.

Definition 1.6.5 We have seen that the median represents a numerical value where 50% of the data is less than or equal to this value. This is also known as the 50^{th} percentile of the data values. The first quartile, denoted by Q_1, is the 25^{th} percentile of the data values, while the third quartile, denoted by Q_3 is the 75^{th} percentile of the data values. This means that 25% of the data is less than or equal to Q_1 while 75% of the data is less than or equal to Q_3.

Definition 1.6.6 The interquartile range, denoted by *IQR* is the difference between the third and first quartiles, so IQR $= Q_3 - Q_1$.

Remark 1.6.7

1. IQR and standard deviation share similar properties. For example, we know that IQR is always non negative since Q_3 is always at least as large as Q_1 and so $Q_3 - Q_1 \geq 0$.

2. If we add a positive constant c to all the data points, not only does the median value increase by c, Q_1 and Q_3 are increased by c as well. Thus, there will be no change in IQR. Of course, IQR also remains unchanged if c is subtracted from all data points.

3. If we multiply all data points by a constant c, then IQR will be multiplied by $|c|$.

Example 1.6.8 Let us consider two simple data sets and compute the first quartile, median, third quartile and interquartile range. The first data set consists of an even number of data points as follows:

$$16,\ 30,\ 5,\ 1,\ 9,\ 22,\ 19,\ 8,\ 10,\ 28.$$

We arrange these 10 data points in increasing order:

$$1,\ 5,\ 8,\ 9,\ 10,\ 16,\ 19,\ 22,\ 28,\ 30.$$

1. Since there are 10 data points, the median is the average of the 5^{th} and 6^{th} ranked data points, so median is $\frac{1}{2}(10 + 16) = 13$.

2. To find the first and third quartiles, we divide the data set into the lower half (1^{st} to 5^{th} ranked data points) and upper half (6^{th} to 10^{th} ranked data points). The first quartile is the median of the lower half

$$1,\ 5,\ 8,\ 9,\ 10,$$

which is the 3rd ranked data point in this lower half, so $Q_1 = 8$. The third quartile is the median of the upper half

$$16,\ 19,\ 22,\ 28,\ 30,$$

which is the 3rd ranked data point in this upper half, so $Q_3 = 22$.

3. The interquartile range is $Q_3 - Q_1 = 22 - 8 = 14$.

Let us consider the second data set which consists of an odd number of data points as follows:

$$5.6,\ 1.5,\ 3.3,\ 8.7,\ -3.1,\ 9.2,\ 15.5,\ 2.6,\ 11.5.$$

We arrange these 9 data points in increasing order:

$$-3.1,\ 1.5,\ 2.6,\ 3.3,\ 5.6,\ 8.7,\ 9.2,\ 11.5,\ 15.5.$$

1. Since there are 9 data points, the median is the 5th ranked data point, so median is 5.6.

2. To find the first and third quartiles, we divide the data set into the lower half (1st to 4th ranked data points) and the upper half (6th to 9th ranked data points). Note that we have **not** included the median in **both** lower and upper halves. The first quartile is the median of the lower half

$$-3.1,\ 1.5,\ 2.6,\ 3.3,$$

which is the average of 1.5 and 2.6, so $Q_1 = 2.05$. The third quartile is the median of the upper half

$$8.7,\ 9.2,\ 11.5,\ 15.5,$$

which is the average of 9.2 and 11.5. So $Q_3 = 10.35$.

3. The interquartile range is $Q_3 - Q_1 = 10.35 - 2.05 = 8.3$.

Remark 1.6.9

1. In the example above, when the data set has odd number of data points, we have not included the median in both the lower and upper halves. This is not the universal practice. You may encounter some texts that includes the median from both halves.

2. In reality, when the number of data points is large, summary statistics like median and quartiles are not computed manually but instead, they are computed using softwares. However, even softwares do not adopt the same algorithm in computing these statistics. The good news is that we do not have to worry too much about finding the exact value of the quartile since for large data sets, all the different methods give pretty close answers and the small difference is not an issue. For small data sets, it is also not really meaningful to summarise the data since we have complete information of the entire data set anyway.

