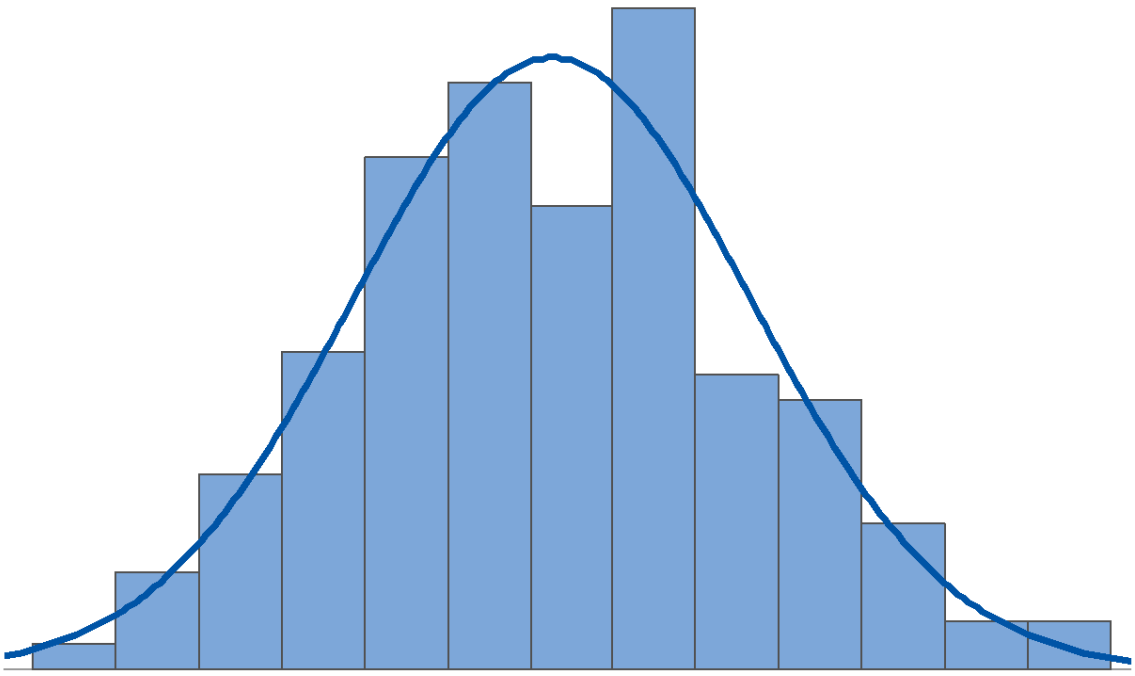
Applied Statistics

FIFTH EDITION



James Reilly

APPLIED STATISTICS

Published by Statistical Solutions

Email address: AppliedStatisticsBook@gmail.com

Caveat:
The datasets in this book have been selected for illustrative purposes and should not be relied upon as a basis for substantive research.

Acknowledgement:
Software output and graphs were produced using Minitab® statistical software.

About the author:
James Reilly is a university lecturer, textbook author, and statistical consultant, and enjoys collaborating with researchers and professionals to gain insight from data.

Table of Contents

6. Making Predictions

7. Designing Experiments

8. Improving Quality

Appendices

1

Exploring Data

Having completed this chapter you will be able to:

– interpret graphs;

– collect data;

– use a calculator in statistical mode.

Numbers tell a story. The story provides information that improves our understanding and leads to better decisions.

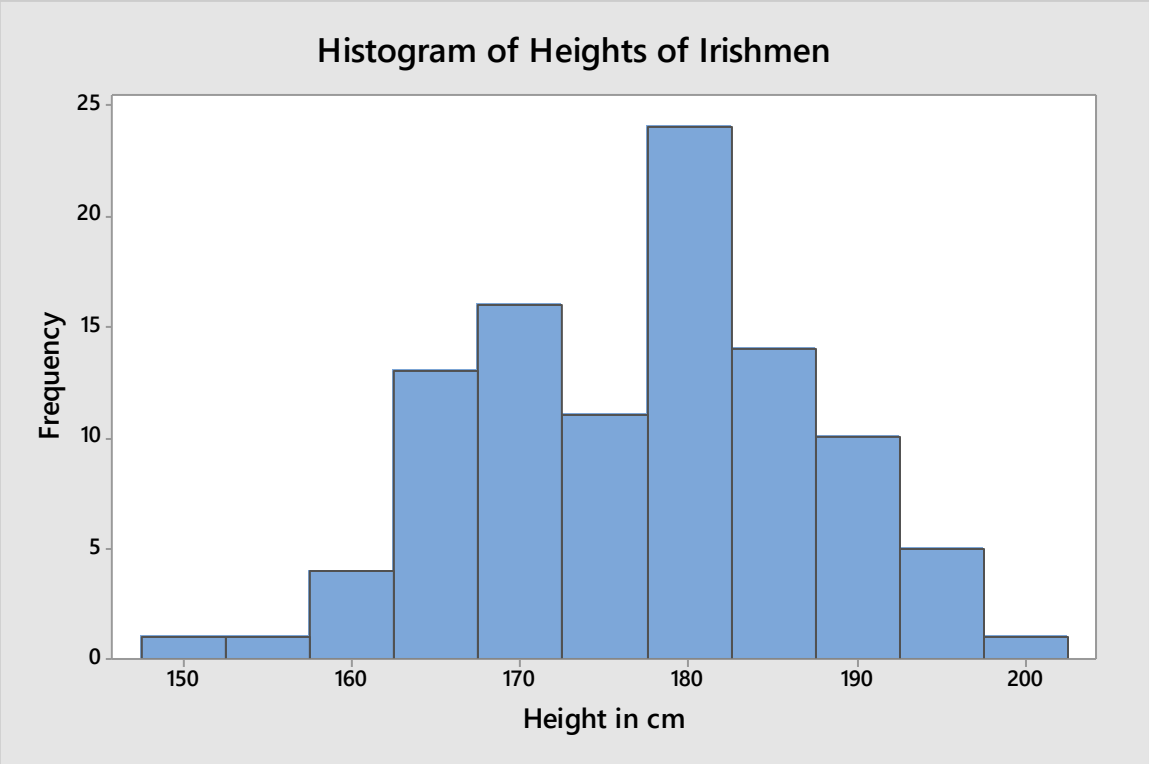
1A. Graphs

Video Lecture https://youtu.be/qBxl-YP5CwY

A graph is a great way to allow a set of data to tell a story. We don't want to get lost in the details but to get an overall impression of what the numbers are saying.

