Quantitative Reasoning with Data

GEA1000 Teaching Team

August 7, 2022

Preface

GEA1000 Quantitative Reasoning with Data is a module that aims to equip students with essential data literacy skills to analyse data and make decisions under uncertainty. It covers the basic principles and practice for collecting data and extracting useful insights, illustrated in a variety of application domains. For example, when two issues are correlated (e.g. smoking and cancer), how can we tell whether the relationship is causal (e.g. smoking causes cancer)? How can we analyse categorical data? What about numerical data? What about uncertainty and complex relationships? These and many other questions will be addressed using data software and computational tools, with real-world data sets.

The framework that we will be making reference to frequently in this course is the PPDAC cycle.[1] The figure below is a representation of the data problem-solving cycle, "**P**roblem, **P**lan, **D**ata, **A**nalysis and **C**onclusion."

The PPDAC cycle is a well-established approach to statistical literacy which is relevant to how we learn data literacy after the transformational change "big data" has had on society.[2] The main features of PPDAC are

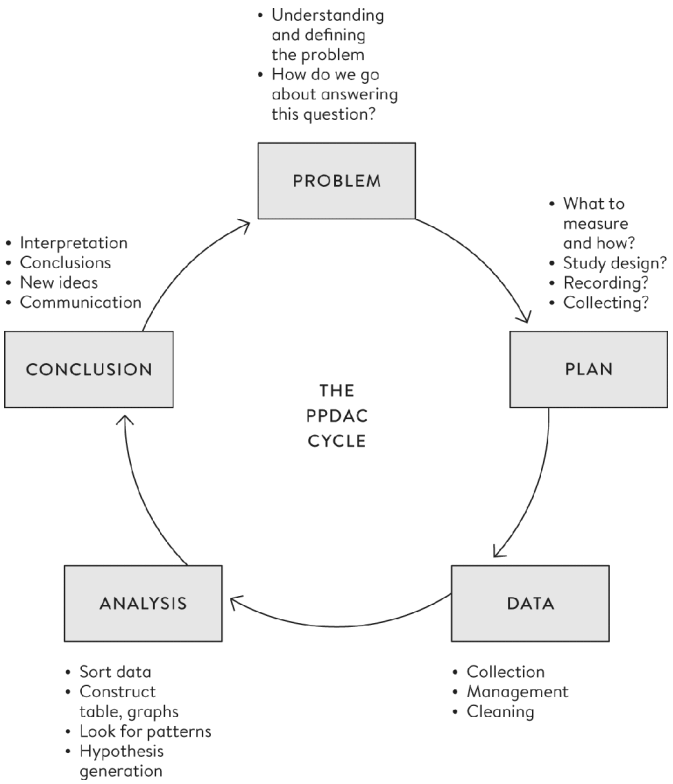
- (to) document the stages a person would undertake when solving a problem using numerical evidence,
- using data which they had collected themselves, or from existing (public) data sets,
- (where) analysis methods can include machine learning algorithms, as well as more traditional statistical techniques.

The following figure briefly describes what happens at each stage of the PPDAC cycle.[3]

[1] Spiegelhalter, David. (2019). *The Art of Statistics.* Penguin/Pelican Books

[2] Wolff, A. et al. (2016). *Creating an Understanding of Data Literacy for a Data-driven Society.* The Journal of Community Informatics, 12(3), 9–26.

[3] Spiegelhalter, David. (2019). *The Art of Statistics.* Penguin/Pelican Books



This set of notes is meant to follow the four chapters of the module closely. The topics covered in the chapters are summarised below.

- Chapter 1: Getting data. Data collection and sampling. Experiments and observational studies. Data cleaning and recoding. Interpreting summary statistics (mode, mean, quartiles, standard deviation etc.)
- Chapter 2: Dealing with categorical data. Bar plots, contingency table, rates and basic rules on rates. Association, confounders and Simpson's Paradox.
- Chapter 3: Dealing with numerical data. Univariate and bivariate data. Histograms, boxplots and scatter plots. Correlation and simple linear regression.
- Chapter 4: Making sense of data. Probability, conditional probability and independence. Discrete and continuous random variables. Interpreting confidence intervals. Hypothesis testing and learning about population based on a sample. Simple simulation.

Exploratory data analysis (EDA) will be incorporated extensively into the content of the module. Students will appreciate that even simple plots and contingency tables can give them valuable insights about data. There will be an emphasis on using suitable real world data sets as motivating examples to introduce content and through the process of problem solving, elucidate techniques/materials in the syllabus.

Contents

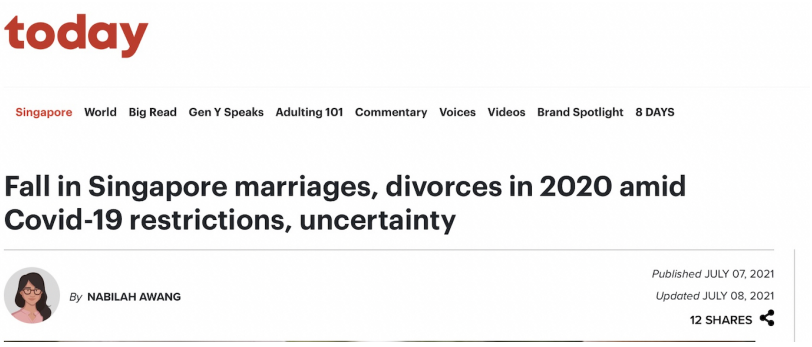
Chapter 1

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 opinions 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 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 two different categories 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.

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

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 there 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 may 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

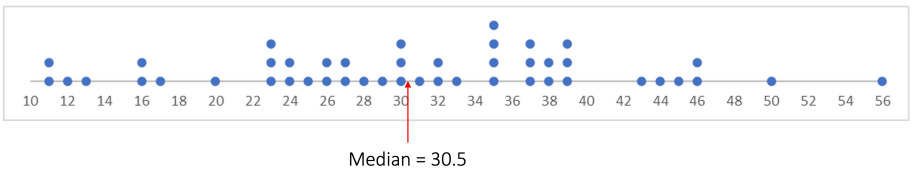
$$3,\ \ 4,\ \ 5,\ \ 6,\ \ 9,\ \ 10,\ \ 11,\ \ 11,\ \ 12,\ \ 12,\ \ 13,\ \ 18.$$

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 the 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 23rd and 24th ranked students' scores. The 23rd ranked student scored 30 marks while the 24th 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 3^{rd} 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 3^{rd} 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 5^{th} 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 (1^{st} to 4^{th} ranked data points) and the upper half (6^{th} to 9^{th} 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 in 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 <u>how</u> 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 rates 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.

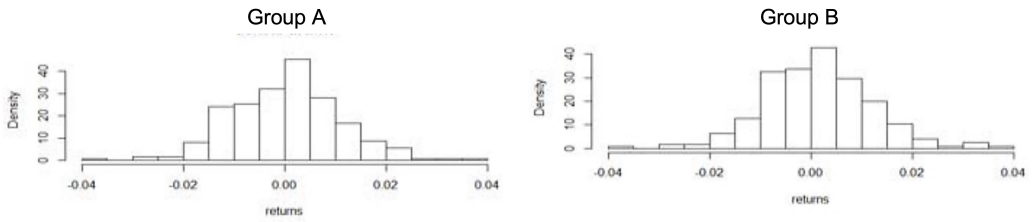
3. Draw the names out one by one until exactly half the total number of students are chosen. The names of the students chosen will form the treatment group.

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 show the distributions 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 rates 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 so that the two treatments appear identical to the subjects. As a result, subjects do not know which group they belong to. 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 need to be addressed. Some immediate questions that need 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 ages 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]previously mentioned in Example 1.3.3, the topic on association between variables will be discussed in Chapter 2.

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.

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 double 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.

Exercise 1

1. Drug A is a new drug created to treat Congenital Amegakaryocytic Thrombocytopenia (CAMT). An experiment was done to evaluate survival outcomes of drug A, against the current standard, Rituximab, for treating CAMT. 150 CAMT patients were assigned to the drug A group, and 150 CAMT patients were assigned to the Rituximab group. A portion of the experimental design is summarised in the table below.

	Drug A	Rituximab
Male	82	85
Female	68	65
Total	150	150

   Determine if the statement below is true or false:

   "The table shows that random assignment was not done, because the two groups (drug A and Rituximab) do not have the same number of females."

2. Which of the following scenarios is an example of random assignment in a controlled experiment?

   (I) For each subject, Peter throws a fair die of six sides. Subjects are assigned based on the number shown on the top surface of the die. Before the start of the assignment, Peter determines that the numbers "1", "2" and "6" will be assigned to the treatment group, and "3", "4" and "5" to the control group.

   (II) James lists all subjects in the experiment by alphabetical order, and selects subjects whose last name starts with "A", "B", "C" till "M" to place in the treatment group, while subjects whose last name starts with "N", "O", "P" till "Z" are placed in the control group.

   (A) Only (I).

   (B) Only (II).

   (C) Neither (I) nor (II).

   (D) Both (I) and (II).

3. Which of the following best describes the function of a placebo in an experiment?

   (A) It is used to assign participants into treatment and control groups.

   (B) It is used to randomly assign the participants into treatment and control groups.

   (C) It is used to blind the participants to which group they belong to.

   (D) It is used to select participants into the experiment.

4. A new drug Y was created to reduce stomach-aches. Researchers wanted to test the effectiveness of drug Y. Thus, they conducted a study containing 2000 subjects. The subjects have the following characteristics:

	Young	Old	Row total
Female	700	700	1400
Male	400	200	600
Column total	1100	900	2000

   Random assignment was conducted to assign the 2000 subjects into the treatment and the control groups. 1200 subjects were assigned to the treatment group, while the remaining 800 subjects were assigned to the control group. How many young females would we expect to see in the treatment group?

(A) About 210.

(B) About 420.

(C) About 500.

(D) About 700.

5. Which of the following statements is/are always true about controlled experiments and observational studies?

(I) There is no control group in observational studies.

(II) Randomised assignment of subjects does not occur in observational studies.

(III) There are no confounders in controlled experiments.

(A) Only (I) and (II).

(B) Only (I) and (III).

(C) Only (II) and (III).

(D) Only (II).

(E) None of the other given options is correct.

6. A researcher in University X wanted to conduct a survey to find out the average amount of time spent studying a week, by students in the university. He obtained the list of email addresses of all 2000 students in the university and sent out a survey form to everyone. As a token of appreciation, students who filled up the form received a "10% off" coupon from the university's bookshop. 300 students responded to the survey.

Which of the following statements is/are correct? Select all that apply.

(A) The study is likely to contain non-response bias.

(B) The study is likely to contain selection bias.

(C) The study uses a census.

7. A publication is estimated to have about 20000 subscribers. A survey was sent to a random sample of 5000 of its subscribers. 300 of them returned the survey. Which of the following statements is true?

(A) The sample results may not be generalisable to the population of subscribers because they used a self-selected sample.

(B) The sample results may not be generalisable to the population of subscribers because there is likely to be non-response bias.

(C) The sample results will be generalisable to the population of subscribers because they used a random sample.

8. Virus X has been known to cause very severe symptoms in its patients. Previously there has been no anti-viral medicine to treat virus X. Recently, researchers have finally managed to produce a trial drug in the form of a tablet. Researchers want to investigate if the trial drug helps to reduce the duration of symptoms (number of days) in patients. 1000 patients were sampled for the study, and all consented to join the study.

Which of the following statements is/are true? Select all that apply.

(A) Random sampling should be done to ensure that the subjects' demographics/characteristics are similar (in the treatment and control groups).

(B) Blinding the researchers to the subjects' assigned groups (treatment or control group) is important because the researchers may have certain bias for/against the drug.

(C) If the study randomly assigns 400 subjects into the treatment group, and 600 subjects into the control group, the result of the study will be biased due to the unequal number in the two groups.

9. (This is a multiple response question.) From the options given, select all possible words that can be used to complete the sentence below.

   Probability sampling refers to a sampling process whereby the probability of selection of individuals within the sampling frame must be _____.

   (A) non-zero

   (B) known

   (C) high

10. A study was conducted to understand the average amount of sleep that current hall ABC students get. The hall has five levels, each with 50 rooms on each floor. Currently in hall ABC, every room is occupied by one student.

    For the study, every room is labelled with a specific number from 1 to 250. Identical slips of paper numbered from 1 to 250 are then placed in a box, and the researcher drew random 60 slips without replacement. These 60 numbers indicate the chosen students for the study. A survey form was then sent out to these 60 students.

    Which of the following must be true about the study?

    (A) There will not be any selection bias in this study.

    (B) There will not be any non-response bias in this study.

    (C) The sample will not be representative of the 250 residents in the hall.

    (D) If the 60 individuals selected did respond, the sample average cannot be used as an estimate of the average studying hours for all 250 residents.

11. Paracetamol company NAS owns a tablet press machine that produces Paracetamol tablets. On one shift, 3000 batches of tablets were manufactured. Each batch contains 10 tablets - a total of 30,000 tablets were manufactured. A researcher wants to ensure the dosage in the tablets is correct but has no time to check every single tablet. Hence, she decides to sample some of the tablets instead.

    Which of the following describes a probability sampling method? Select all that apply.

    (A) Select 3000 tablets at random.

    (B) Label all the tablets in each batch from 1 to 10, select a number from 1 to 10 at random, and select the unit from every batch that corresponds to that number.

    (C) Select 300 batches at random, and then sample all tablets in every selected batch.

    (D) Select the first 3000 tablets that were manufactured.

12. A recent study revealed that Singapore is "the most tired country in the world, due to work and internet." A researcher decided to conduct a further study on internet usage behaviour and working hours among all Singaporean adults in Singapore. Data was collected by interviewing commuters alighting from Pasir Ris MRT (East), Woodlands MRT (North), Redhill MRT (South) and Jurong East MRT (West) from 8am to 11pm over a period of 7 days.

    Which of the following statements is **necessarily true**?

    (A) As data was collected from different parts of Singapore, it is generalisable to the population of Singapore.

    (B) Due to the equal representation of Northern, Southern, Eastern and Western parts of Singapore, selection bias is minimised.

(C) In this example, non-response bias exists because of a bad sampling plan.

(D) None of the other options.

13. Patch Z is a new medicine created to remove muscle soreness. A study was done to investigate the effectiveness of patch Z. The population of interest was Singaporean adult males. For this study, the researchers requested the Singapore Sports Association to sample all male athletes who reported for training over the week. 200 male athletes were sampled. There was no non-response. The 200 subjects had their identity tags randomly shuffled in a box. The first 100 tags picked from the box were assigned to the treatment group - administered patch Z. The remaining 100 were administered a placebo. 72% of the group that received patch Z had their muscle soreness alleviated, while 34% of the other group had their soreness alleviated. Which of the following is/are true?

(I) We are not able to generalise these results to the population of interest.

(II) Random assignment was not conducted.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

14. The United States government conducts a Census of Agriculture every five years. The census comprises of farmland usage in all the 50 states in the country. John generated a sample of 3000 counties across all states from this census. He then collected data on the number of acres of land space these counties in the sample devoted to farms, and summarised his findings in a report as follows:

(I) Of the 3000 counties selected, 25 counties were selected more than once in the sampling process.

(II) 18% of the counties selected in this sample were from the state of Virginia, while none were from the states of Alaska, Arizona, Connecticut, Delaware, Hawaii, Rhode Island, Utah or Wyoming.

John claimed that he obtained the sample of 3000 counties by Stratified Sampling **with replacement**, with the stratum being every state in the United States. Assuming that statements (I) and (II) are true, which of the statements do not/does not support John's claim on his sampling method?

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

15. For the following two cases, determine which sampling plan was used.

Case 1: In an opinion poll, an airline company made a list of all its flights on 1 Jan 2022 and then selected a simple random sample of 30 flights. All the passengers on those flights selected were asked to fill out a questionnaire form.

Case 2: A departmental store wanted to find out if customers would be willing to pay slightly higher prices for their products in order to have a smartphone app which customers can use to help them locate items in the store. The store hired an interviewer John and placed him at the only entrance on a particular day. John was asked to collect a sample of 100 opinions by interviewing the next person who came through the entrance each time he finishes an interview.

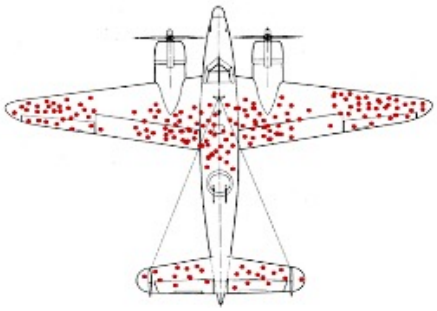
(A) Case 1: Cluster sampling plan; Case 2: Systematic sampling plan.

(B) Case 1: Stratified sampling plan; Case 2: Systematic sampling plan.

(C) Case 1: Stratified sampling plan; Case 2: Non-probability sampling.

(D) Case 1: Cluster sampling plan; Case 2: Non-probability sampling.

16. A military officer was interested in reducing the number of casualties sustained in aerial battle. His population of interest was all planes under his charge. He tasked his men to examine the planes that returned from the war front, and then take note of which parts of the planes sustained ammunition damage. He collated all the data and presented it on a single blueprint of the plane, as shown below (the dots denote where ammunition damage occurred):



The officer then concludes: "Based on my sample data, I propose to fortify the plane armour for regions where ammunition damage was concentrated (using the above blueprint as a guide), so as to help these planes survive better." Would you agree with his assessment and why?

(A) Yes. The sample collected came from a good sampling frame.

(B) No. The sample collected came from an imperfect sampling frame.

(C) Yes. The sample size is big enough.

(D) No. The sample size is too small.

17. If a sampling frame is _____ the target population, it will not lead to a loss in the generalisability of the results from the sample to the population.

Which of the following can be used to fill the blank appropriately? Select all that apply.

(A) equal to

(B) smaller than

(C) larger than

18. Clothes retailer G&N wishes to find out from all their visitors how receptive they are in terms of recycling used clothing. The management decides to survey a sample from customers paying for purchases at the cashier. They interview every fifth paying customer during retail hours from 11am to 9.30pm. Which of the statements is/are correct?

(I) The above is an example of systematic sampling.

(II) The sampling frame is the same as the target population.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

19. A researcher is interested in drawing a sample from town X that has a population of 2000. He has a sampling frame of all 2000 townfolks' names. Which of the following methods can be used to select a simple random sample of 100 people from this population? Select all that apply.

(A) Sort the peoples' names by alphabetical order (A to Z) and place the names in a list. Choose the people whose names appear at the top 100 of the list.

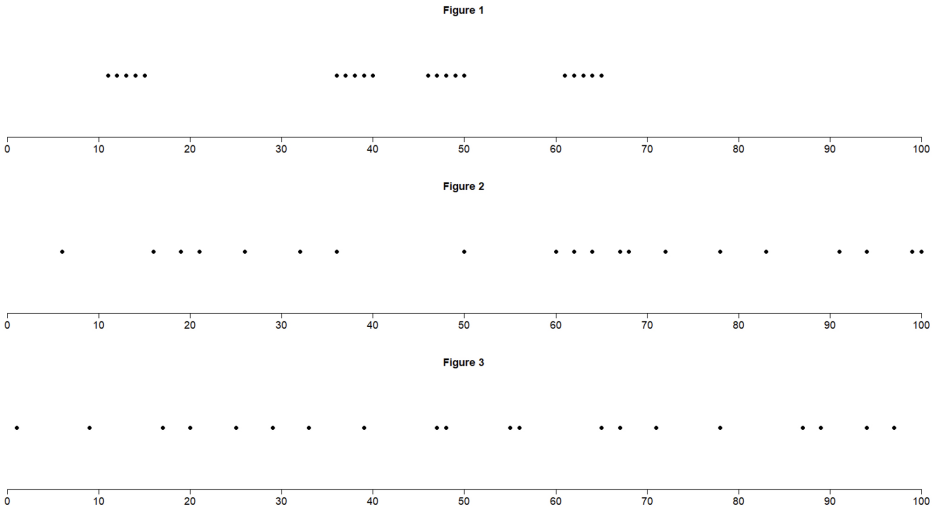
(B) Assign each townsfolk a unique random integer from 1 to 2000. Choose the people assigned numbers 1 to 100.

(C) Write the 2000 peoples' names on equal sized pieces of paper, mix the papers in a box, and draw out 100 pieces of paper in one go. Choose the people whose names appear on the drawn papers.

20. Tom selected 4 samples of 20 integers from the population $\{1, 2, \ldots, 100\}$ using 4 different methods. They are

    1. simple random sampling (SRS).
    2. stratified sampling: the population was divided into the 10 strata $\{1, 2, \ldots, 10\}$, $\{11, 12, \ldots, 20\}$, $\ldots$, $\{91, 92, \ldots, 100\}$; and a SRS of 2 numbers was drawn from each of the 10 strata.
    3. cluster sampling: the population was divided into 20 clusters $\{1, 2, 3, 4, 5\}$, $\{6, 7, 8, 9, 10\}$, $\ldots$, $\{96, 97, 98, 99, 100\}$; and a SRS of 4 of these clusters was selected.
    4. systematic sampling: a random starting point between 1 to 5 was selected; and every 5th unit thereafter was selected too.

    He created dot plots for exactly 3 of the samples generated. Identify the sampling method depicted by each of the following plots.



Sampling method depicted by Figure 1: _____

Sampling method depicted by Figure 2: _____

Sampling method depicted by Figure 3: _____

21. Select the correct word from the list for the respective blank in the sentence.

    "The __(1)__ is used to measure how widely spread the data points are about its __(2)__."

    List: *interquartile range,* *standard deviation,* *mean,* *mode.*

22. The Registry of Marriages is interested to see the relationship between the ages of husbands and wives in City X. They randomly sampled 1000 pairs of husbands and wives from the population of City X and obtained data of their ages (in years). Looking through the data, they found that men always marry women who are younger than them.

    Based only on the information given above, which of the following statements must be true?

(I) The average age of the husbands is more than the average age of the wives.

(II) The standard deviation of husband's age is more than the standard deviation of wife's age.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

23. A teacher has just finished marking the final examination scripts for her class of 50 Secondary 1 students. She informs the students that the class average is 67.3. The maximum mark for the examination is 100 and the passing mark is 50. A student receives his examination script and realises his score is 65 which is lower than the average score. Based only on the information given above, which of the following statements must be true?

(I) The student has performed worse than half the class.

(II) Everyone in the class has passed the test.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

24. The CEO of a company wishes to find out the level of job satisfaction that his employees have. His company has 876 employees. He administers an anonymised survey in which there are questions that ask about the employees' satisfaction with regards to various aspects of their jobs, including welfare, remuneration, career progression, learning etc.

For each question, employees are to rate their satisfaction on a scale of 1-9. In this scale, 1 represents the lowest level of satisfaction whilst 9 represents the highest level of satisfaction. Broadly, any score below 5 for a question is regarded as "not satisfied" and any score above 5 for a question is regarded as "satisfied" while a score of 5 represents being "neutral".

Every employee fills in the survey based on how he/she feels about the job and the data is collected. Assume that every employee is honest in the response and is fully aware of what the values on the scale represent.

What type(s) of analysis is/are considered appropriate for the data on the satisfaction scores?
Select all that apply.

(A) For each question, one can perform a summary statistics calculation on the satisfaction scores to obtain the mean, standard deviation, median, as well as the interquartile range. From these numerical summary statistics, we can conclude meaningful information on the employees' satisfaction as a whole.

(B) For each question, one can plot a histogram to depict the distribution of the satisfaction scores and perform a calculation to determine the exact degree of skewness of the histogram.

(C) For each question, one can calculate the proportion of employees who gave each satisfaction score and determine what percentage of employees are "not satisfied" and what percentage are "satisfied".

25. We have learnt that the standard deviation and interquartile range (IQR) are examples of summary statistics that help us to quantify the spread of data points. However, they are not the only ways of quantifying spread and there are other summary statistics that can also help us to do this. For a

numerical variable x, we can define the Mean Absolute Deviation (commonly abbreviated as MAD) using the formula

$$\text{Mean Absolute Deviation of } x = \frac{|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n},$$

where $x_1, x_2, \ldots, x_n$ are values for the variable in a data set and n is the number of data points in the data set. The MAD is sometimes used in place of the standard deviation as a measure of quantifying the spread of the data. Based on the above formula, which properties **must** the MAD possess? Select all that apply.

(A) The MAD cannot take a negative value.

(B) The MAD does not change when a constant is added to all the data points.

(C) The MAD does not change when a constant is multiplied to all the data points.

(D) If the MAD is zero, then all the values of $x_1, x_2, \ldots, x_n$ in the data set are the same.

26. A teacher has finished marking her students' test scripts for a Mathematics test. The maximum mark attainable for the test is 50. She records the following summary statistics for her class.

- Mean: 37.4
- Median: 35
- Standard deviation: 17.22
- Quartile 1: 23
- Quartile 3: 43
- Highest mark: 48
- Lowest mark: 16
- Range: $48 - 16 = 32$.

She returns the test papers to her class and goes through the answers. Whilst going through the answers, she realises that she has marked a question incorrectly for the whole class. She collects her students' scripts back and corrects her mistake. As a result, everyone in the class gets 2 additional marks. Which of the following summary statistics will change for the class? Select all that apply.

(A) Median.

(B) Standard deviation.

(C) Highest mark.

(D) Quartile 1.

27. Suppose X is a numerical variable and the following are 10 data points for this variable.

$$4,\ 7,\ 4,\ 14,\ 10,\ 11,\ 17,\ 3,\ 8,\ r,$$

where r is a positive whole number that is unknown. Which of the following statements is/are always correct? Select all that apply.

(A) If the mean is greater than 8 then r must be greater than 2.

(B) If r is greater than 2, then the median must be greater than 8.

(C) The mean is always greater than the median regardless of the value of r.

(D) The mode is always greater than the median regardless of the value of r.

28. A school consists of 53 classes. During a budget meeting, school board members decided to review class size information to determine budgeting for the classes. Let x be the numerical variable whose values are the number of students among the 53 classes. Summary statistics for x are shown in the table below.

$\overline{x}$	33.39 students
s_x	5.66 students
min	17
Q_1	29
median	33
Q_3	40
max	40

During the meeting, the following budget is set for classroom stationery supplies. Every class receives \$12 plus an additional \$0.75 for each student in the class. For example, a class with one student receives \$12.75, while a class of 40 students receives \$12 + 40(\$0.75) = \$42. Define a numerical variable y where

$$y = \$(12 + 0.75x).$$

Basically, y takes values that correspond to the amount of money that classes receive for their stationery supplies. Based on the summary statistics for x, which of the following statements must be true regarding the summary statistics of y? Select all that apply.

(A) The maximum value of y is higher than the third quartile of y.

(B) The median of y is $\$12 + 0.75(33) = \$36.75$.

(C) $\overline{y} = 12 + 0.75(33.39) = \$37.04$ (correct to 2 decimal places).

(D) The standard deviation of y is lower than the standard deviation of x.

(E) The IQR for y is the same as the IQR for x.

29. A telecommunication company is interested in understanding how many mobile phones people own. Their population of interest is all 2000 people in town X. They took a random sample of 100 people from town X. Assuming there is 100% response rate, which of the following statements is/are correct?

(I) If among the 100 people sampled, every person has 2 or more mobile phones, then the mean number of mobile phones in the sample will be greater than or equal to 2.

(II) If the mean number of mobile phones in this sample is greater than or equal to 2, then everyone among the 100 people sampled has 2 or more mobile phones.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

30. City planners wanted to know how many people lived in a typical housing unit so they compiled data from hundreds of forms that had been submitted in various city offices. Summary statistics are shown in the table below.

Mean	Standard Deviation	Min	Q_1	Median	Q_3	Max
2.53	1.4	1	1	2	3	8

The city bases their garbage disposal fee on the occupancy level of the home or apartment. The annual fee is \$50 plus \$4 per person, so a single-occupant home pays \$54 and homes with 10 people pay \$50 + \$4 ×10 =\$90 a year.

The median fee paid is _____(1)_____ and the IQR of the fee paid is _____(2)_____ .

Fill in the blanks for the statement above, give your answers correct to 2 decimal places.

This page is blank

Chapter 2

Categorical Data Analysis

Section 2.1 Rates

In Section 1.3, we learnt that there are two main types of variables, namely categorical variables and numerical variables. For categorical variables, there are two sub-types, namely ordinal variables and nominal variables. Ordinal variables are those whose categories come with some natural ordering. On the other hand, there is no intrinsic ordering for the nominal variables. For numerical variables, there are those that are continuous and those that are discrete. The focus of this chapter is on categorical variables and we will discuss numerical variables in the next chapter.

Much of the discussion in this section is centred around the following example.

Example 2.1.1 Suppose a patient newly diagnosed with kidney stones visits his urologist for the first time since diagnosis to discuss what are some of the best possible treatments that he should undergo. In preparation, the urologist took out some historical records of the various patients he had previously and summarised the data into a table. Part of the table is shown below.

Size of stone	Gender of patient	Treatment type (X or Y)	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

Each row of the table is a particular patient that the urologist had seen previously and the columns are the variables related to each patient. While the table only shows the first 6 cases, the data in actual fact contains 1050 observations (or data points). The four variables are

1. The size of the kidney stone. This is an ordinal categorical variable that has two categories. The kidney stones can be classified as either small or large.

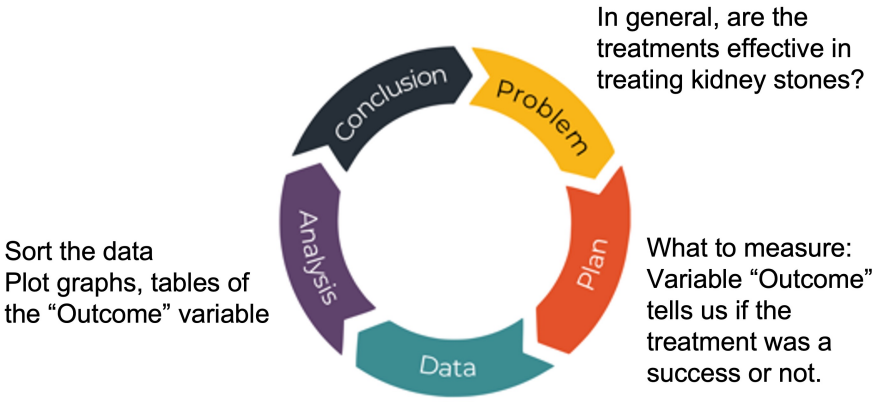
2. The gender of the patient. This is a nominal categorical variable that has two categories, male or female.

3. The treatment that the patient underwent. Again, this is a nominal categorical variable and there are two categories, namely treatment X and treatment Y.

4. The outcome of the treatment is also a nominal categorical variable. The categories are success and failure.

How should the urologist use the 1050 observations to assist in the decision for this new patient?

Before we continue, let us recall the PPDAC cycle that was introduced as the main process behind the approach to a data driven problem.



The overarching question faced by the urologist is simply how to treat his patient better. In particular, **this** new patient. What kind of insights does the data set give the urologist that will enable him to better advise his patient?

To apply the PPDAC cycle to this context, let us start with a question that we want to answer. A simple question to start with is:

- ***Question 1****: Are the treatments given to the patients successful? In other words, should this new patient receive treatment?*

Moving on from "Problem" to "Plan", we next determine what are the variables that needs to be measured and then proceed to obtain data on 1050 previous cases where the *outcome* of the treatment was recorded as either a success or failure. The PPDAC cycle is a continuous process where after looking at the data, drawing some preliminary conclusions might lead to more questions, some of which were even considered from the start. This stage of analysis involves sorting the data, tabulating and plotting graphs of the outcome variable. We may observe interesting trends and this leads us to asking more questions on those trends, leading us back to the top of the cycle again. Some of the new questions that we can ask include

- Do males undergoing treatment X have a higher *rate*[1] of success than females?
- Does treating large kidney stones with treatment X have a higher rate of success than treatment Y?

We will now discuss some of the tables and charts that can be generated from the data that will give us useful information.

Example 2.1.2 (Analysing 1 categorical variable using a table.) Suppose out of the 1050 previous patients, there were 831 records of **success** and 219 records of **failure** after treatment was given. Thus from this simple collation, a preliminary conclusion is that we should generally recommend the new patient to go for treatment since there are more successful outcomes than failed outcomes. We can present this information on the number of success and failures in a table, together with two other columns, namely *rate* and *percentage*.

Categories of the "Outcome" variable	Count	Rate	Percentage
Success	831	$\text{rate(Success)} = \frac{831}{1050} = 0.791$	$0.791 \times 100\% = 79.1\%$
Failure	219	$\text{rate(Failure)} = \frac{219}{1050} = 0.209$	$0.209 \times 100\% = 20.9\%$
Total	1050	$\frac{1050}{1050} = 1$	$1 \times 100\% = 100\%$

[1]The concept of rates will soon be discussed in this section.

The *rate* of successful treatments is simply

$$\frac{\text{Number of successful treatments}}{\text{Total number of treatments}} = \frac{831}{1050} = 0.791.$$

We can also represent this as a *percentage* of total treatments that were successful, which is 79.1%. Similarly, the rate of failed treatments is

$$\frac{\text{Number of failed treatments}}{\text{Total number of treatments}} = \frac{219}{1050} = 0.209.$$

When represented as a percentage, the percentage of failed treatments is 20.9%.

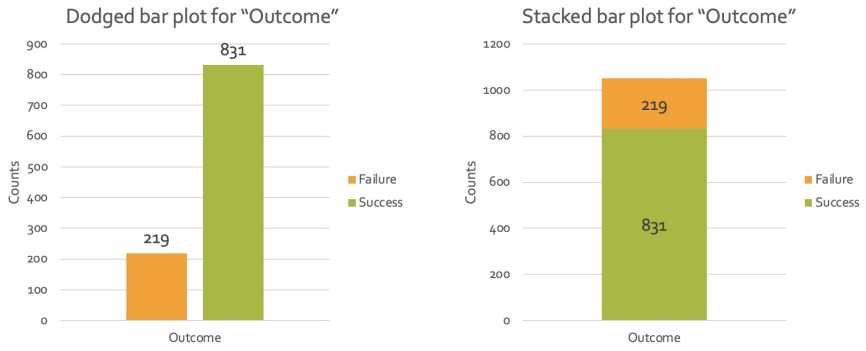
For much of this chapter, we will be using rates in our discussion of the behaviour of categorical variables. Intuitively, we can also think of rate as a fraction, proportion or a percentage. This is useful for understanding some of its properties. For example, we note that

$$0\% \leq \text{rate}(X) \leq 100\% \qquad \text{(if we think of rate as a percentage); or}$$

$$0 \leq \text{rate}(X) \leq 1 \qquad \text{(if we think of rate as a fraction).}$$

(Here, X is some variable of interest.)

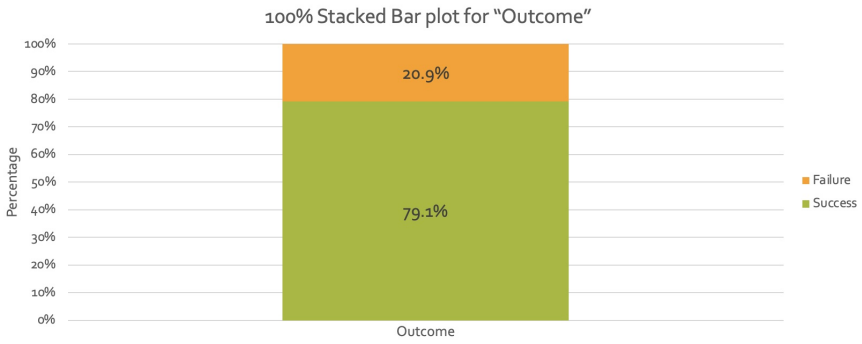
Example 2.1.3 (Analysing 1 categorical variable using a plot.) As an alternative to a table, we can also use easily available softwares to create a plot that presents the data.



Since we are interested in the variable **Outcome**, which is a categorical variable, we can illustrate the *counts* in each of the categories "Success" and "Failure" in the form of a bar plot. The two bar plots above are created using Microsoft Excel.

The bar plot on the left is known as a *dodged bar plot*. The x-axis indicates the variable Outcome whereas the y-axis shows the number (that is, the count) of successes and failures in the variable Outcome. Two bars, one for success counts and the other for failure counts, are placed next to each other. Such an illustration is useful in comparing the relative numbers in the categories.

The bar plot on the right is known as a *stacked bar plot*. The x and y-axes are similar to the dodged bar plot but instead of two bars, we now have only one bar where the counts of failure (219) is stacked on top of the counts of success (831). Such an illustration is useful in comparing the occurrences of each category as a *percentage* or *fraction* of the total number of responses. Instead of showing the absolute numbers in each category, it is also possible to show the percentages directly in the plot itself, as seen in the figure below. However, it should be noted that the y-axis is now giving the percentages rather than the actual numbers.



Regardless of which bar plot is used, we can see that there are many more successes than failures and based on this, it is reasonable to recommend our patient to go for some form of treatment based on the information that we have at this stage.

Remark 2.1.4 In this example on treatment of kidney stones, the success of any treatment is defined as having the kidney stones removed or reduced significantly so that it does not pose any further threat to the patient. On the other hand, failure means that the stones were not able to be removed. In general, kidney stones cause little morbidity and mortality. It is useful to note that for other kinds of illness, where treatments have higher stakes, the conclusion may be different.

Now that we are rather convinced that the new patient should receive treatment, the PPDAC cycle brings us back with new questions that arise from our investigation into the data set of 1050 previous patients. It is reasonable to ask the next question as follows:

- ***Question 2:*** *There are two types of treatment, namely X and Y. Which treatment type is better for our new patient?*

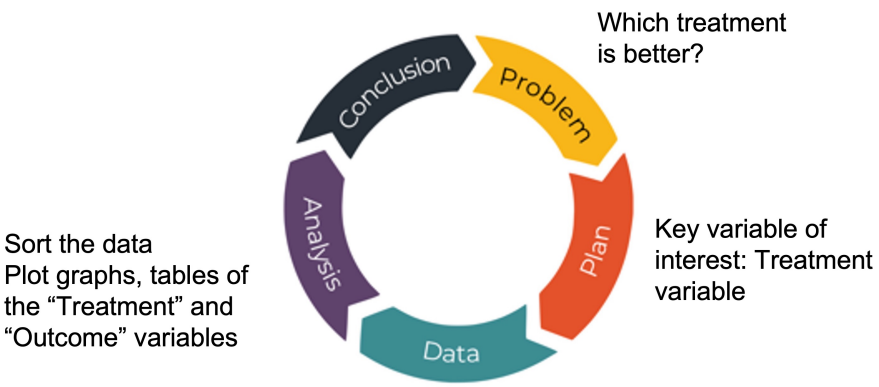
To answer this question, we can revisit the PPDAC cycle and define a new **problem** and **plan** to look at new variable(s) of interest and **analyse** the data again using plots that we have introduced previously.

1. The new problem is as stated above, namely, which treatment is better for our new patient.

2. This means that the key variable that we should look at is the *treatment type* categorical variable, which has two categories, treatment X and treatment Y.

3. This does not mean that treatment type is the only variable of interest, but rather, it should be investigated together with the outcome variable. This is because we want to know how the treatment type affects the outcome.

This leads us to our discussion of how to analyse two categorical variables.



Example 2.1.5 (Analysing 2 categorical variables using a table.) When we used a table to analyse 1 categorical variable (Outcome), the table showed only the number of successes and failures among the 1050 previously treated patients. When we introduce a second categorical variable (Treatment type), we have a 2×2 *contingency table* that will summarise the two variables across the 4 (that's why it is called 2×2) possible combinations of (Treatment, Outcome).

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

Recall that out of 1050 previous patients, 831 underwent successful treatments while the other 219 were failed treatments. The 2×2 table breaks down the 831 successful treatments according to the treatment type. As seen from the *Success* column, 542 were given treatment X while 289 were given treatment Y. Similarly, for the 219 failed treatments, 158 were given treatment X while 61 were given treatment Y.

If we look across a row instead of down a column, we could, for example, see that there were 700 previous patients given treatment X, of which 542 were successful and 158 failed. Similarly, looking at the row for treatment Y, we see that out of 350 people who underwent treatment Y, 289 of them had successful treatments while 61 did not.

Remark 2.1.6

1. It should be noted that by convention, the dependent variable *Outcome* is placed on the columns on the table while the independent variable *Treatment type* is placed on the rows.
2. The column total values for the success (831) and failure (219) columns should add up to the same values as the sum of the row total values for Treatment X (700) and Treatment Y (350), which obviously should both add up to the total number of data points in the data set which is 1050.

Discussion 2.1.7 In order to answer ***Question 2***, it will be useful to ask other related questions, for example:

1. ***Question 2a***: What proportion of the total number of patients were given treatment Y (or X)?
2. ***Question 2b***: Among those patients given treatment X, what proportion were successful?
3. ***Question 2c***: What proportion of patients were given treatment Y and had a failed treatment outcome?

To answer ***Question 2a***, we note that there were 350 previous patients who underwent treatment Y. The *proportion* of the total number of patients that underwent treatment Y is

$$\frac{350}{1050} = \frac{1}{3} = 33\frac{1}{3}\%.$$

We can also denote this as

$$\text{rate(Y)} = \tfrac{1}{3} \text{ or } 33\tfrac{1}{3}\%.$$

We have seen earlier that out of 1050 patients, there were 831 successful treatments, so we can write rate(Success) = 0.791 or 79.1%. We know that

$$\text{rate(X)} = \tfrac{700}{1050} = \tfrac{2}{3} \text{ or } 66\tfrac{2}{3}\%.$$

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	**350**
Column Total	**831**	219	**1050**

Notice that in the calculations above, we have used two numbers in the margin of the table (for example, **831** and **1050**) that relate to just one of the categorical variables (Outcome) each time, we call these *marginal rates*. Similarly, rate(Y) $= \frac{350}{1050} = \frac{1}{3}$ is also a marginal rate.

How should we answer ***Question 2b***? In this case, we need to zoom in onto the patients who had undergone treatment X and figure out what proportion of them have had a successful treatment.

Referring to the table again, we see that out of 700 patients who were given treatment X, 542 of them were successfully treated. Hence the proportion of successful treatments was

$$\frac{542}{700} = 0.774 = 77.4\%.$$

This *rate* of success is computed based on only those patients who were under treatment X, which sets the *condition* for the calculation of the rate. Once such a condition is set, those patients on treatment X will be considered as the population and those on treatment Y will not be part of any consideration. Such a rate is known as a *conditional rate*, which is one that is based on a given condition.