Remark 1.6.10 For a numerical variable, we can always use the mean and standard deviation as a pair of summary statistics to describe the central tendency as well as the dispersion and spread of the data. Similarly, the median and IQR can also be used. Which choice is more appropriate? There is no clear cut answer but very often, the choice depends on the distribution of the data. Generally speaking, the median and IQR is preferred if the distribution of the data is not symmetrical or when there are outliers.

We will conclude this section with a final summary statistic that can be used for both numerical and categorical variables.

Definition 1.6.11 The *mode* of a numerical variable is the numerical value that appears most often in the data. For categorical data, a mode is the category that has the highest occurrence in the data. The mode is generally interpreted as the peak of the distribution and this means that the mode has the highest probability of being observed if a data point is to be selected randomly from the entire data set.

Example 1.6.12 In the following set of numbers,

$$11,\ \ 12,\ \ 5,\ \ 10,\ \ 11,\ \ 11,\ \ 9,\ \ 4,\ \ 12,\ \ 11,\ \ 3,\ \ 13,$$

the mode is 11, since it appears 4 times, the highest among all the other numbers.

Section 1.7 Study Designs - Experimental Studies and Observational Studies

Recall that we introduced three types of research questions earlier in the chapter.

1. To make an estimate about the population.
2. To test a claim about the population.
3. To compare two sub-populations / to investigate a relationship between two variables in the population.

In this section, we will focus on the third type of question, where we investigate a relationship between two variables in the population. For example, consider the question "does drinking coffee help students pass the mathematics examination?" The two variables here are drinking coffee (yes or no) and passing the mathematics examination (yes or no). Here, both variables are nominal categorical variables. Commonly, a researcher looking at this

situation may want to define "drinking coffee" as the independent variable as it can be controlled and adjusted while "passing the mathematics examination" is the dependent variable. In order to investigate this relationship, we need to design a study and for this course, we will discuss two main study designs, namely *experimental studies* and *observational studies*.

Definition 1.7.1 In an *experimental study* (sometimes also known as *controlled experiment* or simply an *experiment*), we intentionally manipulate one variable (the independent variable) to observe whether it has an effect on another variable (the dependent variable). The primary goal of an experiment is to provide evidence for a *cause-and-effect* relationship between two variables.

Example 1.7.2 Returning to the experiment to investigate the relationship between drinking coffee and passing the mathematics examination, we can set up an experimental study by dividing the subjects, that is, the students taking the examination, into two groups. The first group will be required to drink exactly one cup of coffee every day for a month. The second group will not drink any coffee for one month. The group who are required to drink one cup of coffee every day for a month is often known as the *treatment group* since they are thought to be put through the "treatment" of drinking coffee. The other group who does not drink coffee is known as the *control group*.

It is important to have a control group to compare against the treatment group. Without a control group (imagine every subject is required to drink coffee for a month), we would not be able to determine if there were indeed any difference between drinking coffee or not. However, it should be noted at this point that sometimes the control group are also subjected to other forms of treatment (not to be mistaken with the treatment of interest in the study) i.e. a control group does not necessarily mean no treatment at all. One example is when we are comparing the effects of a **new treatment** with an **existing treatment**. For such instances, the treatment group will be formed by subjects receiving the new treatment while the control group will be those who continue to receive the existing treatment.

A natural question now is how the subjects are to be divided into the two groups. Can we do it anyway we like? Can we let the odd numbered subjects be in the treatment group and the even numbered subjects be in the control group? Does it matter? The problem of how to assign subjects to the two groups is our next topic of discussion.

Discussion 1.7.3 Continuing on with the coffee drinking experiment, suppose one month after the experiment started, the subjects from both groups took the mathematics examination and the number of passes in each group is shown below.