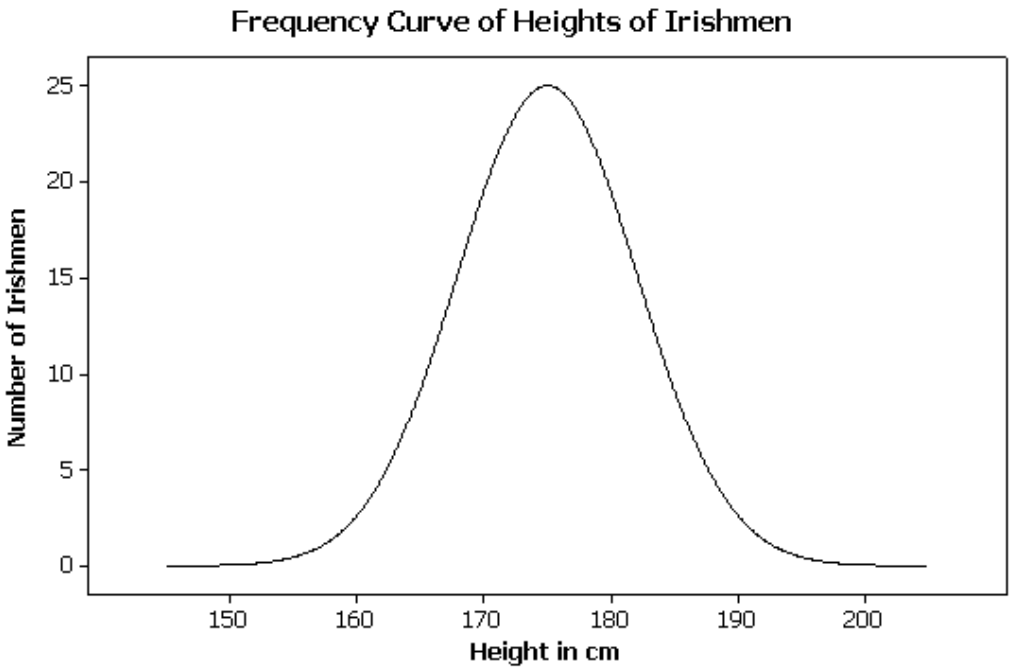
Histograms

Fig 1.1



When reading a histogram, focus on the shape rather than the numbers. Also, ignore small irregularities and pay attention to the more important features. Therefore, when you look at the histogram above you should see the following **frequency curve**. There is no need to draw the frequency curve – you can see it in your imagination.

Fig 1.2



QUESTION What story is this histogram telling? **ANSWER** There is a typical height that is common, and most heights are close to this value. Some individuals are much taller or much shorter than this, but the more extreme values are increasingly rare. We can tell all this by noting the peak in the middle, and the similar tails on either side. This bell-shaped pattern is very common and is referred to as **normal**. Also, notice that we told the story without quoting any numbers. We allowed the numbers to speak.