A note on the notation used for conditional rates is that we replace the word "given" by a vertical bar so that rate(Success given treatment X) is written as

$$\text{rate(Success} \mid \text{X)}.$$

Let us consider ***Question 2c***. From the table, we can see easily that there are 61 cases where treatment Y was given but had an unsuccessful outcome.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	**61**	350
Column Total	831	219	1050

So the rate of patients who were given treatment Y **AND** had a failure was

$$\frac{61}{1050} = 0.0581 = 5.81\%.$$

This rate is known as a *joint rate* and it is not a conditional rate since we are looking at all the 1050 patients as our baseline. In other words, we are now considering patients on treatment X, as well as patients on treatment Y as the population. One should be careful with the implicit difference in the phrasing of the two statements:

- What proportion of patients were given treatment Y and had an unsuccessful outcome?

  Answer: rate(Unsuccessful and Y) $= \frac{61}{1050} = 0.0581$.

- What proportion of patients given treatment Y had an unsuccessful outcome?

  Answer: rate(Unsuccessful | Y) $= \frac{61}{350} = 0.174$.

The first question refers to the joint rate/proportion/percentage while the second question refers to the conditional rate/proportion/percentage.

Discussion 2.1.8 It should be clear at this point from our discussion of rates and proportions that our decision on which treatment to suggest to our new patient cannot be based simply on the absolute number of successes and failures for each treatment type. If we had based the decision on absolute numbers, we would have gone for treatment X since there were 542 success cases compared to only 289 for treatment Y.

The reason why we should look at rates rather than absolute numbers is because the number of patients undergoing each treatment is **different**, so it would not be surprising if there were more successful cases for treatment X because there were just more patients given this treatment, rather than because it is more effective. Finding the rate of success for each treatment before comparing them is a form of *normalisation*. At this stage of our analysis, when using the success rates to compare the treatment types, our conclusion is to recommend treatment Y to our patient.

- The rate of success *given* treatment X is the conditional rate we have already calculated in answering ***Question 2b***, which is
$$\text{rate}(\text{Success} \mid \text{X}) = \frac{542}{700} = 0.774 = 77.4\%.$$
Similarly, we can calculate the rate of success *given* treatment Y, which is
$$\text{rate}(\text{Success} \mid \text{Y}) = \frac{289}{350} = 0.826 = 82.6\%.$$

- We can also look at the conditional rates in another way. For treatment X, having the conditional rate of success to be 77.4% means that out of 100 patients who underwent treatment X, 77 of them had a successful outcome. For treatment Y, the numbers were 83 successes out of 100 patients receiving this treatment.

- As the rate of success for treatment Y is higher, we can now say that treatment Y is better than treatment X and advise the patient appropriately. Notice that we would have given the opposite advice if we were looking at absolute numbers instead of rates, which is incorrect.

- We can now add in the rates to the 2×2 contingency table given at the beginning of Example 2.1.5.

Treatment \ Outcome	Success	Failure	Row Total
X	542 (77.4%)	158 (22.6%)	700 (100%)
Y	289 (82.6%)	61 (17.4%)	350 (100%)
Column Total	831(79.1%)	219 (20.9%)	1050 (100%)

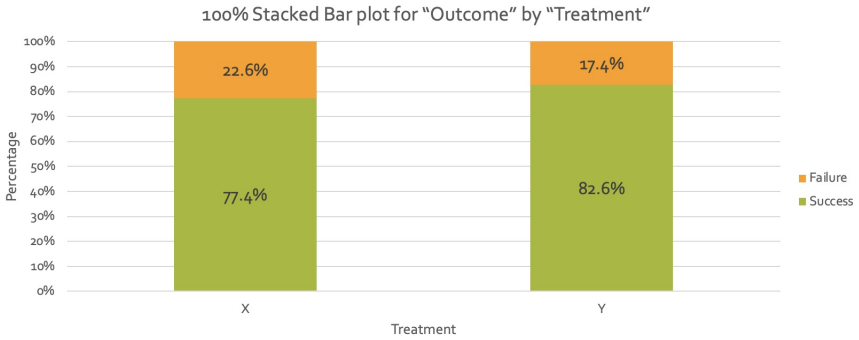
Example 2.1.9 (Analysing 2 categorical variables using a plot.) In Example 2.1.3, we introduced dodged bar plots and stacked bar plots to present the data on a single variable **Outcome**. We can also use these plots to present the counts of **Outcome** broken down by **Treatment**. These were the two variables we analysed using a table in Example 2.1.5.

The dodged bar plot on the left shows the success and failure counts for both treatments X and Y. The numbers above each bar is the success or failure count for that particular treatment type.

The stacked bar plot on the right shows the same information but with the success and failure bars under the same treatment stacked instead of being placed side by side.

Both bar plots tell us that there are a lot more successful treatments in treatment X than in treatment Y, which may lead to the conclusion that treatment X is more effective (since the green bars for treatment X are bigger than the green bars for treatment Y). However, it is also obvious from the stacked bar plot that these two treatments have very different number of patients (represented by the height of the two bars).

Similar to our analysis using tables, we can also create the plots using rates instead of absolute numbers.



In this plot, notice that both the treatment X and treatment Y bars have been *normalised* to the same *height* (which is 100%). We are no longer comparing absolute numbers, but instead comparing the rates of success (the height of the green bars, as a proportion of the total height) between the two treatments. We can see immediately that treatment Y has a higher rate of success (taller green bar) compared to treatment X.

To summarise this section, we have discussed how we can analyse two categorical variables. This can be done either using a 2×2 contingency table, or bar plots (dodged or stacked) which makes it easier for us to observe any differences between the categories. We also introduced the concept of rates, as a means of fair comparison when group sizes are unequal. To formally discuss the relationship between two categorical variables, we will introduce the concept of *association* in the next section.

Section 2.2 Association

Definition 2.2.1 In Section 2.1, we considered the example of two different treatments for patients with kidney stones. Let's say that initially, we guessed that the treatment type involved does not affect the outcome of the treatment, meaning that we could advise our patient to undergo either treatment because the outcome would not be affected. If this was the case, then we can say that the treatment type is not related to the outcome of the treatment.

After analysing the data using rates, we found that this was not the case. There was a higher success rate observed for patients under treatment Y compared to those under treatment X. Due to the difference in success rates, we say that there is a *relationship* between the type of treatment and the outcome of the treatment.

To formalise the notion of such a *relationship*, we say that treatment type is *associated* with the outcome of the treatment. More specifically, treatment Y is *positively associated* with the success of the treatment. What this means is that treatment Y and successful treatments tend to occur together.

On the other hand, we say that treatment X is *negatively associated* with the success of the treatment. This is because we tend to see treatment X and failed treatments go hand in hand.

Remark 2.2.2

1. Note that treatment X being negatively associated with the success of the treatment does not mean that a significant proportion of patients undergoing treatment X will see the treatment fail (77.4% of them still recorded success). The negative association is stated as a **comparison** between the two treatment types X and Y, where in this case treatment Y tends to produce **more successful outcomes**.

2. We should be conscious of the choice of the word **associated** because we do not know if the outcome of the treatment was entirely **due to** the treatment type received. The data we had came from an observational study hence it might be erroneous for us to say that the type of treatment and the outcome of the treatment have a *causal* relationship. It is important to see the distinction between *association* and *causation* and for the rest of this chapter, we will be focussing on discussing associative relationships between categorical variables rather than causal relationships.

Discussion 2.2.3 So how do we identify an association between two variables? Suppose the two variables we are considering represent two characteristics in a population. Let us call these two characteristics A and B. For example, A could be *smoker* (so one categorical variable could be smoking habit, with two categories *smoker* and *non-smoker*) while B could be *male* (so the other categorical variable could be gender, with two categories *male* and *female*). The population can be a well-defined group of people. In the population, those "with A", refers to smokers, while "without A", denoted by NA refers to non-smokers. Similarly, those "with B" refers to male and those "without B", denoted by NB, refers to female.

So if the rate of A given B (proportion of smokers among males) is the same as the rate of A given NB (proportion of smokers among females), then it means that the rate of A is not affected by the presence or absence of B. Thus in this case, there is no difference in the proportion of smokers between both gender groups and we write

$$\text{rate}(\text{A} \mid \text{B}) = \text{rate}(\text{A} \mid \text{NB}).$$

However, if the rate of A given B is not the same as the rate of A given NB, then there are two possible situations.

- The first possibility is the rate of A given B is more than the rate of A given NB. This means that the presence of A **when B is present** is stronger compared to **when B is absent**. Hence we say that there is *positive association* between A and B. In this case, we write

  $$\text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB})$$

  and for the gender/smoking example, this means that there is a higher proportion of smokers among males than the proportion of smokers among females. So being male and smoking are positively associated.

- The other possibility is the rate of A given B is less than the rate of A given NB. This means that the presence of A **when B is present** is weaker compared to **when B is absent**. Hence we say that there is *negative association* between A and B. In this case, we write

  $$\text{rate}(\text{A} \mid \text{B}) < \text{rate}(\text{A} \mid \text{NB})$$

  and for the gender/smoking example, this means that there is a lower proportion of smokers among males than the proportion of smokers among females. So being male and smoking are negatively associated.

The inequality $\text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB})$ (resp. $\text{rate}(\text{A} \mid \text{B}) < \text{rate}(\text{A} \mid \text{NB})$) is not the only one that allows us to conclude that there is positive (resp. negative) association between A and B. The table below provides three other comparisons between rates that leads to the same conclusion of a positive (resp. negative) association between A and B. These different comparisons are mathematically equivalent to each other and their equivalence will be established using the *symmetry rule* in Discussion 2.3.1.

Establishing association	
Positive association between A and B: (any of the following)	Negative association between A and B: (any of the following)
rate(A | B) > rate(A | NB)	rate(A | B) < rate(A | NB)
rate(B | A) > rate(B | NA)	rate(B | A) < rate(B | NA)
rate(NA | NB) > rate(NA | B)	rate(NA | NB) < rate(NA | B)
rate(NB | NA) > rate(NB | A)	rate(NB | NA) < rate(NB | A)

Example 2.2.4 Let us revisit the earlier example on two different treatments for kidney stones. Recall that the two variables were treatment outcome and treatment type. For the treatment outcome variable, let us split the patients into group A, which is the group of patients with successful outcomes and the group NA will be those with unsuccessful outcomes. For the other variable, we will also split the patients into group B for those given treatment X and group NB for those given treatment Y.

Let us revisit some conditional rates that were computed previously.

1. Rate(A | B) = rate(Success | X) = $\frac{542}{700} = 0.774$.
2. Rate (A | NB) = rate(Success | Y) = $\frac{289}{350} = 0.826$.

Since

$$\text{rate(A | B)} < \text{rate(A | NB)},$$

we can say that the presence of A when B is present is weaker than the presence of A when B is absent. Thus there are fewer successful treatments when looking at treatment X compared to treatment Y and hence success of the treatment is *negatively* associated with treatment X.

Conversely, since there are more successful treatments when looking at treatment Y compared to treatment X, we can conclude that success of the treatment is *positively* associated with treatment Y.

Section 2.3 Two rules on rates

Discussion 2.3.1 In this section, we will discuss two important rules regarding rates. Suppose we have a population with two population characteristics A and B. Among the population there are those who possess characteristic A and those who do not.

For ease of notation, we will denote those who possess characteristic A simply as "A" and those who do not as "NA". Similarly for characteristic B, those in the population who possess this characteristic will be denoted as "B" and those who do not as "NB".

(Symmetry rule)

The first rule that we will be discussing is known as the *symmetry rule*. Although there are three parts to this rule, once we can understand the first part, the second and third parts will follow naturally.

Symmetry Rule Part 1:

$$\text{rate(A | B)} > \text{rate(A | NB)} \Leftrightarrow \text{rate(B | A)} > \text{rate(B | NA)}.$$

The above rule states that the rate of A given B is **more than** the rate of A given NB (call this statement 1) **if and only if** the rate of B given A is **more than** the rate of B given NA (call this statement 2). The *if and only if* here, denoted by $\Leftrightarrow$, means that statements 1 and 2 happen together, meaning that if one of the statements is true, the other one will also be true. In other words, the two statements are either *both correct* or *both incorrect*. Another way of understanding (statement 1) if and only if (statement 2) is

- If (statement 1) holds, then (statement 2) must hold; AND
- If (statement 2) holds, then (statement 1) must hold.

Suppose we know that

$$\text{rate(A} \mid \text{B) is more than rate(A} \mid \text{NB)}, \tag{1}$$

then we can safely say that

$$\text{rate(B} \mid \text{A) is more than rate(B} \mid \text{NA)}. \tag{2}$$

Why is this so? Let us try to explain this logically.

1. If rate(A | B) is more than rate(A | NB), which is (1), then this means that there is a positive association between A and B;
2. This means that we are more likely to see A when B is present, compared to when B is absent;
3. This in turn means that we are more likely to see B when A is present, compared to when A is absent;
4. Hence rate(B | A) is more than rate(B | NA), which is (2).

This is the same as saying that A and B are positively associated. Conversely, suppose we know that

$$\text{rate(B} \mid \text{A) is more than rate(B} \mid \text{NA)}, \tag{2}$$

then we can safely say that

$$\text{rate(A} \mid \text{B) is more than rate(A} \mid \text{NB)}. \tag{1}$$

The logical explanation is similar in nature.

1. If rate(B | A) is more than rate(B | NA), which is (2), then this means that there is positive association between B and A;
2. This means that we are more likely to see B when A is present, compared to when A is absent;
3. This in turn means that we are more likely to see A when B is present, compared to when B is absent;
4. Hence rate(A | B) is more than rate(A | NB), which is (1).

We have now seen Part 1 of the Symmetry Rule. Parts 2 and 3, as shown below can be similarly explained.

Symmetry Rule Part 1:

$$\text{rate(A} \mid \text{B)} > \text{rate(A} \mid \text{NB)} \Leftrightarrow \text{rate(B} \mid \text{A)} > \text{rate(B} \mid \text{NA)}.$$

Symmetry Rule Part 2:

$$\text{rate(A} \mid \text{B)} < \text{rate(A} \mid \text{NB)} \Leftrightarrow \text{rate(B} \mid \text{A)} < \text{rate(B} \mid \text{NA)}.$$

Symmetry Rule Part 3:

$$\text{rate(A} \mid \text{B)} = \text{rate(A} \mid \text{NB)} \Leftrightarrow \text{rate(B} \mid \text{A)} = \text{rate(B} \mid \text{NA)}.$$

We will now present a mathematical derivation for Part 1 of the Symmetry Rule. You are encouraged to go through the same process for Parts 2 and 3. Consider a 2×2 contingency table shown below representing variables A and B. Let w, x, y, z denote the counts in each of the 4 cells.

	B	Not B	Row total
A	w	x	$w+x$
Not A	y	z	$y+z$
Column total	$w+y$	$x+z$	$w+x+y+z$

Symmetry Rule Part 1 states that

$$\text{rate(A} \mid \text{B)} > \text{rate(A} \mid \text{NB)} \Leftrightarrow \text{rate(B} \mid \text{A)} > \text{rate(B} \mid \text{NA)}.$$

From the contingency table, the inequality on the left means

$$\begin{aligned} & \frac{w}{w+y} > \frac{x}{x+z} \\ \text{which implies} \quad & w(x+z) > x(w+y), \\ \text{or equivalently} \quad & wx + wz > xw + xy, \\ \text{and thus} \quad & wz > xy. \end{aligned}$$

On the other hand, the inequality on the right means

$$\begin{aligned} & \frac{w}{w+x} > \frac{y}{y+z} \\ \text{which implies} \quad & w(y+z) > y(w+x), \\ \text{or equivalently} \quad & wy + wz > yw + yx, \\ \text{and thus} \quad & wz > xy. \end{aligned}$$

Hence, both inequalities are in fact equivalent to $wz > xy$ and thus they are equivalent.

Example 2.3.2 Let us revisit our kidney stones treatment example. The 2×2 contingency table below gives us the number of patients in each treatment type as well as the number of success and failure outcomes for each treatment type.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

We have earlier shown that A (representing treatment outcome) is associated with B (representing treatment type) since

$$\begin{aligned} \text{rate(A} \mid \text{B)} &= \text{rate(Success} \mid \text{X)} \\ &= 0.774 \\ &< 0.826 = \text{rate(Success} \mid \text{Y)} = \text{rate(A} \mid \text{NB)}. \end{aligned}$$

By symmetry rule part 2, we should have rate(B | A) < rate(B | NA). Let us verify that this is indeed the case.

$$\begin{aligned} \text{rate(B} \mid \text{A)} &= \text{rate(X} \mid \text{Success)} \\ &= \frac{542}{831} \quad \text{(since there are 831 successful cases of which 542 came from treatment X)} \\ &= 0.652 \\ \text{rate(B} \mid \text{NA)} &= \text{rate(X} \mid \text{Failure)} \\ &= \frac{158}{219} \quad \text{(since there are 219 failure cases of which 158 came from treatment X)} \\ &= 0.721. \end{aligned}$$

Since $0.652 < 0.721$, we have thus verified that rate(B | A) < rate(B | NA) as predicted by symmetry rule part 2. This also confirms that there is negative association between success of treatment (A) and treatment X (B).

Discussion 2.3.3 (Basic rule on rates.) The second rule on rates is known as the *basic rule on rates*. The main rule, as well as three consequences of the main rule are shown below.

Basic rule on rates:

The overall rate(A) will always lie between rate(A | B) and rate(A | NB).

Consequence 1:

The closer rate(B) is to 100%, the closer rate(A) is to rate(A | B).

Consequence 2:

If rate(B) = 50%, then rate(A) = $\frac{1}{2}$[rate(A |B) + rate(A | NB)].

Consequence 3:

If rate(A | B) = rate(A | NB), then rate(A) = rate(A | B) = rate(A | NB).

1. The basic rule on rates states that the overall rate(A) is always between the conditional rates of A given B and A given not B.

2. The first consequence gives us a little more indication of **where** the overall rate(A) is going to be. If rate(B) is closer to 100% (than rate(NB)), then rate(A) is going to be closer to rate(A | B) compared to rate(A | NB).

3. The second consequence specifically states that if rate(B) is **exactly** 50%, then rate(A) will be **exactly** the mid point between rate(A | B) and rate(A | NB).

4. Finally, the third consequence states that if the two conditional rates, namely rate(A | B) and rate(A | NB) are the same, then the overall rate(A) will also take the same value of the two conditional rates.

At this point, the significance of the basic rule and its consequences are not immediately apparent or intuitive. The best way towards understanding them is through an example.

Example 2.3.4 Suppose a school has two different classes (call them class Bravo and class Charlie) of students who took the same data analytics examination. We are interested in studying the passing rate of students at the entire school level and also at each individual class level. Suppose we are given the following information:

(1) The passing rate of students from Bravo is 75%.

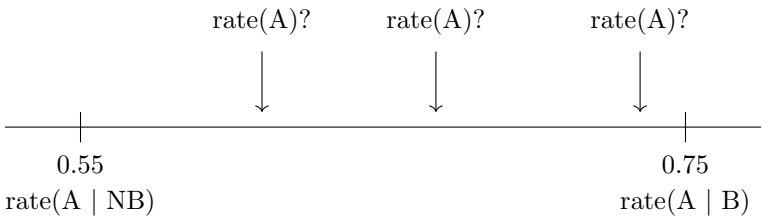
(2) The passing rate of students from Charlie is 55%.

For convenience, let us denote class Bravo as "B" and class Charlie as "NB" (not B). Similarly, denote passing the examination as "A" and not passing as "NA". So the two pieces of information we have above are:

$$\text{rate(A | B)} = 0.75 \quad \text{and} \quad \text{rate(A | NB)} = 0.55.$$

By the basic rule on rates, the overall passing rate of all students from the school, that is, rate(A) is between the two conditional rates,

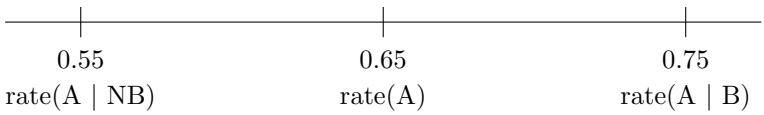
$$0.55 = \text{rate(A | NB)} \leq \text{rate(A)} \leq \text{rate(A | B)} = 0.75.$$



However, without any further information, we will not be able to determine the exact value of rate(A). What about the three consequences of the basic rule?

1. The first consequence states that the closer rate(B) is to 100%, then the closer rate(A) will be to rate(A | B). In our example, it means that if the number of students in Bravo is far more than the number of students in Charlie (thus rate(B) is closer to 100%), then the overall passing rate of the school (that is, rate(A)) will be closer to the passing rate of class Bravo (that is, rate(A | B)).

2. The second consequence states that if rate(B) = 50% (which also means that rate(NB) = 50%), then
$$\text{rate(A)} = \frac{1}{2}\left[\text{rate(A} \mid \text{B)} + \text{rate(A} \mid \text{NB)}\right].$$
That is, rate(A) will be right in between the two conditional rates. In our example, this means that if the number of students in Bravo and Charlie are exactly the same, then the overall passing rate of the school will be exactly in between 0.55 and 0.75, that is 0.65.



3. The third consequence states that if the two conditional rates rate(A | B) and rate(A | NB) are the same, then the overall rate(A) will be the same value as the two conditional rates. In our example, if the passing rates of class Bravo and class Charlie are the same, then the overall passing rate of the school will be the same as the passing rate in either class.

Example 2.3.5 Let us continue with Example 2.3.4 and validate the basic rule on rates and the consequences by considering actual numbers.

1. Suppose the total number of students and the number of passes in each of the two classes are given in the table below.

	Total number of students	Number of passes	Passing rate
Bravo	600	450	$\frac{450}{600} = 0.75$
Charlie	80	44	$\frac{44}{80} = 0.55$
School	680	494	$\frac{494}{680} = 0.73$

Notice that the passing rates of both classes are what they are supposed to be, but the number of students in Bravo far exceeds the number in Charlie (so rate(B) is $\frac{600}{680}$ which is closer to 100%). While the overall school passing rate is between 0.55 and 0.75 (in accordance to the basic rule on rates), it is much closer to the passing rate of Bravo, as predicted by consequence 1.

2. Suppose the total number of students and the number of passes in each of the two classes are as given below instead:

	Total number of students	Number of passes	Passing rate
Bravo	600	450	$\frac{450}{600} = 0.75$
Charlie	600	330	$\frac{330}{600} = 0.55$
School	1200	780	$\frac{780}{1200} = 0.65$

Again, the passing rates of both classes are what they are supposed to be, but in this case, the number of students in Bravo and Charlie are the same (so rate(B) = rate(NB) = 0.5). As predicted by consequence 2, the overall school passing rate will be 0.65, which is right in between the two class passing rates.

3. To illustrate consequence 3, suppose the passing rates for both classes are the same, as shown below.

	Total number of students	Number of passes	Passing rate
Bravo	600	450	$\frac{450}{600} = 0.75$
Charlie	400	300	$\frac{300}{400} = 0.75$
School	1000	750	$\frac{750}{1000} = 0.75$

Now the two conditional rates, namely rate(A | B) and rate(A | NB) are equal. By consequence 3, rate(A) will be the same value as the two conditional rates. This is indeed the case as we see that the two classes have the same passing rate which will result in the school having the same passing rate of 0.75. It is important to note that we **do not require** rate(B) to be the same as rate(NB) for consequence 3 to hold. For our example, this means that we do not require classes Bravo and Charlie to have the same number of students. As long as the two class passing rates are the same, consequence 3 will hold.

Finally, let us verify the rule on rates using our kidney stones data set.

Example 2.3.6 We have seen the following table from Example 2.3.2.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

- The conditional rates of success among the two treatment types are:

$$\text{rate(Success} \mid \text{X)} = \frac{542}{700} = 0.774,$$

$$\text{rate(Success} \mid \text{Y)} = \frac{289}{350} = 0.826.$$

The overall rate of success is

$$\text{rate(Success)} = \frac{831}{1050} = 0.791,$$

which is closer to the rate of success among patients with treatment X. This agrees with the basic rule on rates since there were more patients with treatment X (66.67%) compared to treatment Y (33.33%).

In the three sections of this chapter we have discussed so far, we have seen how we can use the concept of rates to investigate relationships, in particular, association between categorical variables. Very often, exact rates (overall or conditional) are unknown to us but if we can apply some general rules like the

symmetry rule, basic rule or the consequences of the basic rule, we can still obtain valuable insights into the data set we have on our hands. Making the best use of limited information is an important skill when analysing data.

In the next section, we will discuss a surprising observation that can be counterintuitive to some but is very important for anyone analysing data to be aware of.

Section 2.4 Simpson's Paradox

Discussion 2.4.1 From earlier sections, when faced with the problem of advising our new kidney stones patient, we have gone through two cycles of the PPDAC process.

The first question we asked was whether having any sort of treatment was better than not having one. By comparing the rate of success (0.791) versus the rate of failure (0.209), we conclude that there are many more successful than failed treatments from past records, so the decision was to advise our new patient that he should be treated.

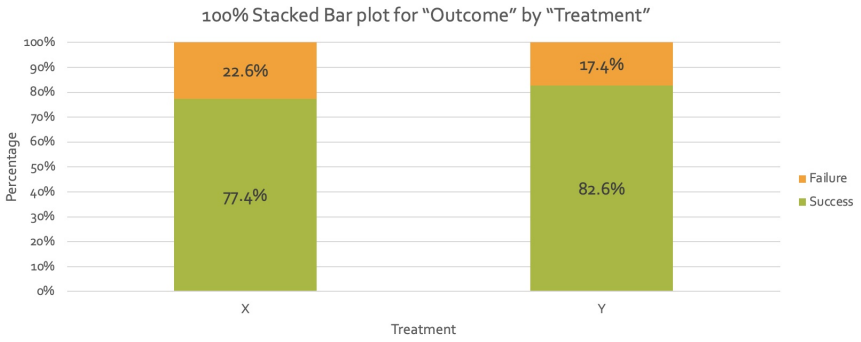
But this led us to the next question, as there are two treatment types available, which type of treatment should we recommend? This made us go back to the data and compare the success rates of those patients who were given treatment X as opposed to the success rates of those given treatment Y. Upon delving deeper into the data, we discovered that treatment Y is positively associated with success rate. This suggests that treatment Y is "better" than treatment X and perhaps we should advise our patient to undergo treatment Y.

Are we done with our analysis? Is there some lingering doubt in our minds that we may be providing wrong advice to our patient? If we are convinced that treatment Y is better, should we **always** send kidney stone patients for treatment Y? If not always, then when do we do so? What should our decision be based on? These are again questions that prompt us to go back to our data and see if more information can be obtained from it.

Size of stone	Gender of patient	Treatment type (X or Y)	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

The table above from Example 2.1.1 shows there are two other variables that we have not used in our analysis thus far, namely the size of the kidney stone and also the gender of the patient. Would these variables be an important factor in our consideration? How should we go about analysing them? Let us begin by exploring the stone size variable.

Example 2.4.2 (Analysing 3 categorical variables using a plot.) In Example 2.1.9, we used a stacked bar plot for "Outcome" by "Treatment" to compare the success rates for treatments X and Y.

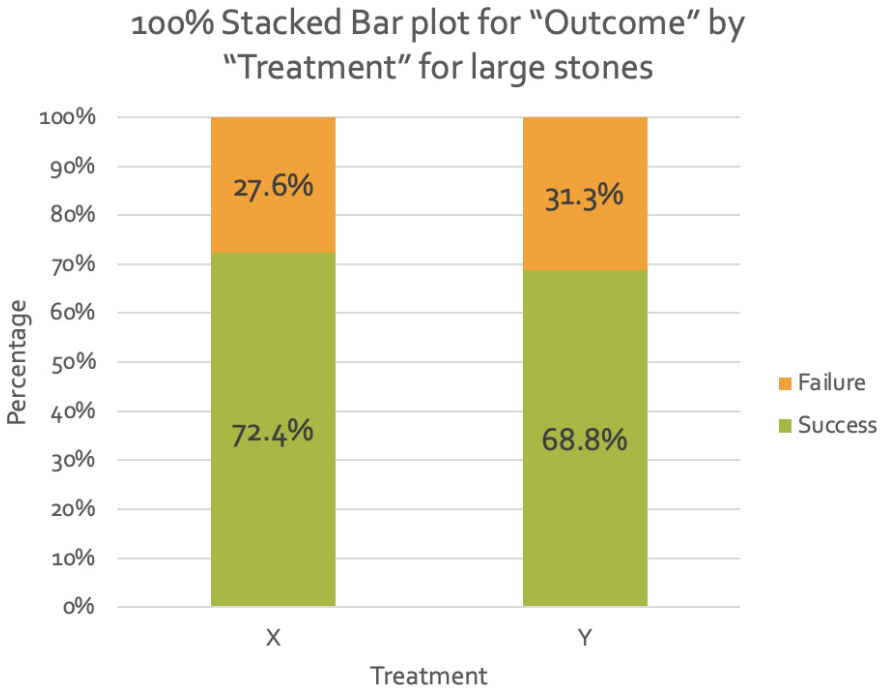


From the stacked bar plot, we have concluded that treatment Y is positively associated with success. We have not taken stone sizes into consideration thus far and the plot was made based on simply counting the number of successes and failures across all stone sizes. In other words, this plot gave us the **overall success rates** of treatments X and Y.

Let us now separate the data by considering the categorical variable of "stone size" which has two categories, namely *large stones* and *small stones*.

Large stones	Success	Failure	Total
X	381	145	526
Y	55	25	80
Total	436	170	606

The table above shows the outcome of treatments given to patients with large stones. For example, out of 526 treatment X patients with large kidney stones, 381 had a successful outcome and 145 were unsuccessful. Similarly, out of 80 treatment Y patients with large kidney stones, 55 were successful while 25 were not. We can present these information using a stacked bar plot like before, as shown below[2].



[2]Notice that for the bar plot for treatment Y, the two percentages do not add up to 100%. This is due to rounding off in Excel, where the success percentage is in fact 68.75% and the failure percentage is 31.25%.

How do the two different treatments compare? Although the margin of difference is not very big, there is no doubt that treatment X has a *higher* success rate of 72.4% compared to treatment Y, which has a success rate of 68.8%. This means that, for treating large kidney stones,

$$\frac{381}{526} = 0.724 = \text{rate}(\text{Success} \mid \text{X}) > \text{rate}(\text{Success} \mid \text{Y}) = 0.688 = \frac{55}{80},$$

and thus treatment X is positively associated with success for treating large stones. This observation is surprising, since we have already concluded that **for all stone sizes combined together**,

$$\text{rate}(\text{Success} \mid \text{X}) < \text{rate}(\text{Success} \mid \text{Y}),$$

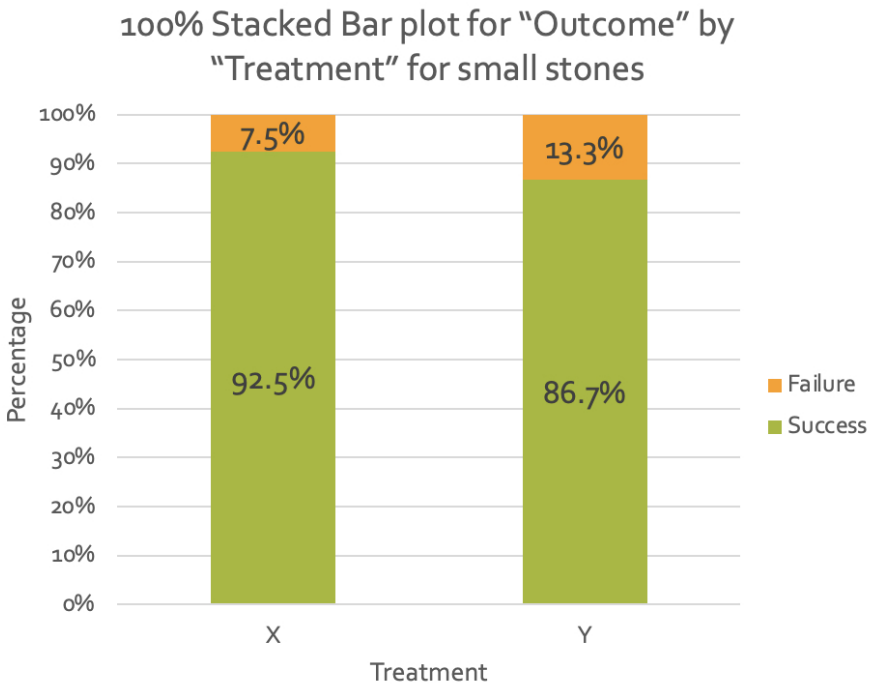
that is, treatment X is negatively associated with success if we do not segregate by stone size.

Why are we observing a different behaviour for large kidney stones as opposed to what we saw earlier when all kidney stone sizes are combined?

Let us consider the data for small kidney stones.

Small stones	Success	Failure	Total
X	161	13	174
Y	234	36	270
Total	395	49	444

The table above shows the outcome of treatments given to patients with small stones. For example, out of 174 treatment X patients with small kidney stones, 161 had a successful outcome and 13 were unsuccessful. Similarly, out of 270 treatment Y patients with small kidney stones, 234 were successful while 36 were not. Let us again present these data using a stacked bar plot.

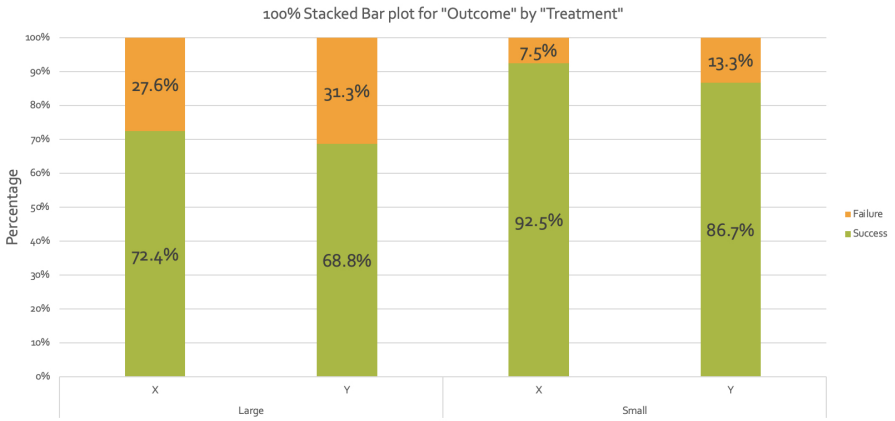


The margin of difference between the two treatment types is again not very big, but again we see that treatment X has a *higher* success rate of 92.5% compared to treatment Y, which has a success rate of 86.7%. This means that, for treating smaller kidney stones,

$$\frac{161}{174} = 0.925 = \text{rate}(\text{Success} \mid \text{X}) > \text{rate}(\text{Success} \mid \text{Y}) = \frac{234}{270} = 0.867,$$

so again treatment X is positively associated with success for treating small stones, which is the opposite from what we had when the data was combined and not segregated by stone size.

We can now combine the two previous plots by putting them side by side as shown below.



Notice that the first two bars from the left are for large kidney stones data while the last two bars are for small kidney stones. This type of plot is sometimes referred to as a *sliced stacked bar plot*. Such a plot can be used for comparing across three categorical variables. The three variables here are stone size, treatment outcome and treatment type.

We are now facing a *paradox*. Although treatment Y is the better treatment overall, when the stone sizes are combined and not segregated, we see that if we focus only on the large stones, or only on the small stones, treatment X is observed to have higher success rate that treatment Y. This is indeed strange!

This phenomenon is known as *Simpson's Paradox*.

> Simpson's Paradox:
>
> Simpson's Paradox is a phenomenon in which a trend appears in more than half of the groups of data but disappears or reverses when the groups are combined. Here, "disappears" means the two variables in question (say A and B) are no longer associated, that is, rate(A | B) = rate(A | NB).

We are now back to the same question which we thought we have already answered: Which treatment is better for our patient? Should we advise him to undergo treatment X or Y?

Remark 2.4.3 In the example of kidney stones, there were only two subgroups for the stone size, namely, small and large. We claim that Simpson's Paradox was observed because the trend in **both** subgroups is different from the trend observed when the subgroups are combined.

In examples where there are more than two subgroups, we will say that Simpson's Paradox is observed as long as a majority of the individual subgroup rates shows the opposite trend to the overall rate. For example, if there are three subgroups, as long as there are at least 2 subgroups showing the opposite trend to the overall rate, we can say that Simpson's Paradox is observed.

Example 2.4.4 (Analysing 3 categorical variables using a table.) Let us put the two tables in Example 2.4.2 for both the large and small kidney stones together into one unified table.

	Large stones			Small stones			Total (Large+Small)		
	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %
X	381	**526**	**72.4%**	161	**174**	**92.5%**	542	700	77.4%
Y	55	**80**	68.8%	234	**270**	86.7%	289	350	**82.6%**

To recap, we have two different treatment types, X and Y. In the row for treatment X, we see that there were 526 large stones cases that were under treatment X, of which 381 were successful. This gives a success rate of 72.4%. Similarly, in the row for treatment Y, we see that there were 270 small stones

cases that were under treatment Y, of which 234 were successful. This gives a success rate of 86.7%. The last 3 columns of the table gives the combined numbers for both stone sizes.

Recall we had initially concluded that treatment Y was the better treatment because **82.6**% of patients who were given treatment Y had a successful outcome, compared to 77.4% for treatment X. We then separated the cases according to the size of the stone, i.e., we created subgroups and this method of subgroup analysis is called *slicing*.

This is when we observed Simpson's Paradox, where the rate of success amongst small (**92.5**%) and large (**72.4**%) stones is higher for treatment X compared to treatment Y. This **reverses** the trend observed when the small and large kidney stones were combined.

Let us look at the numbers highlighted in blue in the table. A crucial observation at this point is that treatment X seems to be used to treat mostly patients with large stones as compared to small stones. Thus, by the **basic rule on rates**, we know that the overall success rate of treatment X will be closer to the large stones success rate of 72.4% than the small stones success rate of 92.5%. Indeed, we have the overall treatment X success rate to be 77.4%.

Turning our attention to the numbers highlighted in orange in the table, we observe the opposite of the above. Treatment Y seems to be used to treat patients with small stones compared to large stones. Again, by the **basic rule on rates**, we would expect the overall success rate of treatment Y to be closer to the small stones success rate of 86.7% than the large stones success rate of 68.8%. Indeed, we have the overall treatment Y success rate to be 82.6%.

Combining these two observations, it is no wonder that we have the overall success rate of X to be lower than the overall success rate of Y.

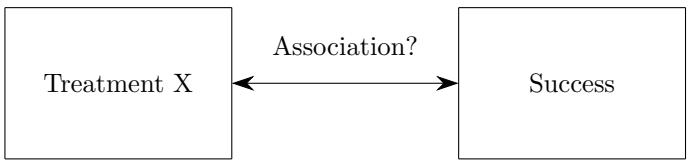
Another very telling observation from the table is that the range of success rates for treating large stones is between 68.8% (treatment Y) and 72.4% (treatment X). Compare this with the range of success rates for treating small stones which is between 86.7% (treatment Y) and 92.5% (treatment X). This tells us that treatments for large stones have a lower rate of success compared to small stones, which is not unreasonable to believe.

In conclusion, we can explain Simpson's Paradox in the following way. Treatment X is in fact a better treatment than Y. However, because patients have been using Treatment X to treat more difficult cases (large kidney stones), this lowers the overall success rate of treatment X. It does not change the fact that in the individual subgroups, regardless of stone size, treatment X achieves a higher success rate than treatment Y. **Slicing the data** into the small and large stone subgroups will reveal that treatment X is indeed a better treatment.

Before we conclude this section, let us recap the story so far.

- We started off with a new kidney stone patient coming to us for advice. Based on past patient records, we were convinced that the success rate of undergoing treatment is higher than the failure rate and thus conclude that the patient should undergo some form of treatment.

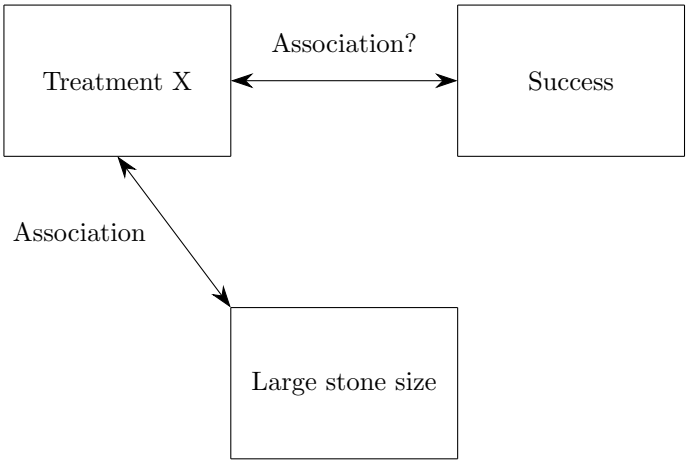
- We were then faced with the decision between two treatment types. Treatment X and Treatment Y. In determining which treatment type to recommend to the patient, we looked at the data on hand of past patients and investigated if there was any association (positive or negative) between Treatment X and Treatment Success.



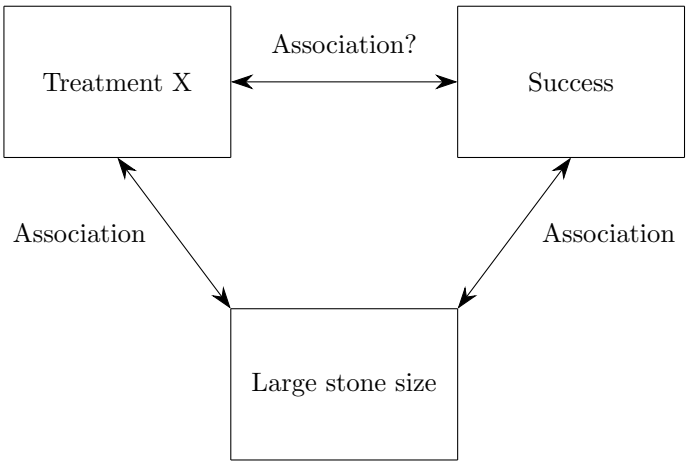
- With further analysis, we found that Treatment X was *negatively associated* with Success. This meant that we should recommend Treatment Y to our new patient. However, through another iteration of the PPDAC cycle, we wondered how another variable like stone size may affect our conclusion.

- By slicing, we segregated our data into past patients with large stone size and others with small stone size. Surprisingly for both subgroups, we found that Treatment X had a higher success rate than Treatment Y. This reversed that trend that we saw when the subgroups were combined.

- More importantly, we observed that Treatment X was used more often in dealing with large stones compared to Treatment Y, which was more frequently used to deal with small stones. This means that large stone size is likely to be associated with Treatment X.



- On the other hand, we also observed that patients with large stones have a lower success rate (regardless of treatment type) compared to patients with small stones. This is perfectly reasonable and thus also suggests that large stone size is likely to be associated with treatment success.