	Treatment group (coffee)	Control group (no coffee)
Pass	900	450
Fail	100	550

We see that 90% (900 out of 1000) of the students in the treatment group passed the examination while only 45% (450 out of 1000) of the students in the control group passed. There seems to be some evidence that drinking coffee may help a student pass the mathematics examination. Is this evidence convincing? Can we go one step further and say that coffee *causes* improvement in passing the examination?

The skeptics among us will probably not be so easily convinced. Possible doubts that could arise and questions that can be asked could be

1. Maybe the students in the "coffee group" just happen to be better in mathematics and thus have a higher chance of passing the examination? Or maybe they just have higher IQ than those in the "no-coffee" group?

2. Maybe many of the students in the "coffee group" had longer revision time before the examination than those in the "no-coffee" group?

These are some of the possible factors that could have contributed to the difference in passing rate between the two groups. In trying to establish a cause-and-effect relationship between two variables, we want to make sure that the independent variable is the **only** factor that impacts the dependent variable.

In the coffee drinking example, we want to ensure that coffee drinking (or not) is the only variable that distinguishes the treatment group from the control group. In other words, we need to ensure that coffee drinking (or not) is the only difference between the subjects in the two groups. All other possible differentiating factors, for example amount of revision time, should be removed.

How can these factors be "removed"? Surely we cannot mandate that all students in both groups are only allowed a fixed number of revision hours before the examination! Even if we could, we most definitely cannot enforce that all students in both groups must have the same IQ! The answer to this is a powerful statistical method known as *random assignment*.

Definition 1.7.4 *Random assignment* is an impartial procedure that uses chance (or probability) to allocate subjects into treatment and control groups.

How do we perform random assignment for our coffee drinking experiment? The following procedure can be considered:

1. Write down the name of each student on a piece of paper.

2. Put all the pieces of paper into a box and mix them up.

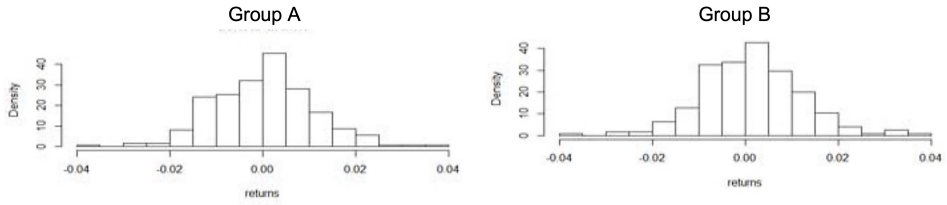
3. Draw the names out one by one until exactly half the total number of students are chosen. The names of the students chosen will form the treatment group.

4. The remainder of the students not in the treatment group will form the control group.

The procedure above is just an example of how random assignment can be done. As long as there is a random element, there can be other procedures to conduct random assignment. It should be noted that at every draw, each name in the box has an equally likely chance of being chosen. Perhaps there are still doubters out there who feels that even with such a chance event of assigning the subjects into treatment and control groups, it may still happen that many of the high IQ students will be assigned to the treatment group. However, we can be assured that:

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

Example 1.7.5 The S&P Index is a stock market index of the largest US publicly traded companies. We are interested in the percentage returns of these S&P companies in 2013. Suppose these percentage returns were written on 1000 tickets and we are aware that the percentage returns range from -4% to 4%. Using the method of random assignment, the 1000 tickets are separated into two groups, each comprising of 500 tickets. The following plots shows the distribution of the percentage returns of companies in both groups.



In each plot, the horizontal axis is the percentage returns and the vertical axis counts the number of companies with the specific percentage returns. We observe the effect of random assignment as the distribution in both groups are rather similar.

Remark 1.7.6

1. While performing random assignment to allocate subjects into treatment and control groups, it is not necessary/possible for both groups to have exactly the same number of subjects. For example, if we have 501 students to be divided into two groups. As long as some form of random assignment is done and the number of subjects in each group is big enough, we can still be assured that the two groups are similar in almost every aspect.