Fig 1.3

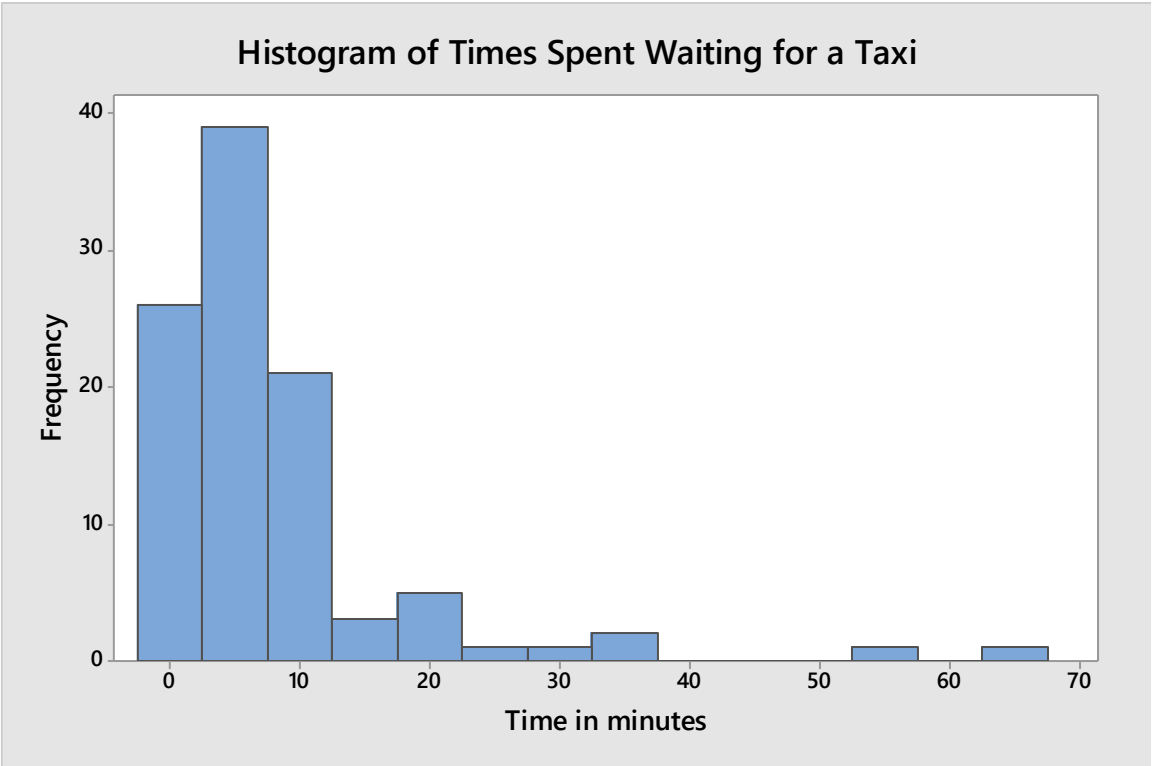
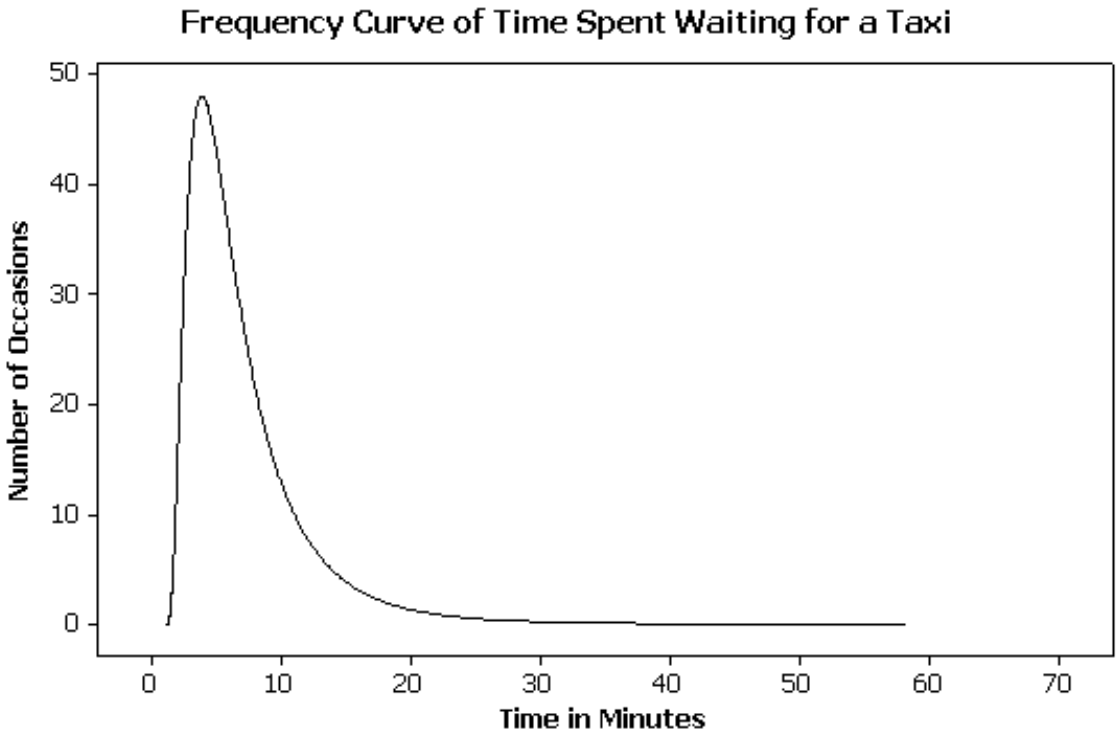


Fig 1.4



This time, the tails are not symmetric. We say that these data are positively **skewed**, or skewed to the right, because there is one long tail on the right. The waiting time for a taxi might be much longer than the typical value, but it could not be much shorter than the typical value, because the waiting time cannot be less than zero.