- This means that stone size is a (third) variable that was associated with the other two variables whose relationship we were initially investigating, thus affecting the conclusion of our initial study. Such a variable is called a *confounder* and they will be the focus of our discussion in the next section.

- For now, we will note that when Simpson's Paradox is observed, it implies that there is definitely a confounding variable present, that is a third variable that is associated with the two variables whose relationship we are investigating. However, the existence of a confounder does not necessarily lead to us observing Simpson's Paradox.

Section 2.5 Confounders

Discussion 2.5.1 Continuing our kidney stones patients example, we were fortunate that the data set contained information that may not seem to be important initially. Without performing further investigation into the size of the kidney stones, we could have ended up giving the wrong recommendation to our new patient.

In data collection, it is often important to collect more information on the subjects in addition to those variables that are immediately apparent to be of importance. This is because we can never be sure whether there would be some other variables that may be confounders that would influence our study of association between two variables of interest. Of course, as the owner of the study, we can ask our subjects (for example, in a survey) as many questions and collect as much data as we want, but practically, we also know that respondents do not like to see a long list of seemingly unrelated questions in surveys. There are also cost considerations if we collect more data than necessary. To design a good study, we need to strike a balance between the two.

Definition 2.5.2 A *confounder* is a third variable that is associated with both the independent and dependent variables whose relationship we are investigating. Note that we do not specify the direction (positive or negative) of association here. As long as the variable is associated in some way to the main variables, we will call it a confounder, or a *confounding variable*.

Example 2.5.3 At the end of the previous section, we explained how the variable kidney stone size is a confounding variable because it is associated with both the (independent) variable Treatment type and (dependent) variable Treatment outcome. Let us now work through the calculations to justify these associations. First, let us show that stone size is associated with treatment type.

Treatment	Large	Small	Total
X	526	174	700
Y	80	270	350
Total	606	444	1050

The table shows the number of large and small stones treated by treatments X and Y respectively. Out of 700 cases treated by treatment X, 526 were large stones and 174 were small stones. Out of 350 cases treated by treatment Y, 80 were large stones and 270 were small stones. Since

$$\text{rate}(\text{Large} \mid \text{X}) = \frac{526}{700} = 0.751 \quad \text{and} \quad \text{rate}(\text{Large} \mid \text{Y}) = \frac{80}{350} = 0.229,$$

we see that

$$0.751 = \text{rate}(\text{Large} \mid \text{X}) > \text{rate}(\text{Large} \mid \text{Y}) = 0.229,$$

and so large stones are positively associated with treatment X. This means that there is a higher proportion of large stones being treated by treatment X compared to treatment Y.

Now let us turn our attention to the association between stone size and treatment outcome.

Stone size	Success	Failure	Total
Large	436	170	606
Small	395	49	444
Total	831	219	1050

This table shows the number of success and failure outcomes for patients with large and small stones. Out of 606 large stones cases, 436 were successfully treated while 170 were not successful. Out of 444 small stones cases, 395 were successfully treated while 49 were not successful. Since

$$\text{rate}(\text{Success} \mid \text{Large}) = \frac{436}{606} = 0.719 \quad \text{and} \quad \text{rate}(\text{Success} \mid \text{Small}) = \frac{395}{444} = 0.890,$$

we see that

$$0.719 = \text{rate}(\text{Success} \mid \text{Large}) < \text{rate}(\text{Success} \mid \text{Small}) = 0.890,$$

and so large stones are negatively associated with success outcome. This means that there is a lower proportion of successful outcomes for large stones cases compared to small stones cases.

As we have now shown that stone size is associated with both the treatment type and the treatment outcome, we are convinced that stone size is a confounding variable that needs to be managed. The way to do it, as shown previously is to use *slicing*, where we segregate the data by the confounding variable. This is done by investigating the association between the dependent and independent variables for large stone cases separately from the small stone cases.

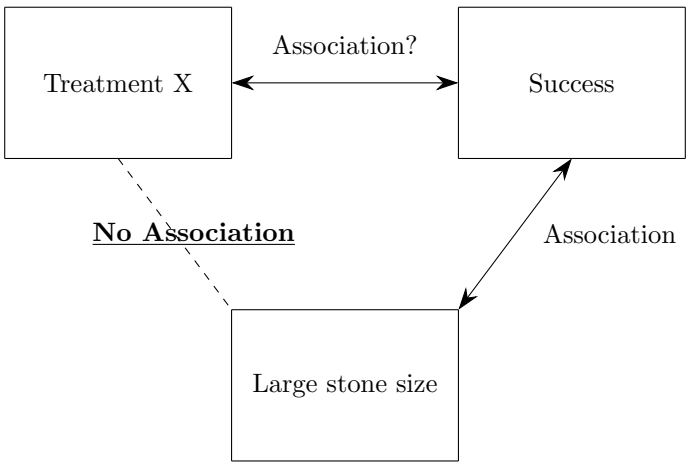
Discussion 2.5.4 We have now seen the benefits of having more information on the subjects because it allows us to identify confounding variables which would not have been possible if, for example, information of stone size was not available or collected. Thus, an important learning point when it comes to designing a study is to measure and collect data on additional variables that we feel may be relevant in our study. Whether these additional variables turn out to be confounders or not we would have to probe further, but we will never know if we do not collect data on them in the first place.

That said, we have to come to terms with the fact that most of the time, collecting information on variables is costly in practice. Even if we do manage to collect all the information we need, the analysis can be complicated if the data needs to be sliced along many different variables.

For non-randomised designs like observational studies, it is usually the case that the two groups that we are comparing are not "identical" except for the treatment. Despite our best efforts, we can never be totally sure that every single confounder has been identified and controlled for. Thus, observational studies offer only a limited conclusion in providing evidence of *association* and not *causation*.

(Randomisation as a preferred solution to confounding.) An alternative approach to address potential confounders is to rely on a strategy that was discussed in Chapter 1: **randomised assignment**. Let us discuss how this in done in detail, using our much developed example on kidney stones treatment.

Example 2.5.5 Fundamentally, confounding variables occur due to association which is a consequence of having unequal proportion of variables in the two groups that we are trying to compare. For the kidney stones example, stone size was a confounder because patients with large stones were disproportionately allocated to treatment X instead of treatment Y. Now, if the allocation of large (and small) stone size cases to the two treatment types was done randomly, which tends to result in an equal proportion across the two groups, there would no longer be any association between stone size and treatment type. In this case, stone size would no longer be a confounder. Note that a confounding variable is associated to **both** the independent and dependent variables, so removing one of the associations is enough to remove the confounding variable.



How can we achieve randomised assignment of patients to the two treatment types? One simple way is, for example, to toss a fair coin when deciding which treatment a patient will be given. Surely, such a method of randomised assignment tends to give us approximately equal proportions of large (and small)

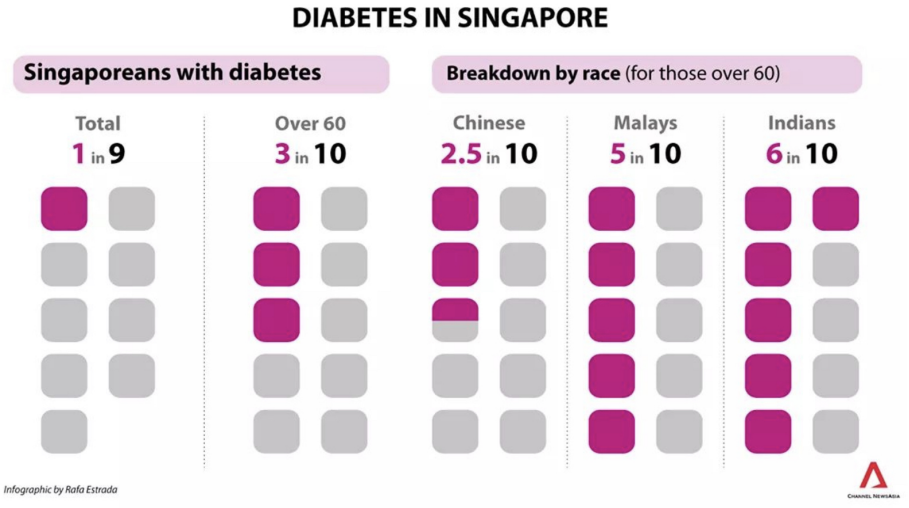
stone cases across the two treatment types. If we have sufficiently many patients to assign to either treatment types, the two groups of patients assigned to treatment X and treatment Y will tend to be similar in all characteristics, including stone size.

Surely this addresses the problem of confounders appropriately right? Unfortunately, randomisation is not always possible in every study. Imagine the scenario where the type of treatment given to each patient is dependent on a coin toss! Would you agree to this if you were one of the patients? Certainly not! Patients usually have the right to choose which treatment group they want to be in and this would make the assignment process non-random. Such ethical issues could very well constrain and prevent us from performing randomised assignment of our subjects. In such a situation, we have no choice but to fall back on the method of slicing for suspected confounders.

With this, we conclude Chapter 2, where we discussed, in detail how we can use rates to study the association between two (or more) **categorical variables**. We learnt about Simpson's Paradox which led us eventually to the issue of confounders and how they can be managed. In the next Chapter, we will turn our attention to the other variable type, namely **numerical variables**.

Exercise 2

1. The figure below shows that out of every 9 Singaporeans, 1 of them has diabetes. Similarly, out of 10 Singaporeans over 60, 3 of them have diabetes. Let us define "over 60" as old and "60 and below" as young. Which of the following statements is/are true? Select all that apply.



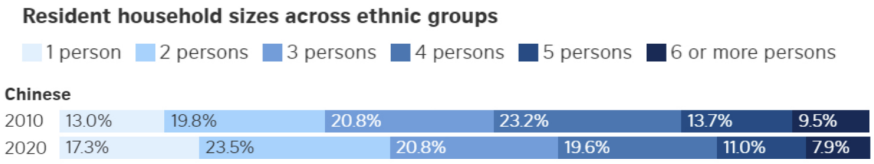
(A) rate(Diabetes | Young) > rate(Diabetes | Old).

(B) rate(Young | Diabetes) < rate(Young | No diabetes).

(C) rate(Old | Diabetes) < rate(Old | No diabetes).

(D) rate(Diabetes | Young) < rate(Diabetes | Old).

2. On 19 June 2021, The Straits Times published the figure below, taken from a population census of Singapore.



Use only the information shown in the figure to answer the following question.

Suppose that households with 1-3 people are considered "small" whereas those with 4 or more people are considered "large". Which of the following statements is/are true among Chinese resident households in the years 2010 and 2020? Select all that apply.

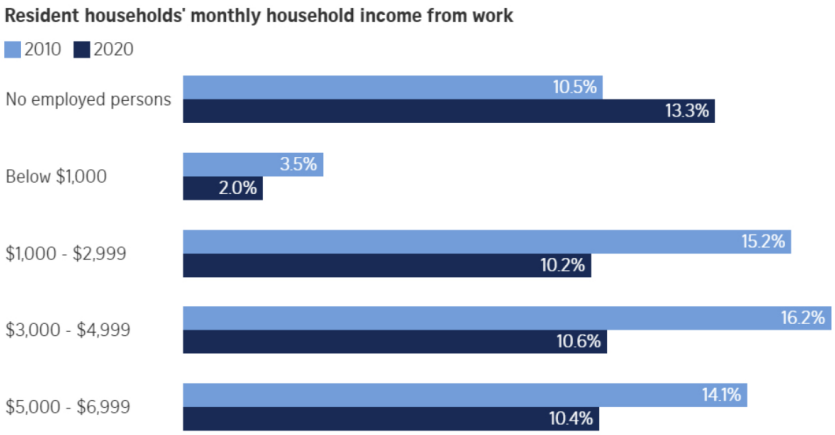
(A) The year 2020 is positively associated with small households.

(B) The year 2020 is positively associated with large households.

(C) The year 2010 is positively associated with small households.

(D) The year 2010 is positively associated with large households.

3. On 19 June 2021, The Straits Times published the figure below, taken from a population census of Singapore. Each household may only belong to a single category.



What can be said about the resident households, earning more than $6,999 from work? From the following statements, select all that apply.

(A) A majority of resident households are earning more than $6,999 from work in 2020.

(B) A larger proportion of resident households are earning more than $6,999 from work in 2020, as compared to 2010.

(C) rate(Income > $6,999 | 2020) > rate(Income > $6,999 | 2010). Here "Income" represents Household monthly income from work.

(D) rate(Income > $6,999 | 2020) < rate(Income > $6,999 | 2010). Here "Income" represents Household monthly income from work.

4. How does "forgiveness" (being forgiving) and empathy go together? The study of Toussaint and Webb on 45 men and 82 women are summarised in the following hypothetical tables:

Distribution of 45 men

	Empathy	No empathy	Row total
Forgiving	10	10	20
Not forgiving	9	16	25
Column total	19	26	45

Distribution of 82 women

	Empathy	No empathy	Row total
Forgiving	30	31	61
Not forgiving	12	9	21
Column total	42	40	82

Which of the following statements is/are true?

(I) Forgiveness and empathy are positively associated among men.

(II) Forgiveness and empathy are positively associated among women.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

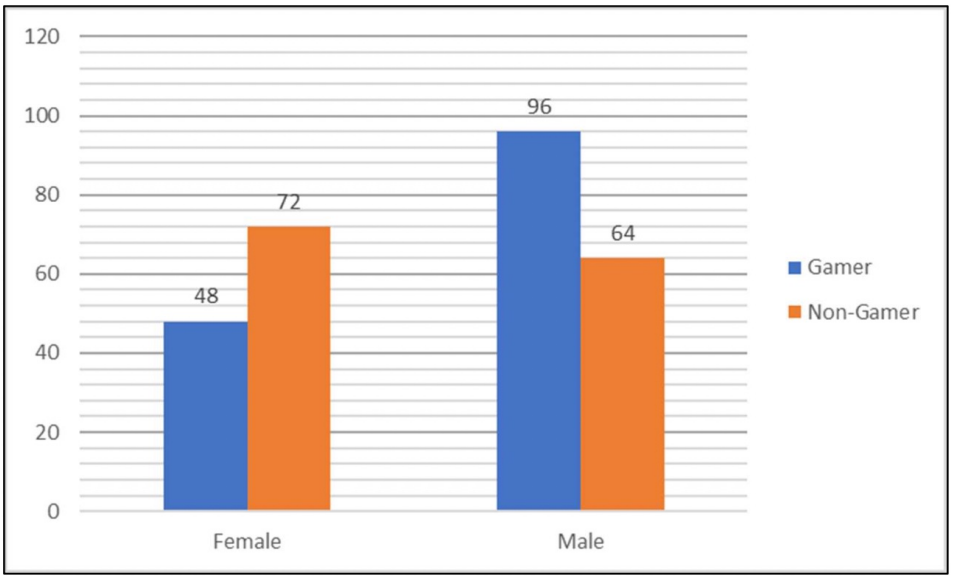
(D) Both (I) and (II).

5. The contingency table below shows the classification of hair descriptions of students studying in an international school in Singapore.

	Hair type				
	Straight		Curly		
Hair colour	Male	Female	Male	Female	**Total**
Red	7	9	8	5	29
Brown	35	20	12	16	83
Blonde	51	55	38	27	171
Black	22	25	19	24	90
Total	115	109	77	72	373

The marginal rate, rate(Curly), is _____%; while the joint rate, rate(non-Black and Female) is _____%. Give each answer as a percentage correct to 2 decimal places.

6. The bar graph below shows the number of gamers and non-gamers among males and females. Which of the following statements is/are true?



(A) There is a negative association between being female and being a gamer since rate(Female | Gamer) = 0.33 is less than rate(Female | Non-Gamer) = 0.53.

(B) There is a negative association between being female and being a gamer since rate(Gamer | Female) = 0.4 is less than rate(Gamer | Male) = 0.67.

(C) There is a negative association between being female and being a gamer since rate(Female | Gamer) = 0.33 is less than rate(Male | Gamer) = 0.67.

7. There are (another) two types (A and B) of possible treatment for kidney stones. The number of patients given each treatment and the number of successful outcomes, according to treatment type and stone size are shown in the table below.

	Treatment A		Treatment B	
Stone size	Number	Success	Number	Success
Small	87	81	270	234
Large	263	192	80	55
Total	350	273	350	289

Which of the following statements is/are correct? Select all that apply.

(A) Treatment A has a higher success rate than treatment B in each of the two groups (small stone size and large stone size).

(B) When groups of patients with small stones and large stones are combined, treatment B has a higher success rate than treatment A.

(C) There is positive association between treatment A and success among patients with small stones, and also among patients with large stones.

(D) When groups of patients with small stones and large stones are combined, treatment A and success have a negative association.

(E) There is no association between stone size and success, as rate(Sucess | Small stones) is equal to rate(Success | Large stones).

8. Joseph conducted a study on night owls (individuals who sleep after 12am on average every night) among staff and students in NUS. He gathered the following result on rates:

$$\text{rate(Night owl} \mid \text{Student)} = 0.7;$$

$$\text{rate(Non-Night owl} \mid \text{Staff)} = 0.4.$$

Which of the following statements is correct?

(A) rate(Night owl) must be between 0.4 and 0.7.

(B) rate(Night owl) must be between 0.6 and 0.7.

(C) rate(Non-Night owl) must be between 0.4 and 0.7.

(D) rate(Non-Night owl) must be between 0.6 and 0.7.

9. A newspaper article had a headline "30% of local university students admitted last year graduated from a polytechnic". Assume there are only 2 universities (Uni A and Uni B). In Uni A, 50% of its local students admitted last year graduated from a polytechnic. In Uni B, the percentage of its local students admitted last year who graduated from a polytechnic must be

(A) 10%.

(B) 40%.

(C) between 30% and 50%.

(D) less than 30%.

(E) more than 50%.

10. The relative frequency table below shows the distribution of annual total personal income for the entire population of 6,402,386 living in City X. It is known that there are 59% males and 41% females in City X.

Income	Percent
\$9,999 or less	2.2%
\$10,000 to \$14,999	4.7%
\$15,000 to \$24,999	15.8%
\$25,000 to \$34,999	18.3%
\$35,000 to \$49,999	21.2%
\$50,000 to \$64,999	13.9%
\$65,000 to \$74,999	5.8%
\$75,000 to \$99,999	8.4%
\$100,000 or more	9.7%

We are told that in City X, 71.8% of females earn less than \$50,000 per year. Which of the following statements is correct?

(A) There is positive association between being male and earning less than \$50,000.

(B) There is negative association between being male and earning less than \$50,000.

(C) There is no association between being male and earning less than \$50,000.

(D) We do not have sufficient information to determine the correctness of the other three statements.

11. The website correlated.org presents the following for December 24th 2018. 352 people are surveyed, of whom 131 find the sound of windshield wipers to be soothing. Among the 352 people, 55% stay in the movie theater until the credits end. But among those who find the sound of windshield wipers to be soothing, 75% stay in the movie theater until the credits end. Among those who do not find the sound of windshield wipers to be soothing, what would be the percentage who stay in the movie theater until the credits end?

(A) More than 75%.

(B) Equal to 75%.

(C) More than 55% and less than 75%.

(D) Equal to 55%.

(E) Less than 55%.

12. By "elderly", we mean a person who is more than 65 years old. In Singapore, the percentage of elderlies among women is higher than the percentage of elderlies among men. Which of the following statements is/are true?

(I) In Singapore, the percentage of women among elderlies is higher than the percentage of women among the non-elderlies.

(II) In Singapore, the percentage of women is higher than the percentage of men among elderlies.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

13. Consider the following statements.

(I) A spokesman was quoted as saying that the proportion of a certain ethnic group among those who contracted Disease X was lower than the proportion of that ethnic group in the general population.

(II) A reporter interpreted the statement and concluded that the members of this ethnic group are less likely to contract Disease X than a random member of the population.

Which of the following is correct?

(A) The two statements are equivalent.

(B) We can infer statement (I) from (II) but not the other way round.

(C) We can infer statement (II) from (I) but not the other way round.

(D) We can neither infer statement (I) from (II), nor infer statement (II) from (I).

14. Su is investigating the association between blood pressure and "workaholism" in a certain population. Someone who works more than 75 hours per week is considered a workaholic.

The income level and blood pressure (high or normal) for each subject and whether or not they are classified as "workaholic" are recorded and summarised in the table below. Here "HBP" denotes "high blood pressure" while "NBP" denotes "normal blood pressure".

	Income Group					
	Low		Middle		High	
	HBP	NBP	HBP	NBP	HBP	NBP
Workaholic	25	75	23	87	26	134
Non-workaholic	25	80	18	72	9	51

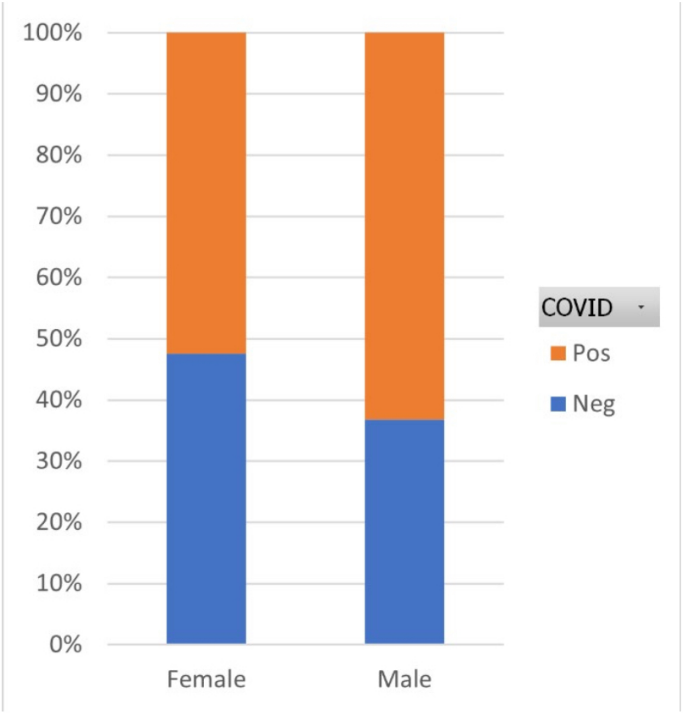
Which of the following statements is true?

(A) For subjects in the "Middle" income level group, there is a positive association between being a "workaholic" and having "high blood pressure".

(B) For subjects in the "Middle" income level group, there is no association between being a ''workaholic" and having "high blood pressure".

(C) For subjects in the "Middle" income level group, there is a negative association between being a "workaholic" and having "high blood pressure".

15. The graph below shows the stacked bar plot for the rate of COVID infection among males and females in Country X. Which of the folowing variables must be positively associated with each other? Select all that are true.



(A) Female and COVID-positive.

(B) Male and COVID-positive.

(C) Female and COVID-negative.

(D) Male and COVID-negative.

16. The Lord of the Rings: The Fellowship of the Ring was released in December 2001. Suppose that

(I) Among the people in Singapore who were born before 2000, 10% watched the film.

(II) Among the people in Singapore who were born during or after 2000, 20% watched the film.

Choose the best option below. Among all the people in Singapore, the percentage who watched the film _____.

(A) must be 15%.

(B) must be between 10% and 20%.

(C) can be less than 10%.

(D) can be more than 20%.

17. Coriander is a common herb used to add flavour to various kinds of dishes. Suppose we have two countries: Country X and Country Y. We have individuals who either dislike coriander, or like coriander, and are either male or female. We know that in country X,

    - rate(Dislike) = 0.1.
    - rate(Dislike | Male) = 0.3.

    Also, in country Y,

    - rate(Female) = 0.8.
    - rate(Female | Dislike) = 0.4.

    Which of the following statements must be true? Select all that apply.

    (A) rate(Male) < rate(Female) in country X.

    (B) There is a positive association between disliking coriander and females in country X and country Y separately.

    (C) Overall rate(Male) is between 0.2 and 0.5, when both countries are combined.

18. Which of the following statements is **true**?

    (A) Confounders will always lead to Simpson's Paradox.

    (B) An observational study can be conducted when the researcher is unable to assign participants into the treatment and control groups.

    (C) Randomised assignment will always result in equal allocation of the number of subjects across the treatment and control groups.

    (D) Observational studies are better at showing causation than experimental studies because random assignment is used in observational studies to minimise the effect of confounding variables.

19. The rate of lung cancer among females in Singapore is 40%, while the rate of lung cancer among males in Singapore is also 40%. Researchers also discovered that the rate of lung cancer among smokers in Singapore is 70%. Which of the following statements is/are true?

    (I) Sex is a confounder when discussing the relationship between smoking and lung cancer.

    (II) Lung cancer is positively associated with smoking.

    (A) Only (I).

    (B) Only (II).

    (C) Neither (I) nor (II).

    (D) Both (I) and (II).

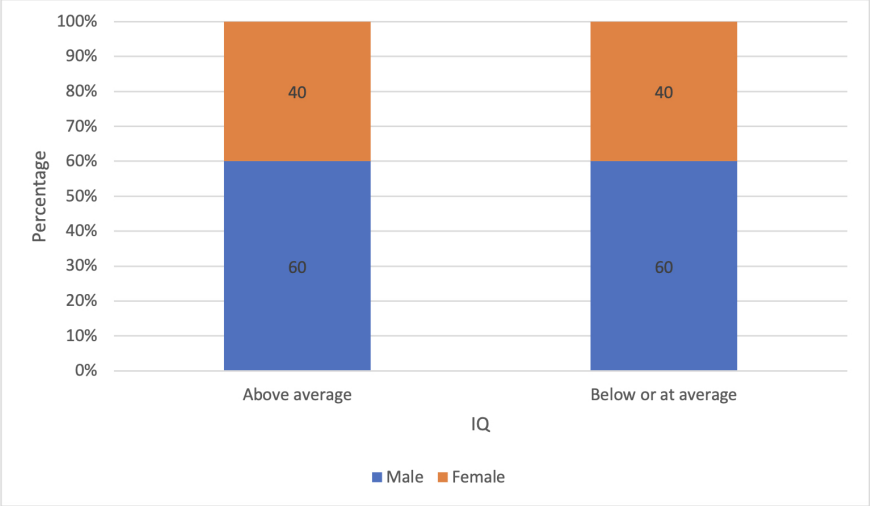
20. A researcher wants to find out if drinking tea helps to reduce memory loss. He interviewed 100 elderly citizens from an Elder Care Center and inquired if they were tea drinkers. 60 of them were classified as tea drinkers, while the remaining 40 were not. He then asked them to play a specific memory game to test their memory. The researcher also noted that a potential confounding variable was "gender". To control for this potential confounder (gender), the researcher could perform

    (A) double blinding.

(B) random assignment.

(C) slicing of the data.

21. A researcher is interested in finding out if gender affects whether a student is left-handed. The information on gender (Male/Female), master hand (Left/Right) and IQ (Above average/Below or At average) of all students was collected. The data was used to plot the figure below.



The researcher makes the following two statements.

(I) IQ can be a confounder when investigating the relationship between gender and master hand.

(II) When finding out if IQ affects whether a person is left-handed, it is possible that Simpson's Paradox is observed in the data due to the third variable, gender.

Which of the following correctly describes the two statements above?

(A) Statement (I) is true but statement (II) is false.

(B) Statement (I) is false but statement (II) is true.

(C) Both statements are true.

(D) Both statements are false.

22. A researcher conducted an observational study to investigate whether smoking is associated with heart disease. 1000 participants were recruited in the study and the researcher obtained the following result:

	Heart disease	No heart disease
Smoker	146	317
Non-smoker	324	213

He also observed that 80% of the smokers were alcoholics, while 85% of the people with heart disease were alcoholics. Consider the following statements:

(I) There is positive association between smoking and heart disease.

(II) Being an alcoholic is a confounder.

Based only on the information given above, which of the two statements must be true?

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

23. Consider a study that intends to examine whether the colour red makes children act impulsively. A group of 500 children were assigned into two groups by the expert opinion of a child psychologist; group Red if the psychologist pointed to the child, and group Green if the psychologist did not. Each child is then led into a room that has a big button in the colour of their group and labelled "DO NOT PRESS ME!". It is then recorded whether the child presses the button within 10 minutes. All the children were each then given a candy for participating.

    Which of the following conclusions from the analysis of data can establish that wearing spectacles (whether the child wears spectacles) is a confounder in this study?

    (A) Wearing spectacles is positively associated with being in group Red, and negatively associated with pressing the button.

    (B) Wearing spectacles is associated with being in group Red, and is not associated with pressing the button.

    (C) Wearing spectacles is not associated with being in group Red, and is not associated with pressing the button.

    (D) None of the other given options is correct.

24. Consider a study that intends to examine whether the colour red makes children act impulsively. A group of 500 children were assigned into two groups by the expert opinion of a child psychologist; group Red if the psychologist pointed to the child, and group Green if the psychologist did not. Each child is then led into a room that has a big button in the colour of their group and labelled "DO NOT PRESS ME!". It is then recorded whether the child presses the button within 10 minutes. All the children were each then given a candy for participating.

    The children were also asked if they like candies. The following table summarises the data. For instance, 22 children from group Red that pressed the button do not like candies.

	Like candies		Does not like candies	
	Red	Green	Red	Green
Pressed button	3	135	22	1
Did not press button	177	60	38	64

    Is liking candy a confounder in this study?

    (A) Yes.

    (B) No.

    (C) There is insufficient information given to determine whether liking candy is a confounder in this study.

25. In a certain year, it is known that the prevalence of diabetes among Singapore residents is 10% and the prevalence of diabetes among old (age 60 and above) Singapore residents is 30%. It was suggested that sex is a possible confounder in the observed association between age and diabetes among Singapore residents. After further analysis, the researchers concluded that sex is not a confounder, and there is an association between sex and age. Which of the following statements is/are true? Select all that apply.

    (A) rate(Diabetes | Male) = rate(Diabetes | Female).

    (B) rate(Male | Diabetes) = rate(Female | Diabetes).

    (C) rate(Diabetes | Female) = 10%.

26. A researcher would like to find out if there is any relationship between age (young and old) and ramen consumption (high and low) among Singaporeans. From the data he obtained, he suspects that sex is a confounder. Which of the following should hold in order to show that his suspicion is correct? Select all that apply.

    (A) The percentage of old people among males is different from the percentage of old people among females.

    (B) The percentage of males among the high ramen consumers is different from the percentage of females among the high ramen consumers.

    (C) The percentage of high ramen consumers among males is different from the percentage of high ramen consumers among females.

27. Su is investigating the association between blood pressure and "workaholism" in a certain population. Someone who works more than 75 hours per week is considered a workaholic.

    The income level and blood pressure (high or normal) for each subject and whether or not they are classified as "workaholic" are recorded and summarised in the table below. Here "HBP" denotes "high blood pressure" while "NBP" denotes "normal blood pressure".

	Income Group					
	Low		Middle		High	
	HBP	NBP	HBP	NBP	HBP	NBP
Workaholic	25	75	23	87	26	134
Non-workaholic	25	80	18	72	9	51

    Which of the following statements is true?

    (A) We have an instance of Simpson's Paradox for this data set, when considering the association between being a "workaholic" and having "high blood pressure", first for individual income levels ("Low", "Middle", "High") and then overall.

    (B) We do not have an instance of Simpson's Paradox for this data set, when considering the association between being a "workaholic" and having "high blood pressure", first for individual income levels ("Low", "Middle", "High") and then overall.

    (C) We are not able to determine if we have an instance of Simpson's Paradox for this data set (or not), when considering the association between being a "workaholic" and having "high blood pressure", first for individual income levels ("Low", "Middle", "High") and then overall. There is insufficient information given.

28. Some scientists have found that drinking coffee is associated to students' ability to sleep (enough vs not enough sleep). Sex was also found to be a confounder. This means that:

    (I) Percentage of coffee drinkers among males is different from the percentage of coffee drinkers among females.

    (II) Percentage of males among students who have enough sleep is different from the percentage of males among students who do not have enough sleep.

    Which of the statements above is/are correct?

    (A) Only (I).

    (B) Only (II).

    (C) Both (I) and (II).

    (D) Neither (I) nor (II).

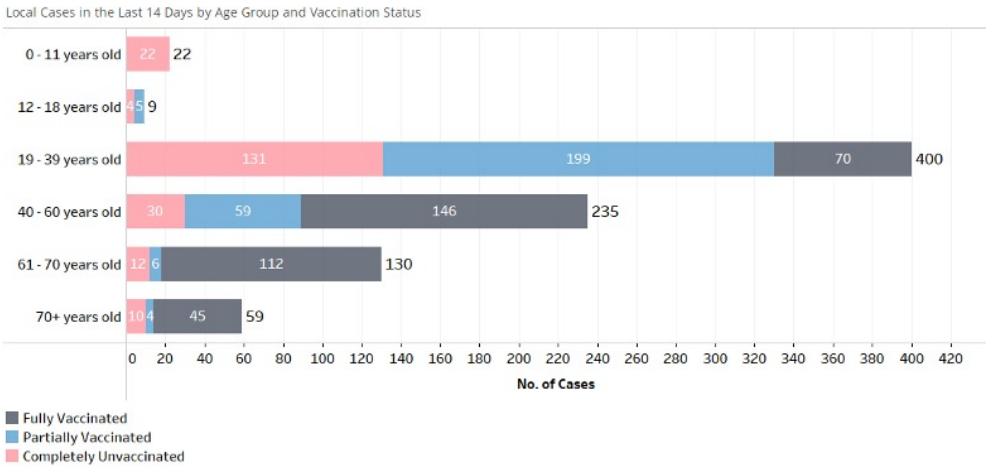
29. A team of researchers are interested in seeing if there is an association between the amount of sleep and memory retention. They have also collected information on each subject's gender. Which of the following statements is correct?

    (A) If Simpson's Paradox is not observed when combining the 2 subgroups of gender, then gender is not a confounder when exploring the association between the amount of sleep and memory retention.

    (B) Suppose that when the 2 subgroups of gender are combined, Simpson's Paradox is observed when checking for association between the amount of sleep and memory retention. Then gender must be a confounder.

    (C) If gender is a confounder when determining the association between the amount of sleep and memory retention, Simpson's Paradox will be observed when combining the 2 subgroups of gender.

30.



The above graph, depicting cases of COVID infection (titled "Local cases in the last 14 days by Age group and Vaccination status"), was published in a press release by Singapore's Ministry of Health on 21 July 2021. Let us designate the age range of those 61 years or older as "seniors" , and all others as "non-seniors". Let us also consider the status of either full or partial vaccination as "vaccinated", and the status of being completely unvaccinated as "unvaccinated".

What can be concluded based on the information given? Select all statements that apply.

(A) The rate of infection among seniors is lower than the rate of infection among non-seniors.

(B) For cases depicted by the graphic, being vaccinated is positively associated with seniors.

(C) Age is a confounder in the association between infection and vaccination status.

(D) There were about twelve cases on average daily (correct to the nearest whole number) for those senior and vaccinated.

This page is blank

Chapter 3

Dealing with Numerical Data

Section 3.1 Univariate EDA

In Chapter 1, we introduced two main types of variables that we will be focussing on, namely *categorical variables* and *numerical variables*. Categorical variables were discussed extensively in Chapter 2 and in this chapter, we will turn our attention to numerical variables and how they can be analysed.

Consider the following table that shows a portion of a data set relating to COVID-19 cases in Singapore.

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

An example of a numerical variable in this data set is **Age**. Can you identify another numerical variable? The analysis of data, more precisely, Exploratory Data Analysis (or EDA) is a process of summarising or understanding the data and extracting insights or main characteristics of the data. This is a critical part of the "Analysis" step of the PPDAC problem solving cycle. In this chapter, we will discuss how numerical variables can be summarised and understood. To begin, the focus of this section will be on data exploration techniques for one variable, or *univariate exploratory data analysis*.

Example 3.1.1 In Chapter 2, the recurring data set that was used to drive the discussion on categorical variables was the patients with kidney stones data set. In this chapter, we will be using a data set closer to home.

	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	1/1/2017	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	1/1/2017	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Generation	1978	60 years 07 months	250000
4	1/1/2017	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	262000
5	1/1/2017	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1980	62 years 01 month	265000
6	1/1/2017	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	265000
7	1/1/2017	ANG MO KIO	3 ROOM	150	ANG MO KIO AVE 5	01 TO 03	68	New Generation	1981	63 years	275000
8	1/1/2017	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	280000
9	1/1/2017	ANG MO KIO	3 ROOM	218	ANG MO KIO AVE 1	04 TO 06	67	New Generation	1976	58 years 04 months	285000
10	1/1/2017	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	285000
11	1/1/2017	ANG MO KIO	3 ROOM	571	ANG MO KIO AVE 3	01 TO 03	67	New Generation	1979	61 years 04 months	285000
12	1/1/2017	ANG MO KIO	3 ROOM	534	ANG MO KIO AVE 10	01 TO 03	68	New Generation	1980	62 years 01 month	288500
13	1/1/2017	ANG MO KIO	3 ROOM	233	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1977	59 years 08 months	295000
14	1/1/2017	ANG MO KIO	3 ROOM	235	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1977	59 years 08 months	295000
15	1/1/2017	ANG MO KIO	3 ROOM	219	ANG MO KIO AVE 1	07 TO 09	67	New Generation	1977	59 years 06 months	297000
16	1/1/2017	ANG MO KIO	3 ROOM	536	ANG MO KIO AVE 10	07 TO 09	68	New Generation	1980	62 years 01 month	298000
17	1/1/2017	ANG MO KIO	3 ROOM	230	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1978	60 years	298000
18	1/1/2017	ANG MO KIO	3 ROOM	570	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1979	61 years 04 months	3.00E+05
19	1/1/2017	ANG MO KIO	3 ROOM	624	ANG MO KIO AVE 4	04 TO 06	68	New Generation	1980	62 years 08 months	301000
20	1/1/2017	ANG MO KIO	3 ROOM	441	ANG MO KIO AVE 10	07 TO 09	67	New Generation	1979	61 years	306000

The data set (Microsoft Excel file partially shown above) that we will be looking at in this chapter corresponds to sales of Housing Development Board (HDB) resale flats within the period of January 2017 to June 2021. The entire data set contains $99,236$ rows and 11 columns. Note that each transaction is a row of the Excel file and each transaction contains information on variables (the columns) like month (of sale), flat's floor area (in square metres), resale price, etc.

The PPDAC cycle starts off with

1. **Problem.** So what is the problem that we are considering and attempting to answer? If you are a potential buyer, perhaps a question that you may be interested in investigating could be

   > What factors may affect the pricing of resale flats sold in Singapore?

2. **Plan.** Here, we need to decide what are some of the variables that are relevant and possible factors that answer the question. Suppose these variables were determined to be the 11 columns of the data set. Some of these variables are
   - "Month" - this is the month/year of the resale transaction;
   - "Town" - this is the town that the resale flat belongs to;
   - "Floor_area_sqm" - this is the floor size of the resale flat;
   - "Resale price" - this is how much the flat was sold for.
3. **Data.** In this stage, data is collected and prepared as shown in the table above.
4. **Analysis.** We are now at this stage where the data is going to be analysed in attempting to answer the **Problem**.

Definition 3.1.2 A *distribution* is an orientation of data points, broken down by their observed number or frequency of occurrence.

Example 3.1.3 Let us look at our HDB resale flats data set. The first few rows of the data set for transactions from January to June 2021, is reproduced in the table below.

Month	Floor area sqm	Age	Resale price
1/1/2021	45	35	225000
1/1/2021	45	35	211000
1/1/2021	73	45	275888
1/1/2021	67	43	316800
1/1/2021	67	43	305000
1/1/2021	68	40	260000
1/1/2021	73	44	351000
1/1/2021	73	44	343000
1/1/2021	75	41	306000

We would like to investigate the distribution of the Age[1] variable. To do this, we would need to collate the number of flats with the same ages when the resale transaction was made and put them in a frequency table. For example the first two rows of the data indicates that the first two HDB flats in the data set had the same age of 35 years when they were sold, while the third flat was 45 years old and so on. Suppose the frequency table collated for the entire data set is as follows:

Age	Frequency
2	9
3	8
4	583
5	1105
6	884
7	295
8	255
⋮	⋮

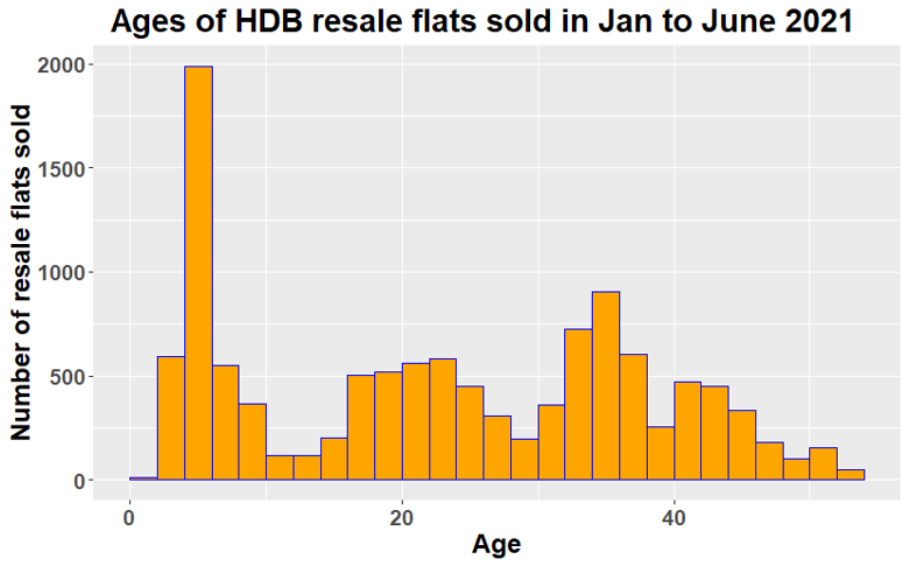
If we simply look at the frequency values in the table, it would be hard to observe any patterns or gain insights into **how** the frequencies are distributed across the different age values. We will introduce two different graphs to present the distribution in better way.

Example 3.1.4 (Histograms for Univariate EDA) A *histogram* is a graphical representation that organises data points into ranges or bins. It is particularly useful when we have large data sets. Let us see how the histogram will look like when we use Microsoft Excel to create one based on the "Age" frequency from Example 3.1.3. To create a histogram, the variable values are "grouped" into equal size intervals called bins. For our "Age" variable, we can use bins with a width of 2 years. The number of flats in each bin are counted and tabulated.

Bins	Frequency
0-2	9
2-4	591 $(8 + 583)$
4-6	1989 $(1105 + 884)$
6-8	550 $(295 + 255)$
8-10	336 $(219 + 47)$
⋮	⋮

You may notice that for the 2-4 Bin, the frequency is obtained by adding the number of flats sold at Age 3 and Age 4 and excludes those sold at Age 2. Thus, the left-end point of the interval 2-4 is excluded. The same is observed for the rest of the bins. The histogram created using Radiant is shown below:

[1]The data set, which can be downloaded from https://data.gov.sg/dataset/resale-flat-prices actually does not contain the "Age" variable. The "Age" variable was created by subtracting lease_commence_date from the year the flat was sold.