2. When we use the term "random" in random assignment, we do not mean that the assignment is haphazard. The term random in this case is used in relation to the use of an impartial chance mechanism that is effected to assign the subjects into two (or more) groups.

Discussion 1.7.7 While random assignment is an important step to take when we divide our subjects into the treatment and control groups, there is another important consideration when it comes to designing a controlled experiment. If we make it known to the control group that they *are indeed* the control group, and therefore not going to receive any form of treatment, this could possibly lead to bias.

To see why this is so, let us return to the coffee experiment. If the subjects in the control group are told that they will not be assigned any coffee for a month, when we are testing if coffee helps a student pass the mathematics examination, students in the control group may feel disadvantaged and therefore lack confidence and motivation to study. This may in turn result in these students not doing well in the examination and perform poorer than their friends in the treatment group who were given coffee. Any observed difference in passing rate between the two groups of students **may not** be the result of coffee at all. If this happens, the effect of coffee may be *overstated*.

On the other hand, to the students in the control group, knowing that they will not be given coffee may actually cause them to take certain measures for their own benefit of passing the examination. For example, they may study harder and spend more time on their revision which may then result in the control group performing better than the treatment group in passing the examination. Again, any observed difference in passing rate between the two groups of students may not be the result of coffee at all. If this happens, the effect of coffee may be *understated*.

One way to reduce the anxiety of the control group which could influence the study on the effects of coffee drinking is to give the subjects in the control group another beverage which tastes and smells the same as coffee but is without the active ingredients in coffee that is believed to improve one's cognitive ability.

Definition 1.7.8 In the previous discussion, the alternative beverage is termed a *placebo*. A placebo is an inactive substance or other intervention that looks the same as, and is given the same way as, an active drug or treatment being tested. In the context of an experiment, a placebo is something given to the control group that in actual fact, has no effect on the subjects in the group.

However, it has been observed in some instances, subjects in the control group upon receiving the placebo still showed some positive effects which is likely caused by the **psychology of believing** that they are actually being "treated". This is known as the *placebo effect*.

Definition 1.7.9

1. One way to prevent the placebo effect from interfering with our experiment and observation on the benefits (if any) of the treatment is to *blind* the subjects involved in the experiment. By *blinding* the subjects, we mean that they do not know whether they belong to the treatment or control group. To do this, a placebo that is "similar" to

the treatment is given to the control group to make them believe that they are indeed receiving treatment. In this way, the treatment and control groups would respond in the same way to the idea of being "treated". If we can do this, we would have achieved *single blinding*.

2. To take blinding one step further, other than blinding the subjects, it may be necessary to consider blinding the researchers conducting the study as well, especially if measuring the effects of the treatment may involve subjective assessments of the subjects. For example, in the coffee experiment, if the assessors marking the students' answers are aware of which group each student belongs to, they may be inclined to award higher marks to students in the treatment group than those in the control group. This is because the assessors may subconsciously believe that the treatment is effective and this could introduce bias in the outcome.

   Thus, we should also blind the assessors so that they do not know whether they are assessing the treatment or the control group. We would have acienved *double blinding* if subjects and assessors are blinded about the assignment.

To conclude this discussion on blinding, we should note that sometimes it may not be possible to blind both the subjects and the assessors (can you think of one such experiment?) but when done right, double blinding can be very effective in reducing bias in the outcome of the experiment.

Discussion 1.7.10 Besides an experimental study, another study design is an observational study. Consider the following research question: Does vaccination help reduce the effects of the coronavirus?

If we were to design a controlled experiment, would the following be a possible and reasonable approach?

- Enrol a group of participants into the study and inject all the participants with low dosages of the virus strain.
- Perform random assignment to divide the group of subjects into the treatment group and control group.
- Inject the treatment group with the vaccine and inject a harmless liquid (similar in colour, smell etc to the vaccine) into the control group, without revealing what they are being injected with.
- Observe the number of participants in each group who develop symptoms similar to a coronavirus patient.