Fig 1.5

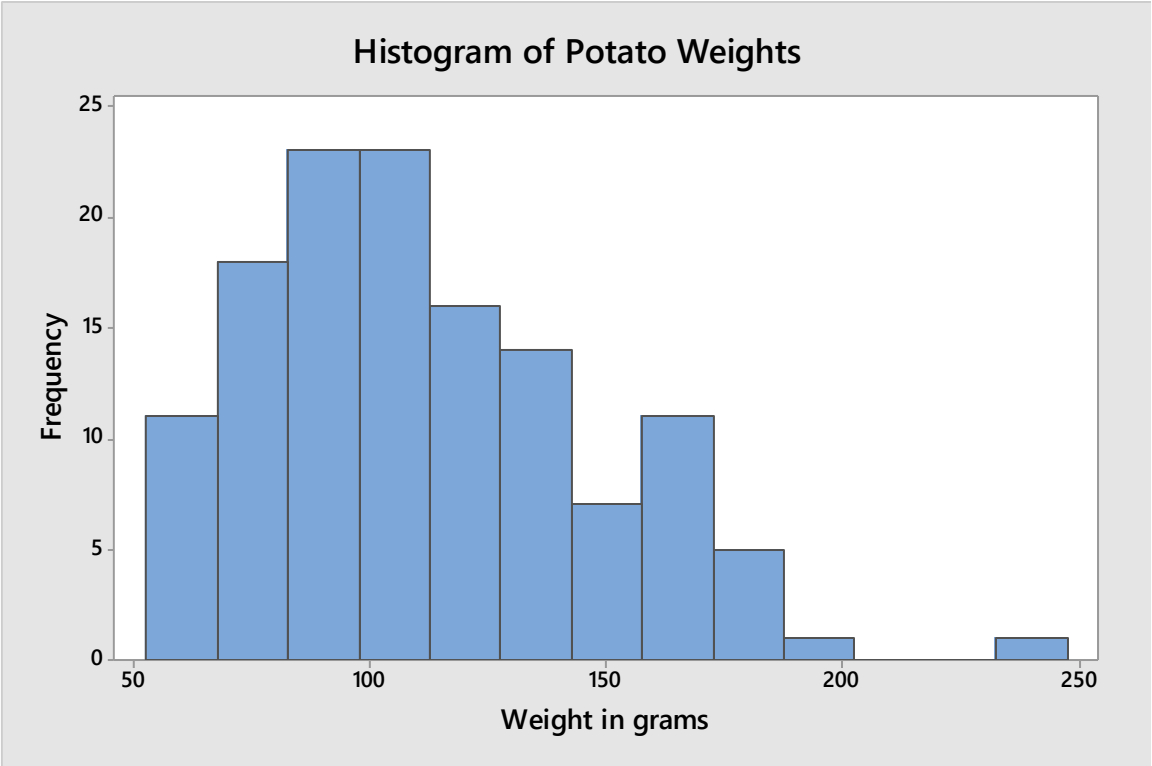
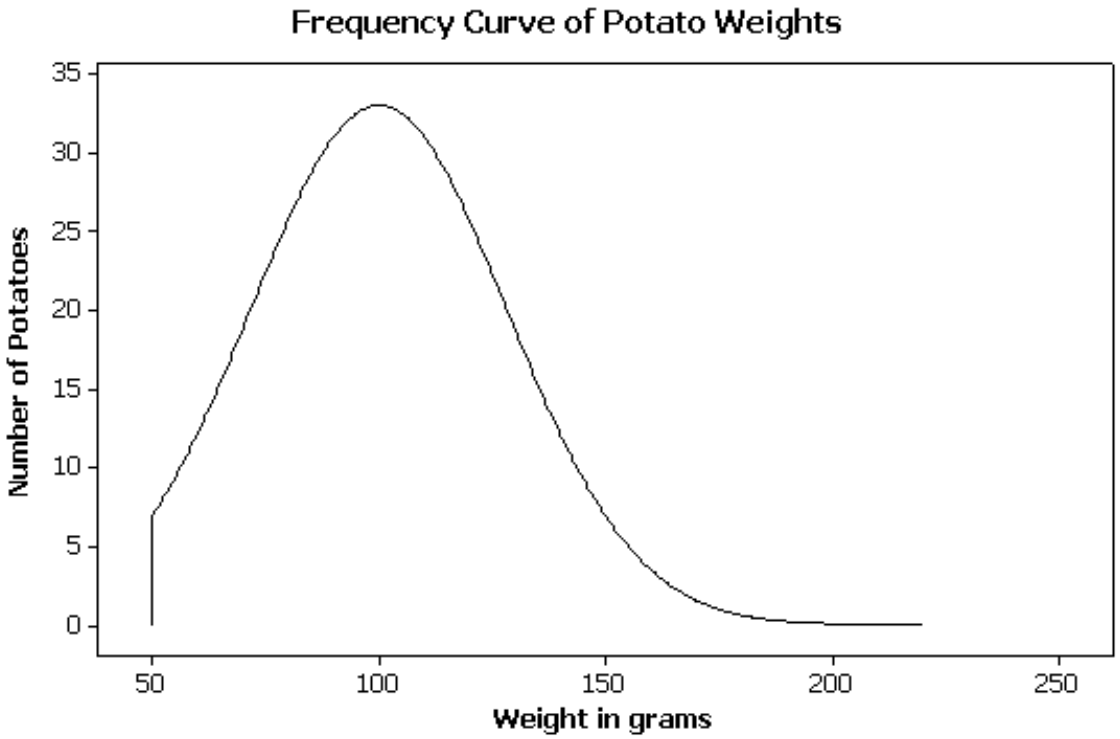


Fig 1.6



Can you tell by looking at the histogram that these potatoes come from a supermarket and not from a farm? The potatoes have been sorted, and any potato weighing less than 50 grams has been removed. We say that the frequency curve is **truncated**. A frequency curve can be truncated on the left, or on the right, or on both sides.