With the height of each bar representing the frequency for that bin range, the highest bar would represent the most frequently occurring range of values.

From the histogram above, we see that the range 4-6 years has the highest frequency as it accounts for 1989 out of the total 11644 transactions, or about 17% of the flats sold.

Remark 3.1.5 You may wonder how we came to the decision to have bin widths of 2 years rather than 3 years (or any bigger number). There is no correct answer for this. Normally, we would construct several histograms with different bin widths before deciding which one is most appropriate.

Once we have obtained and visualised the distribution of a numerical variable, we would like to describe the overall pattern of the distribution as well as whether there are any deviations from the overall pattern. To describe the overall pattern of the distribution, we will focus on the

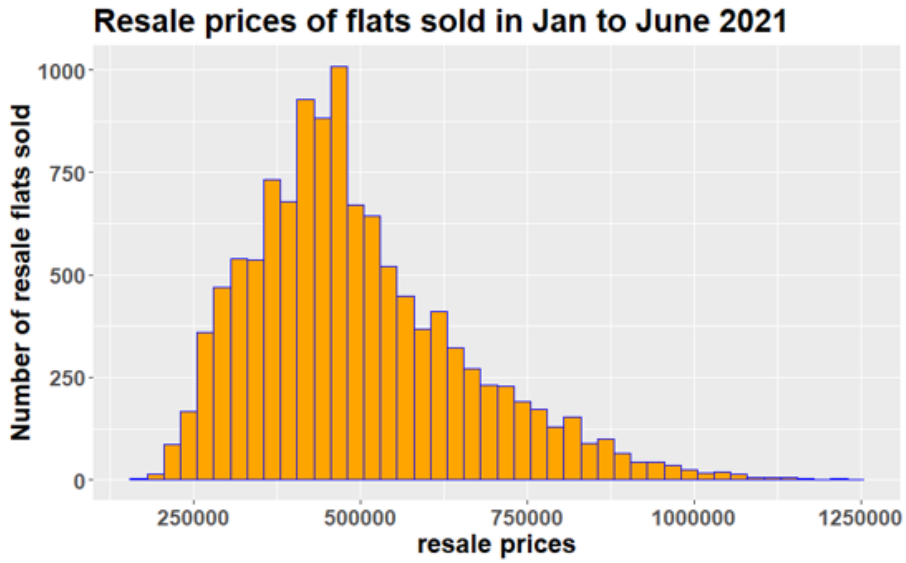
1. Shape;

2. Center; and

3. Spread of the distribution.

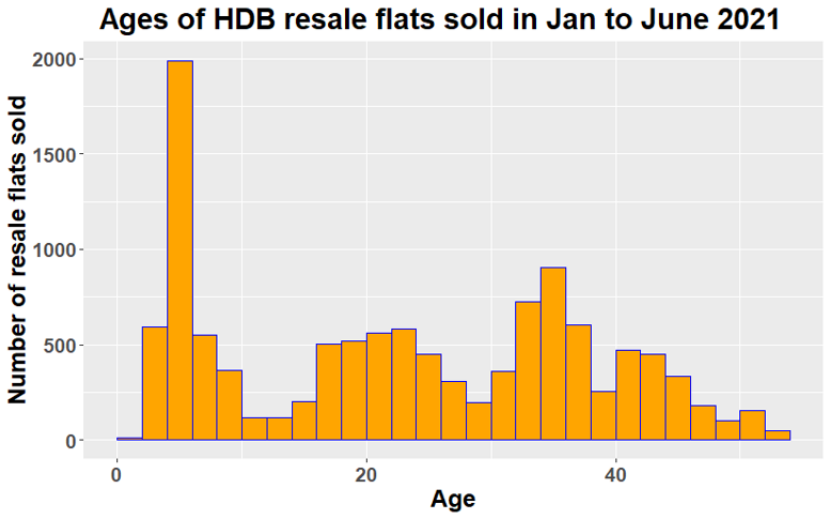
For deviations from the overall pattern, this usually refers to identifying *outliers* which will be discussed later on in this Chapter. Let us start by looking at how we can describe the shape of a distribution.

Discussion 3.1.6 (Shape - peaks and skewness). There are two important descriptors when we discuss the shape of a distribution, namely the *peaks* and the *skewness*. Let us look at another histogram plot obtained from the HDB resale data set. Rather than the age of the flat at the point of resale, we consider another numerical variable of interest, which is the "Resale Price". The following histogram was obtained when we set a bin size of 25,000.



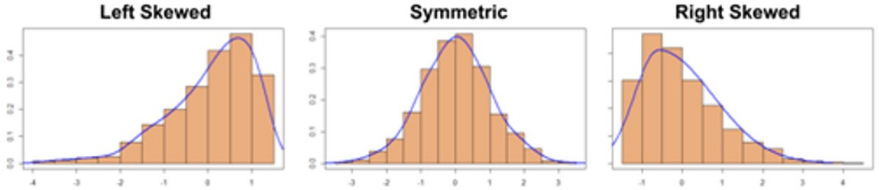
There is a peak in the interval [455000, 480000]. The distribution is *unimodal*, which means that it has one distinct peak. This tells us that the most frequent resale flat prices lies between $455,000 and $480,000.

Distributions are not always unimodal. Looking at the histogram we plotted earlier for the Age of the resale flats, we see that there is more than one distinct peak. In such a situation, we say that the distribution is *multimodal*. If a distribution has exactly two distinct peaks, we say it is *bimodal*.



In the histogram above, we see the highest peak in the 4-6 years range and the second highest peak occurring in the 34-36 years range. It should be noted that we say these are peaks because they occur most frequently in their immediate neighbourhoods of age ranges.

For a unimodal distribution, we can use another descriptor to describe the shape of the distribution, that is, whether the distribution is *symmetrical* or *skewed*.

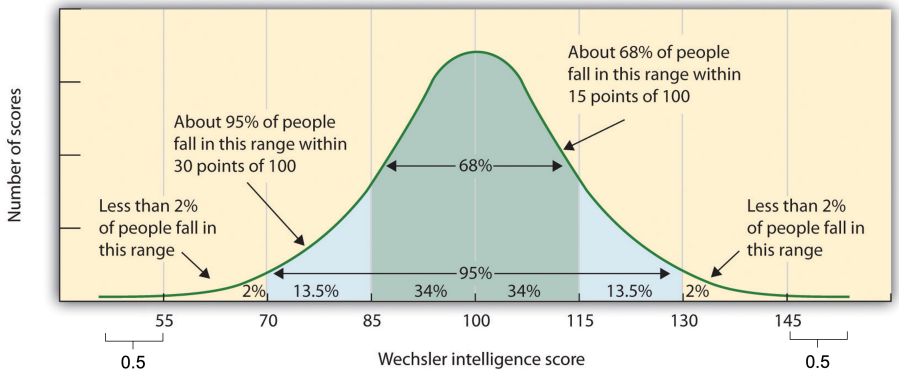


In a *symmetrical* distribution (middle picture above), the left and right halves of the distribution are approximate mirror images of each other, with the peak in the middle.

For the picture on the left, the distribution is **left skewed**, with the peak shifted to the right and a relatively long "tail" on the **left**.

The picture on the right shows a distribution that is **right skewed**. Such a distribution has the peak shifted to the left and a relatively long "tail" on the **right**. Referring back to the distribution of resale prices of HDB flats, we see that the distribution is right skewed, meaning that there are some (but few) flats sold at very high prices. These data points gave rise to the long tail to the right of the peak.

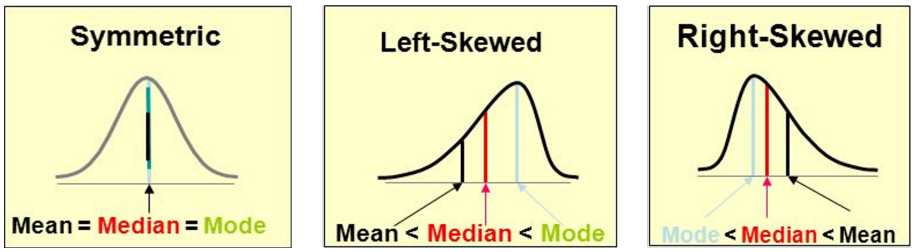
Example 3.1.7 (Symmetrical distribution - Bell curve) One of the most well-known symmetrical distributions is the *normal distribution* or what is commonly known as the bell curve. A famous example of the normal distribution is that of the IQ scores in a population, based on the Wechsler Intelligence scale.



From the figure, we see that the peak happens at 100, which means that the average IQ of a person in the population is 100. We also see that about 68% of the population has IQ scores in the range between 85 and 115, whereas about 95% of the population has IQ scores between 70 and 130.

Discussion 3.1.8 (Central tendency - mean, median and mode). Besides describing the shape of the distribution, we can also describe the characteristics of a distribution more precisely using measures of central tendency. The three most common measures of central tendency are *mean*, *median* and *mode*, which were all introduced in Chapter 1.

The three possible shapes of a distribution have different relative positions of the mean, median and mode.



1. For a symmetrical distribution, the mean, median and mode will be very close to each other near the peak of the distribution.

2. For a left skewed distribution, we usually (but not always) have

$$\text{mean} \ < \ \text{median} \ < \ \text{mode} \ .$$

   To see why this is the case, notice that the small number of extremely small values which contributes to the long tail on the left, will push down the mean/average, as compared to the median which is less affected by these extremely small values. The mode, found at the peak of the distribution is naturally the largest among the three measures of central tendency.

3. For a right skewed distribution, we have the opposite of the left skewed distribution, which is

$$\text{mode} < \text{median} < \text{mean} .$$

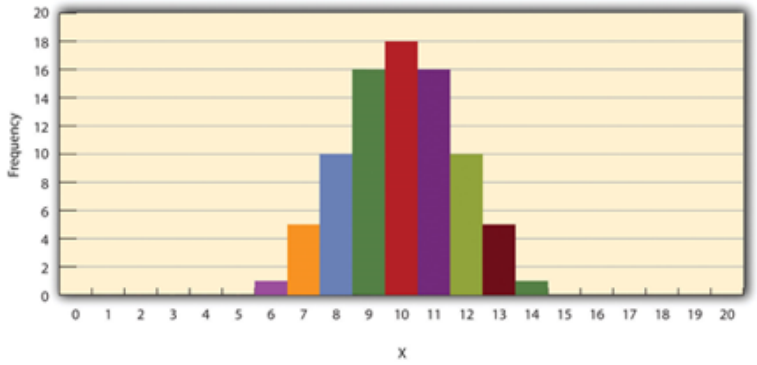
In this case, there are a small number of extremely big values which contributes to the long tail on the right. These big values will push up the mean/average as opposed to the median which is less affected by these extremely large values. The mode in such a distribution would be the smallest among the three measures of central tendency.

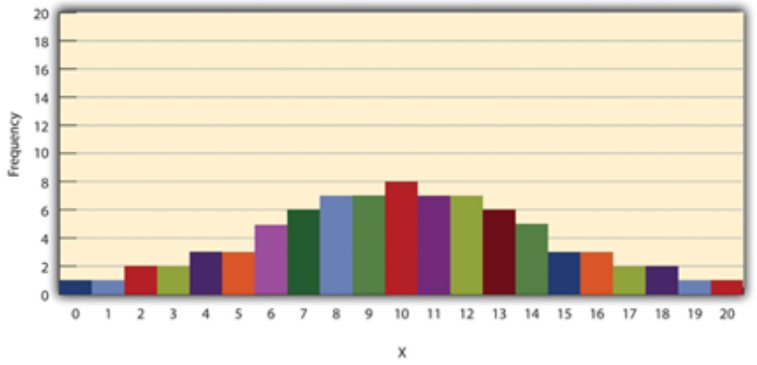
Example 3.1.9 Referring again to the resale prices distribution, we have seen the shape of the distribution and concluded that the distribution is right skewed.

The mean, median and mode of this distribution were found to be \$496,870.40, \$468,000 and \$420,000 respectively. This indeed agrees with

$$\text{mode} < \text{median} < \text{mean} .$$

Discussion 3.1.10 (Spread - standard deviation and range). Besides the shape and center of the distribution, we can also describe the *spread* of a distribution. This refers to how the data vary around the central tendency.





Take a look at the two distributions above, both of which have the same central tendencies. In fact, the mean, median and mode of both distributions are 10. However, the top distribution has a relatively lower variability compared to the distribution below. This means that the data in the top distribution are all relatively close to the center while the data in the bottom distribution are more spread out, or has more variability. We can also say that the data in the bottom distribution is spread across a much wider range.

The most commonly used measure of variability is *standard deviation* which was introduced in Section 1.5. For the two distributions shown here, the top distribution has standard deviation 1.69 while the bottom distribution has standard deviation 4.30.

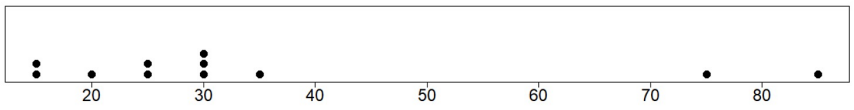
A simpler measure of variability is the *range* of the distribution. This is defined to be the difference between the largest and the smallest data points in the distribution. The range is simple to compute

but sometimes it can be misleading. For example, if we look at the range of the HDB resale prices data, we obtain

$$\text{Range} = \text{ Highest resale price } - \text{ Lowest resale price } = \$1,250,000 - \$180,000 = \$1,070,000.$$

The range is very large and is due to the existence of a few extremely high resale prices. It is not really the case that there is great variability in resale prices as we see that most of the resale prices are actually much lower and the variability is not as big as the range indicates it to be.

Definition 3.1.11 An *outlier* is an observation that falls well above or below the overall bulk of the data.



Consider the data set with 11 data points shown above. We can consider 75 and 85 as outliers since they are way larger than the rest of the data points. At this point, we use our judgement to identify values that appear to be exceptions to the general trend in the data. Later on, we will be introducing a more precise method (boxplot) to identify outliers.

Identifying outliers can be useful when we wish to identify any strong skewness in a distribution. Sometimes the outliers are caused by erroneous data collection or data entry but this may not always be the case. It is also possible that outliers are legitimate data points that provide us interesting insights into the behaviour of the data. A general rule when we investigate a data set is that outliers should not be removed unnecessarily as they do tell us something about the behaviour of the variable and prompt us to investigate further why such extreme values can happen.

Example 3.1.12 Consider the data set below:

$$4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 300.$$

It is not difficult to be convinced that 300 is an outlier in the data set. The table below shows the three different central tendencies as well as the standard deviation for the entire set and also when the outlier is removed from the data set.

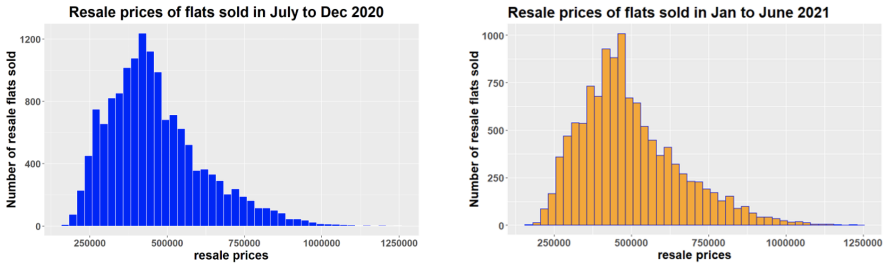
	Mean	Median	Mode	Standard deviation
Without removing 300	30	5.5	5	85.03
With 300 removed	5.45	5	5	1.04

We see that between the three central tendencies, the mean seems to be the most affected by the removal of the outlier, while both the median and the mode either remained the same or only changed slightly. Without removing the outlier, the mean is pulled away in the direction of the skew (in this example, the distribution is skewed to the right). In such cases, mean may no longer be a good measure of the central tendency of the distribution. We call the median and the mode *robust statistics*.

In addition, the standard deviation also increases greatly from 1.04 to 85.03 because of the outlier. This is expected because the standard deviation measures the spread of the data points and with the outlier being far away from the other data points, the variability of the distribution is understandably high.

As mentioned above, we need to treat outliers with care. If they have minimal effect on the conclusions and if we cannot figure out why they are there, such outliers may possibly be removed. However, if they substantially affect the results, then we should not drop them without justification.

Example 3.1.13 Suppose we are interested to find out if there are significant differences in the distribution of HDB resale prices for different time periods. For example, would the distributions differ significantly if we compare the period July to December 2020 with January to June 2021? The two distributions are shown below.

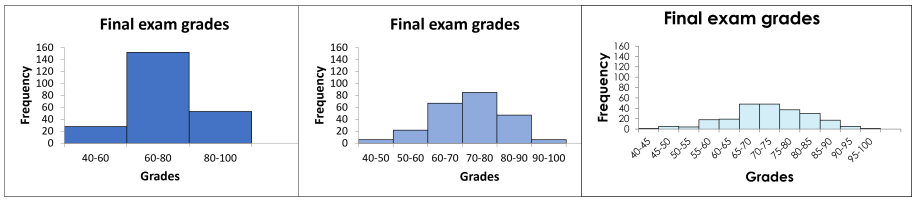


The distribution on the left corresponds to the period of resale from July to December 2020. The distribution for January to June 2021 is shown on the right. We observe that both distributions have a similar shape which is right skewed with a single peak. Taking it one step further, we compare the central tendencies and variabilities of the data points in both periods. The values in the table can be computed using the Microsoft Excel Data Analysis Toolpak.

	Mean	Median	Mode	Range	Standard deviation
July to December 2020	\$462,827	\$435,000	\$400,000	\$1,098,000	\$155,955
January to June 2021	\$496,870	\$468,000	\$420,000	\$1,070,000	\$162,107

Observe that all measures of mean, median and mode are higher in the time period January to June 2021 compared to those in the time period July to December 2020. The range of the resale prices is lower in January to June 2021 while the standard deviation is actually higher. In conclusion, we can say that resale prices in January to June 2021 are higher, but more spread out (in terms of standard deviation) compared to the resale prices in July to December 2020.

Example 3.1.14 In Example 3.1.4, we described the setting of bin widths when creating a histogram. Deciding the bin width to use can have a big impact on how the histogram looks like and thus affect our observation and conclusion on the shape of the distribution.



The three histograms above are constructed using the same data set of 233 students' final exam scores with the only difference being the bin width settings. The histogram on the left has a bin width of 20, while the one in the middle has bin width of 10. The last histogram has bin width set at 5. What conclusions can be made on the distribution based on these histograms?

Based on the first histogram, we may make the conclusion that most students score between 60 to 80 marks, and the distribution is rather symmetric. However, with a slightly smaller bin width, the second histogram reveals that most students actually scored between 70 to 80 marks. This does not contradict the observation made earlier based on the first histogram but because of the smaller bin width, we are able to narrow the range of marks that are scored by most students. With an even smaller bin width, the third histogram suggests that most students scored between 65 and 75 marks. How do you rationalise this conclusion with the one from the second histogram?

In general, we should bear in mind the following when determining bin widths for histograms.

1. Avoid histograms with bin widths that are too large. This will result in only a few bins and information in the data will be lost when data points are grouped together into a small number of groups/bins.

2. Avoid histograms with bin widths that are too small. If we do this, there may be bins that have very few data points (or none) that does not give us a sense of the distribution.

3. Our initial choice of bin width may not be the most appropriate. Different histograms with various bin widths should be created before deciding which one is the most useful and informative.

Remark 3.1.15 We should not confuse histograms with bar graphs introduced in Chapter 2. A histogram shows the distribution of a numerical variable across a number line. So one of the axes (usually the horizontal) will display the range of values taken on by the numerical variable. On the other hand, the horizontal axis of a bar graph will show the different categories of a categorical variable.

In addition, the ordering of the bars in a histogram cannot be changed, as it progresses through the range of values, usually in an ordered manner, taken on by the numerical variable. On the other hand, the ordering of the bars in a bar graph can be switched around with little consequence. There are also usually no gaps between the bars in a histogram.

Discussion 3.1.16 (Boxplots for Univariate EDA) Besides a histogram, another way to visualise the distribution of a numerical variable is to use a *boxplot*. To construct a boxplot, we will use the five-number summary, consisting of

1. Minimum;

2. Quartile 1 (Q_1);

3. Median (Q_2);

4. Quartile 3 (Q_3);

5. Maximum.

The median and quartiles have already been introduced in Definition 1.6.1 and Definition 1.6.5. Furthermore, we have also introduced the Interquartile range

$$\text{IQR} = Q_3 - Q_1.$$

While the median can be viewed as the center of a data set, the IQR is a way to quantify the spread of a data set. We have defined an outlier in Definition 3.1.11 but did not provide an explicit way to classify a data point as an outlier. For our purpose we will adopt the following consideration to classify a data point as an outlier.

> A data point is considered an *outlier* if it satisfies one of the following conditions:
>
> - The value of the data point is **greater than** $Q_3 + 1.5 \times \text{IQR}$;
> - The value of the data point is **less than** $Q_1 - 1.5 \times \text{IQR}$.

To construct a boxplot, we do the following:

1. Draw a box from Q_1 to Q_3.

2. Draw a vertical line in the box where the median (Q_2) is located.

3. Identify all the outliers by using the consideration above.

4. Extend a line from Q_1 to the smallest value that is not an outlier and another line from Q_3 to the largest value that is not an outlier. These lines are called *whiskers*.

5. Mark each of the outliers with dots or asterisks.

Example 3.1.17 Consider the following data set, with the data points already sorted in increasing order.

$$18,\ 44,\ 47,\ 55,\ 61,\ 62,\ 78,\ 79,\ 83,\ 145.$$

There are 10 data points. The median (Q_2) is the average of the fifth and sixth data points, so

$$Q_2 = \frac{1}{2}(61 + 62) = 61.5.$$

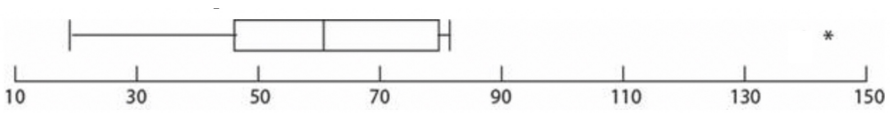
The first quartile is the median of the first five data points: $18, 44, 47, 55, 61$, so $Q_1 = 47$. The third quartile is the median of the last five data points: $62, 78, 79, 83, 145$, so $Q_3 = 79$. Following Remark 1.6.9, it should be pointed out that you may encounter slightly different ways of finding quartiles for a data set in other texts. For this course, we will adopt what is presented here.

The Interquartile Range is

$$\text{IQR} = Q_3 - Q_1 = 79 - 47 = 32.$$

To determine if we have outliers, note that $1.5 \times \text{IQR} = 48$. Since there are no data points smaller than $Q_1 - 48$, there are no small-valued outliers. On the other end, since $145 > Q_3 + 48 = 79 + 48 = 127$, we see that 145 is the only big-valued outlier.

The boxplot constructed is shown below.



Example 3.1.18 Let us return to the HDB resale flats data set. The boxplot below is based on the resale prices of flats sold in January to June 2021.

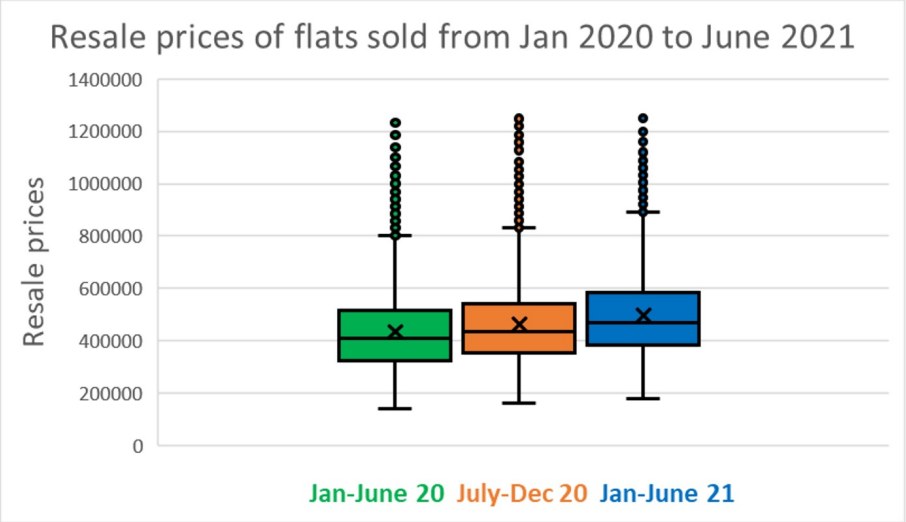
The boxplot confirms our earlier conclusion that there are outliers that correspond to very high resale prices. Note that the cross in the box, just above the median line represents the mean resale price. Recall that we have discussed the shape, center and spread of the distribution using a histogram. What can we say based on the boxplot?

1. **(Shape)** From the boxplot, we see that the variability in the upper half of the data, given by (Max − Median) is significantly larger than the variability in the lower half of the data which is equal to (Median − Min). This confirms our earlier observation that the distribution is skewed to the right and there is a relatively long tail to the upper end of the distribution due to the existence of outliers.

2. **(Center)** The center, described by the median is easily observed in the boxplot, unlike in a histogram. We can also compare the relative positions of the median and the mean from the boxplot.

3. **(Spread)** The IQR of 204, 000 gives us an idea of the spread for the middle 50% of the data set. On its own it may not be immediately informative but this would be a meaningful measure to compare across different distributions (see next example).

Example 3.1.19 The three boxplots below show the distributions of resale flat prices in three different time periods, namely January to June 2020 (call this period P1), July to December 2020 (call this period P2) and January to June 2021 (call this period P3). What can we say about the three distributions after comparing the three boxplots?



1. All three distributions are right skewed as the upper halves of the data have greater variability than the lower halves, due to (large-valued) outliers. However, upon a closer look, it is also apparent that the upper half variability in period P1 is greater than the upper half variability in P2 which in turn is greater than the upper half variability in P3.

2. The middle 50% (that is, the IQR) box of resale prices is lowest in P1, followed by P2 and then P3. Hence, the overall resale prices have increased over time. The spread (given by the height of the boxes) appears to be similar between P1 and P2 while slightly higher in P3.

3. There appears to be more outliers in P1 and P2 compared to P3.

To conclude this section, we summarise the comparison between using histograms and boxplots to represent a distribution.

1. A histogram typically gives a better sense of the shape of the distribution of a variable, compared to a boxplot. When there are great differences among the frequencies of the data points, a histogram will be able to illustrate this difference better than a boxplot.

2. If we wish to compare the distributions of different data sets, putting the different boxplots side by side is more illustrative than using histograms.

3. To identify and indicate outliers, boxplots do a better job than histograms.

4. The number of data points we have in a data set is better shown in a histogram than in a boxplot. In fact, two distributions with very different number of data points can have almost identical boxplots. On the other hand, this difference is apparent by comparing the histograms.

The bottom line is that different graphics and summary statistics have their advantages and disadvantages and they are often used together to complement each other.

Section 3.2 Bivariate EDA

In this section, we will focus on how we can investigate a relationship between two variables in a population.

Discussion 3.2.1 We start off with a relationship between two variables that is *deterministic*. This means that the value of one variable can be determined exactly if we know the value of the other variable. Perhaps the most common type of deterministic relationship is the one that involves the conversion of units of measurement from one metric to another. For example:

1. The relationship between Fahrenheit (F) and Degree Celsius (C) in the measurement of temperature. We know that F and C are related by
$$C = (F - 32) \times \frac{5}{9}.$$
This is a deterministic relationship between F and C. For example, if the temperature in the oven now is 450 degrees Fahrenheit (so $F = 450$), then the temperature in the oven now, measured in Degree Celsius is
$$C = (450 - 32) \times \frac{5}{9} = 232.22.$$
2. Meters (M) and Feet (F) are both measurements of length (or height) and they are related (approximately) by
$$F = 3.2808 \times M.$$
So, if Johnny's height is 5.9 Feet (so $F = 5.9$), then his height in meters will be
$$M = \frac{F}{3.2808} = \frac{5.9}{3.2808} \approx 1.8 \text{ meters}.$$

Discussion 3.2.2 The main focus of this section is on a relationship between two variables that is not deterministic in nature. We say such a relationship is *statistical* or non-deterministic. Recall that in a deterministic relationship, given the value of one variable, we can find a unique value of another variable. However, this is not possible for a statistical relationship, where given the value of one variable, we can describe the average value of the other variable. Such relationships between variables, called *associations* occur quite often in our daily life.

Example 3.2.3 In a Medical News Today article published in November 2020, it was reported that in a study involving more than 150,000 participants, a clear link was observed between low physical fitness and the risk of experiencing symptoms of depression, anxiety, or both.


Large study finds clear association between fitness and mental health

New research from a large study demonstrates that low cardiorespiratory fitness and muscle strength have a significant association with worse mental health.

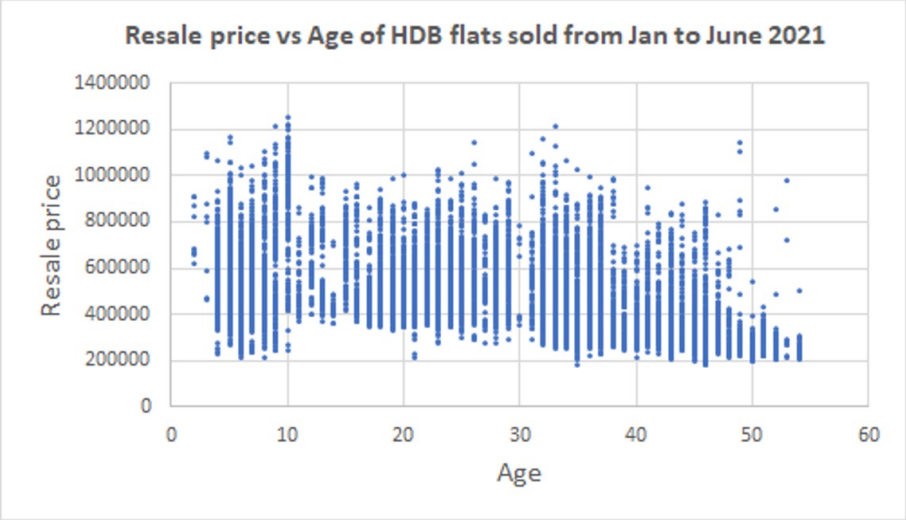
This association between physical fitness and mental health may not be surprising but we wonder if it could be due to other factors, like a confounder. More interestingly, does having better fitness make a person mentally healthier or having better mental health make a person exercise more resulting in better physical fitness? We will not only measure the association (if one exists) between variables but also attempt to interpret any observed associations.

Bivariate data is data involving two variables. For example, in the HDB resale flat data set, we can study the two variables Age and Resale Price .

Month	Floor area sqm	Age	Resale price
1/1/2021	45	35	225000
1/1/2021	45	35	211000
1/1/2021	73	45	275888
1/1/2021	67	43	316800
1/1/2021	67	43	305000
1/1/2021	68	40	260000
1/1/2021	73	44	351000
1/1/2021	73	44	343000
1/1/2021	75	41	306000

In Section 3.1, we saw two ways to display univariate data, using either a histogram or a boxplot. For bivariate data, it is clear that using a table like the one above is not really useful if we wish to investigate if the two variables are associated. Instead, we will use a *scatter plot* to give us an idea of the pattern formed by the data between the two variables in question. After looking at the scatter plot, we use a quantitative measure called the *correlation coefficient* to quantify the level of linear association (if any) between the two variables. Finally, we will attempt to fit a line or a curve through the points in the data set which will enable us to make predictions on the values of the variables. This process is known as *regression analysis*. For now, we will focus on scatter plots and defer the discussion on correlation coefficients and regression analysis to the next few sections.

Example 3.2.4 Returning to our HDB resale flats prices data set, we will focus on the bivariate data with the variables Age and Resale price. Suppose we wish to know if the age of the flat affects the resale price, with the ultimate intention to make a prediction, based on the past resale prices, of how much a 38 year old resale flat is likely going to cost. In this case, we can treat age as the independent (or explanatory) variable and resale price as the dependent (or response) variable.



Our scatter plot shown above has the age (independent) variable on the x-axis and the resale price (dependent) variable on the y-axis. Each resale transaction would be represented by an *ordered pair*

$$(x, y)$$

where x is the age of the resale flat and y is the resale price of that flat. For example, the ordered pair $(35, 225000)$ corresponds to the first resale flat listed in the table above. With a point plotted for each ordered pair, since there are $11,644$ resale transactions in the data set, there will be $11,644$ points on the scatter plot. Observe that in the scatter plot, each value of x (age of flat) corresponds to many different values of y (the resale price). This is to be expected because there are many different transactions involving flats of the same age and all these transactions are made at different resale prices.

How do we describe the relationship between two numerical variables using a scatter plot?

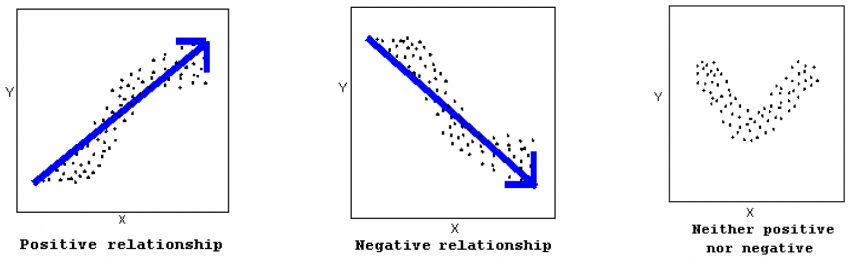
Univariate data		Bivariate data	
Overall pattern	Deviation from the pattern	Overall pattern	Deviation from the pattern
1) Shape 2) Center 3) Spread	Outliers	1) Direction 2) Form 3) Strength	Outliers

We have seen that for univariate data, we discussed the shape (symmetrical or skewed), center (median, mean and mode) and spread (interquartile range, standard deviation and range) of the distribution. For bivariate data, we will use descriptors like the direction, form and strength to describe the relationship between the two variables. For both univariate and bivariate data, data points that deviate significantly from the pattern of the main bulk of data points are called outliers.

Definition 3.2.5 The *direction* of the relationship can be either positive, negative or neither. We say that there is a positive relationship between two variables when an increase in one of the variables is associated with an increase in the other variable.

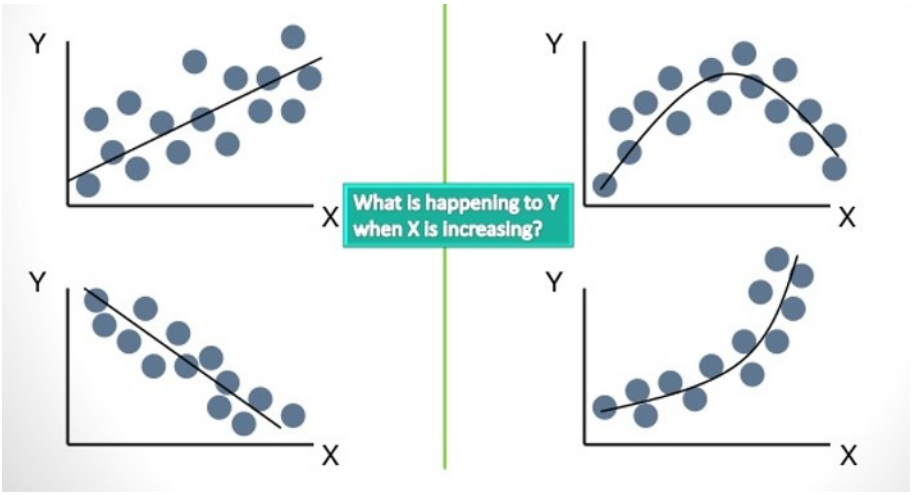
On the other hand, a negative relationship between two variables means that an increase in one variable is associated with a decrease in the other.

Not all relationships can be classified as either positive or negative and there are those that do not behave in one way or the other.



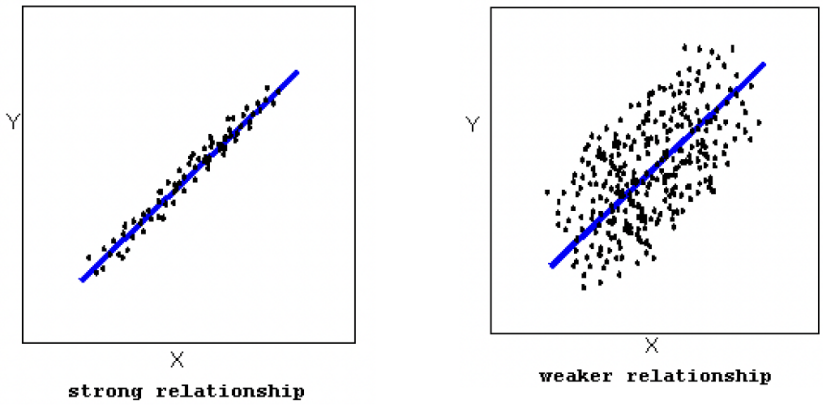
The *form* of the relationship describes the general shape of the scatter plot. In general, we can classify the form of the relationship as either linear or non-linear. The form of the relationship is linear when the data points appear to scatter about a straight line. Later in the chapter, we will use a mathematical equation to describe the straight line when the form of the relationship between two variables is linear.

When the data points appear to scatter about a smooth curve, we say that the form of the relationship is non-linear. It is beyond the scope of this course to summarise curve patterns in the data but it is useful to note that quadratic and exponential equations are examples of non-linear forms of relationship.



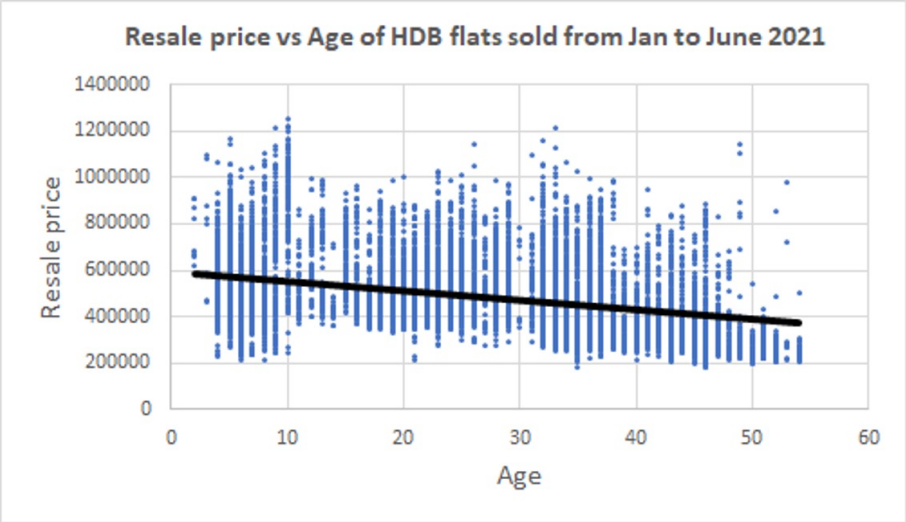
The two scatter plots on the left shows a linear form of the relationship between the two variables while the two scatter plots on the right shows non-linear forms.

The *strength* of the relationship indicates how closely the data follow the form of the relationship.



Both scatter plots above suggests that there is **positive**, **linear** relationship between the two variables. However, the scatter plot on the left shows the data points lying very close to the straight line. This indicates that the strength of the relationship is *strong*. The scatter plot on the right shows the data points scattered loosely around the straight line and thus the strength of the relationship is *weaker* than that in the scatter plot on the left.

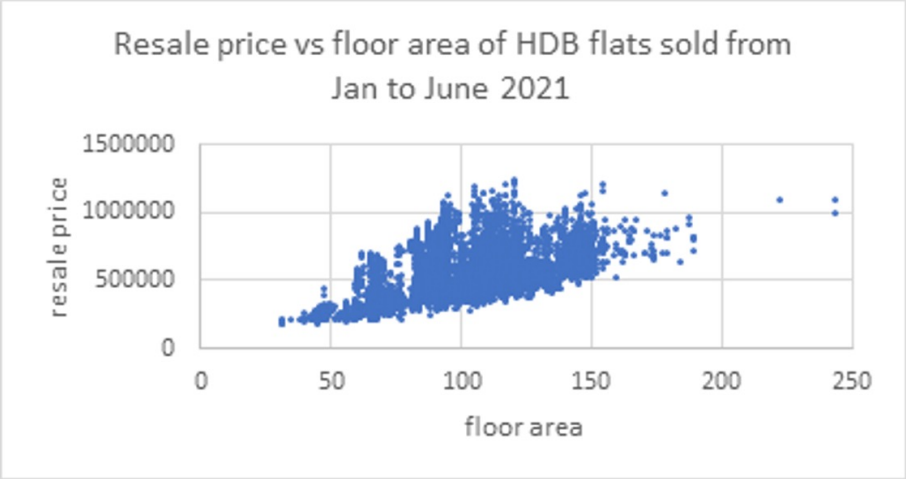
Example 3.2.6 Let us look at the scatter plot from the HDB resale flats data again. The scatter plot below is similar to the one from Example 3.2.4 except for an additional *trendline* drawn in black.



The trendline suggests that as the age of the HDB flat increases, the resale price decreases linearly on average, in the period of January to June 2021. Is this relationship strong or weak? In fact, one can argue that without the trendline, one may not even observe that there is a linear relationship between age and resale price.

At this point, we cannot really tell if there is indeed a linear relationship and if there is, whether the relationship is strong or weak. Nevertheless, in the next section, we will discuss a more precise measure of the strength of a relationship.

As mentioned earlier, outliers are data points that deviate significantly from the pattern of the relationship. Consider the scatter plot shown below that plots the resale price against the floor area of the HDB resale flats. Do you observe any outliers?



Recall that for univariate data, using a boxplot, we can determine if a data point is an outlier by checking if its value is greater than $Q_3 + 1.5 \times \text{IQR}$ or smaller than $Q_1 - 1.5 \times \text{IQR}$. What about for bivariate data? We will discuss more about outliers in the next section.

Section 3.3 Correlation coefficient

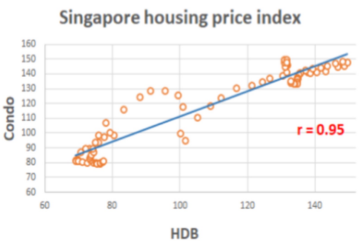
In the previous section, using the HDB resale flats data set, we have observed that a flat's resale price is associated with the age of the flat. From the scatter plot, we concluded that the relationship between the age of the flat and the resale price of the flat was negative. This means that flats whose ages were higher tended to have a lower resale price. This is not surprising. However, can we say anything about whether this relationship is strong or weak? If possible, can we measure the strength of this relationship using a number?