It is interesting to note that this is not a hypothetical situation. In fact, in 2020, during the COVID-19 pandemic, a Dublin-based commercial clinical research organisation was

reported to be planning an experiment to test the effectiveness of a COVID-19 vaccine. The plan was similar to the approach described above.

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.

You probably realise by now that it is not so straightforward to design a controlled experiment like this. There are obvious *ethical issues* that needs to be addressed. Some immediate questions that needs to be answered are

1. Should we inject such a virus into humans in the first place?
2. How should we decide who is to be assigned to the treatment group and who is to be assigned to the control group?
3. Is it fair not to let the subjects know if they are injected with the vaccine or with a placebo? Should we obtain consent from the subjects at the beginning of the study?

Experiments can give us useful evidence for a cause-and-effect relationship. However, not all research questions are suitable to be investigated using an experiment, sometimes due to ethical issues like those listed above. Therefore, we need to consider the pros, cons and feasibility of an experimental study before deciding if we should proceed.

Definition 1.7.11 An *observational study* observes individuals and measures the variables of interest, usually without any direct/deliberate manipulation of the variables by the researchers.

Remark 1.7.12 Observational studies are alternatives to experiments that can be used when we are faced with ethical issues in experiments. An observational study observes individuals and measures the variables of interest. As researchers usually do not attempt to directly manipulate or change one variable to cause an effect in another variable, observational studies do not provide convincing evidence of a cause-and-effect relationship between two variables.

Example 1.7.13 We would like to investigate whether exercising regularly (defined as exercising at least 3 times a week, at least 30 minutes of strenuous exercise each time) is associated[2] with having a healthy body mass index (BMI) (defined as between 18.5 to 22.9 kg/m^2) for Singaporean men between the age of 30 to 40 years old.

Participants were recruited into the study and by their own declaration, they were classified into either the "treatment" group (those who exercise regularly) or the "control group" (those who do not). Participants were then told to proceed with their usual lifestyle habits and their body mass index were measured after 3 months. The following table summarises the findings at the end of the study.

	Treatment (Exercise regularly)	Control (Do not exercise regularly)
Healthy BMI range	320	127
Outside Healthy BMI range	101	191

This is an example of an observational study. Do you think there is sufficient evidence of association between exercising regularly and having a healthy BMI? We will discuss more questions like this in subsequent chapters.

Let us conclude this chapter with some final remarks on study designs.

Remark 1.7.14

1. Not all research questions can be studied practically using an experiment. For example, if we would like to investigate if long term smoking is linked to heart disease, it is extremely difficult to design an experiment and put subjects into the treatment group where they will be **required** to smoke for the long term, even if this is against their will. This is challenging and unethical. An observational study may be more suitable for such an investigation.

2. For observational studies, there is no actual treatment being assigned to the subjects but we normally still use the term *treatment* and *control* in the same way as though we are dealing with an experiment. For the investigation on smoking and heart disease, smokers who are observed to be smoking over a long period of time will be in the treatment group while non smokers are in the control group. Sometimes, we may use the term *exposure group* instead of treatment group and *non-exposure group* instead of control group.

3. For experimental studies, subjects are assigned into either the treatment or control group by the researcher. For observational studies, subjects assign themselves into either the treatment or control group.

[2]previously mentioned in Example 1.3.3, the topic on association between variables will be discussed in Chapter 2.

4. Observational studies cannot provide evidence of cause-and-effect relationships. On the other hand, experimental studies can provide such evidence if it has the features of randomised assignment and blinding (preferably doubling blinding).

5. The question of *generalisability* is often asked. That is,

If an experiment is well-designed, can the conclusion of the experiment based on a sample be generalised to the population from which the sample was drawn?

   Having a good design is not the only important piece of the puzzle. In order to generalise the results from a sample to a bigger population, there are other factors that are equally important, for example, the sampling frame, sampling method, sample size and response rate.