Fig 1.7

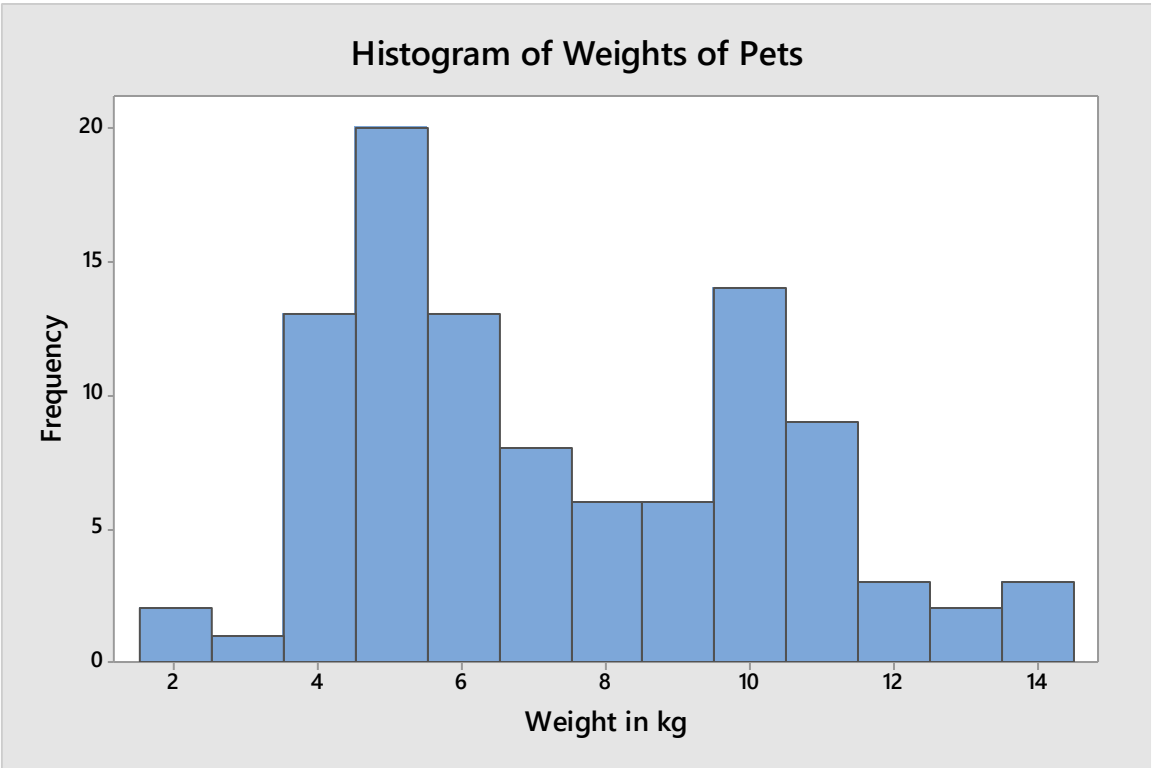
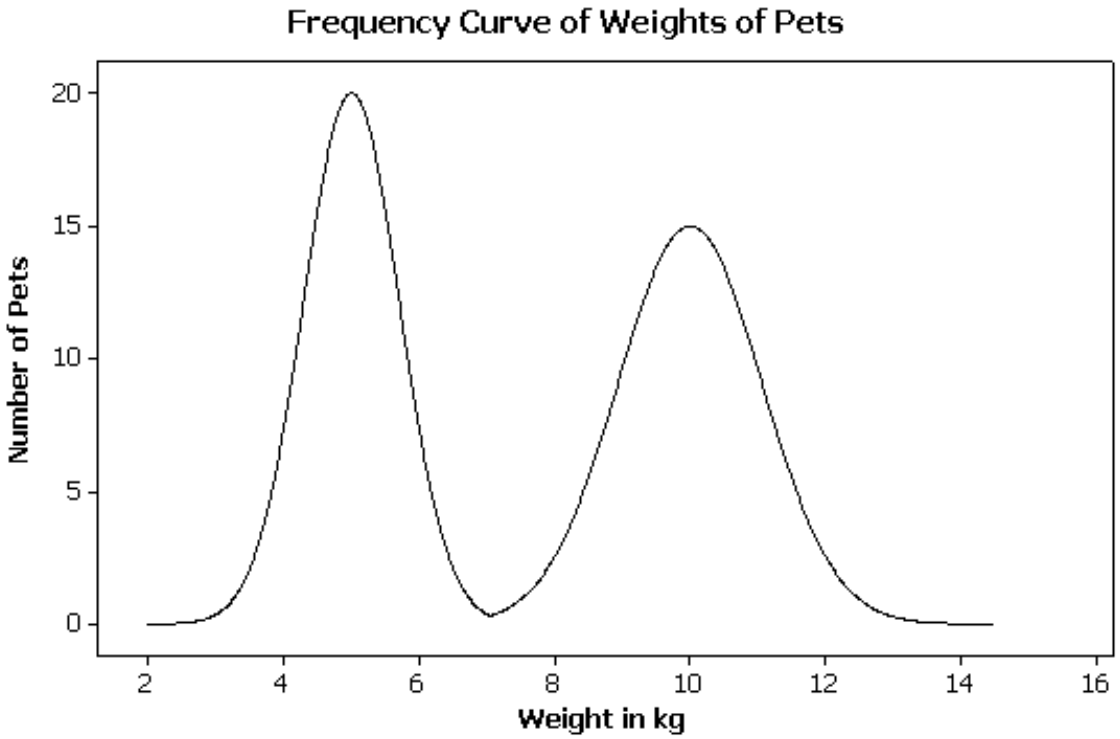


Fig 1.8



Although these animals are all called pets, we can see from the histogram that they are not a homogeneous group. Maybe there is a mixture of cats and dogs. We say that these data are **bimodal**, because instead of having one common value, the **mode**, we have two.