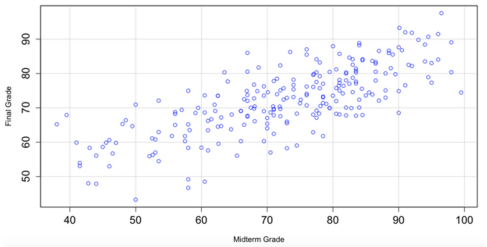
More generally, given two numerical variables, is it possible for us to measure the relationship between the two variables quantitatively?

Definition 3.3.1 The *correlation coefficient* between two numerical variables is a measure of the **linear** association between them. The correlation coefficient, denoted by r, always ranges between -1 and 1. We can use this number to summarise the direction and strength of linear association between two variables.

The **sign** of r tells us about the **direction** of the linear association. If $r > 0$, then the association is *positive*, which means that when one of the variables increase, the other variable will tend to increase as well. On the other hand, if $r < 0$, then the association is *negative*, which means that when one of the variables increase, the other variable will tend to decrease. In the event that $r = 1$ (resp. $r = -1$), we say that there is *perfect* positive association (resp. negative association). When $r = 0$, we say there is *no linear association*. Thus, while the sign of r tells us the direction of the linear association, the **magnitude** of r (that is, how close r is to 1 or -1) will tell us the strength of the linear association between two numerical variables.

Example 3.3.2 The two scatter plots below are examples of positive linear association between two variables.

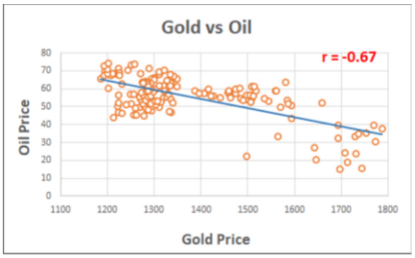


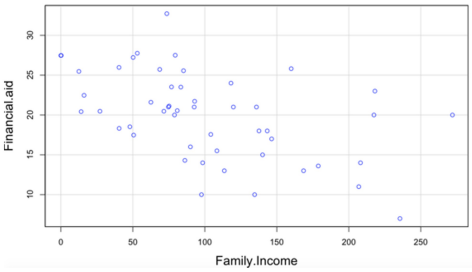


The plot on the left plots the price index of HDB flats against the price index of condominiums. We observe that there is positive linear association between the two indices, which means that as the price of HDB flats increase, it is likely that the price of condominiums would increase as well. The value of r in this case is 0.95 which indicates that the association is strong.

The plot on the right shows the midterm mark of students against the final mark. Again, we observe that there is positive linear association between the two marks and in this case, r was found to be 0.75.

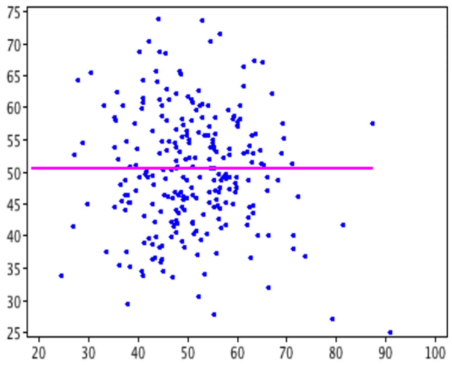
The next two scatter plots are examples of negative linear association between two variables.

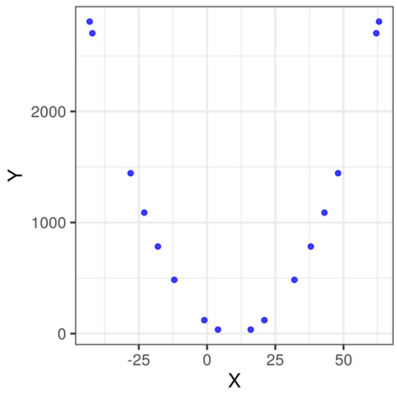




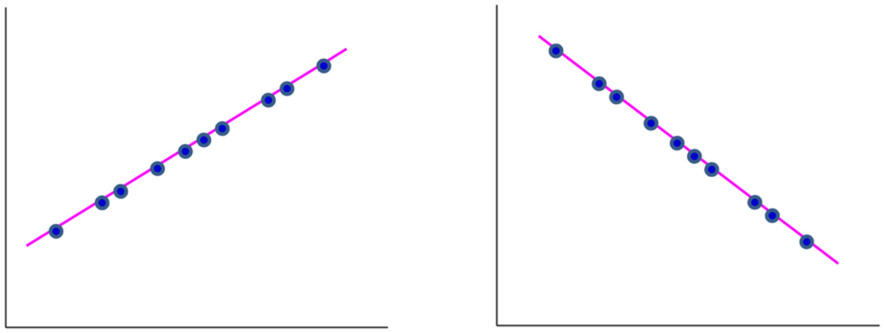
The plot on the left shows the price of oil against the price of gold. In this case, we observe that the trend is that when the price of gold increases, the price of oil tends to decrease. The value of r was found to be -0.67 and this indicates that there is negative linear association between gold and oil prices.

The plot on the right shows the amount of financial aid received by students against the students' family income. It is not surprising to find that as the family income increases, the amount of financial aid received by students would tend to decrease. The value of r in this case is -0.49 and there is negative linear association between the two variables.



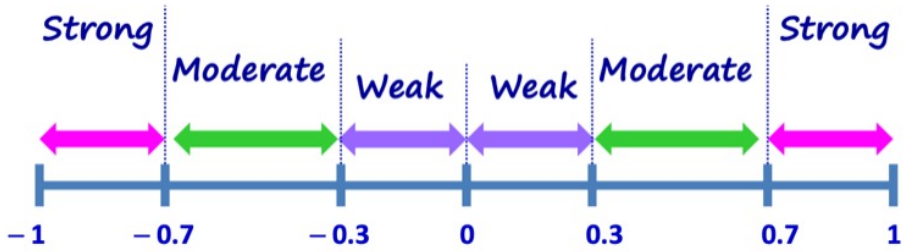


The two scatter plots above are examples where $r = 0$. This means that there is no linear association between the two variables. However, note that while $r = 0$ for the second plot, we can see that the data points fit very well onto a curve and there is a clear non-linear relationship between X and Y. More generally, no *linear association* between variables does not necessarily mean no association between variables.



The two plots above show situations where there is perfect (positive or negative) linear correlation between the two variables. In such cases, all the data points are connected by (and thus lie on) a straight line. There is however, one exception, which is when the straight line joining all the data points is actually a straight horizontal (or vertical) line. In such instances, the value of r is 0 and there is no association between the two variables. This is because when the data points are connected by a vertical or horizontal line, a change of value in one of the variables does not relate to a change in the other variable.

When describing the strength of a linear relationship, we usually follow the rule of thumb as given in the diagram below.



When the magnitude of r is between 0.7 and 1, we say that the two variables have a *strong linear association*. If the magnitude is between 0.3 and 0.7, the two variables have a *moderate linear association*. If the magnitude is between 0 and 0.3, the two variables have a *weak linear association*. Do note that other sources may differentiate strong/moderate/weak linear associations at other "cut-off" points that are different from 0.3 and 0.7.

In general, as the value of r becomes closer to 1 or -1, the data points will increasingly fall more closely to a straight line. Scatter plots where the data points are loosely dispersed typically mean that correlation is weak (or non-existent). We will now discuss how to compute the value of r numerically.

Example 3.3.3 We will go through the steps required to compute the correlation coefficient using an example. Consider the following table that shows a total of 10 data points of bivariate data (x, y):

x	9	4	5	10	6	3	7	2	8	1
y	41	17	28	50	39	26	30	6	4	10

1. First compute the mean and standard deviation of x and y. (Refer to Definition 1.4.1 and Definition 1.5.1 if you have forgotten how these are computed.) For this data set, we find the mean and standard deviation of x to be 5.5 and 3.03 respectively while the mean and standard deviation of y are 25.1 and 15.65 respectively.

2. Convert each value of x and y into *standard units*. To convert x (resp. y) into its standard unit, we compute
$$\frac{x-\overline{x}}{s_x} \qquad \left(\text{resp. } \frac{y-\overline{y}}{s_y}\right),$$
where s_x and s_y are the standard deviations of x and y respectively. The table below shows the values of x and y after they have been converted to standard units.

x	1.16	−0.50	−0.17	1.49	0.17	−0.83	0.50	−1.16	0.83	−1.49
y	1.02	−0.52	0.19	1.59	0.89	0.06	0.31	−1.22	−1.35	−0.96

3. Compute the product xy in their standard units for each data point. The table below has an additional row for the value xy for each data point.

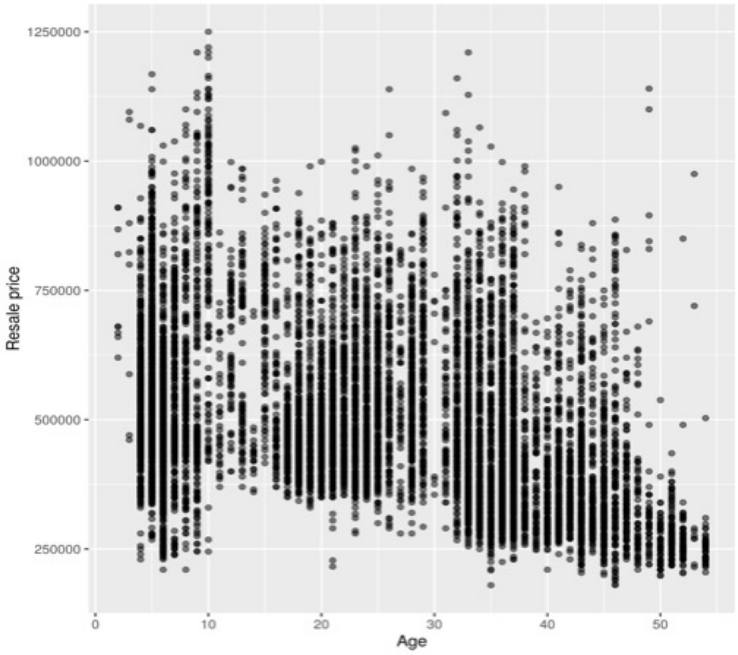
x	1.16	−0.50	−0.17	1.49	0.17	−0.83	0.50	−1.16	0.83	−1.49
y	1.02	−0.52	0.19	1.59	0.89	0.06	0.31	−1.22	−1.35	−0.96
xy	1.17	0.26	−0.03	2.36	0.15	−0.05	0.15	1.41	−1.11	1.43

4. Sum the products xy obtained in the previous step over all the data points and then divide the sum by $n-1$, where n is the number of data points. The result is the correlation coefficient r. For the data set above,

$$r = \frac{1}{9}(1.17 + 0.26 - 0.03 + 2.36 + 0.15 - 0.05 + 0.15 + 1.41 - 1.11 + 1.43) = 0.64.$$

Remark 3.3.4 For the purpose of this module, you will not be required to compute r manually, instead you should be familiar with the method of how r is computed and thereby develop some basic intuition on the properties of r.

Example 3.3.5 Let us revisit Example 3.2.6, where the scatter plot of HDB resale flat prices against the ages of the flat shown below does indeed suggest that these two variables are negatively associated.



Indeed, upon computing the correlation coefficient between these two variables, we find that $r = -0.356$, confirming that there is moderate negative linear association between the age and resale price of HDB flats from the period January to June 2021.

We will now present three properties of correlation coefficients.

1. From the "Age" vs. "Price" of HDB resale flats example, we saw that $r = -0.36$ when we consider the scatter plot with Age as the x-axis and Resale price as the y-axis. What would happen to r if we had done the plot with Resale price as the x-axis and Age as the y-axis? In other words, what happens to r when we interchange the x and y variables? If we revisit the process that describes how r is computed from a bivariate data set, you would realise that regardless of which variable is x (or y), the computation of r would not be affected in any way.

> The correlation coefficient r is not affected by interchanging the x and y variables.

2. What would happen to the value of r if we add a constant to all the values of a variable? For example, suppose it was discovered that there was an error in the recording of all the resale prices of HDB flats and that the actual resale prices were all \$1000 higher than what was given in the data set. To correct this error, we would have to add \$1000 to all the resale prices in the data set. It turns out that such a change **does not** affect the value of r.

> The correlation coefficient r is not affected by adding a number to all values of a variable.

   While this may not be immediately obvious, you are encouraged to verify this result by using the data set in Example 3.3.3 and adding some number to all the values of x (or y).

3. Instead of adding the same number to all the values of a variable, what would happen to the value of r if we multiply a positive number to all the values of a variable instead? For example, if the resale prices were converted to US dollars instead? This means that we have to multiply a factor of 0.73 (assuming an exchange rate of 1 Singapore dollar is to 0.73 US dollars) to all the resale prices in the data set. It turns out that such a change again **does not** affect the value of r.

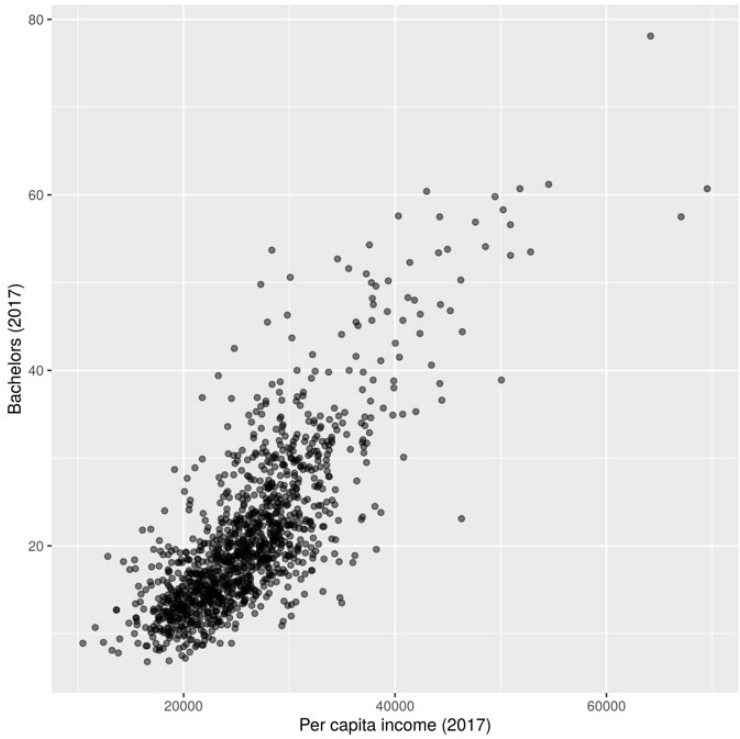
> The correlation coefficient r is not affected by multiplying a positive number to all values of a variable.

   You are again encouraged to verify this result by adjusting the data set in Example 3.3.3 and recalculating the correlation coefficient.

While the correlation coefficient between two numerical variables is insightful, there are certain limitations.

Discussion 3.3.6

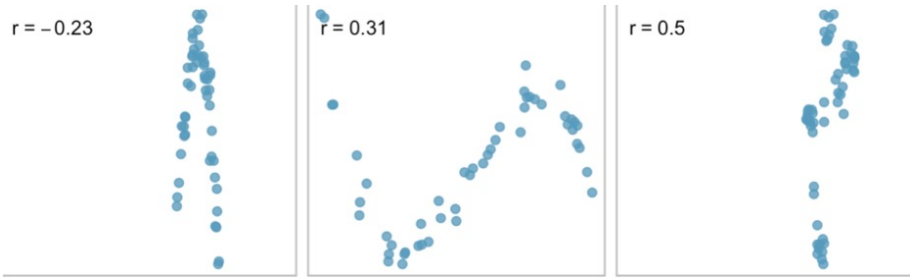
1. **Association is not causation.** To confuse association with causation is a common mistake that is made by many. Very often when there is a strong association between two variables, with a correlation coefficient of r that is close to 1 or -1, it is mistakenly concluded that any change in the explanatory variable, say x, will cause the response variable y to change. This is incorrect as what we can conclude is only a *statistical relationship* between x and y and not a *causal relationship*.



Consider the example above of a scatter plot that came from a data set containing information on the percentage of people that earned a Bachelor's Degree in 2017 across 3142 counties in the United States, as well as the per capita income of these counties in 2017.[2] Each data point in the scatter plot represents a county. The x-axis is the per capita income in the past 12 months while the y-axis is the percentage of the population in the county that earned a Bachelor's Degree in 2017. The correlation coefficient for the two variables is 0.79, which indicates that there is strong and positive association between the two variables.

It would be tempting to conclude that the higher the per capita income of a county, the higher the percentage of the county's population would have earned a Bachelor's Degree. This is not necessarily true. The data here merely suggests association of the two variables and does not establish any causal relationship.

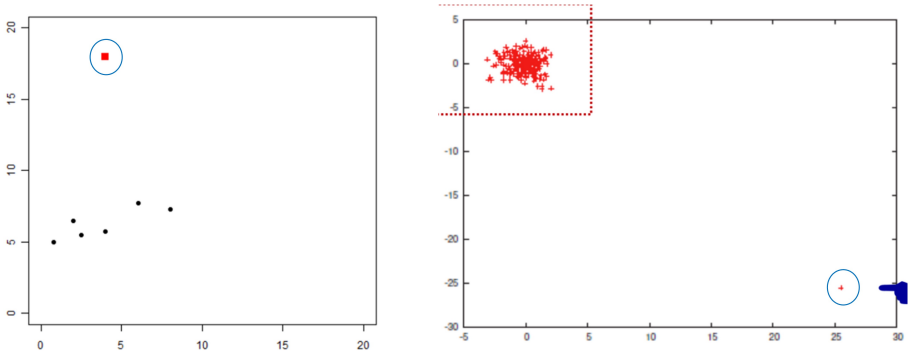
2. r **does not tell us anything about non-linear association.** The correlation coefficient r, as defined and described in this section, measures the degree of linear association between two numerical variables. Whatever the computed value of r is, it does not give any indication of whether the two variables could be associated in a non-linear way.



The correlation coefficients for the three scatter plots above are small but yet there is actually a strong relationship between the variables. The value of r is small because the relationship between the variables is not a linear one. It is always a good practice to look at a scatter plot of the data set and not just deduce any relationship between the variables from the computed value of r.

[2]Data set can be downloaded from www.openintro.org/data/?data=county_complete.

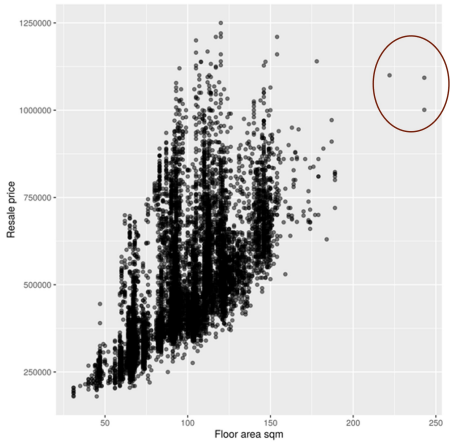
3. **Outliers can affect the correlation coefficient significantly.** Outliers are observations that lie far away from the overall bulk of the data. How do outliers affect the value of the correlation coefficient? The removal of outliers from a data set can have different effects on the correlation coefficient, depending on how the outlier is positioned in relation to the rest of the data points.



Consider the scatter plot on the left, where the outlier is circled, the correlation coefficient is 0.22 based on the data set that includes the outlier. However, when we remove the outlier, we see that there is a strong positive linear association between the remaining data points. Thus, in this case, the presence of the outlier decreases the strength of the correlation, compared to when the outlier is removed.

Consider the scatter plot on the right where again the outlier is circled. In this case, the correlation coefficient is -0.75 based on the data set that includes the outlier. When the outlier is removed, the remaining data points give a correlation coefficient of 0.01. Thus, in this case, the presence of the outlier actually increases the strength of the correlation, compared to when the outlier is removed.

Example 3.3.7 For the HDB data set that we introduced earlier, the scatter plot below shows the relationship between the resale price and the floor area of the flat. There are three outliers (circled) and these are resale flats whose floor areas are larger than 200 square meters.



Using a statistical software, it was found that the correlation coefficient was 0.626 before the outliers were removed. After the outliers are removed, the correlation coefficient becomes 0.625, which is practically the same as before.

Section 3.4 Linear regression

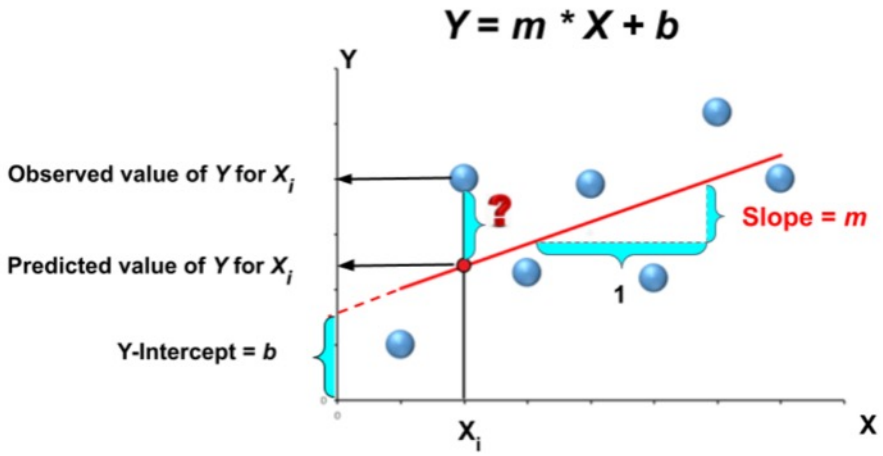
Now that we have seen that the age of a HDB resale flat is negatively associated with the resale price, it is reasonable to wonder if we can make some predictions on the resale price of a flat given the age of

the flat. For example for a flat that is 40 years old, what is our guess for its resale price?

Definition 3.4.1 If we believe that two variables X and Y are linearly associated, we may model the relationship between the two variables by fitting a straight line to the observed data. This approach is known as *linear regression*. Recall that the equation of a straight line is given by

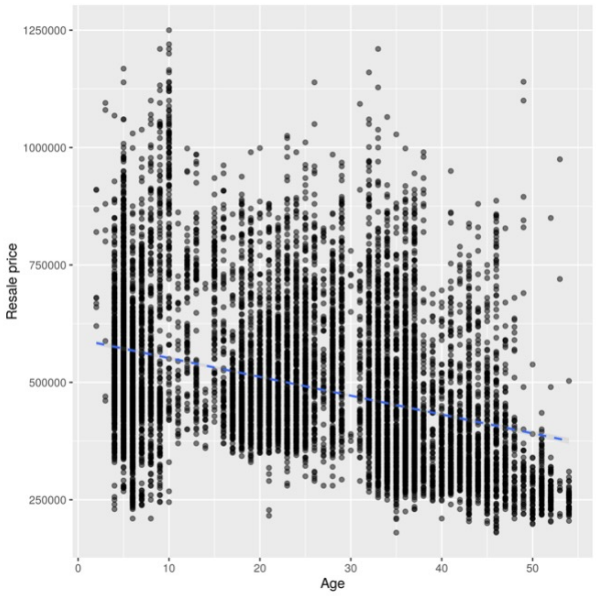
$$Y = mX + b,$$

where b is the *y-intercept* and m is the *slope* or *gradient* of the line. The y-intercept is the value of Y when the value of X is 0. The slope of the line is the amount of change in Y when the value of X increases by 1.



In the figure above, the straight line in red is the regression line that is fitted to the observed data, represented by the blue dots. Consider the i-th observation (X_i, Y_i). The "?" in the figure represents the *residual* of the i-th observation, which is the observed value of Y for X_i (that is, Y_i) minus the predicted value of Y for X_i (predicted by the straight line). This residual, denoted by e_i, is sometimes also called the *error* of the i-th observation as it measures how far the predicted value is from the observed value.

Example 3.4.2 Let us return to the question we posed at the beginning of this section. What is our prediction for the resale price of a HDB flat that is 40 years old?



With X representing the age of the resale flat and Y being the resale price, the regression line obtained from the data set is

$$Y = -4007X + 591857.$$

This means that when $X = 40$, (age of resale flat is 40),

$$Y = -4007 \times 40 + 591857 = 431577.$$

So the **predicted** resale price of a 40 year old flat is \$431,577. It is important to note that we are **not concluding** that

> A 40 year old resale flat will be sold at \$431,577.

But instead our linear regression model predicts that

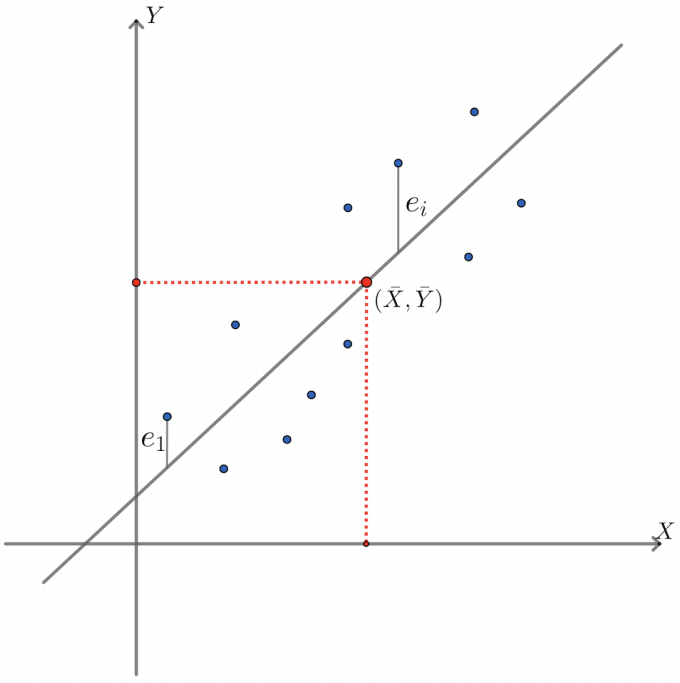
> The average resale price of 40 year old HDB flats is \$431,577.

Furthermore, as the correlation between resale flat price and age of the flat is weak, the prediction obtained from the linear regression above may not be as accurate compared to the scenario where the correlation is stronger.

Now that we have seen how a regression line can be used, the question is how do we obtain such a line given bivariate data? What method and principle is used to determine the regression line? Among the many different straight lines that we can use to fit the data points, which one is the "best"?

Discussion 3.4.3 There are several ways to assess which straight line fits the observed data better. One of the most common way is the *method of least squares*. For this module, we will not go into the technicalities of this method but instead we will briefly describe the idea behind this method.

Recall that when we fit a straight line through a set of observed data points (x_i, y_i), the difference between the observed value y_i and the predicted outcome, predicted by the straight line, is known as the residual of the i-th observation. This residual, denoted by e_i is also known as the error of the i-th observation that measures how far is the observed from the predicted.



In the plot above, we see that each data point gives rise to an error term and it is reasonable to say that a line of good fit is one that keeps the error terms (considered over all data points) small. However, instead of looking at the overall error by summing up

$$e_1 + e_2 + \cdots + e_n,$$

where n is the total number of data points, the method of least squares seek to find a straight line that minimises the overall **sum of squares of errors**,

$$e_1^2 + e_2^2 + \cdots + e_n^2.$$

You may wonder why minimising $e_1^2 + e_2^2 + \cdots e_n^2$ is more appropriate than minimising $e_1 + e_2 + \cdots + e_n$. We will leave you to ponder about this question before having a discussion with your friends or instructor.

Remark 3.4.4

1. It is useful to note that the least squares regression line obtained from a set of observed data points (x_i, y_i) will **always** pass through that point of averages for that data set, that is, $(\overline{x}, \overline{y})$. This fact can be established mathematically, but is beyond the scope of this course.

2. It is important to note that while we have obtained the least squares regression line that allows us to predict the average resale price for a given age of the resale flat, the same regression line cannot be used to predict the average age of resale flats for a given resale price. The reason is essentially because of the way the regression line was obtained.

   In obtaining the regression line with the independent variable (x) as age and the dependent variable (y) as the resale price, the line was fitted to minimise the square of error terms between the observed and predicted resale prices.

   If the intention was to use a given resale price to predict the average age of the resale flats, then we would be looking at another regression line that minimises the square of error terms between the observed and predicted ages of resale flats.

   The two regression lines are different and thus not interchangeable.

3. The correlation coefficient r between the variables X and Y is closely related to the regression line

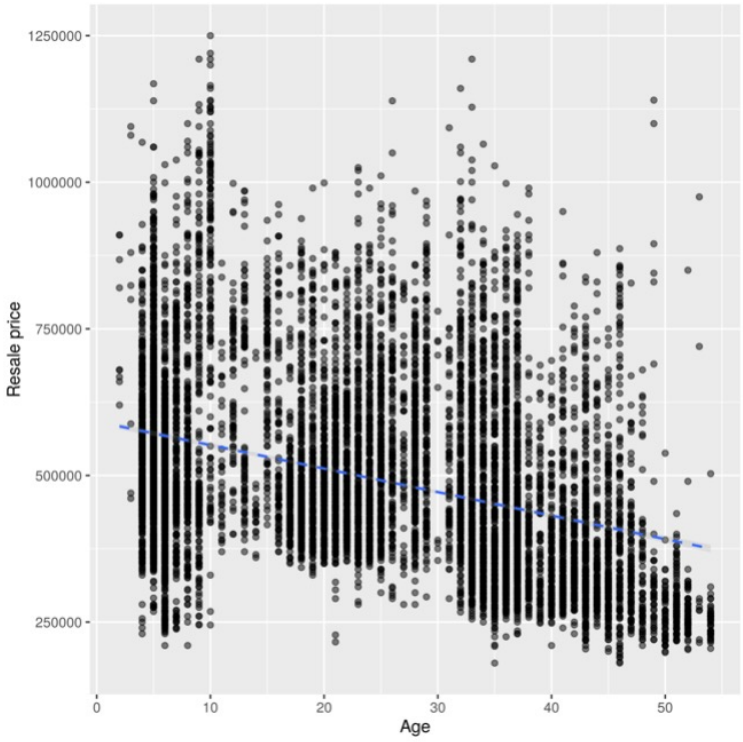
   $$Y = mX + b$$

   obtained using the method of least squares. More precisely, we have

   $$m = \frac{s_Y}{s_X} r,$$

   where s_X (resp. s_Y) is the standard deviation of X (resp. Y). With this relationship, we see that if the correlation coefficient r is positive, then the gradient of the regression line is also positive. Similarly, if the correlation coefficient is negative, then the gradient of the regression line will also be negative. However, it is important to remember that the correlation coefficient is not necessarily **equal** to the gradient of the regression line.

4. Another important point to note about the linear regression line obtained using a data set is with regards to the range of the independent variable in the data set.



Recall that we have obtained the linear regression line for the purpose of predicting the average resale price based on the age of the resale flat. From the data set, the value of the independent variable (in this case, this is the age of the resale flat) ranges from 2 to 54 years. Thus the prediction that can be arrived at using the regression line is only applicable for HDB flats whose age is between 2 and 54 years old. Outside this range, we should not use the regression line to make our prediction as the best fit regression line may change outside this range. For example, we should not use the regression line to predict the average resale price of flats that are 60 years old as our data set does not contain any information on resale flats that are more than 54 years old.

Discussion 3.4.5 To conclude this section and also the chapter, we will describe a method to study the relationship between two variables if the relationship is not linear. The following table shows part of a data set that provides the total number of confirmed COVID-19 cases in South Africa since 5 March 2020.[3].

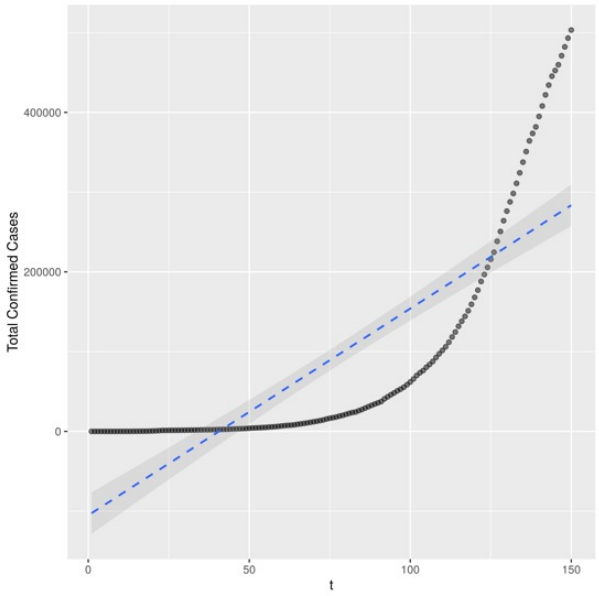
t	Total confirmed cases
76	17200
77	18003
78	19137
79	20125
⋮	⋮
95	48285
96	50879
⋮	⋮

In this data set, t is the variable representing the number of days since 5 March 2020.

It can be computed using Microsoft Excel or other statistical software that the correlation coefficient between the total number of confirmed cases and t is 0.812, which indicates that there is a strong positive linear association between the two variables. Is this indeed the case? Perhaps we may make

[3]Data set can be downloaded from www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset.

such a conclusion but as stated earlier, correlation coefficient alone does not give the entire picture. We should create a scatter plot using our bivariate data and verify if there is really a linear relationship.



Are the two variables associated linearly? It is quite clear visually that the total number of confirmed cases increases exponentially when t increases. Thus, if we let y be the variable representing the total number of confirmed cases, y and t are not linearly associated but instead the relationship between them seems to be exponential. For such a situation, can we apply our linear regression technique to make predictions on the total number of confirmed cases? The answer is yes, but it would have to be done indirectly.

Now, if the relationship between y and t is indeed exponential in nature, we can model this relationship using the equation

$$y = cb^t,$$

where c and b are some constants that we will determine. Using the property of the logarithmic function, we see that

$$y = cb^t \quad \text{is equivalent to} \quad \ln y = \ln(cb^t) \quad \text{is equivalent to} \quad \ln y = \ln c + t \ln b.$$

Thus, instead of making a scatter plot with y plotted against t, we will make a scatter plot with $\ln y$ plotted against t. If there is indeed an exponential relationship between y and t, then we would expect to see a linear relationship between $\ln y$ and t, as indicated by the equivalent equations above. Let us go through the steps:

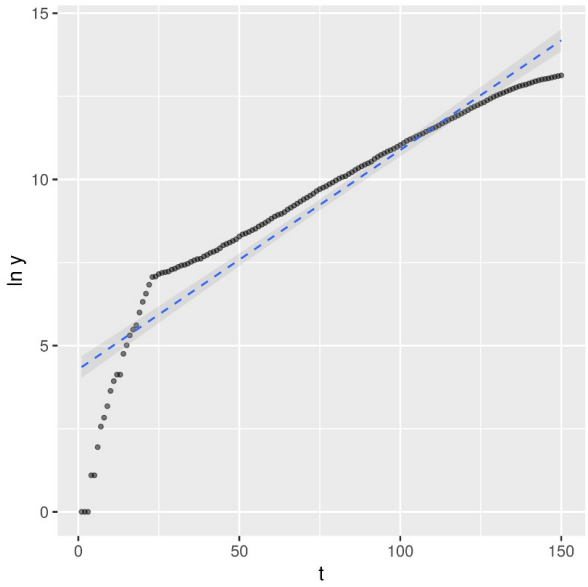
(a) Step 1: For each data point (t, y), compute $(t, \ln y)$. For our data set on COVID-19 cases in South Africa, we have the following table:

t	Total confirmed cases (y)	$\ln(y)$
76	17200	9.753
77	18003	9.798
78	19137	9.859
79	20125	9.910
⋮	⋮	⋮
95	48285	10.785
96	50879	10.837
⋮	⋮	⋮

We then plot $\ln y$ against t.

(b) Step 2: Find the linear regression line for $\ln y$ vs t. For our example, the regression line was found to be

$$\ln y = 4.287 + 0.066t.$$



This means that $\ln c = 4.287$ and $\ln b = 0.066$.

(c) Step 3: Since $\ln c = 4.287$ and $\ln b = 0.066$, we have

$$c = e^{4.287} \quad \text{and} \quad b = e^{0.066}.$$

We are now able to write down the exponential equation relating y and t:

$$y = cb^t = e^{4.287}e^{0.066t} = e^{4.287+0.066t}.$$

Exercise 3

1. The results of 10 students in a test are given below, where L indicates that a student obtained a score below 55, and the two H's (not necessarily the same) indicate scores above 90. The maximum possible score is 100.

   60 80 L 70 H 83 59 H 70 65.

   Which of the following statements must be true about the scores given above?

   Select all that apply.

   (A) The median score is 70.

   (B) There is only one mode among them.

   (C) The range is at least 35.

   (D) The mean score is greater than 72.2.

2. Outliers are observations that fall well above or below the overall bulk of the data. Consider a set of 50 (univariate) data points with a single outlier. Suppose the outlier is removed from the data set, which of the following is/are always true? Select all that apply.

   (A) The removal will cause the mean to decrease.

   (B) The removal will cause the interquartile range to decrease.

   (C) The removal will cause the standard deviation to decrease.

   (D) The removal will cause the range to change.

3. Suppose that there are 76 pairs of siblings living in a particular block in Ang Sua, where the older sibling is always heavier than the younger sibling. Consider a scatter plot using the younger sibling's weight to predict the older sibling's weight, where each point in the scatter plot represents the weights of a pair of two siblings in the block. Which of the following statements must be true?

   (I) There is a positive association between the older and younger siblings' weights.

   (II) All the points lie above the line $y = x$ in the scatter plot.

   (A) Only (I).

   (B) Only (II).

   (C) Neither (I) nor (II).

   (D) Both (I) and (II).

4. The Registry of Marriages is interested to see the relationship between husband's and wife's age in City X. They randomly sampled 1000 pairs of husbands and wives from the population of City X and obtained data of their ages (in years). Looking through the data, they found that men always marry women who are younger than them. Which of the following statements about the sample is/are definitely correct based on the information given?

   (I) The average age of the husbands is greater than the average age of the wives.

   (II) The standard deviation of husband's age is greater than the standard deviation of wife's age.

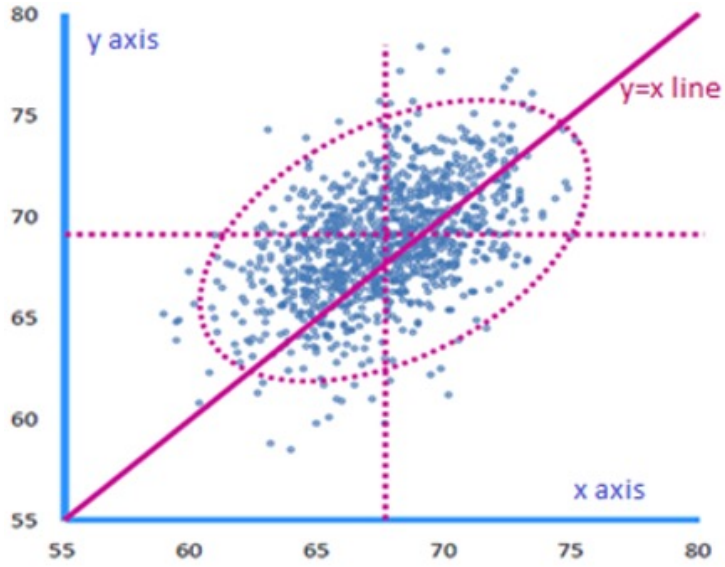
   (A) Only (I).

   (B) Only (II).

   (C) Both (I) and (II).

   (D) Neither (I) nor (II).

5. In the scatter plot below, the dotted straight lines mark the average values of X and Y. What can we say from the information given in the plot?



Which of the following statements is/are correct?

(I) The line $Y = X$ cuts through the data points in half, with 50% of the data points on either side of the line.

(II) The average of Y is larger than the average of X.

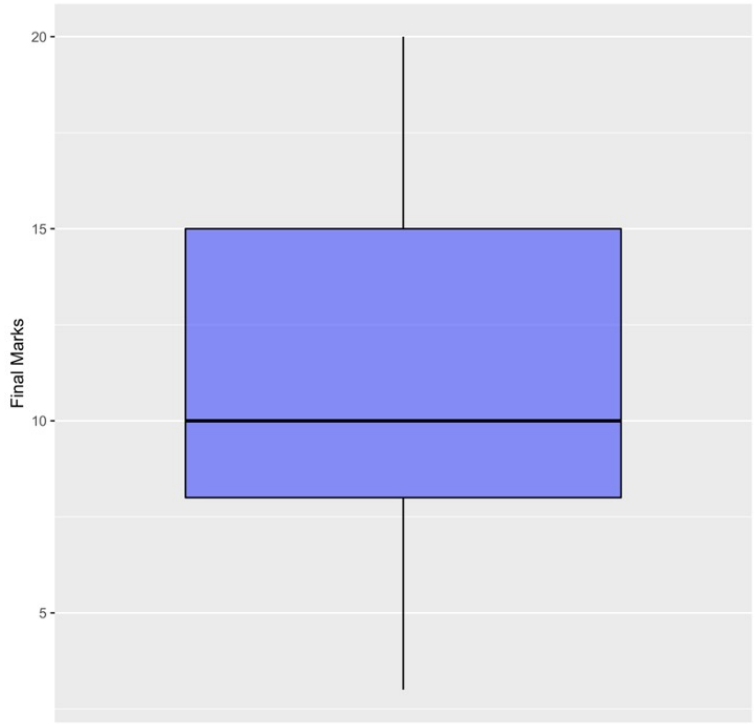
(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

6. The following boxplot shows the final examination marks of students from class A.



The passing mark for the final examination is 12 out of 20. Suppose that the boxplot of the final examination marks for class B is the same as the boxplot for class A. What can be said about the final examination marks of students from class B? Select all the correct statements.

(A) The proportion of students in class B who passed the examination must be the same as that for class A.

(B) At least 50% of the students from class B failed the final examination.

(C) The standard deviation of the students' marks in class B is equal to the standard deviation of those in class A.

(D) Based on the boxplot, there are no outliers in the grades of students from class B.

(E) The average of the students' marks for class B must be equal to that for class A.

7. The five-number summary of a numerical variable with 47 values is:

Min	Q_1	Median	Q_3	Max
12.0	15.0	16.5	18.0	24.0

Which of the following statements must be true? Select all that apply.