Fig 1.9

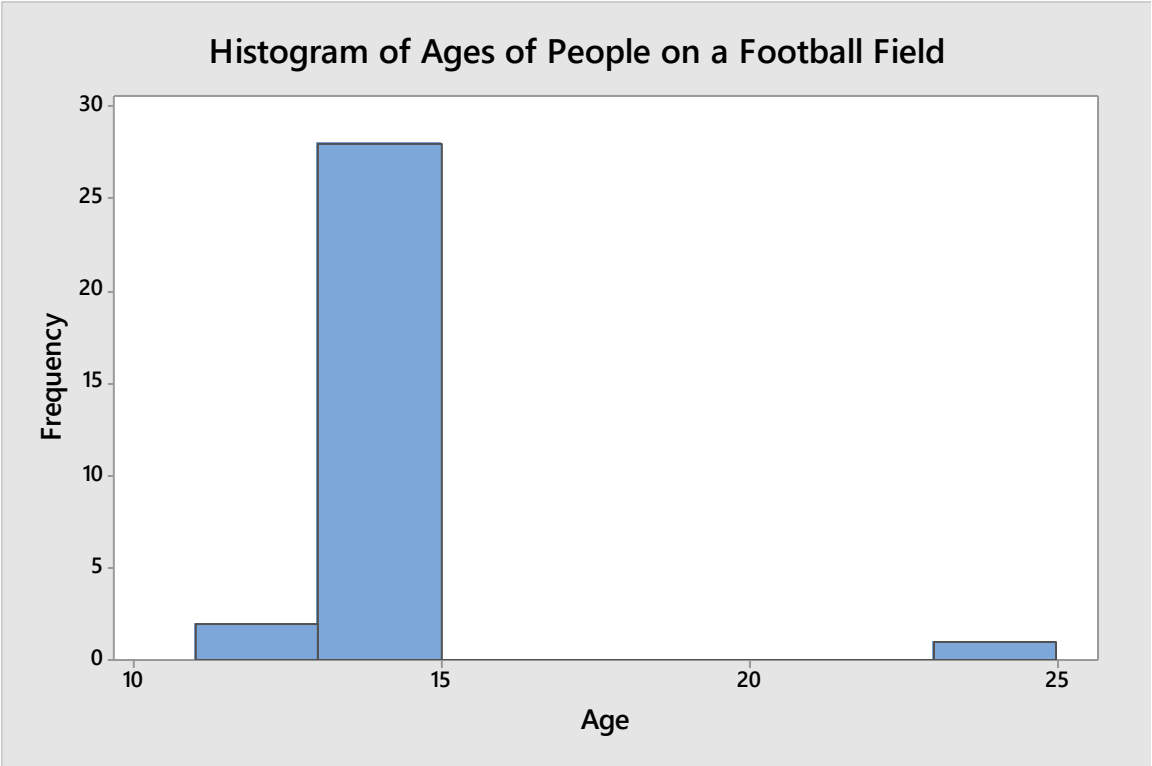
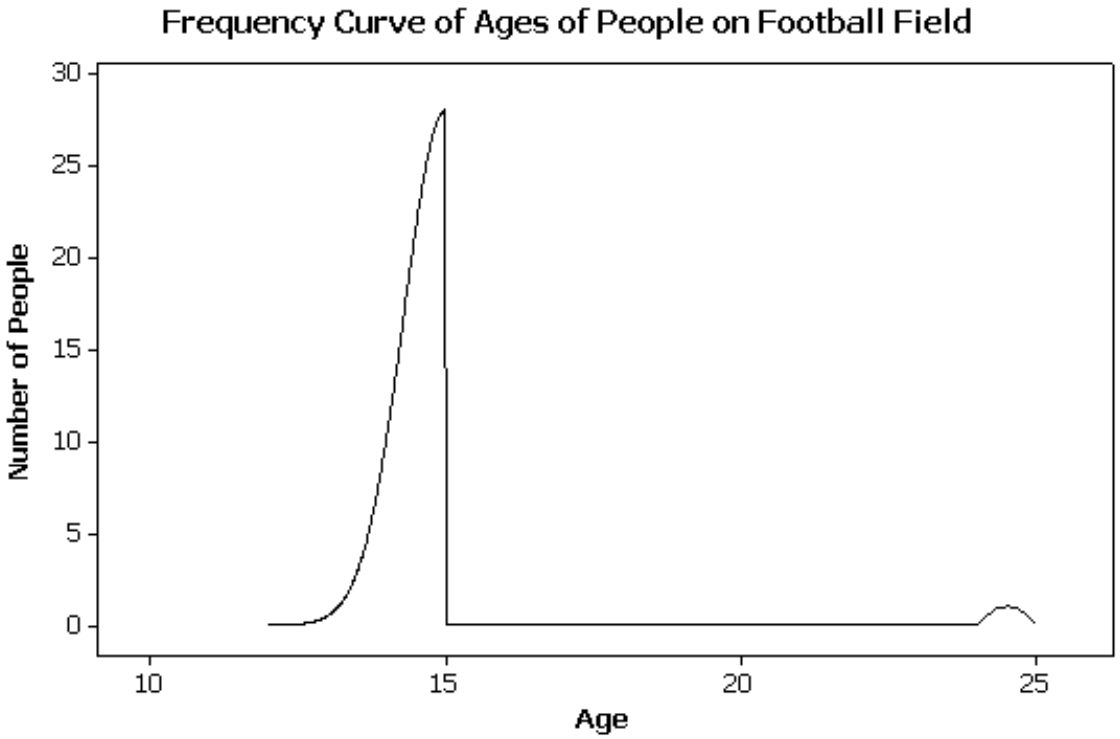


Fig 1.10



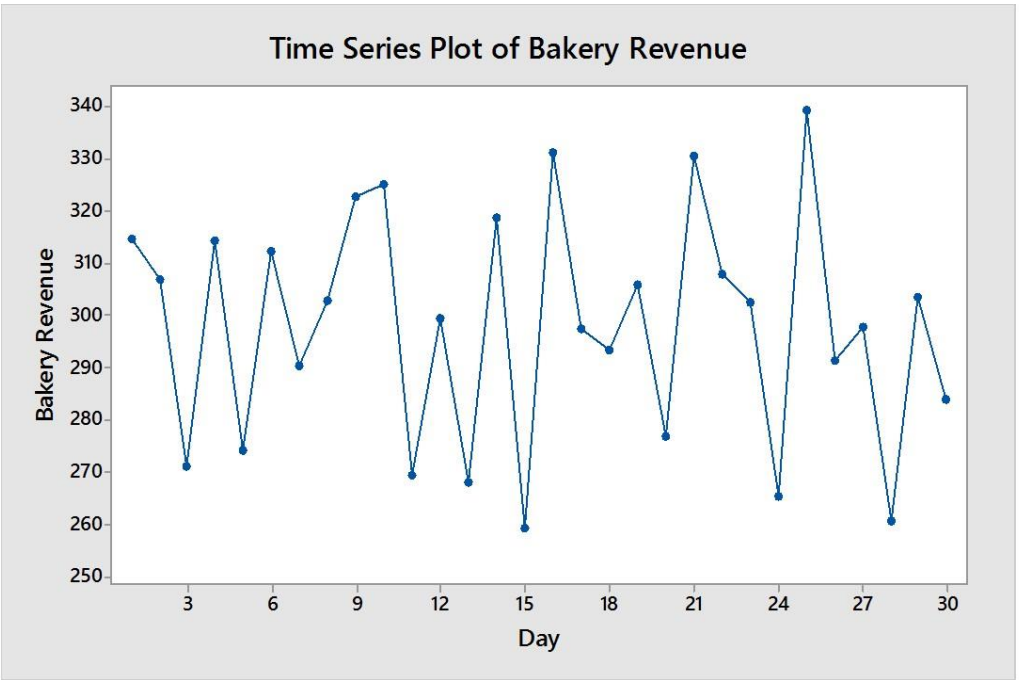
One of these numbers does not seem to belong with the others. It is called an **outlier**. Outliers can represent the most interesting feature of a data set. On the other hand, an outlier may simply be the result of someone making a mistake when typing the numbers. What do you think the outlier represents in the histogram above?