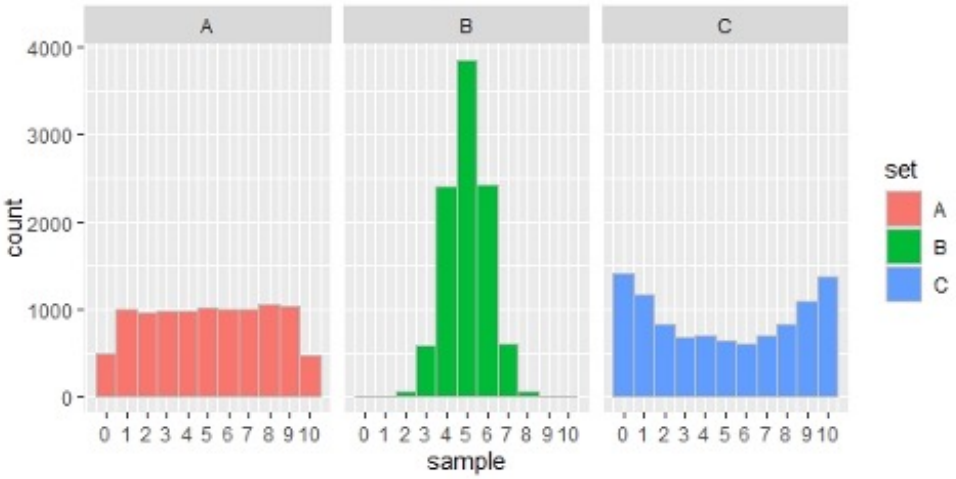
(A) There are no outliers in the data.

(B) There is at least one low outlier in the data.

(C) There is at least one high outlier in the data.

(D) There are both low and high outliers in the data.

8. Consider data sets A, B and C, each consisting of 10,000 numbers with mean 5. The histograms for A, B and C are shown below.



Order the data sets according to the values of their standard deviations, from the smallest to the largest.

(A) A, B, C.

(B) A, C, B.

(C) B, A, C.

(D) B, C, A.

(E) C, A, B.

(F) C, B, A.

9. Suppose that the following are 10 data points for a numerical variable X:

$$3, \quad 50, \quad r, \quad 8, \quad 20, \quad 1, \quad 32, \quad 58, \quad 10, \quad 138,$$

where r is an unknown whole number and $r \neq 138$. Based on the definition of an outlier for a boxplot, if 138 is the only outlier in this data set, the **maximum** possible value of r is _____.

10. Of the five values below, which would be that of a correlation coefficient with the strongest correlation?

(A) -1.4.

(B) -0.9.

(C) 0.

(D) 0.3.

(E) 0.7.

11. A researcher predicts the total number of bacteria in an experiment (denoted by y) using simple linear regression on $\ln y$ vs x. The regression equation is given by

$$\ln y = 0.5x + 2.5,$$

where the logarithm here is taken to base e, and x is the number of hours since 12PM. Which of the following statements is/are always correct?

(I) According to the researcher's model, there will be an average of 33 (rounded to the nearest integer) bacteria in 2 hours.

(II) The average number of bacteria is predicted to increase by the same amount every hour.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

12. In a study of 100 mother-daughter pairs, their heights were measured and plotted in a scatter diagram - mothers' heights at the horizontal axis, and their respective daughters' heights at the vertical axis. The horizontal and vertical axes are drawn on the same scale. All the data points are above the 45-degree line passing through the origin. Which of the following statements must be true?

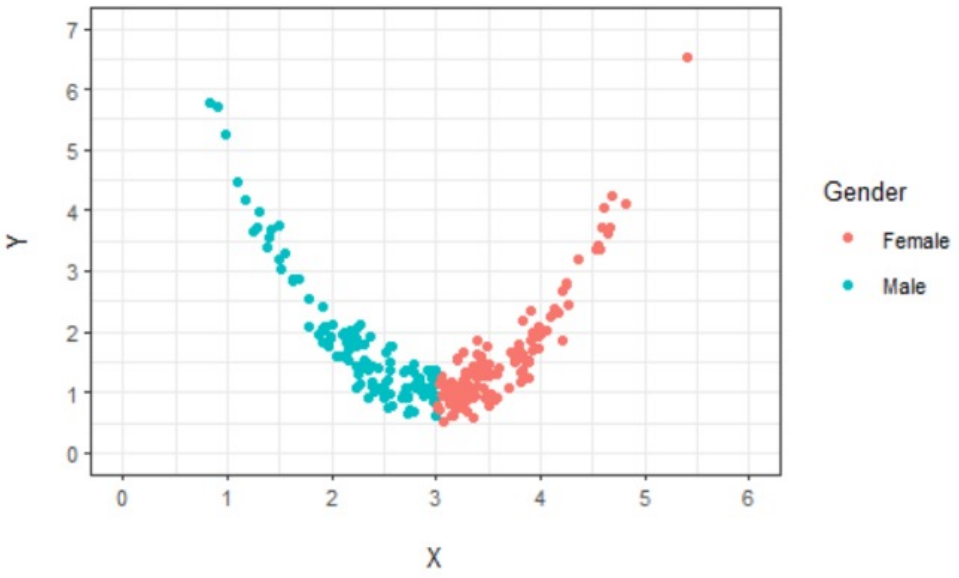
(A) The correlation between mothers' height and daughters' height is negative.

(B) The correlation between mothers' height and daughters' height is positive.

(C) The correlation between mothers' height and daughters' height is zero.

(D) None of the other given options is correct.

13. A researcher examined the relationship between variables X and Y among 250 male and female subjects. He graphed the relationship in the scatter plot shown below. Let r be the correlation coefficient for all 250 subjects, r_1 be the correlation coefficient among male subjects only and r_2 be the correlation coefficient among female subjects only.



Which of the following correctly describes the relationship between r, r_1 and r_2?

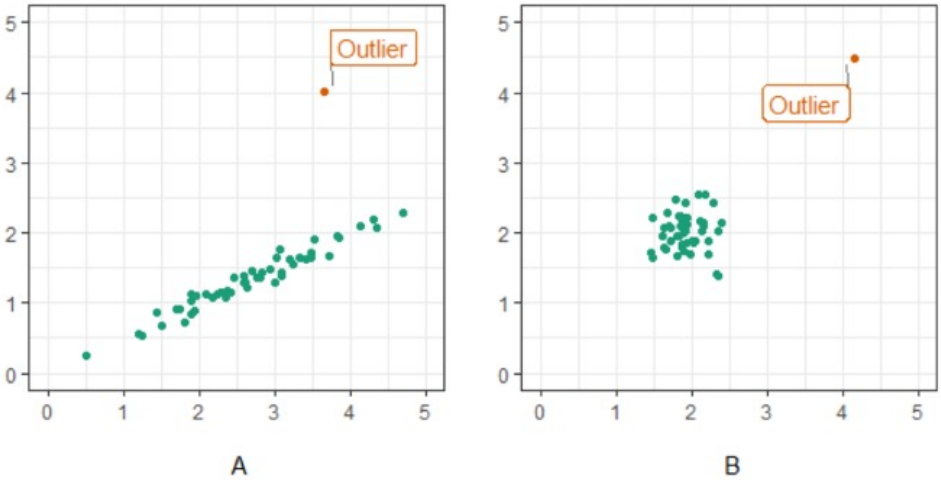
(A) $r_1 < r < r_2$.

(B) $r_1 > r > r_2$.

(C) $r > r_1 > r_2$.

(D) $r < r_1 < r_2$.

14. Consider plots A and B shown below. What will happen to the correlation coefficients for both plots after removing the outliers indicated?



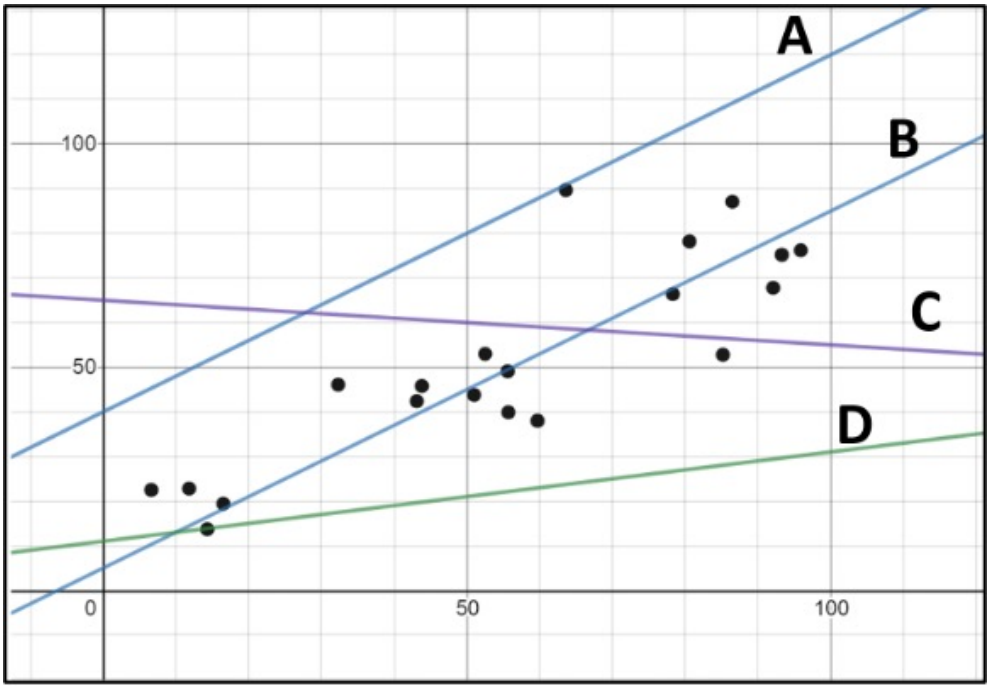
(A) The correlation coefficient in plot A will increase and correlation coefficient in plot B will decrease.

(B) The correlation coefficient in plot A will decrease and correlation coefficient in plot B will increase.

(C) The correlation coefficients in plots A and B will both increase.

(D) The correlation coefficients in plots A and B will both decrease.

15. A researcher examined the relationship between variables X and Y among 20 male subjects, and he graphed a scatter plot as shown below. One of the lines in the graph (A, B, C or D) is the actual best-fit regression line. Which one is it?



(A) Line A.

(B) Line B.

(C) Line C.

(D) Line D.

16. A researcher is interested in the correlation between the amount of time an individual spends on social media and the individual's level of happiness. Suppose that she observed that the correlation coefficient r_1 for males only is 0.8, and that the correlation coefficient r_2 for females only is also 0.8. Which of the following statements must be true for r, the correlation coefficient when the data for males and females are combined?

(A) $0 \leq r \leq 0.8$.

(B) $r = 0.8$.

(C) $0.8 < r \leq 1$.

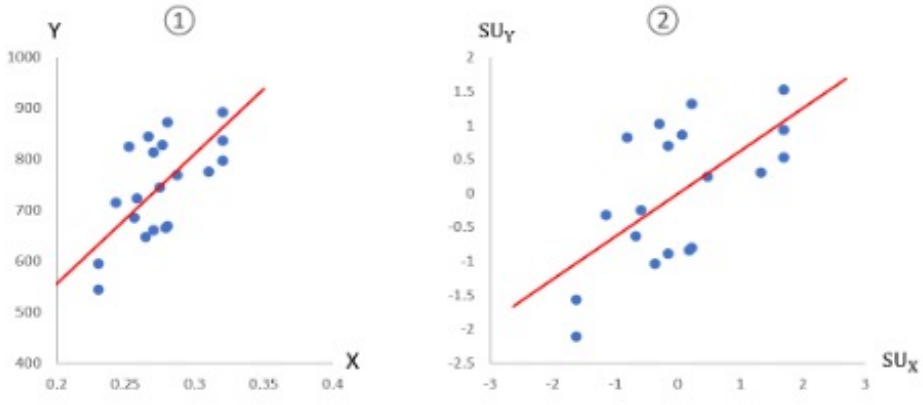
(D) None of the other given options is correct.

17. Let X and Y denote two variables measured on 19 subjects. Let the mean and standard deviation for the X values be written as $\overline{X}$ and s_X; similarly, the mean and standard deviation for the Y values are written as $\overline{Y}$ and s_Y. Plot "1" on the left shows the scatter plot for the 19 points (X, Y).

After converting the values of X and Y to standard units using the formulas:

$$SU_X = \frac{X - \overline{X}}{s_X} \qquad\qquad SU_Y = \frac{Y - \overline{Y}}{s_Y},$$

the scatter plot for the 19 points (SU_X, SU_Y) is shown as plot "2" on the right. The lines on both plots are the respective linear regression lines.



Which of the following statements is correct?

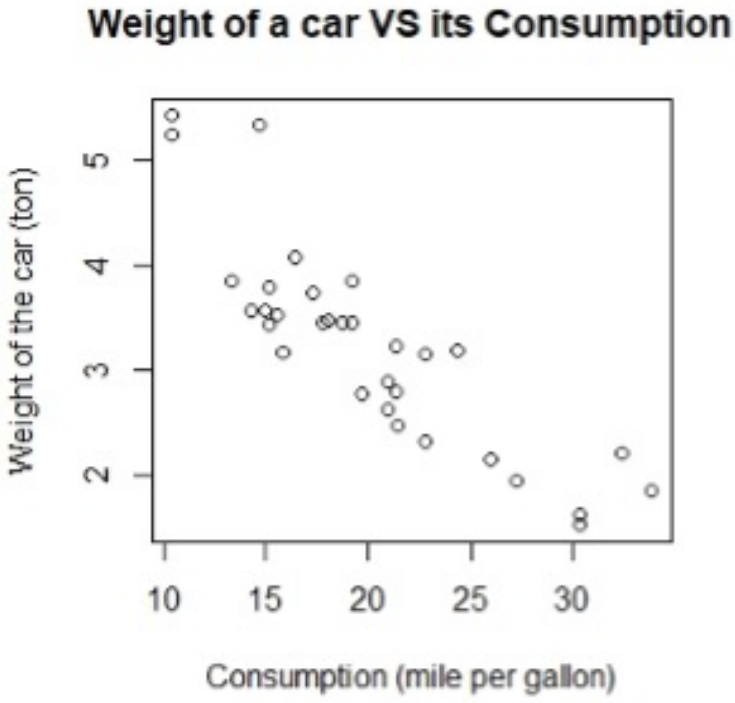
(A) The correlation coefficients in both plots are the same.

(B) The correlation coefficient decreased after conversion of X and Y to standard units.

(C) The correlation coefficient increased after conversion of X and Y to standard units.

(D) We do not have sufficient information to determine if the correlation coefficient has increased, decreased or remained the same after conversion of X and Y to standard units.

18. Based on the scatter plot shown below, which of the following is closest to the equation for the regression line? Here, W is the weight of the car and C is the consumption.



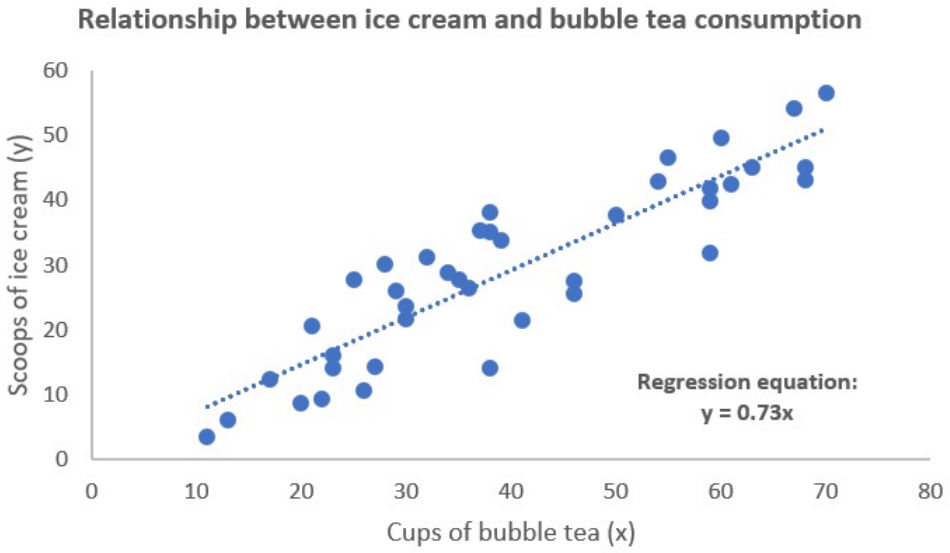
(A) $W = 3 - 0.8C$.

(B) $W = 5 - 0.8C$.

(C) $W = 3 + 0.8C$.

(D) $W = 5 + 0.8C$.

19. A group of students wanted to investigate the relationship between bubble tea and ice cream consumption. They conducted a survey to find out the number of cups of bubble tea and scoops of ice cream consumed in a year. From the data collected, they drew a scatter plot and fitted a linear regression line to the data with the equation of $y = 0.73x$.



Based on the information given in the scatter plot, which statement(s) must be true?

(I) The correlation between bubble tea and ice cream consumption is 0.73.

(II) People who consume 100 cups of bubble tea are predicted to consume 73 scoops of ice cream on average.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

20. There are two primary six classes in a tuition center. Class A and Class B each has 100 students and all students sat for a mathematics midterm test as well as a final examination. In Class A, every student scores 1 point higher in the final examination than in the midterm. In Class B, every student scores 1 point lower in the final examination than in the midterm. For the midterm test, the average score is 50 and standard deviation is 20 for both classes A and B. Which of the following statements is/are correct?

(I) The correlation coefficient between the final examination score and the midterm score in Class A is 1.

(II) The correlation coefficient between the final examination score and the midterm score in Class B is -1.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

21. There are two primary six classes in a tuition center. Class A and Class B each has 100 students and all students sat for a mathematics midterm test as well as a final examination. In Class A, every student scores 1 point higher in the final examination than in the midterm. In Class B, every student scores 1 point lower in the final examination than in the midterm. For the midterm test, the average score is 50 and standard deviation is 20 for both Classes A and B. Suppose now Class C is formed by combining all the students from Classes A and B. Which of the following statements is/are correct?

    (I) The correlation coefficient of Class C is smaller than the correlation coefficient of Class A.

    (II) The correlation coefficient of Class C is larger than the correlation coefficient of Class B.

    (A) Only (I).

    (B) Only (II).

    (C) Both (I) and (II).

    (D) Neither (I) nor (II).

22. Which of the following is/are true about a non-zero correlation coefficient? Select all that apply.

    (A) The correlation coefficient does not change when we add 5 to all the values of one variable.

    (B) The correlation coefficient is positive when the slope of the regression line is positive.

    (C) The correlation coefficient does not change when we multiply all the values of one variable by 2.

    (D) A correlation of -0.3 is stronger than a correlation of -0.8.

23. For the 40 students in a class, the results of their second English test are plotted against the results of their first English test. It was found that for each student, the result of the second test is better than that of the first test (i.e., the second test score is higher). Which of the following must be true about the relationship between the students' second test results and their first test results?

    (I) If student A scores better than student B in the first test, then student A also scores better than student B in the second test.

    (II) The correlation between students' second test results and first test results is positive.

    (A) Only (I).

    (B) Only (II).

    (C) Both (I) and (II).

    (D) Neither (I) nor (II).

24. A student produces a scatter plot whereby the y values range from 1 to 3, and he observes a linear association, with regression equation $y = 1.5x + 5$. Which of the following statements is/are correct?

    (I) The correlation coefficient must be positive.

    (II) The average y value of the data set collected is 2.

    (A) Only (I).

    (B) Only (II).

    (C) Both (I) and (II).

    (D) Neither (I) nor (II).

25. There is a weak positive linear association between numerical variables X and Y, where X ranges from 0 to 5 (inclusive). Based on the data from X and Y, the regression line is given by the equation $Y = 0.25X + 2$. Which of the following statements must be true? Select all that apply.

    (A) We can obtain the exact value of Y when $X = 4$.

(B) The predicted average value of Y is 4 when $X = 8$.

(C) The correlation coefficient is 0.25.

(D) The equation corresponds to a non-deterministic relationship between X and Y.

26. A researcher wanted to find the correlation between heights of 100 father-and-son pairs. After collecting and analysing his data, he realised that the device he had been using to measure height suffered from significant bias causing every measurement to be too high by 10cm. He then corrected the values of all his analyses. After the correction was done, which of the following will change? Select all that apply.

    (A) The correlation coefficient between the heights of father-and-son pairs.

    (B) The standard deviations of son's height and father's height.

    (C) The average son's height and the average father's height.

27. The relationship between the number of glasses of beer consumed daily (x) and blood alcohol content in percentage (y) was studied in young adults. The equation of the regression line is $y = -0.015+0.02x$ for $1 \leq x \leq 10$. The legal limit to drive in Singapore is having a blood alcohol content below 0.08%. Des, a young adult, had just finished 5 glasses of beer. After that, he wanted to take his car out for a drive. Is it legal for him to drive in Singapore?

    (A) Yes.

    (B) No.

    (C) Unable to determine.

28. You are given that the variables X and Y are negatively correlated. Which of the following statements must be true? Select all that apply.

    (A) If we multiply all the values of X and Y by -1, then the correlation coefficient between X and Y will change.

    (B) The gradient of the regression line for Y vs X is the same as the gradient of the regression line for X vs Y.

    (C) If we remove an outlier from the data set, the correlation coefficient will change.

    (D) If we add 6 to each value of X and subtract 3 from each value of Y, the correlation coefficient does not change.

    (E) If there are only 2 points, the correlation coefficient between X and Y must be -1.

29. 4 students take a midterm examination and a final examination. The minimum and maximum midterm scores are 20 and 40 respectively. The minimum and maximum final scores are 60 and 80 respectively. All midterm and final scores of the 4 students are plotted on a scatter diagram: the midterm scores on the horizontal x-axis, and the final scores on the vertical y-axis.

    Consider the following statements:

    (I) All points in the scatter plot lie above the line $y = x$.

    (II) The correlation coefficient between the midterm scores and final scores **must be** nonzero.

    Which of the above statements is/are true?

    (A) Only (I).

    (B) Only (II).

    (C) Both (I) and (II).

    (D) Neither (I) nor (II).

30. Suppose that there are 40 male students in a class and each student scored 5 less marks for his maths test than what he scored for his science test. What can we say about their maths and science test marks? Select all that apply.

    (A) The interquartile range of science test marks is higher than that for maths test marks.

    (B) If student A scored a higher mark for the maths test than student B, then he must have scored a higher mark than student B for the science test.

    (C) The science test marks and maths test marks are perfectly negatively correlated.

    (D) The standard deviation of maths test marks is equal to that of science test marks.

Chapter 4

Statistical Inference

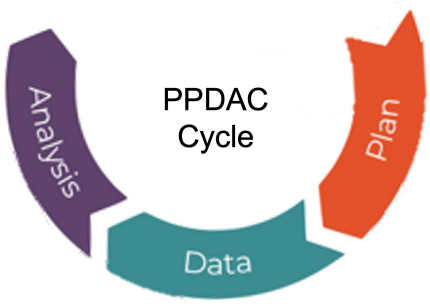
Section 4.1 Probability

In Chapter 1, we introduced the following types of research questions that are of interest.

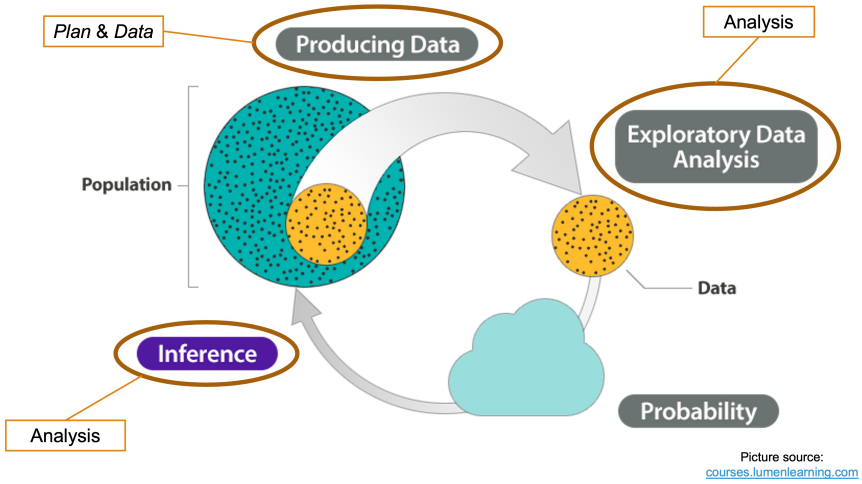
1. To make an estimate about the population.
2. To test a claim about the population.
3. To compare two sub-populations / to investigate a relationship between two variables in the population.

You would have noticed that a common term that recurs in the three research questions is the word **population**. Indeed these are all questions pertaining to the population. In order to answer these questions, we would need to have complete information about the entire population. This is usually not possible due to the sheer size of the population.

In order to give an approximate answer to the research questions, we need to use a **sample** of the population. The process of drawing conclusions about the population from sample data is known as *statistical inference*.



Recall the PPDAC cycle, first introduced in Chapter 1. In particular, when we focus on the second to fourth phases of the cycle, "Plan", "Data" and "Analysis", these phases involve specialised tools and techniques. These tools and techniques lead us to take a closer look at how these three phases are inter-related.



The "Plan" and "Data" phases were discussed in Chapter 1. How is a sample obtained from a population? What are the different methods of sampling and what are the types of biases we need to avoid? From summary statistics introduced in Chapter 1, to how categorical variables can be analysed in Chapter 2 and likewise for numerical variables in Chapter 3, these are all tools that we can use under the Analysis phase.

To conclude the "Analysis" phase, we need to look at the results of our analysis of the sample and subsequently draw conclusions on the population. This is where *statistical inference* comes into the picture. In order to have a meaningful discussion on statistical inference, we need to acquire some knowledge about *probability*. Probability and inference will form the main thrust of this chapter. To begin, we will introduce some basic results in probability which allow us to discuss the tools required in statistical inference. The two kinds of tools most common in statistical inference are *confidence intervals* and *hypothesis tests*, both of which will be discussed in some detail in the rest of the chapter.

Let us lay the groundwork for probability by defining some basic terms that are necessary in this subject.

Discussion 4.1.1 In previous chapters, we have occasionally touched on the notion of uncertainty. When we use the word "chance" it is understood intuitively that *something* is not definite, or not certain to hold or occur. In order to compare the *likelihood* of occurrence, we use terms like "more likely" or "less likely". These terms are common and adequate for everyday use. However, they are not precise and as we deal with data at a deeper level, we need a rigorous framework to ground the concept of uncertainty. *Probability* is a mathematical means that we can use to reason about uncertainty.

Consider a coin, with one side called "heads", and the other called "tails". Let's say the coin is tossed twice and the side that faces up when the coin lands is observed for both tosses. What are the possible outcomes after two tosses? If we represent observing heads as H and observing tails as T, then the four possible outcomes are:

$$HH, \quad TT, \quad HT, \quad TH.$$

Here HT is differentiated from TH as HT means heads was observed in the first toss and tails in the second toss, while TH means tails was observed first, followed by heads. In this example, the procedure of **tossing the coin twice** is called a *probability experiment*. The set $\{HH, TT, HT, TH\}$ contains the *outcomes* of the probability experiment.

It should be noted that the probability experiment defined here is narrower than the type of experiments described in Chapter 1. A probability experiment must be repeatable and allows for the exact listing of all the possible outcomes, like the way we have listed down all the four possible outcomes of the probability experiment of tossing a coin twice.

A *sample space* is the collection of all possible outcomes of a probability experiment. A sub-collection of the sample space is called an *event*. Referring back to our coin tossing probability experiment:

1. The sample space, as indicated earlier is:

$$HH, \quad TT, \quad HT, \quad TH.$$

2. An event could be
$$HH, \quad TT.$$
We can describe this event as *two in a row* or *two identical observations*.

3. Another event could be
$$TT, \quad TH, \quad HT.$$
This event can be described as *at least one tail*.

4. Another event could be
$$HT.$$
We can describe this event as *first toss is heads, second toss is tails*.
Note that an outcome can also be considered an event but not all events are outcomes!

Having understood a probability experiment, the sample space of the experiment and an event of the sample space, we are now ready to give context to the mathematical discussion of probability.

> For a probability experiment with an associated sample space, the probability of an event of the sample space is the total probability that the outcome of the experiment is an element of the event.

For example, in our coin tossing experiment, the probability of at least one tail is how likely the outcome of the experiment will be TT, TH or HT.

Example 4.1.2 Another common example of a probability experiment is the rolling of a six-sided die. It is obvious that such an experiment can be repeated as many times as we wish to and it is also easy to list down all the outcomes of a single roll of the die.

1. Probability experiment: rolling (once) a six-sided die and observe the top facing side.

2. Sample space:
$$1, \quad 2, \quad 3, \quad 4, \quad 5, \quad 6.$$
Rather than to use a list, we can put all the outcomes into a **set**,
$$\{1, 2, 3, 4, 5, 6\}.$$

3. An example of an event:
$$2, \quad 4, \quad 6.$$
This event can be described as "die shows an even number". As an event is a sub-collection of the sample space, when we represent the sample space as a set, an event will be a **subset** of the sample space, for example
$$\{2, 4, 6\}.$$

As an exercise, you may wish to write down the sample space (as a set) for the probability experiment of rolling a six-sided die twice and observing the top facing side on both rolls. Follow this by describing the event and writing down the subset of the sample space that represents the event.

Notation 4.1.3 Suppose E is an event, then $P(E)$ is the probability of event E. Probabilities are numerical values between 0 and 1 (both inclusive), so $P(E)$ takes on a numerical value between 0 and 1 and this is the probability assigned to event E.

Discussion 4.1.4 The question now is how do we know which numerical value between 0 and 1 to assign to an event E? In other words, how do we know the probability $P(E)$? Mathematically, we can define $P(E)$ as the long run proportion of observing E when a large number of repetitions of the experiment is being performed. Thus, we can repeat the probability experiment a large number of times (say N times) and each time we observe if the outcome is an element of the event E. Suppose the first experiment's outcome is in E, then we mark that experiment with a "YES". Repeat the experiment again, suppose now the outcome is not in E, then we mark this experiment with a "NO". Continue this, till we have done the experiment N times and each time we have either a "YES" or a "NO".

We now count the number of "YES" we have, out of the N times the experiment was done. Then the probability of event E, $P(E)$ is estimated by

$$\frac{\text{number of "YES"}}{N}.$$

It should be noted that

1. The estimate of $P(E)$ we obtain from these N repetitions of the experiment is **likely to be different** if we repeat the experiment (and get another estimate) another N times.

2. Such estimates get more accurate and closer to the true value of $P(E)$ as the number N becomes larger.

Example 4.1.5 Let us return to our die rolling experiment.

1. Probability experiment: rolling (once) a six-sided die and observe the top facing side.

2. Sample space: $\{1, 2, 3, 4, 5, 6\}$.

3. Event $E =$ die shows an even number, that is $E = \{2, 4, 6\}$.

What would be an estimate of $P(E)$? Suppose we repeatedly roll (and observe) the die 500 times and recorded the "YES" and "NO" as follows:

Roll	1	2	3	$\cdots$	499	500
Outcome	2	3	1	$\cdots$	6	2
Outcome belong to E?	YES	NO	NO	$\cdots$	YES	YES

Suppose the total number of "YES", out of the $N = 500$ times the experiment was carried out, is 268, then an estimate of $P(E)$ is

$$P(E) = \frac{268}{500} = 0.536.$$

Rules of Probabilities

It is virtually impossible to verify what is the true probability for an event of a probability experiment. For example, even if we say that a coin is "fair" does it mean that the probability of "heads" is *exactly* 0.5 and that for 'tail' is *exactly* 0.5? Probably (pun intended) not! The probabilities that we encounter in everyday life are just estimates of what the true probability is but in the analysis of data, it is sufficient to treat the estimates as if it is the true probability. What is important and relevant in the use of such estimates is that in the assignment of probabilities to events of a probability experiment, the following *rules of probabilities* must be obeyed.

1. The probability of each event E, denoted by $P(E)$ is a number between 0 and 1 (inclusive).

2. If we denote the entire sample space by S, then the probability of S, $P(S)$ is 1.

3. If E and F are mutually exclusive events (meaning both events cannot occur simultaneously), then the probability of E union F is equal to the sum of the probabilities of E and F. That is, $P(E \cup F) = P(E) + P(F)$.

When the sample space contains only a **finite** number of outcomes, we only need to assign probabilities to the outcomes so that these probabilities sum up to 1. The probabilities of all other events can then be derived from there.

Example 4.1.6 Suppose we have a *biased* six-sided die being rolled once. The following probabilities are assigned to the six possible outcomes.

Outcome	1	2	3	4	5	6
Probability	0.1	0.1	0.1	0.1	0.1	0.5

Check that the probabilities add up to 1. We are now able to derive the probabilities of certain events by applying the third rule of probability as stated above. For example, if E is the event "an odd-numbered face" and F is the event "an even-numbered face", it is easy to see that

1. $P(E)$ is the sum of $P(1)$, $P(3)$ and $P(5)$, so $P(E) = 0.3$. (Here "1", "3", "5" are mutually exclusive events.)
2. $P(F)$ is the sum of $P(2)$, $P(4)$ and $P(6)$, so $P(F) = 0.7$. (Here "2", "4", "6" are mutually exclusive events.)
3. E and F are mutually exclusive events, so $P(E \cup F) = P(E) + P(F) = 0.3 + 0.7 = 1$.

Definition 4.1.7 *Uniform probability* is the way of assigning probabilities to outcomes such that equal probability is assigned to every outcome in the finite sample space. Thus, if the sample space contains a total of N different outcomes, then the probability assigned to each outcome is

$$\frac{1}{N}.$$

As an example, if the sample space S contains the outcomes of flipping a coin twice, then

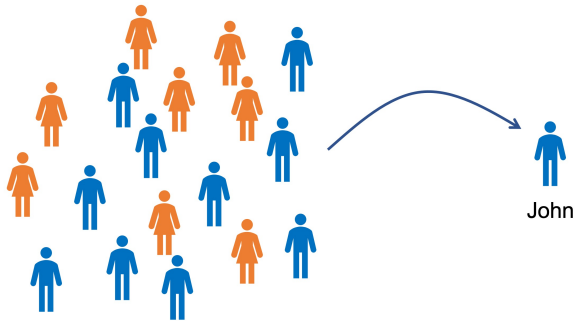
$$S = \{HH,\ HT,\ TH,\ TT\}.$$

Using uniform probability, we will assign the probability of $\frac{1}{4}$ or 0.25 to each of the four outcomes.

Example 4.1.8 We have in fact seen uniform probability in action much earlier in Chapter 1 when simple random sampling was introduced. Recall that in simple random sampling, r units from the sampling frame are randomly selected to be included in the sample. We are conducting a probability experiment where the sample space is the sampling frame that contains all the units that could possibly be selected. The probability of selecting a particular unit at the first draw from the sampling frame is thus $\frac{1}{N}$ where N is the size of the sampling frame.

Furthermore, for any subset of the sample space (an event) denoted by A, the probability of this event, $P(A)$ is interpreted as the likelihood of selecting a unit belonging to A into the sample. This is equal to the rate of A in the sampling frame.

As a concrete example, suppose the sampling frame consists of 500 adults, comprising of 280 males and 220 females. By simple random sampling, each adult, for example, John, has a probability of $\frac{1}{500}$ to be selected at the first draw. So the probability of a person selected at the first draw being male would be $\frac{280}{500}$.



If we define A to be the subset of the sample space consisting of the male adults, then the probability of event A is the rate of A in the sampling frame. That is,

$$P(A) = \text{rate(male)} = \frac{280}{500} = 0.56.$$

Section 4.2 Conditional Probability and Independence

Let us begin this section by using the same example that concluded the previous section. Suppose we have 500 adults, comprising of 280 males and 220 females, as participants in a lucky draw where there is only one prize to be won. Under uniform probability, each person, for example, John has a probability of $\frac{1}{500}$ to be the winner of the prize.

Now suppose it is known that the winning ticket will be drawn from the male participants, what is John's probability of being the winner of the prize now?

Definition 4.2.1 The scenario described above involves the concept of *conditional probability*. Conditional probability is normally written using the notation

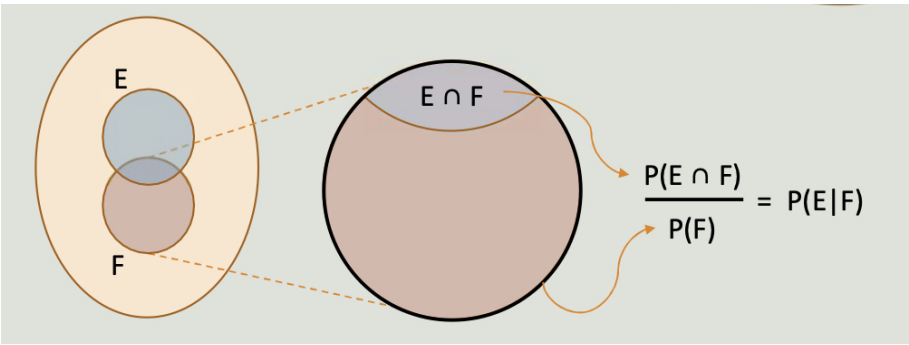
$$P(E \mid F)$$

and is read as "probability of E given F". Here, E and F are events of a particular sample space. With reference to our lucky draw example above, events E and F are:

E: Winner of the prize is John;

F: Winner of the prize is a male.

So the conditional probability $P(E \mid F)$ is the probability that John is the winner given that it is a male who won. Intuitively, the probability of E given F measures how likely the outcome of the probability experiment is an element of E, **if we already know that it is an element of** F. To compute conditional probabilities, we use the idea of restricting the sample space based on the condition that event F is known to have occurred.



More precisely, to compute the probability of E given F, we restrict our focus on the given event F as our restricted sample space (rather than to look at the entire sample space). The event F may or may not contain overlap with event E. The overlap is denoted by $E \cap F$, to be read "E intersect F";. The probability of E given F is obtained by dividing the probability $P(E \cap F)$ by the probability $P(F)$ which acts as the baseline (restricted sample space). Thus,

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}.$$

Remark 4.2.2

1. It is perfectly possible that there is no overlap between events E and F, meaning that it is not possible that E and F happen simultaneously. In such a situation, it is clear that the probability that event E occurs given that event F is **known** to have occurred is certainly 0. Indeed, with $P(E \cap F) = 0$, we see that
$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = 0.$$

2. If event F itself cannot occur, that is $P(F) = 0$, then we will stipulate by convention that $P(E \mid F)$ is also equal to 0.

Example 4.2.3 In our lucky draw example, we have

$$\begin{aligned}
P(F) &= \text{ probability that the winner is a male} \\
&= \frac{280}{500} \\
P(E \cap F) &= \text{ probability that the winner is John and winner is a male} \\
&= \text{ probability that the winner is John} \\
&= \frac{1}{500} \\
P(E \mid F) &= \text{ probability that the winner is John given that the winner is a male} \\
&= \frac{P(E \cap F)}{P(F)} \\
&= \frac{1}{500} \times \frac{500}{280} = \frac{1}{280}.
\end{aligned}$$

So the conditional probability that John is the winner, given that the winner is a male is

$$\frac{1}{280} = \frac{1}{\text{number of males in total}}.$$

Example 4.2.4 Suppose there are two bags, each containing 10 colored balls. **Bag A** contains 7 red balls and 3 green balls while **Bag B** contains 4 red balls and 6 green balls. One bag is randomly selected and a ball is then randomly selected from the chosen bag. What is the probability that the selected ball chosen is green?

Let us consider the events E, F and G such that

- E is the event that **Bag A** is selected.
- F is the event that **Bag B** is selected.
- G is the event that the selected ball is green.

Note that E and F are mutually exclusive and either E or F must occur. The probability required $P(G)$ is the probability of (G and E) plus the probability of (G and F). That is,

$$P(G) = P(G \cap E) + P(G \cap F).$$

By the definition of conditional probability, this is equivalent to

$$P(G) = P(G \mid E) \times P(E) + P(G \mid F) \times P(F).$$

In our example, this means that the probability that a green ball is selected is the sum of the probabilities

$$P(\text{green ball selected} \mid \text{Bag A is selected}) \times P(\text{Bag A is selected}) \quad \text{and}$$

$$P(\text{green ball selected} \mid \text{Bag B is selected}) \times P(\text{Bag B is selected}).$$

Formally, the *law of total probability* states that if E, F and G are events from the same sample space S such that

(1) E and F are mutually exclusive; and

(2) $E \cup F = S$.

Then,

$$P(G) = P(G \mid E) \times P(E) + P(G \mid F) \times P(F).$$

Discussion 4.2.5 (**Conditional probabilities as rate**) We have seen earlier that uniform probabilities are manifested as the probability experiment of randomly selecting a unit from a fixed sampling frame. The table below draws the analogy between the two interpretations.

Random sampling	Corresponds to	Probability experiment
Sampling frame	Corresponds to	Sample space
A subgroup A of the sampling frame	Corresponds to	An event A of the sample space
The rate of A, rate(A)	Corresponds to	The probability of A, $P(A)$

What about for conditional probabilities? Will there be a similar correspondence to conditional rates? More specifically, is the conditional probability of A given B equal to the rate of A given B whenever A and B are subgroups of the sampling frame? The following derivation shows that they are indeed equal.

$$\begin{aligned}
P(A \mid B) &= \frac{P(A \cap B)}{P(B)} && \text{(by using the idea of restricted sample space)} \\
&= \frac{\text{rate}(A \cap B)}{\text{rate}(B)} && \text{(by the correspondence between probability and rates)} \\
&= \frac{\text{size of } (A \cap B)}{\text{size of sampling frame}} \div \frac{\text{size of } B}{\text{size of sampling frame}} \\
&\quad \text{(by the definition of rates as ratios of two sizes)} \\
&= \frac{\text{size of } (A \cap B)}{\text{size of } B} \\
&= \text{rate}(A \mid B) && \text{(by the definition of rates as ratios of two sizes)}
\end{aligned}$$

So indeed, this derivation shows that just as probabilities are equivalent to rates in this probability experiment, conditional probabilities are also equivalent to *conditional rates*.

Example 4.2.6 Conditional probability shows up frequently in medical diagnosis. Most of us should be familiar with the term ART, or Antigen Rapid Test, by now. This test is used to test for the presence of COVID-19 infection in humans. For most medical diagnostic tests, there are four possible scenarios that can happen when the test is administered to an individual to assess if the individual is infected. The possible scenarios are:

1. Scenario 1: Individual is known to be infected and test shows positive.
2. Scenario 2: Individual is known to be infected and test shows negative.
3. Scenario 3: Individual is known to be not infected and test shows positive.
4. Scenario 4: Individual is known to be not infected and test shows negative.

Scenario 1 is concerned with the *conditional probability* of an individual being tested positive, *given* that the individual is infected. This is known as the *true positive rate*. This probability

$$P(\text{Test positive} \mid \text{Individual is infected})$$

is known as the *sensitivity* of the test. For this example, let us assume that this probability is 0.8.

On the other hand, scenario 4 is concerned with the *conditional probability* of an individual being tested negative, *given* that the individual is not infected. This is known as the *true negative rate*. This probability

$$P(\text{Test negative} \mid \text{Individual is not infected})$$

is known as the *specificity* of the test. For this example, let us assume that this probability is 0.99.