Note that a graph needs to have a sensible title, and to have labels and units on both axes.

Time Series Plots

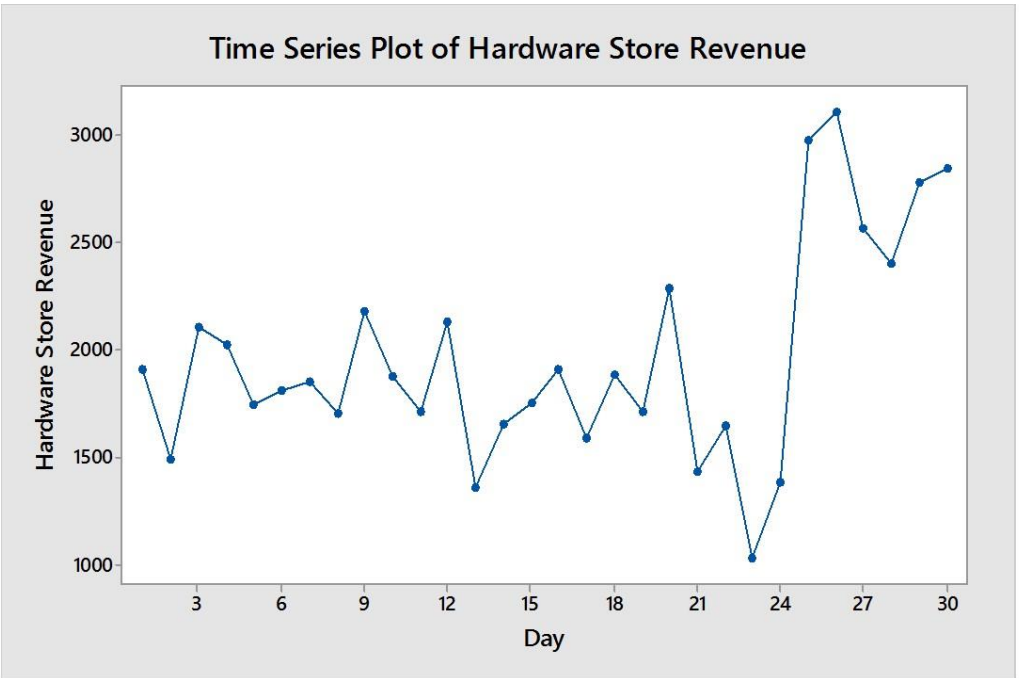
A time series plot is useful for showing how a **variable** behaves over time. Time is represented on the horizontal axis, and the variable in question is represented on the vertical axis. Each observation is plotted as a point, and these points are joined together with line segments. Time series plots of exchange rates, share prices, air temperature, and many other variables, are common. When you scan a time series plot with your eyes, remember that time progresses from left to right, and we wish to see whether the figures remain steady, rise or fall suddenly, rise or fall gradually, follow repetitive cycles, or display some other pattern. Some typical time series patterns are illustrated in the following graphs. These examples are based on the daily revenue of a number of different businesses.

Fig 1.11



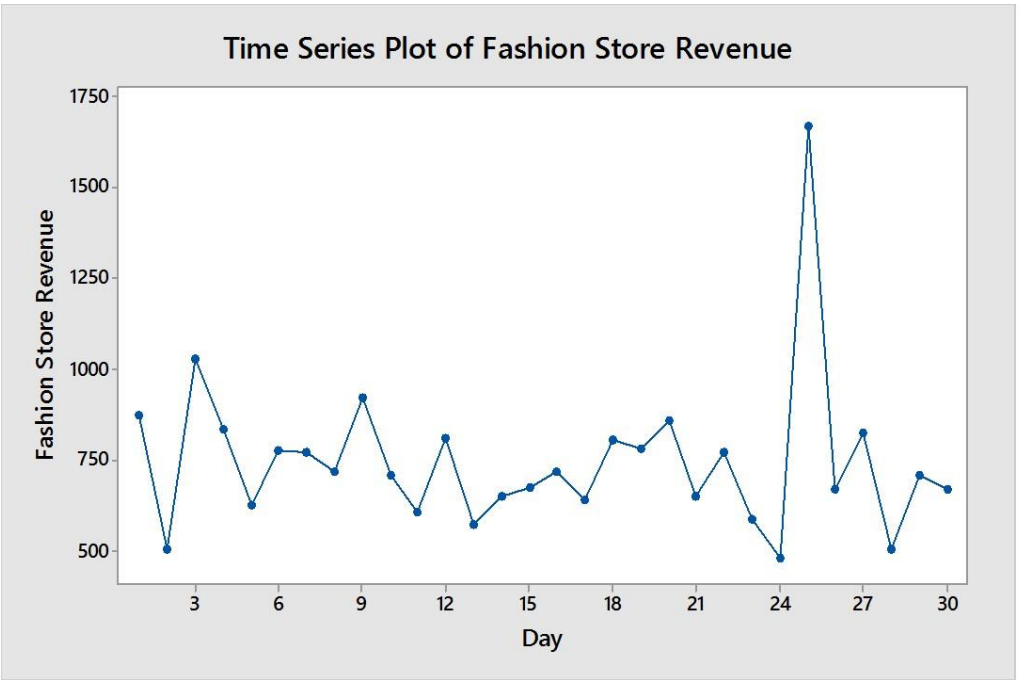
In the first time series plot above, there is a good deal of **random variation**, but there is no obvious pattern. We conclude that the variable remained more or less steady throughout the period. This is the most common kind of time series plot, and it corresponds to a normal frequency curve. If the graph was rotated anticlockwise through 90 degrees, and the points fell onto the variable axis, they would form a bell-shaped mound with most of the points in the middle.