In reality, these two conditional probabilities are not helpful to average users like ourselves because we do not really know whether we are indeed infected or not. What we do know, with certainty is whether the individual's test returns positive or negative. Therefore, instead of the conditional probability

$$P(\text{Test positive} \mid \text{Individual is infected})$$

which is difficult to ascertain if the "condition" is fulfilled, we look at the conditional probability

$$P(\text{Individual is infected} \mid \text{Test positive}).$$

It is important to gain insight into this conditional probability as it can cause an individual much distress after being tested positive only to find out later that the person involved is actually not infected. To determine this conditional probability, having only the sensitivity and specificity of the test is insufficient. We require one additional piece of information, which is the *base rate* of infection in the population. This is the infection rate in the population and we can interpret this as the probability of a person selected at random from the population is infected with COVID-19. For this example, let us assume that 1% of the population is infected with COVID-19, so

$$P(\text{Individual is infected}) = 0.01.$$

Since conditional probabilities correspond to conditional rates in probability experiments involving random sampling, we can use a contingency table that was introduced in Chapter 2 to study rates here. To start, we choose a large enough number to represent the total population such that our calculations would result in whole numbers. Let us assume that the population consists of 100,000 individuals.

	Tested positive	Tested negative	Row total
Infected with COVID-19			
Not infected with COVID-19			
Column Total			100,000

Using the information we have for the base rate of infection, we can now fill in the row total for those that are infected ($= 1\% \times 100,000 = 1,000$) and those that are not infected ($= 99\% \times 100,000 = 99,000$).

	Tested positive	Tested negative	Row total
Infected with COVID-19			1,000
Not infected with COVID-19			99,000
Column Total			100,000

Next, using the true positive rate (sensitivity) of 0.8, we see that 80% of those infected would be tested positive, that is,

$$\text{Number of tested positive and Infected } = 0.8 \times 1,000 = 800, \quad \text{and}$$

$$\text{Number of tested negative and Infected } = 0.2 \times 1,000 = 200.$$

Similarly, using the true negative rate (specificity) of 0.99, we see that 99% of those not infected would be tested negative, that is,

$$\text{Number of tested negative and Not infected } = 0.99 \times 99,000 = 98,010 \quad \text{and}$$

$$\text{Number of tested positive and Not infected } = 0.01 \times 99,000 = 990.$$

	Tested positive	Tested negative	Row total
Infected with COVID-19	800	200	1,000
Not infected with COVID-19	990	98,010	99,000
Column Total			100,000

The table can now be completed by summing up the column totals for those tested positive and those tested negative.

	Tested positive	Tested negative	Row total
Infected with COVID-19	800	200	1,000
Not infected with COVID-19	990	98,010	99,000
Column Total	1,790	98,210	100,000

By now, you should appreciate the choice of $100,000$ as the total population, as we did not have to deal with the awkward situation of not having whole numbers when we are dealing with human individuals. We are now able to calculate the rate of COVID-19 infection among those tested positive. Since there are $1,790$ individuals tested positive, and 800 of them are infected, the rate is

$$\text{rate(Infected} \mid \text{Tested positive)} = \frac{800}{1790} = 0.447 \text{ (rounded to 3 significant figures).}$$

Using the correspondence between conditional rates and conditional probabilities, we are now able to say that if an individual is tested positive for COVID-19 infection using an ART, the probability of him actually being infected is about 0.45. This conditional probability is rather low so typically, more rigorous tests need to be conducted to ascertain if the individual is indeed infected with COVID-19.

Definition 4.2.7 When we say that two events A and B are *independent*, it means that

$$P(A) = P(A \mid B),$$

that is, the probability of A is the same as the probability of A given B. So, the fact that event B has occurred does not affect the probability of A occurring. Now, if we express the conditional probability $P(A \mid B)$ as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

then A and B being independent means that

$$P(A) = \frac{P(A \cap B)}{P(B)} \quad \text{which implies} \quad P(A) \times P(B) = P(A \cap B).$$

We thus have an equivalent definition of what it means for two events to be independent.

(Independence as non-association) Consider studying a population along two categorical variables, one with categories "A" and "Not A" and the other variable with categories "B" and "Not B". Such studies have been discussed in Chapter 2 and one of the main questions concern whether there is association between the two categorical variables. We used a 2×2 contingency table similar to the one below, to compute and compare the conditional rates.

	B	Not B
A		
Not A		

Recall that by the basic rule on rates, we say that the two variables are not associated if

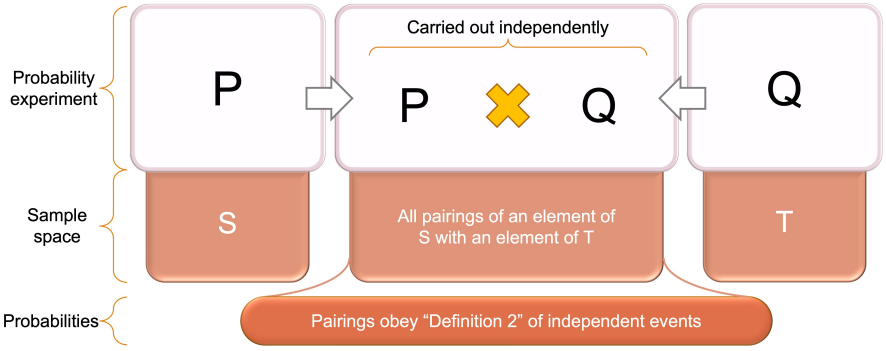
$$\text{rate(A)} = \text{rate(A} \mid \text{B)}.$$

Since we have drawn the correspondence between rates and probabilities, and conditional rates as conditional probabilities, this leads us to conclude that A and B are independent events whenever A and B are not associated with each other.

Here, the relevant probability experiment that draws the correspondence is one that involves randomly selecting one unit from the population we are studying and checking the values (presence or absence) of the selected unit with regards to the two categorical variables (with/without A and with/without B).

(Independent Probability Experiments) Let us return to the first example of the chapter, where a coin is tossed twice. Such probability experiments are common and we often see statements like "independent probability experiments" or "two independent tosses of a coin" or "three independent rolls of a die". Although we have understood what it means for two *events* to be independent, what does it mean to say that two *probability experiments* are independent? To understand this, we will view the two probability experiments as two parts of one combined probability experiment.

We start with two probability experiments, one of which is labelled **P** and the other is labelled **Q**. Suppose the sample space of experiment **P** is S and the sample space for experiment **Q** is T. It should be noted that **P** and **Q** do not have to be the same probability experiment (like tossing a coin) and thus the sample spaces S and T can be totally different. For example, **P** could be the probability experiment of tossing a coin while **Q** could be the probability experiment of rolling a die.



If these two probability experiments are independent, we can view them as two components of a larger experiment where **P** is coupled with **Q**. This *combined probability experiment* has a sample space that contains all pairings of possible outcomes of the two components. For example, if **P** is the probability experiment of tossing a coin and **Q** is the probability experiment of rolling a die,

$$(\text{Heads}, 6), \quad (\text{Tails}, 3)$$

are examples of outcomes in the sample space of the combined probability experiment. When we say the two components are independent, it means that the probabilities are assigned to each outcome in the sample space of the combined experiment such that Definition 2 of Independent events,

$$P(A \cap B) = P(A) \times P(B)$$

is obeyed. In layman terms, it means that the occurrences of events in Experiment **P** do not influence the chances of occurrences of events in Experiment **Q** and vice versa. For example, the probability assigned to the outcome (Heads, 6) is

$$P(\text{Heads}) \times P(6)$$

and the probability assigned to the outcome (Tails, 3) is

$$P(\text{Tails}) \times P(3).$$

Example 4.2.8 Suppose **P** is the experiment of rolling a particular six-sided die and **Q** is the experiment of two tosses of a coin.

- The sample space of **P** is $\{1, 2, 3, 4, 5, 6\}$.
- The sample space of **Q** is $\{HH, HT, TH, TT\}$.

- The sample space of the combined probability experiment is

$$\{(1, HH), (1, HT), (1, TH), (1, TT), (2, HH), (2, HT), \ldots, (6, TH), (6, TT)\}.$$

The assigned probabilities of outcomes now follow, for example,

$$P(1, HH) = P(1) \times P(HH).$$

Section 4.3 Random Variables

Definition 4.3.1 A *random variable* is a numerical variable with probabilities assigned to each of the possible numerical values taken by the numerical variable.

Example 4.3.2 Consider a probability experiment where each outcome is given a numerical value. Some examples are:

- The game of roulette, played in a casino. With each spin of the roulette wheel, the ball will land on one of the numbers $0, 1, 2, \ldots, 36$.
- Rolling a six-sided die. The faces of the die are denoted $1, 2, 3, \ldots, 6$.
- Randomly selecting a person from a population, checking the person's COVID-19 infection status and assigning 1 if the person is infected and 0 if the person is not infected.

For the three examples above:

- If we let X be the numerical variable that represents the outcome of a spin of the roulette wheel, assuming that the roulette wheel is fair, then the assigned probabilities to each of the outcomes i, $i = 0, 1, 2, \ldots, 36$ is

$$P(X = i) = \frac{1}{37}.$$

  Now X is a random variable.

- If we let Y be the numerical variable that represents the outcome of rolling a six-sided die with the assigned probabilities

$$P(Y = 1) = \frac{1}{3}, \qquad P(Y = 2) = \frac{1}{3}, \qquad P(Y = 3) = \frac{1}{12},$$

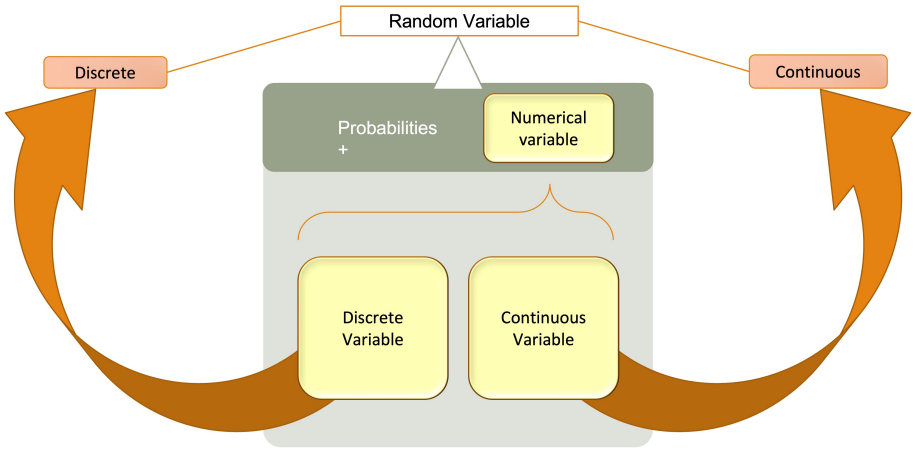
$$P(Y = 4) = \frac{1}{12}, \qquad P(Y = 5) = \frac{1}{12}, \qquad P(Y = 6) = \frac{1}{12},$$

  then Y is a random variable.

- If we let Z be the numerical variable that represents whether the selected person is infected with COVID-19, assuming that the population infection rate is 0.1, then

$$P(Z = 0) = 0.9, \quad P(Z = 1) = 0.1.$$

  Now Z is a random variable.



Definition 4.3.3 If the numerical variable is a discrete variable, we call the random variable a *discrete random variable*. On the other hand, if the numerical variable is a continuous variable, then the random variable is a *continuous random variable*. The three random variables listed above are all discrete random variables.

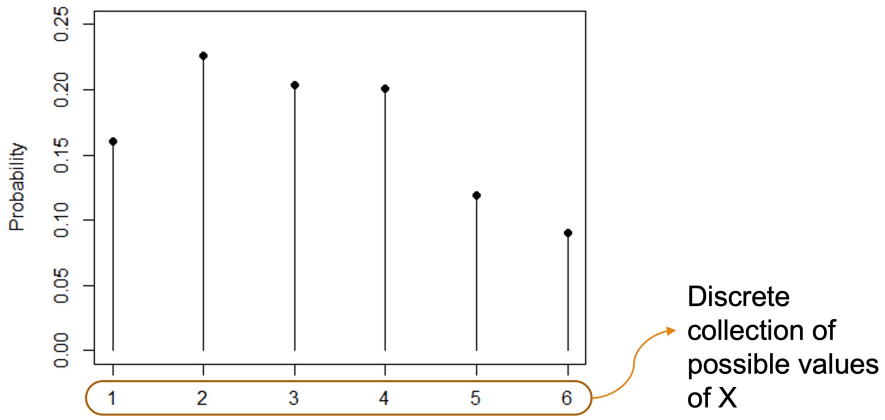
Remark 4.3.4 Random variables were conceived as a mathematical way to model data distributions. While we have discussed summary statistics like central tendency (e.g., mean) and dispersion (e.g., standard deviation) for a data set, these statistics can also be computed for random variables.

Example 4.3.5 Suppose we have a data set containing all HDB households in Singapore, together with the household size of each household. Here, the household size refers to the number of individuals living in the household. Consider the probability experiment of randomly selecting one household from all the households in the data set and observing the household size of the selected household. By equating probability with rate, the probability of the selected household having a particular household size x is the rate of households with size x in our data set.

The table below shows the possible household sizes and their respective probabilities. It is now obvious that household size is a discrete random variable.

Household size	1	2	3	4	5	6
Probability	0.16	0.226	0.204	0.201	0.119	0.09

Let us label this discrete random variable as X. To visualise the probability distribution of X, we can use a set of points (x, y) where x represents a possible value of X and y is the probability of X taking on that particular value. i.e, $y = P(X = x)$. We can use vertical line segments to connect the points in the plot to the x-axis.



It is useful to note that since the points in the plot are separated by gaps, the possible values that the random variable X can take are *discrete*, indicating that X is a discrete random variable.

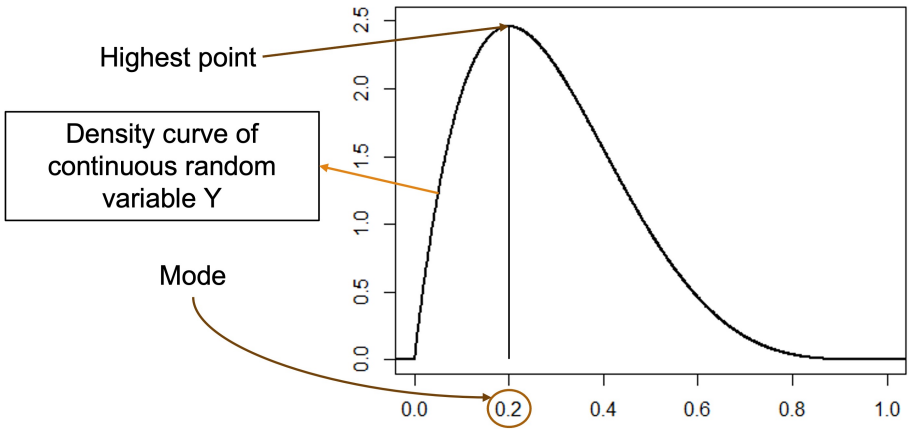
1. Recall the rule of probability where the sum of the probabilities assigned to each of the possible outcomes must be equal to 1. Indeed, the representation above agrees with this property.

2. The *mode* of a discrete random variable is the value of x that attains the highest y-value. In this example, when $x = 2$, the value of y $(= 0.226)$ is the greatest. So 2 is the mode of this discrete random variable.

3. Instead of the probability of X taking on a particular value, we sometimes ask for the probability of X taking on a range of values. For example, we may wish to know what is the probability of X taking a value greater than or equal to 5. In this example, this is the probability that the household size of a randomly chosen household is 5 or 6. To compute this probability, we simply add the probabilities (y-values) of the points corresponding to $x = 5$ and $x = 6$. That is,

$$P(X \geq 5) = P(5) + P(6) = 0.119 + 0.09 = 0.209.$$

Example 4.3.6 Example 4.3.5 showed how a discrete random variable X can be visualised, both in terms of the discrete values that X can take (the x-axis) and the probabilities of X taking each value (the y-axis). We now turn our attention to continuous random variables. It is not difficult to extend the idea of discrete values of x, separated by gaps, to one where there are no more gaps separating the values that a continuous random variable can take. Any continuous random variable Y can therefore be visualised by a "continuous series of points" which forms a *density curve* on the standard x and y-axes.

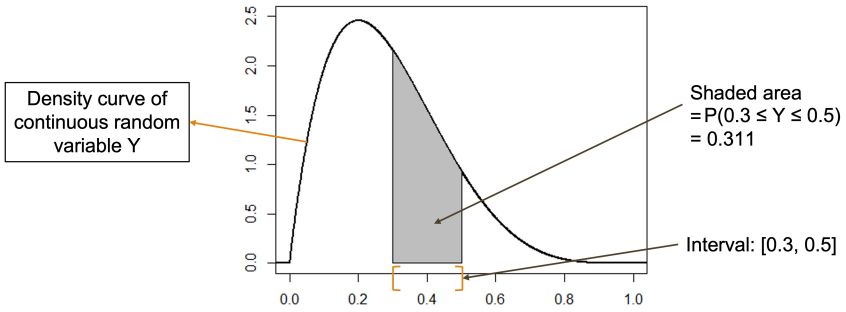


1. The values on the x-axis correspond to the possible values that the continuous random variable Y can take. This is analogous to that in a discrete random variable.

2. However, it should be noted that the y-axis in the density curve **does not** represent probability (unlike for discrete random variables), but instead it is the **probability density**. For the purpose of this module, we will not go into details of the interpretation of probability density. Nevertheless, it is important to remember that for a continuous random variable, the **area** under the density curve is always equal to 1.

3. The value of x that corresponds to the highest point of the density curve is the *mode* of the continuous random variable. This is again analogous to that in a discrete random variable. In this example, 0.2 is the (unique) mode of the continuous random variable Y.

4. Similar to discrete random variables, we are often interested in knowing the probabilities of a continuous random variable taking on a range of values. For example, what is the probability that Y assumes a value between 0.3 and 0.5? Consistent with the fact that the **area** under the entire density curve is equal to 1, the probability that Y takes on a value between 0.3 and 0.5 is the area under the density curve in the interval 0.3 to 0.5. In this example, the area shaded in the plot evaluates to 0.311.



In general, the probability that a continuous random variable takes on a value in an interval $[a, b]$ is equal to the area under its density curve from a to b.

Definition 4.3.7 *Normal distributions* are a class of continuous random variables that are often featured and have interesting properties. We use the notation

$$N(x, y)$$

to denote the normal distribution with mean x and variance y.

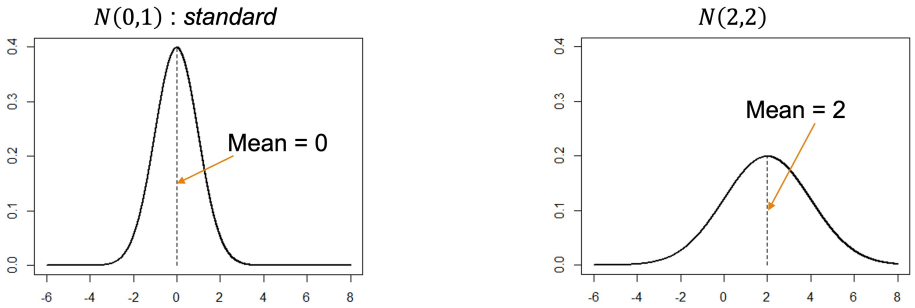
1. A particular normal distribution is completely described by its mean and variance. Therefore, any two normal distributions can only differ by their means and/or variance.

2. The density curve of a continuous random variable that is normally distributed is always bell-shaped.

3. The peak of the curve occurs at the mean. This implies that the mode is equal to the mean.

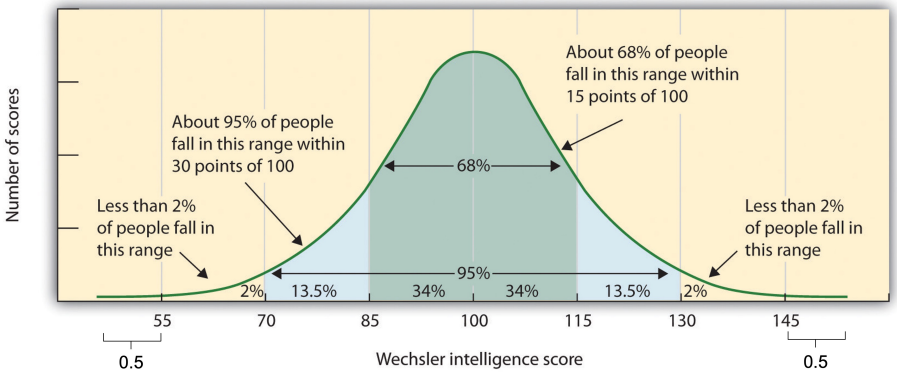
4. The density curve is symmetrical about the mean. This implies that the median is also equal to the mean (and the mode). Thus, the mean, mode and median of any normal distribution are the same.

To gain some insights into the visualisation of normal distributions, consider the two density curves below.



The density curve on the left is that of the normal distribution $N(0, 1)$. Evidently, the mean (as well as the mode and median) occurs at 0 and the variance, which measures the *spread* of the distribution is 1. The normal distribution $N(0, 1)$ is commonly known as the *standard normal distribution*. The density curve on the right is that of the normal distribution $N(2, 2)$. Here, the mean occurs at 2 and the variance is 2. Note that the spread of the distribution is larger than that for $N(0, 1)$ and thus the curve is flatter with a lower peak. This is a consequence of the required property that for any continuous random variable, the area under the density curve must always be equal to 1.

Example 4.3.8 A famous real life example of a normal distribution is observed in the Wechsler Adult Intelligence Scale that measures intelligence quotient (or IQ). It turns out that IQ scores follow a normal distribution with mean 100 and standard deviation 15 (hence a variance of $15 \times 15 = 225$).



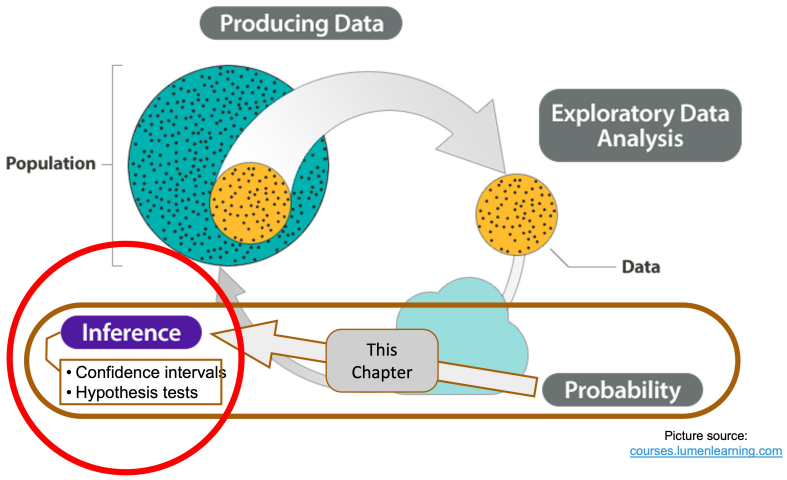
This translates to approximately 68% of IQ scores falling within the range of 85 to 115 and approximately 95% of the scores falling within the range of 70 to 130. Near the two ends, we have less than 2% of the scores falling in either of the ranges 55 to 70 and 130 to 145.

Remark 4.3.9 Before we move on to the next few sections, it is useful to summarise what we have discussed so far in this chapter. We have set the foundation in discussing uncertainty by starting with the description of a probability experiment which always comes with a sample space containing all possible outcomes of the experiment. Events of the experiment are simply sub-collections of the sample space and we assign probabilities to each outcome (and event) of the experiment according to some rules of probability.

Once we have the probabilities of events, it allows us to discuss conditional probability as well as the concept of independence. The probabilities of events are not restricted to a single probability experiment but can also be combinations of multiple experiments. There are also obvious connections to topics studied in earlier chapters like rates and sampling. Finally, we conceptualise random variables, both discrete and continuous, as numerical variables with associated probabilities. Normal distributions, probably the most well known class of continuous random variables, are also introduced. These concepts and constructs have now set the stage for us to continue our discussion of statistical inference in the forthcoming sections.

Section 4.4 Confidence Intervals

In the first part of this chapter, we have discussed probability and laid the foundation that is necessary for us to move towards statistical inference.



In Chapter 1, through the discussion of sampling and random assignment, we saw that using a sample statistic to estimate the population parameter is subjected to inaccuracies. These inaccuracies primarily come under two categories, namely bias and random error. So we typically have

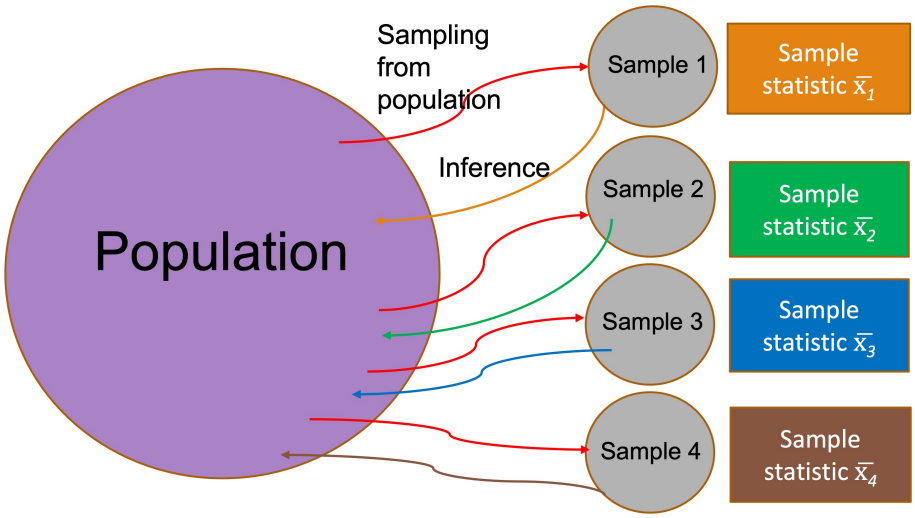
$$\text{Sample statistic} \;=\; \text{population parameter} \;+\; \text{bias} \;+\; \text{random error}.$$

By adopting good sampling methods (e.g. using simple random sampling) and practices (e.g. having a good sampling frame), bias can be reduced to an insignificant level. This would allow us to say

$$\text{Sample statistic} \;=\; \text{population parameter} \;+\; \text{random error}.$$

For the remainder of this chapter, we will assume that our samples are simple random samples taken from a perfect sampling frame with 100% response rate. Hence we assume that there will be no selection bias or non-response bias.

Definition 4.4.1 *Statistical inference* refers to the use of samples to draw inferences or conclusions about the population in question.



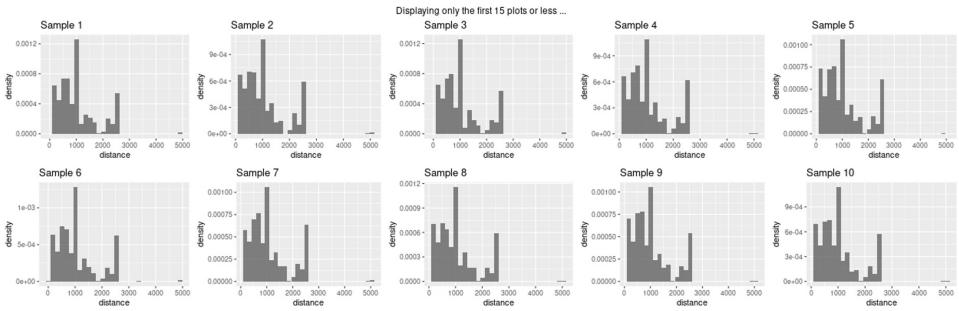
For example, suppose we are interested in the population mean (say, the average height of all twelve year olds in Singapore). We will draw random samples from the population and each sample will give rise to a sample statistic, in this case, the sample mean. Based on our definition of statistical inference, the goal is to state with some level of confidence on what the population mean is, based on the sample means. In what follows, we will discuss two types of statistical inference, namely *confidence intervals* and *hypothesis testing*.

Example 4.4.2 As a more concrete example, consider the following screenshots that show 10 simple random samples, each of size 2500, drawn from a data set containing information on the distances covered by various airplane flights[1] .

```
Mean & SD for samples generated

          S1      S2      S3      S4      S5      S6      S7      S8      S9
mean 1042.98 1075.58 1048.25 1061.51 1040.78 1017.94 1019.73 1045.18 1039.53
sd    731.88  756.74  712.79  729.99  723.69  720.84  711.04  736.02  716.08
         S10
mean 1061.15
sd    732.36
```

[1] You are encouraged to explore the R shiny app at https://david-chew.shinyapps.io/WhySRS/



Notice that the sample means (average distances covered by the 2500 flights) of the 10 samples are all different. What can we infer the population mean to be? We require the concept of confidence intervals.

Definition 4.4.3 A *confidence interval* is a range of values that is likely to contain a population parameter based on a certain degree of confidence. This degree of confidence is called the *confidence level* and is usually expressed as a percentage (%).

Some examples of population parameters that we would like to construct confidence intervals for are proportion, mean and also standard deviation. For this course, we will focus on the construction of confidence intervals for population proportion and mean. We will use two examples to explain the idea behind the construction of the confidence intervals.

Example 4.4.4 The figure below shows part of the data set "2020 Resale Price Data". This data set provides information on the resale transactions of HDB resale flats in the year 2020. There are a total of 23334 transactions (the population) in 2020 and there are 14 variables in this data set.

	year	month	town	flat_type	block	street_na	lstorey	ustorey	floor_are	flat_mode	lease_con	remaining	resale_pri	price_psm
2	2020	1	ANG MO K	3 ROOM	208	ANG MO K	4	6	73	New Gene	1976	55.58333	265000	3630.137
3	2020	1	ANG MO K	3 ROOM	307C	ANG MO K	19	21	70	Model A	2012	91.66667	470000	6714.286
4	2020	1	ANG MO K	3 ROOM	319	ANG MO K	1	3	73	New Gene	1977	56.33333	230000	3150.685
5	2020	1	ANG MO K	3 ROOM	216	ANG MO K	4	6	73	New Gene	1976	55.25	280000	3835.616
6	2020	1	ANG MO K	3 ROOM	556	ANG MO K	7	9	68	New Gene	1980	59.08333	220000	3235.294
7	2020	1	ANG MO K	3 ROOM	536	ANG MO K	10	12	68	New Gene	1980	59.08333	280000	4117.647
8	2020	1	ANG MO K	3 ROOM	560	ANG MO K	4	6	67	New Gene	1980	59.08333	240000	3582.09
9	2020	1	ANG MO K	3 ROOM	463	ANG MO K	4	6	82	New Gene	1980	59.16667	301000	3670.732
10	2020	1	ANG MO K	3 ROOM	476	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97
11	2020	1	ANG MO K	3 ROOM	442	ANG MO K	1	3	67	New Gene	1979	58.58333	233000	3477.612
12	2020	1	ANG MO K	3 ROOM	578	ANG MO K	7	9	67	New Gene	1980	59	250000	3731.343
13	2020	1	ANG MO K	3 ROOM	442	ANG MO K	4	6	67	New Gene	1979	58.58333	245000	3656.716
14	2020	1	ANG MO K	3 ROOM	445	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97
15	2020	1	ANG MO K	3 ROOM	435	ANG MO K	4	6	67	New Gene	1979	58	288000	4298.507
16	2020	1	ANG MO K	3 ROOM	442	ANG MO K	1	3	67	New Gene	1979	58.58333	235000	3507.463
17	2020	1	ANG MO K	3 ROOM	466	ANG MO K	7	9	67	New Gene	1984	63.66667	268000	4000
18	2020	1	ANG MO K	3 ROOM	570	ANG MO K	4	6	67	New Gene	1979	58.41667	240000	3582.09
19	2020	1	ANG MO K	3 ROOM	345	ANG MO K	4	6	88	New Gene	1978	57.33333	320000	3636.364
20	2020	1	ANG MO K	3 ROOM	213	ANG MO K	4	6	67	New Gene	1976	55.25	250000	3731.343
21	2020	1	ANG MO K	3 ROOM	126	ANG MO K	1	3	67	New Gene	1978	57.75	250000	3731.343
22	2020	1	ANG MO K	3 ROOM	587	ANG MO K	1	3	67	New Gene	1979	58.41667	265000	3955.224
23	2020	1	ANG MO K	3 ROOM	121	ANG MO K	10	12	67	New Gene	1978	57.75	3.00E+05	4477.612
24	2020	1	ANG MO K	3 ROOM	345	ANG MO K	4	6	73	New Gene	1978	57.25	320000	4383.562
25	2020	1	ANG MO K	3 ROOM	212	ANG MO K	10	12	67	New Gene	1977	56.25	255000	3805.97
26	2020	1	ANG MO K	3 ROOM	301	ANG MO K	4	6	88	New Gene	1978	57.16667	355000	4034.091
27	2020	1	ANG MO K	3 ROOM	120	ANG MO K	4	6	67	New Gene	1978	57.75	270000	4029.851

To illustrate the construction of the confidence interval for population proportion, we will consider the variable "flat_type". This variable indicates whether the resale flat is of the type 1-room, 2-rooms, 3-rooms, 4-rooms, 5-rooms, executive or multi-generational. It is clear that "flat_type" is a categorical data with 7 categories. Suppose we ask the following question on the population parameter:

> Among the HDB resale transactions in 2020, what proportion (denoted by p) of them is for 5-room flats?

Now, let's say that a simple random sample of 2000 resale transactions are taken and the breakdown of the 2000 transactions according to flat_type is shown in the table below.

	Flat type						
	1-rm	2-rm	3-rm	4-rm	5-rm	Executive	Multi-Gen.
Frequency	2	41	464	819	508	165	1
Proportion	0.001	0.0205	0.232	0.4095	0.254	0.0825	0.0005

Notice that for this sample, the proportion of resale transactions that are 5-room flats is $\frac{508}{2000} = 0.254$. The *population proportion* p is unknown to us and can only be found if we take a census of all the 23334 transactions. What we are interested to know is how good an estimate is our sample proportion of 0.254. If we assume that there is no bias in our sample, then

$$0.254 = \text{ population proportion } + \text{ random error.}$$

It should be noted at this point that random error can be positive or negative. If the random error is positive, then the sample proportion of 0.254 is larger than the population proportion. On the other hand, if the random error is negative, then the sample proportion is smaller than the population proportion. To construct a confidence interval for the population proportion, we use the following formula

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}},$$

where

$$\begin{aligned} p^* &= \text{sample proportion} \\ z^* &= \text{“}z\text{-value” from standard normal distribution} \\ n &= \text{sample size} \end{aligned}$$

The exact value of z^* depends on the *confidence level* of the confidence interval we are constructing. For a 90% confidence interval, the value of z^* is 1.645 while for a 95% confidence interval, the value of z^* is 1.96. Thus, for this example, the 95% confidence interval for the population proportion is

$$0.254 \pm 1.96 \times \sqrt{\frac{0.254(1-0.254)}{2000}} = 0.254 \pm 0.0191.$$

So the interval is 0.254 ± 0.0191.

Remark 4.4.5

- While the computation of the confidence interval is simple, for this course, we will use software to help us perform these computations.

- In particular, the value of z^* that is dependent on the chosen confidence level can be found from statistical tables similar to the one shown below. However, when we use software for the computation of confidence intervals, these values will be appropriately chosen by the software when we specify the confidence level we wish to use.

Standard Normal Probabilities

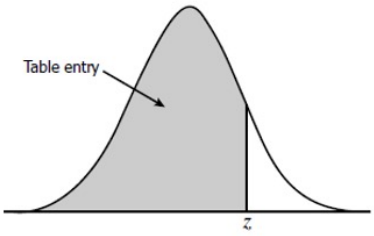


Table entry for z is the area under the standard normal curve to the left of z.

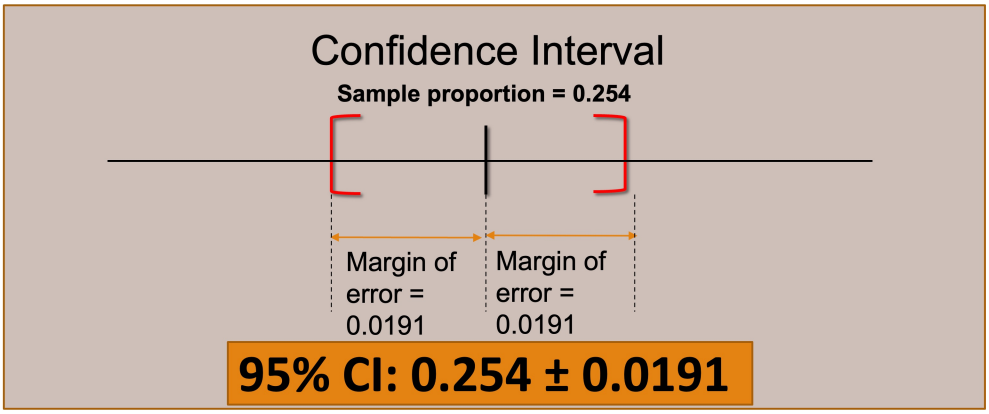
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Discussion 4.4.6 Now that we have seen how the computation of a confidence interval for population proportion is done, what is also important is the **interpretation** of the interval. What does it mean to say that the 95% confidence interval for the population proportion of 5-room resale flat transactions in 2020 is 0.254 ± 0.0191?

A confidence interval is reported in 2 parts, namely:

- The confidence level (for example, 95% in the example above); and
- The interval (0.254 ± 0.0191 for the example above).

The value 0.0191 is known as the *margin of error* which directly impacts the width (how wide/narrow) of the confidence interval.

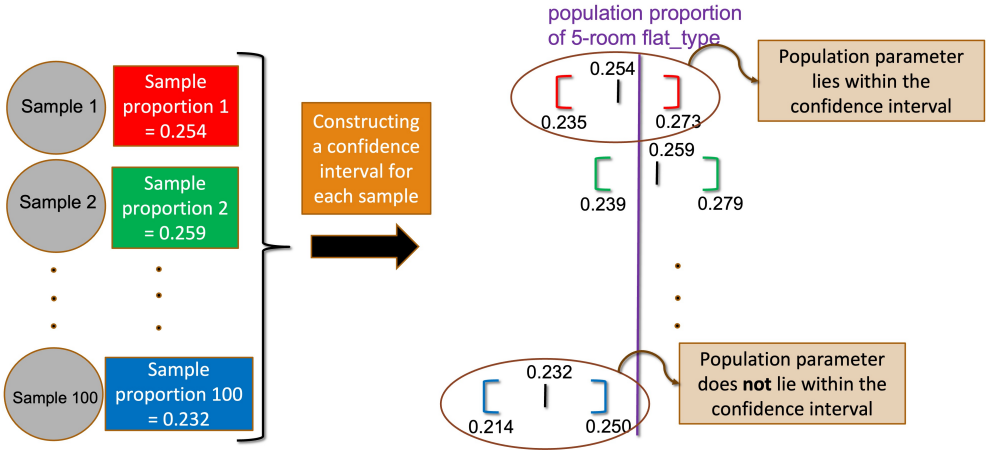


In reporting the confidence interval computed above, we say

> We are 95% confident that the population proportion (the parameter in this case) of resale flat transactions in 2020 that are 5-room, lies within the confidence interval.

It is natural to ponder what we mean by "95% confident". In the context of confidence intervals, this has a specific meaning which can be explained by *repeated sampling*. Recall that the sample statistic of 0.254 was computed from a single sample (collected via Simple Random Sampling) of 2000 resale transactions. It was from this sample statistic that the confidence interval was constructed.

The idea of repeated sampling is based on the supposition that many simple random samples of the same size are taken and with the different sample statistics obtained from the different samples, different confidence intervals are constructed using the same method as above.



Using the idea of repeated sampling, the interpretation of "95% confident" is that if many simple random samples of the same size are taken, and a confidence interval is constructed for each of them, then about 95% of the confidence intervals constructed would contain the population parameter. Thus, if we collected 100 simple random samples and their 95% confidence intervals were computed in the same manner, then about 95 out of the 100 confidence intervals will contain the population parameter. So in the figure above, assuming that the purple line is the actual population proportion of 5-room resale flats, the confidence intervals constructed for samples 1 and 2 would actually contain the population parameter while the confidence interval derived from sample 100 would not.

It is important to remember that in actual fact, we do not know what is the exact value of the population parameter. Confidence intervals certainly give us a better idea of *where* this parameter lies but they can never tell us its exact value.

Remark 4.4.7 Going back to Example 4.4.4, based on the first sample of 2000 households, it is a **common mistake** to say that there is a *95% chance* that the population proportion of 5-room resale flats lies between 0.235 and 0.273. It is actually incorrect to make a statement like this because

- The population proportion p is "fixed", although unknown to us. There is no probabilistic element in what this proportion is going to be.
- For any particular sample, the confidence interval constructed only depends on the sample proportion and the value of z^* corresponding to a chosen confidence level. Thus, the confidence interval is also "fixed" and there is also no probabilistic element in it.

Thus either the population parameter **IS** in the interval or it **IS NOT**. It is wrong to say there is a 95% chance that it is in the interval (and 5% chance that it is not)! The element of chance (or probability) comes from the uncertainty of sampling rather than the uncertainty in the value of the population parameter. Therefore, we should always remember the interpretation as the percentage of samples of the same size collected repeatedly, using the same method of simple random sampling, that give rise to confidence intervals containing the unknown population parameter.

Remark 4.4.8 (Properties of confidence intervals.)

1. Recall that in Example 4.4.4 we computed the confidence interval using the sample estimate of 0.254, confidence level of 95% and sample size $n = 2000$:

$$0.254 \pm 1.96 \times \sqrt{\frac{0.254(1-0.254)}{2000}} = 0.254 \pm 0.0191.$$

What happens when another sample is taken, using the same sampling frame, same sampling method (simple random sampling) but a smaller sample size of 1000? If this new sample also resulted in the sample estimate of 0.254, the 95% confidence interval would be

$$0.254 \pm 1.96 \times \sqrt{\frac{0.254(1-0.254)}{1000}} = 0.254 \pm 0.0270.$$

This confidence interval is wider than the previous one **because the sample size is smaller**. Similarly, if yet another sample is taken, under exactly the same conditions with the only difference being that the sample size is 5000, then if the sample estimate is again 0.254, the confidence interval will now be

$$0.254 \pm 1.96 \times \sqrt{\frac{0.254(1-0.254)}{5000}} = 0.254 \pm 0.0121.$$

What we are seeing here is that the larger the sample size, the smaller the random error, which will then result in a narrower confidence interval. This is not surprising as we have seen in Chapter 1 that increasing sample size can result in reducing random error.

2. Other than the size of the sample, the other factor that affects the width of the confidence interval is the confidence level. Recall that when we set the confidence level to be 95%, the confidence interval obtained, based on $n = 2000$ and the sample proportion of 0.254 was 0.254 ± 0.0191. What happens if we set the confidence level to be 90%? In this case, the value of z^* is 1.645 and the 90% confidence interval for the population proportion is

$$0.254 \pm 1.645 \times \sqrt{\frac{0.254(1-0.254)}{2000}} = 0.254 \pm 0.0160.$$

So the interval is 0.254 ± 0.0160. This interval is *narrower* than the interval obtained when the confidence interval was 95%. Using the idea of repeated sampling, this makes sense since having a narrower interval would imply that a smaller percentage (90% and not 95%) of repeated samples would contain the population parameter. Generally speaking, the higher the confidence level at which the confidence interval is constructed, the wider the confidence interval.

Example 4.4.9 Let us consider another example on the construction of a confidence interval, where we would like to estimate the **population mean** based on a *sample mean*. Using the same data set previously, containing all the resale transactions of HDB flats in 2020, we would like to investigate the mean resale price of all the transactions by constructing confidence intervals.

	year	month	town	flat_type	block	street_na	lstorey	ustorey	floor_are	flat_mode	lease_con	remaining	resale_pr	price_psm
2	2020	1	ANG MO K	3 ROOM	208	ANG MO K	4	6	73	New Gene	1976	55.58333	265000	3630.137
3	2020	1	ANG MO K	3 ROOM	307C	ANG MO K	19	21	70	Model A	2012	91.66667	470000	6714.286
4	2020	1	ANG MO K	3 ROOM	319	ANG MO K	1	3	73	New Gene	1977	56.33333	230000	3150.685
5	2020	1	ANG MO K	3 ROOM	216	ANG MO K	4	6	73	New Gene	1976	55.25	280000	3835.616
6	2020	1	ANG MO K	3 ROOM	556	ANG MO K	7	9	68	New Gene	1980	59.08333	220000	3235.294
7	2020	1	ANG MO K	3 ROOM	536	ANG MO K	10	12	68	New Gene	1980	59.08333	280000	4117.647
8	2020	1	ANG MO K	3 ROOM	560	ANG MO K	4	6	67	New Gene	1980	59.08333	240000	3582.09
9	2020	1	ANG MO K	3 ROOM	463	ANG MO K	4	6	82	New Gene	1980	59.16667	301000	3670.732
10	2020	1	ANG MO K	3 ROOM	476	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97
11	2020	1	ANG MO K	3 ROOM	442	ANG MO K	1	3	67	New Gene	1979	58.58333	233000	3477.612
12	2020	1	ANG MO K	3 ROOM	578	ANG MO K	7	9	67	New Gene	1980	59	250000	3731.343
13	2020	1	ANG MO K	3 ROOM	442	ANG MO K	4	6	67	New Gene	1979	58.58333	245000	3656.716
14	2020	1	ANG MO K	3 ROOM	445	ANG MO K	1	3	67	New Gene	1979	58.58333	255000	3805.97

We will describe how a confidence interval for population mean resale price is constructed. The properties and interpretations of the confidence interval for a population mean are similar to those for a population proportion that we have discussed in Example 4.4.4. Again, we will not be computing these confidence intervals by hand but instead use software to help us perform these computations.

Suppose we have a sample, obtained via simple random sampling, with sample size 2000. The sample mean resale price is found to be $\overline{x} = \$448,727$. Let μ be the population mean resale price, a population

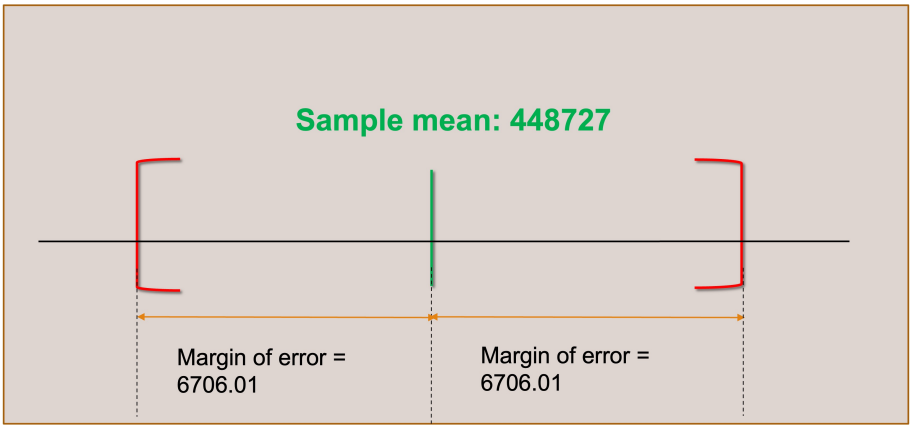
parameter of interest that is unknown to us unless we take a census of the population. A 95% confidence interval for the population mean μ is constructed using the formula

$$\overline{x} \pm t^* \times \frac{s}{\sqrt{n}},$$

where

$$\begin{aligned} \overline{x} &= \text{sample mean} \\ t^* &= \text{“}t\text{-value” from } t\text{-distribution} \\ s &= \text{sample standard deviation} \\ n &= \text{sample size} \end{aligned}$$

The exact value of t^* depends on the sample size n and the *confidence level* of the confidence interval we are constructing. Without going into the computation details, we will simply state that the 95% confidence interval for the population mean is found to be $448,727 \pm 6706.01$.



The margin of error, $\$6,706.01$ is a way of quantifying the random error and as discussed previously, this error can be reduced by increasing the sample size n (everything else being equal). The width of the confidence interval can also be narrowed if we reduce the confidence level to one that is lower than 95%.

To summarise this section on confidence intervals, recall the following:

1. The use of confidence intervals is a way for us to quantify random error that is present in every sample, even in those obtained via simple random sampling where the level of bias can be reduced or assumed negligible.

2. Confidence intervals and the confidence level used to compute the intervals can be understood via the idea of repeated sampling. We should avoid using the word “chance” or “probability” when we discuss whether the population parameter lies inside the confidence interval constructed from a single sample.

3. We have discussed some properties of confidence intervals, in particular how the interval varies according to the sample size and the confidence level applied.

4. We saw how confidence intervals are constructed for two population parameters, namely the population proportion and the population mean. It is useful to understand how the construction is done although for the purpose of this module, we will rely on software to assist in the computation.

Section 4.5 Hypothesis Testing

Discussion 4.5.1 Recall that many research questions are of the type where we wish to test a claim about a population. It is often that such questions involve a claim that has a "YES" or "NO" answer. One example of such a question could be "*Is vaccine X effective and safe enough to be administered to the entire Singapore population aged 10 and above?*". One way we can begin to answer such a question is through the collection of sample data, followed by performing a *hypothesis test* on that data collected.

The process of hypothesis testing comprises of a few key steps, which we will describe below.

(Four steps of hypothesis testing)

Step 1: The first step of hypothesis testing is to identify the question and state the *null hypothesis* and *alternative hypothesis*.

- The null hypothesis usually asserts the stand of **no effect** or **no difference**. Indirectly, this means that whatever differences or variances that is observed in the sample data is not inherent in the population and had occurred by random chance when we were choosing the sampling units.
- The alternative hypothesis, on the other hand, is typically what we wish to confirm and pit against the null hypothesis. In many research questions, we often hope that the sample data provides sufficient evidence for us to reject the null hypothesis in favour of the alternative hypothesis.
- It is important to note that the null hypothesis and the alternative hypothesis must be *mutually exclusive*, meaning that they cannot be true simultaneously.

Example 4.5.2 For example, suppose we believe that a particular coin is "loaded" and thus biased towards one of the two sides, say heads. This means that any toss of the coin will **more likely** show heads rather than tails. How can we formulate our hypotheses in this case? As mentioned above, the null hypothesis takes the stand of "no difference" so in our case, it means that for any toss, "there is no difference in the likelihood of showing heads as compared to showing tails". We write

> **Null hypothesis** H_0: "The coin is fair", that is, $P(H) = 0.5$ where $P(H)$ is the probability of observing heads in a toss.

The alternative hypothesis would be asserting our belief that the coin is biased, more precisely written as

> **Alternative hypothesis** H_1: "The coin is biased towards heads", that is $P(H) > 0.5$.

It should be noted that H_0 (resp. H_1) are standard notations to denote the null (resp. alternative) hypothesis.

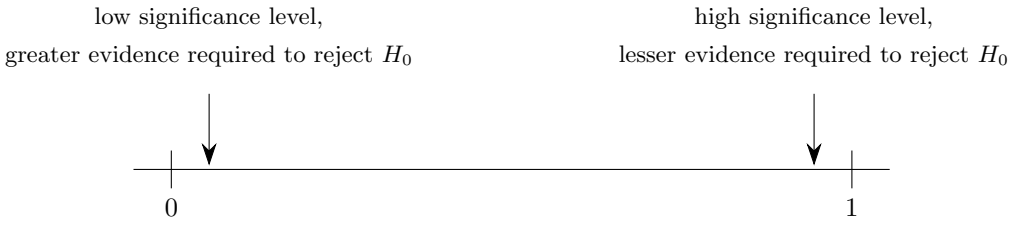
Step 2: After the hypotheses have been formulated, the second step involves the collection of relevant data that is necessary for the test. In our example of testing if a coin is biased, we can simply toss the coin several times (say 8 times) and observe the outcomes (heads or tails) of the tosses. These observations will form our sample data that will subsequently be used for the test. If we are testing whether vaccine X is effective and safe enough for Singaporeans aged 10 and above, the relevant data to be collected would be a sample of Singaporeans aged 10 and above, and records of their well-being and vaccine efficacy after being administered the vaccine.

The process of testing a hypothesis usually involves a *random variable*, previously introduced in Section 4.3, and its probability distribution. For the example of testing the fairness of a coin, the random variable is the number of heads out of 8 independent coin tosses, assuming the coin is fair. If we represent this random variable by X, then we see that X is a *discrete random variable* and X can

take integer values from 0 to 8. Going hand in hand with the random variable is the *test statistic* which is a value computed/observed from the sample data that will be used to determine whether the null hypothesis is to be rejected or not. For the coin testing example, the test statistic is simply the number of heads observed from the 8 tosses. Thus, the test statistic will also be an integer from 0 to 8. The concept of a test statistic will be discussed in detail subsequently (see Discussion 4.5.8, Example 4.5.9 and Example 4.5.11).

Step 3: In this step, we have to set the *significance level* of our test. The significance level of a hypothesis test can be thought of as **how "convincing" we need our evidence to be before we reject the null hypothesis in favour of the alternative hypothesis**.

- The significance level of a hypothesis test is always a number between 0 and 1.
- The lower (closer to 0) the significance level, the "greater" the evidence needs to be (the more convincing) before we reject the null hypothesis in favour of the alternative.



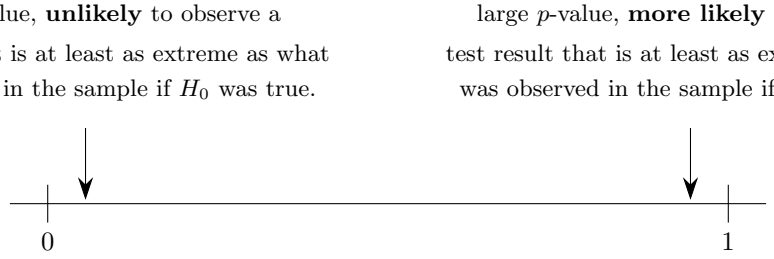
- A commonly used level of significance is 0.05 (also referred to as 5% *level of significance*). Other levels used frequently are 0.10 (10% level of significance) and 0.01 (1% level of significance).

Let us provide a more intuitive explanation of what is meant by "greater evidence". If we look at our test of whether a coin is fair, suppose the 8 tosses we performed resulted in 7 heads and 1 tail, this would certainly provide "greater evidence" for the alternative hypothesis (that the coin is biased towards heads) against the null hypothesis, as compared to if the outcome of the 8 tosses were 5 heads and 3 tails..

To continue with Step 3, we will compute the *p-value* of the hypothesis test.

> The p-value is the probability of obtaining a test result at least as extreme as the result we observed, **assuming the null hypothesis is true**. An easy way to remember the meaning of the p-value is the probability of observing a test result that favours the alternative hypothesis at least as much as what is observed in the current sample, while assuming that the null hypothesis is true.

The following figure aids in the interpretation of the p-value.



It is now clear that some form of comparison between the relative values of the significance level and the p-value will be done before a conclusion of the test can be made. More about that in a while.

Example 4.5.3 Let us continue our coin toss test from Example 4.5.2. So suppose in the 8 independent tosses of the coin, we did in fact observe 7 heads and 1 tail. What is the p-value in this case? Recall

that the p-value is computed under the assumption that the null hypothesis is true, so in this case, we assume that the coin is fair and therefore, the probability of obtaining "head" (or "tail") on any toss is 0.5. Next, we need to establish the meaning of "at least as extreme as our observation" under the null hypothesis.

If the coin is fair, observing 7 heads from 8 tosses seems pretty extreme. What can be more extreme than this? Observing 8 heads of course! We are now ready to compute the p-value for the hypothesis test.

$$\begin{aligned} p\text{-value} &= P(\text{Obtaining a result at least as extreme as observed} \mid H_0 \text{ is true}) \\ &= P(\text{Exactly 7 out of 8 tosses result in heads} \mid H_0 \text{ is true}) \\ &\quad + P(\text{Exactly 8 out of 8 tosses result in heads} \mid H_0 \text{ is true}) \\ &= 8\left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^8 = 9\left(\frac{1}{2}\right)^8 = 0.035156. \end{aligned}$$

Note that the probability of observing exactly 7 heads out of 8 tosses is the sum of

1. the probability of the first toss is tails, the rest of the tosses are heads, which is $\frac{1}{2} \times \left(\frac{1}{2}\right)^7 = \left(\frac{1}{2}\right)^8$.
2. the probability of the second toss is tails, the rest of the tosses are heads, which is also $\left(\frac{1}{2}\right)^8$.

⋮

8. the probability of the eighth toss is tails, the rest of the tosses are heads, which is also $\left(\frac{1}{2}\right)^8$.

Thus,

$$P(\text{Exactly 7 out of 8 tosses results in heads} \mid H_0 \text{ is true}) = 8 \times \left(\frac{1}{2}\right)^8.$$

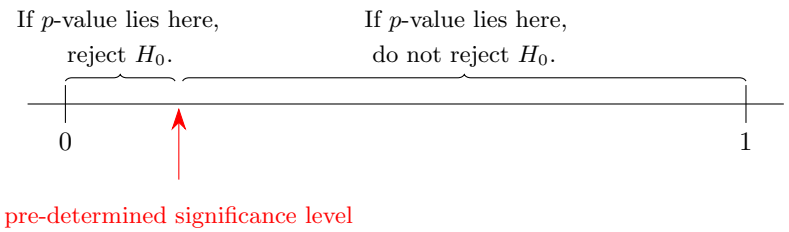
Step 4: We are now ready to make a conclusion on whether to reject or not reject the null hypothesis. There are only two possible conclusions that we can make:

- The null hypothesis is rejected in favour of the alternative, that is "Reject H_0". We make this conclusion if the p-value computed in step 3 is less than the significance level set for the test in step 3.
- The null hypothesis is not rejected, that is "Do not reject H_0". This happens if the p-value computed is greater than or equal to the significance level.

It is important to note that the following are **NEVER** conclusions from any hypothesis test:

- "Accept H_0." We never accept the null hypothesis. Our test only attempts to reject this hypothesis based on the observation in the sample.
- "Reject H_1." We should also avoid saying that the alternative hypothesis is rejected when the p-value is greater than the significance level. It is more appropriate to say that based on the observation in the sample, there is insufficient evidence to support the alternative hypothesis, at the significance level that we have set for our test.

The relative values of the level of significance and the p-value, resulting in the two possible outcomes of the test is summarised in the figure below.



Example 4.5.4 Completing our hypothesis test from Example 4.5.3, since the computed p-value is 0.035156, we concluded that an observation of 7 heads out of 8 independent tosses will lead to the rejection of H_0 at 5% significance level and conclude that there is evidence to support the alternative hypothesis that the coin is biased towards heads.

On the other hand, it should be noted that if we had set the level of significance of the test to be 3%, observing 7 heads out of 8 independent tosses would not be enough evidence to reject H_0 in favour of H_1. This is because the p-value would be the same at 0.03516, which is greater than 0.03 (= 3%). One should remember that the significance level for a hypothesis test is decided before the p-value is calculated and not the other way round.

Example 4.5.5 Let us look at the coin toss example again but with slight variation. Again, we would like to test if a coin is biased towards heads, so the null and alternative hypotheses remain the same as per Example 4.5.2. Now suppose in 8 tosses, we observe 3 heads. What would be the p-value in this case? Recall that the p-value is the probability of obtaining a test result at least as extreme as the one observed, assuming that the null hypothesis is true. Thus, if X is the number of heads we observe occurring out of 8 tosses, then the p-value will be

$$P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8),$$

that is, the sum of the probabilities of observing 3, 4, 5, 6, 7 or 8 heads in 8 tosses.

Example 4.5.6 Let us consider a slightly different hypothesis test. Suppose instead of suspecting that the coin is biased towards showing heads, we now believe that it is biased towards showing tails. In other words, our null and alternative hypothesis can now be

Null hypothesis H_0: "The coin is fair", that is, $P(H) = 0.5$.

Alternative hypothesis H_1: "The coin is biased towards tails", that is $P(H) < 0.5$.

Now, suppose we once again observe $X = 3$ **heads** out of 8 coin tosses, meaning that **5 tails** are observed. What is the range of results as extreme as our observation? Since our alternative hypothesis is now $P(H) < 0.5$, results at least as extreme as observing 5 tails would be if **5 or more (that is, 6, 7 or 8) tails** are observed. This would mean $X = 0, 1, 2$ **or** 3 where X is defined as the number of heads observed in 8 tosses. Thus the p-value will be

$$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3).$$

It is important to note the difference between this and Example 4.5.5 where the p-values are different despite having the same observation of 3 heads out of 8 tosses. The reason for this difference lies in the difference in the alternative hypothesis. Thus, when we are computing p-values, it is imperative that we know what the null and alternative hypotheses are.

Remark 4.5.7

1. When the p-value is greater than the level of significance, we do not reject the null hypothesis. This means that the test is inconclusive and we do not know if the observation is due to chance or not.

2. By not rejecting H_0, it does not mean that H_0 is true. It really just means that our test is inconclusive and nothing further can be said about the null hypothesis **based on this test**.

3. Hypothesis tests are only done when we have sample data and not information on the entire population. In the unlikely event that we have population level data, all can be determined and there is no need for any hypothesis test!

4. Let us summarise the four steps of hypothesis testing as follows:

    - Step 1 is to identify the question and state the null and alternative hypotheses.

- Step 2 is to collect the relevant data and determine the test statistic based on the data collected.
- Step 3 is to set the level of significance followed by computing the p-value.
- Step 4 is to compare the p-value with the level of significance and determine if we should reject the null hypothesis (in favour of the alternative hypothesis) or not.

It is useful to note that in practice, the collection of data (Step 2) may sometimes happen *before* a researcher examines the data, poses a question and states the hypotheses (Step 1).

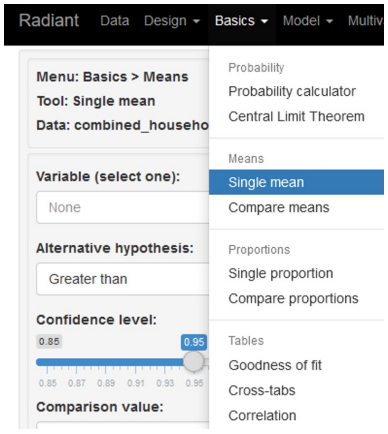
Discussion 4.5.8 Now that we have given the overall framework of a hypothesis test, there are some details that require more explanation. We need to decide on the hypothesis that we actually want to test and to be clear with regards to what type of variable we are performing a hypothesis test on. This is because the type of hypothesis test we use depends on the type of variable we are dealing with. In the next two examples, we will discuss two common hypothesis tests that you are most likely to come across. The computation for these examples are done using Radiant[2].

Example 4.5.9 One type of hypothesis that is commonly tested is about the **mean** of a numerical variable in a population. For example, we may be interested in the average (valuation) price of HDB flats in the year 2020. Suppose there are two researchers, A and B who are making claims on the average price of HDB flats in 2020. Researcher A claims that the average price is \$600,000 while researcher B believes that it is higher than \$600,000. If we let the population of interest to be all HDB flats in 2020 and denote the population mean price as μ, then we can set up a hypothesis test with

Null hypothesis H_0: $\mu = 600,000$.

Alternative hypothesis H_1: $\mu > 600,000$.

We draw a simple random sample of 1000 HDB flats from the 1.3 million flats in Singapore in 2020. We will use Radiant to perform a *one-sample t-test*. After loading the data set into Radiant, from the toolbar, we will select "Basics" followed by "Single mean".



Make the appropriate selection as follows:

- Under "Variable", select "flat_price".
- Under "Alternative hypothesis", select "Greater than". Notice that this must agree with the intended H_1 of $\mu > 600,000$.
- Under "Comparison value", enter "600,000".

Radiant now computes the p-value and the output is shown below.

[2]See https://radiant-rstats.github.io/docs/install.html

```
Single mean test
Data      : combined_household_SRS
Variable  : flat_price
Confidence: 0.95
Null hyp. : the mean of flat_price = 600000
Alt. hyp. : the mean of flat_price is > 600000

        mean     n n_missing          sd        se        me
 616,326.799 1,000         0 155,981.331 4,932.563 9,679.372

    diff       se t.value p.value  df       5% 100%
 16326.8 4932.563    3.31  < .001 999 608205.9  Inf ***
```

Notice that the **confidence level** is 0.95, which means that the significance level is 0.05 (5%). The computed p-value is less than 0.05, in fact it is less than 0.001 and thus we conclude that at 5% level of significance, we have sufficient evidence to reject the null hypothesis in favour of the alternative hypothesis that the mean resale flat price is greater than \$600,000.

Remark 4.5.10 The one-sample t-test conducted in Example 4.5.9 can be done when the following criteria are satisfied.

1. If the size of our sample is smaller than 30, we must have sufficient reason to believe that the population distribution of the variable is approximately normal. On the other hand, if the size of our sample is larger than or equal to 30, we need not make this assumption.
2. The sample was randomly produced, for example, obtained via simple random sampling.

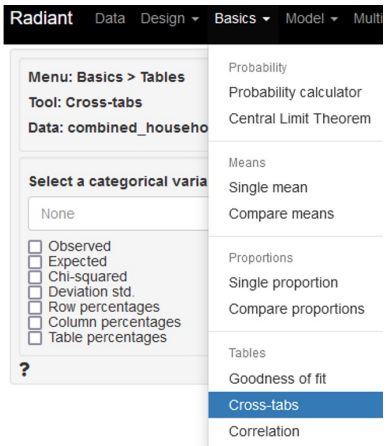
Example 4.5.11 The second type of hypothesis test we will introduce is known as the *chi-squared test*. This test is commonly used to test whether two categorical variables, say A and B, are associated at the population level. Let us use the same data set from Example 4.5.9. For the purpose of this example, we will create two categorical variables, namely, *flat value* and *Large small household*. For each flat in the population, the *flat value* is said to be "**Expensive**" if its value is greater than or equal to \$600,000 and "**Not Expensive**" otherwise. Thus, the categorical variable *flat value* has two categories, "Expensive" and "Not Expensive".

Similarly, for each flat in the population, the *Large small household* of the flat is said to be "**Large**" if the household has more than 3 members and "**Small**" otherwise. Thus, the categorical variable *Large small household* also has two categories, "Large" and "Small". We would like to test if the two categorical variables are associated with each other at the population level. To do this, we set up our null and alternative hypotheses as follows:

Null hypothesis H_0: *Large small household* is not associated with *flat value* in Singapore.

Alternative hypothesis H_1: *Large small household* is associated with *flat value* in Singapore.

We will again use Radiant to perform a chi-squared test. After loading the data set into Radiant, from the toolbar, we will select "Basics" followed by "Cross-tabs".



We are now ready to select two categorical variables that we are interested in testing. After selecting "flat_value" and "Large_small_household" (the names of the variables), we select the options "Observed", "Expected" and "Chi-squared". Radiant now computes the p-value and the output is shown below.

```
Observed:
               Large_small_household
flat_value      Large Small Total
  Expensive       193   313   506
  Not Expensive   205   289   494
  Total           398   602 1,000

Expected: (row total x column total) / total
               Large_small_household
flat_value      Large    Small    Total
  Expensive       201.39   304.61   506.00
  Not Expensive   196.61   297.39   494.00
  Total           398.00   602.00 1,000.00

Contribution to chi-squared: (o - e)^2 / e
               Large_small_household
flat_value      Large Small Total
  Expensive     0.35  0.23  0.58
  Not Expensive 0.36  0.24  0.59
  Total         0.71  0.47  1.17

Chi-squared: 1.175 df(1), p.value 0.278
```

Observe that in our sample of 1000 flats, 506 are in the "Expensive" category while 494 are "Not Expensive". Among the 506 flats in the "Expensive" category, 193 are "Large" households while 313 are "Small" households. Similarly, among the 494 "Not Expensive" flats, 205 are "Large" households while 289 are "Small" households. The p-value computed for this chi-squared test is 0.278. Comparing the p-value with the significance level of 0.05, we observe that there is not enough evidence to reject the null hypothesis. Thus the conclusion is that we cannot conclude that *Large small household* is associated with *flat value* in Singapore.

Remark 4.5.12

1. To use the chi-squared test, naturally the data must be counts for the categories of the categorical variables that we are testing. Similar to the one-sample t-test, the sample taken from the population should be a random sample before the test can be deployed.

2. We have only given very brief descriptions of the t-test and the chi-squared test. Further explanations on them, as well as the technicalities involved can be found in standard books on statistics, but are beyond the scope of this module.

3. The table below provides a summary of the main features of both tests.

One-sample t-test	Chi-squared test
Mainly used to test difference between sample mean and a known or hypothesised mean.	Mainly used to test for association between two categorical variables.
Population distribution should be approximately normal if sample size is small.	Data required for the test is the count for the categories of a categorical variable.
Data used should be acquired via random sampling.	Data used should be acquired via random sampling.

Exercise 4

1. There are 5 red balls and 5 blue balls in a bag. Mary picks two balls from the bag, one after another without replacement. She wins a prize if the two balls she picks are of the same colour. The first ball she picks is red. What is the probability that Mary wins a prize?

   (A) $\frac{2}{9}$
   (B) $\frac{1}{2}$
   (C) $\frac{4}{9}$
   (D) $\frac{1}{9}$

2. A glass-making machine produces tempered glass for mobile phones. The tempered glass produced are either grade A (good), grade B (minor defects) or grade C (unacceptable). The glass produced are put through an inspection machine that is able to detect any grade C glass and discard it. We may assume that the inspection machine makes no mistakes in the detection process, meaning that given a piece of grade C glass, the probability of it being discarded is 1. At the same time, given a piece of grade A or B glass, the probability of it being discarded is 0.

   Suppose the glass-making machine produces grade A glass 90% of the time, grade B glass 2% of the time and grade C glass 8% of the time. If a piece of glass is not discarded, what is the probability (computed to 3 significant figures) that it is grade A?

   (A) 0.900
   (B) 0.920
   (C) 0.978
   (D) There is not enough information for the probability to be computed.

3. Which of the following statements is/are true of normal distributions? Select all that apply.

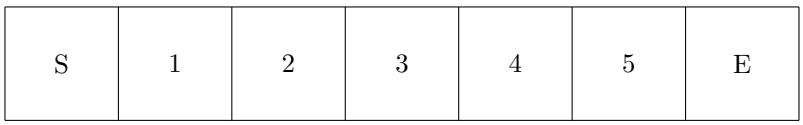
   (A) The density curve of a normal distribution is symmetrical about its mode.
   (B) $N(1, 4)$ has a standard deviation of 2.
   (C) A normal distribution can be left-skewed.
   (D) A normal distribution can be right-skewed.

4. A study was conducted on 1000 students in a secondary school on 11th January 2021. On that day, of these 1000 students, 80% of the students took the MRT (mass rapid transit) to school. Among those students that took the MRT to school, 10% were late for school. Among those students who did not take the MRT to school, 20% were late for school.

   Among those students who participated in the study and were late for school, a student was randomly chosen, with every student having the same chance of being selected. What is the probability (correct to 1 decimal place) that the chosen student did not take the MRT?

   (A) 20.0%
   (B) 10.0%
   (C) 33.3%
   (D) 30.0%

5. A game is played using a fair six-sided die, a pawn and a simple board as shown below. (A pawn is a chess piece.)



Initially, the pawn is placed on square S. The game is played by throwing the die and moving the pawn back and forth in the following manner:

S 1 2 3 4 5 E 5 4 3 2 1 2 3 .....

Thus, for example if the first and second throws of the die give a "5" and "4" respectively, the final position of the pawn will be on square "3", because the first throw would send the pawn to square 5, and the second throw would then send the pawn from square "5" to square "3".

The game will stop only when the pawn stops at square "E" after a die roll, passing by "E" **does not** end the game.

Let X denote the number of throws of the die required to move the pawn such that it stops at square "E". Which of the following statements is/are true?

(I) $P(X = 2) = \frac{5}{36}$.

(II) The events $X = 1$ and $X = 2$ are mutually exclusive.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

6. Tom selects a child at random from a population of children. Let

   - A be the event a child of age < 3 is selected;
   - B be the event a child of age < 5 is selected.

   It is known that $P(A) > 0$. Which of the following must be true?

   (I) $P(A \text{ or } B) < P(A) + P(B)$.

   (II) $P(A) \leq P(B)$.

   (A) Only (I).

   (B) Only (II).

   (C) Neither (I) nor (II).

   (D) Both (I) and (II).

7. Benny is a messy student who keeps all his coloured socks in a box. The box contains a total of 4 blue and 2 yellow socks. While running late for class, he randomly selects (without replacement) two socks out of the box to wear before leaving the house. Assume the socks are indistinguishable from one another in all respects other than their color.

   What is the probability that Benny will end up wearing a pair of matching coloured socks when leaving the house?

   (A) $\frac{1}{3}$.

   (B) $\frac{2}{5}$.

   (C) $\frac{7}{18}$.

   (D) $\frac{7}{15}$.

8. A standard deck of 52 playing cards comprises of 4 suits (Clubs, Diamonds, Hearts, Spades), each suit with 13 cards of distinct ranks (A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K).

   Let P_1 be the probability that a card randomly selected from the deck is a "2".

   Let P_2 be the probability that a card randomly selected from the deck is a "2", **given that it is a "Spade"**.

   Which of the following statements is correct?

   (A) $P_1 = P_2$.
   (B) $P_1 < P_2$.
   (C) $P_1 > P_2$.

9. Systematic sampling with interval length $k = 20$ is to be utilised for an exit poll at an event. Suppose a total of 410 people exit the event. What is the probability that the 57th person to exit the event is selected for the poll?

   (A) $\frac{1}{317}$.
   (B) $\frac{1}{20}$.
   (C) $\frac{1}{410}$
   (D) Cannot be determined with the information provided.

10. Adrian takes an instant test for a viral disease. On the box of the test kit, it is mentioned that both the sensitivity and specificity of the test is 0.99. Upon checking the Ministry of Health website, he also finds that the disease affects 0.1% of Singapore residents. What is the probability (rounded to 2 decimal places) that Adrian has the disease if he obtains a positive test result from the test kit?

    (A) 0.01.
    (B) 0.99.
    (C) 0.09.
    (D) 0.91.

11. There are 3 families X, Y and Z. The families have 2 children each. Family X has 1 boy and 1 girl. Family Y has only 2 girls and Family Z has only 2 boys. 1 child is randomly selected among the 6 children. If the selected child is a boy, what is the probability that he is from family X?

    (A) 0.
    (B) $\frac{1}{6}$.
    (C) $\frac{1}{3}$.
    (D) $\frac{1}{2}$.

12. We wish to deploy a certain number of sensors around a particular area so as to detect intruders moving through the area. We may assume that the sensors function independently and each has probability 0.9 of detecting an intruder in the area. We would like to achieve at least 99.5% success rate of detecting an intruder using the sensors. What is the minimum number of sensors we need to deploy in order to achieve this target?

    (A) Two.
    (B) Three.
    (C) Four.
    (D) Target cannot be achieved.

13. Suppose A and B are two events. Which of the following statements is/are true?

(I) $P(A \text{ and } B)$ is always less than or equal to $P(A)$.

(II) $P(A \mid B)$ is always less than or equal to $P(A)$.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

14. Let A and B be events of a sample space. Consider the following statements.

(I) $P(A) + P(\text{not } A) = 1$.

(II) $P(A) = P(A|B) + P(A \mid \text{ not } B)$.

Which of the statements must be true?

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

15. A test comprises three multiple choice questions. The first question provides two options, and the remaining two questions provide 3 options each. Each question has a unique answer. Suppose that a student (who did not study at all for the test) picks an option randomly for each question, and that these picks are independent of each other. Which of the following is closest to the probability that this student gets exactly 1 out of 3 questions correct?

(A) 0.22.

(B) 0.33.

(C) 0.44.

(D) 0.55.

16. A bag contains four balls numbered 1, 2, 3 and 4. In a game, a ball is drawn once at random from the bag to have its number read. Next, a fair coin is tossed that number of times independently. The discrete random variable X is the number of heads observed from the coin tosses.

Fill in the blank in the following statement:

The probability of X being 0 is _____ (give your answer correct to 2 decimal places).

17. For two events A and B, $P(A) = 0.5$ and $P(B) = 0.2$. Which of the following statements is/are always true? Select all that apply.

(A) $P(A \text{ or } B) = 0.7$ if A and B are independent events.

(B) $P(A \text{ or } B) = 0.7$ if A and B are mutually exclusive events.

(C) $P(A \text{ and } B) = 0.1$ if A and B are independent events.

(D) $P(A \text{ and } B) = 0.1$ if A and B are mutually exclusive events.

(E) $P(A \text{ and } B) = 0$ if A and B are mutually exclusive events.

18. A bag consists of 10 balls. 4 of the balls are yellow while the remaining are green. 2 balls are drawn at random from the bag, one at a time with every ball having the same chance of being chosen on each draw. Let A be the event that the first ball drawn is yellow. Let B be the event that the second ball drawn is green. Which of the following statements is true?

(A) If the balls are drawn **without replacement**, events A and B are independent and mutually exclusive.

(B) If the balls are drawn **without replacement**, events A and B are independent but not mutually exclusive.

(C) If the balls are drawn **with replacement**, events A and B are independent and mutually exclusive.

(D) If the balls are drawn **with replacement**, events A and B are independent but not mutually exclusive.

19. Based on a random sample of 200 staff members in NUS, the 95% confidence interval for the proportion of all NUS staff who went on vacation for at least 5 days in 2018 is $(0.33, 0.59)$. Which of the following statements must be true?

(I) If another sample of size 500 is drawn using the same sampling method, for the same confidence level, the confidence interval will be wider than $(0.33, 0.59)$.

(II) A maximum of 59% of all NUS staff went on vacation for at least 5 days in 2018.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

20. A researcher takes a random sample from Country X's population to estimate its unemployment rate. From the sample, the researcher obtains the 95% confidence interval for the population unemployment rate to be between 0.18 and 0.22.

Which of the following statements correctly interprets the results?

(A) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the sample unemployment rate lie between 0.18 and 0.22.

(B) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the population unemployment rate lie between 0.18 and 0.22.

(C) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the sample unemployment rate lie within the samples' respective confidence intervals.

(D) If many samples of the same size were collected using the same procedure, and their respective confidence intervals calculated in the same way, about 95% of these samples will have the population unemployment rate lie within the samples' respective confidence intervals.

21. A 95% confidence interval, constructed from a random sample, for the population mean number of children per household in Country Z is $(1.21, 4.67)$. Which of the following statements is/are true? Select all that apply.

(A) The probability that the population mean number of children per household in Country Z lies between 1.21 and 4.67 is 0.95.

(B) We are 95% confident that the population mean number of children per household in Country Z lies between 1.21 and 4.67.

(C) 95% of all samples of the same size and sampling procedure should have sample mean number of children per household between 1.21 and 4.67.

(D) 95% of all households in Country Z have between 1.21 and 4.67 children.

(E) If we take 100 different samples of the same size using the same sampling procedure and compute the confidence interval for each sample in the same way, approximately 95 of the intervals will contain the true population mean.

22. Two different random samples (call them Sample 1 and Sample 2) of size 100 each were chosen from a population of 10000 people. For Sample 1, the 95% confidence interval (call this Interval 1) for the population mean height was calculated. For Sample 2, the 99% confidence interval (call this Interval 2) for the population mean height was calculated. Which of the following statements is/are correct? Select all that apply.

(A) If the population mean height lies in Interval 1, then it must lie in Interval 2.

(B) If the population mean height lies in Interval 2, then it must lie in Interval 1.

(C) The population mean height must lie in at least one of the two confidence intervals.

(D) It is possible that the population mean height does not lie in both Intervals 1 and 2.

(E) It is possible that the population mean height lies in both Intervals 1 and 2.

23. A random sample of size 500 is taken from a population of 10000 people of age 50. From the sample, a 95% confidence interval for the population mean weight is constructed. Which of the following statements is/are correct?

(I) The confidence interval will always contain the sample mean weight.

(II) If many samples of the same size are collected using the same sampling method, about 5% of the confidence intervals from these samples will not contain the population mean weight.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

24. A 99% confidence interval for the mean height (in meters) of NUS students is $[1.58, 1.80]$. It is constructed using a random sample of 100 students. Using the same sample, which of the following is a plausible 95% confidence interval for the mean height?

(A) $[1.64, 1.86]$.

(B) $[1.61, 1.77]$.

(C) $[1.48, 1.89]$.

(D) $[1.63, 1.85]$.

25. A coin manufacturer claims that he has produced a biased coin with $P(H) = 0.4$ and $P(T) = 0.6$, where $P(H)$ denotes the probability of the coin landing on heads and $P(T)$ denotes the probability of the coin landing on tails. Out of 10 independent tosses, Brad observes 8 heads and 2 tails. Based on these data, he decides to do a hypothesis test to see if there is enough evidence to reject the manufacturer's claim. Which of the following statements should he adopt as his null hypothesis?

(A) $P(H) = 0.4$.

(B) $P(H) = 0.5$.

(C) $P(H) = 0.8$.

(D) $P(H) = 0.6$.

26. Suppose we want to test if a coin is biased towards heads. We decide to toss the coin 10 times and record the number of heads. We shall assume the independence of coin tosses, so that the 10 tosses constitute a probability experiment.

    Let X denote the number of heads occurring in 10 tosses of the coin. We will carry out a hypothesis test with X as the test statistic. Let H be the event that the coin lands on head, in a single toss. We set our hypotheses to be

    - $H_0 : P(H) = 0.5$,
    - $H_1 : P(H) > 0.5$.

    Suppose in our execution of the 10 tosses, we observe 4 heads. This means $X = 4$ is the test result we observe.

    Recall the definition of p-value to be the probability of obtaining a test result at least as extreme as the one observed, assuming the null hypothesis is true. What is the range of test results "at least as extreme as the one observed", in this scenario?

    (A) $0 \leq X \leq 4$.

    (B) $4 \leq X \leq 10$.

    (C) $0 \leq X \leq 5$.

    (D) $5 \leq X \leq 10$.

27. A group of students wants to find out if there is any association between staying in a hall and being late for class in NUS in a particular month. If students are late for at least 5 classes, they are considered "late for class" in that month. After collecting a random sample of 1000 students, they found that 200 out of 350 students who stay in a hall are late for class, while 390 out of 650 students who do not stay in a hall are late for class.

    A chi-squared test was done to test for association between staying in a hall and being late for class at 5% level of significance. The p-value derived from the chi-squared test is 0.3809.

    Which of the following statements is/are true? Select all that apply.

    (A) There is a positive association between staying in a hall and being late at the sample level.

    (B) There is a negative association between staying in a hall and being late at the sample level.

    (C) Since the p-value is more than 0.05, we can conclude that there is an association between staying in a hall and being late at the population level.

    (D) Since the p-value is more than 0.05, we cannot conclude that there is an association between staying in a hall and being late at the population level.

28. 25 mothers were each allowed to smell two articles of infant's clothing. Each of them was then asked to pick the one which belongs to her infant. They were successful in doing so 72% of the time. You want to show that this has not happened by chance and mothers can indeed recognise the smell of their children. To test such a hypothesis, what should the null and alternative hypotheses be?

    (A) H_0: P(Success)= 0.5; H_1: P(Success)> 0.5.

    (B) H_0: P(Success)= 0.72; H_1: P(Success)> 0.72.

    (C) H_0: P(Success)= 0.5; H_1: P(Success)< 0.5.

    (D) H_0: P(Success)= 0.72; H_1: P(Success)< 0.72.

29. A hypothesis test is done to find out whether vaccine X prevents cancer in a population of dogs, where cancer affects 10% of dogs. A random sample of 100 puppies was selected for the study. All 100 puppies received vaccine X and we observed them for their entire lifetimes. 5 of these puppies evertually had cancer. The null hypothesis is

H_0: Vaccine X has no effect on cancer in the population.

Then the p-value is

(A) the probability that vaccine X is effective.

(B) the probability that vaccine X is not effective.

(C) the probability that the hypothesis will be rejected.

(D) the probability that 5 puppies out of 100 have cancer, given that the probability of cancer is 0.1.

(E) the probability that 5 or less puppies out of 100 have cancer, given that the probability of cancer is 0.1.

30. Mandy and Sue were playing a game - Sue draws a random card from a deck of five cards (red, blue, yellow, green and black) and hides it out of sight from Mandy. Mandy will try to guess the colour of that drawn card. Mandy wins if she can correctly guess the colour of the drawn card. Otherwise, Mandy loses.

They played 5 rounds of the game, and Mandy won 4 out of 5 games. Sue is surprised that Mandy won so many times, and suspects Mandy may have some method to detect the colour of the cards instead of just guessing the colour. Based on the above, she wants to conduct a hypothesis test, with the null hypothesis:

$$P(\text{Mandy guesses a card colour correctly}) = 0.2.$$

What event(s) need to be considered in the calculation of the p-value? Select all that apply.

(A) Event: Mandy losing 5 games out of 5.

(B) Event: Mandy losing any 4 games out of 5.

(C) Event: Mandy losing any 3 games out of 5.

(D) Event: Mandy losing any 2 games out of 5.

(E) Event: Mandy losing any 1 game out of 5.

(F) Event: Mandy losing 0 games out of 5.

Index