Fig 1.12



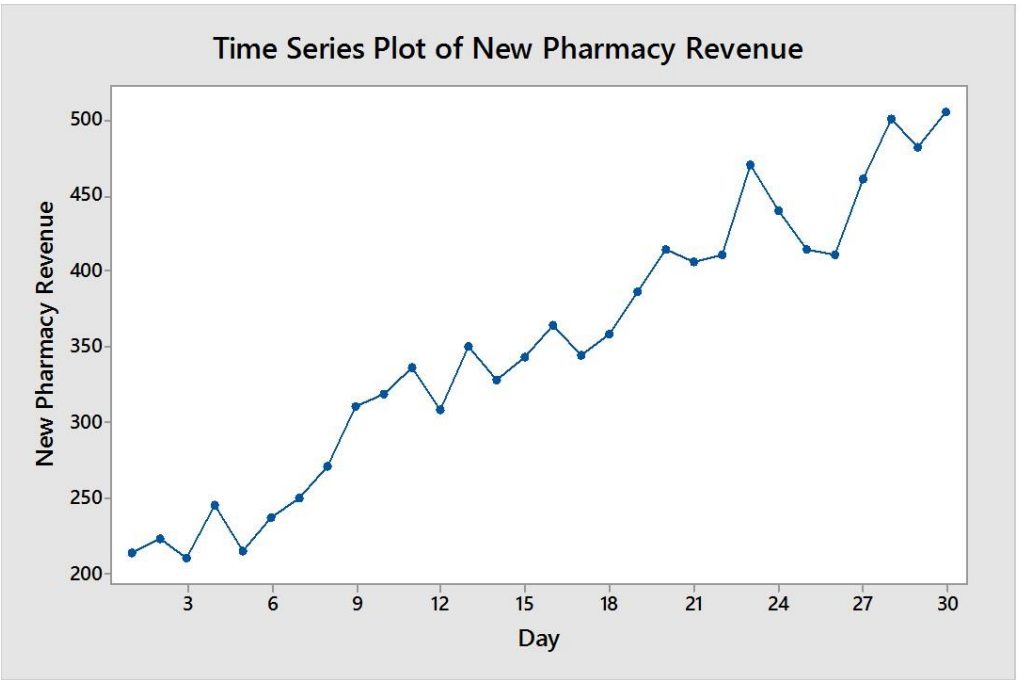
The striking feature of the second time series plot is the **shift**, i.e. the sudden and sustained change in the figures. There must be some explanation. This **explained variation** could be as a result of the hardware store opening a new department.

Fig 1.13



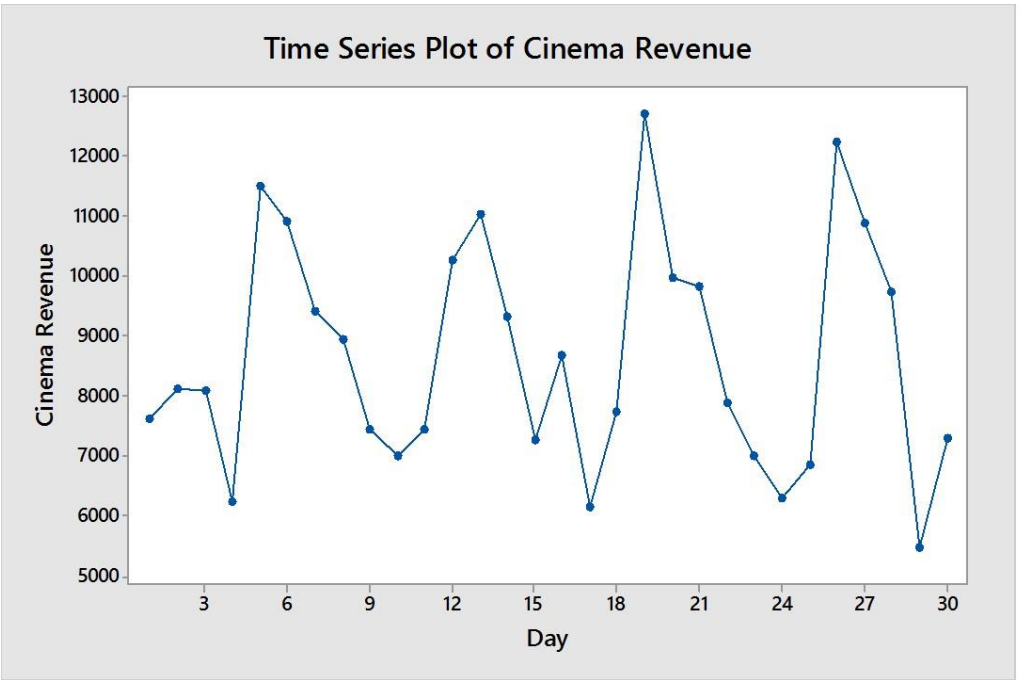
This time series plot shows a **spike**, i.e. there is a sudden change but then the figures return to their former level again. Perhaps there was a one-day sale in the fashion store. This feature corresponds to an outlier in a histogram.

Fig 1.14



A new pharmacy has opened recently. The time series plot of daily revenue shows a **trend**, i.e. a gradual movement in the figures. As the new pharmacy becomes better known, its customer base increases. Because there is some random variation present, this trend will not be obvious until a number of observations have been plotted.

Fig 1.15

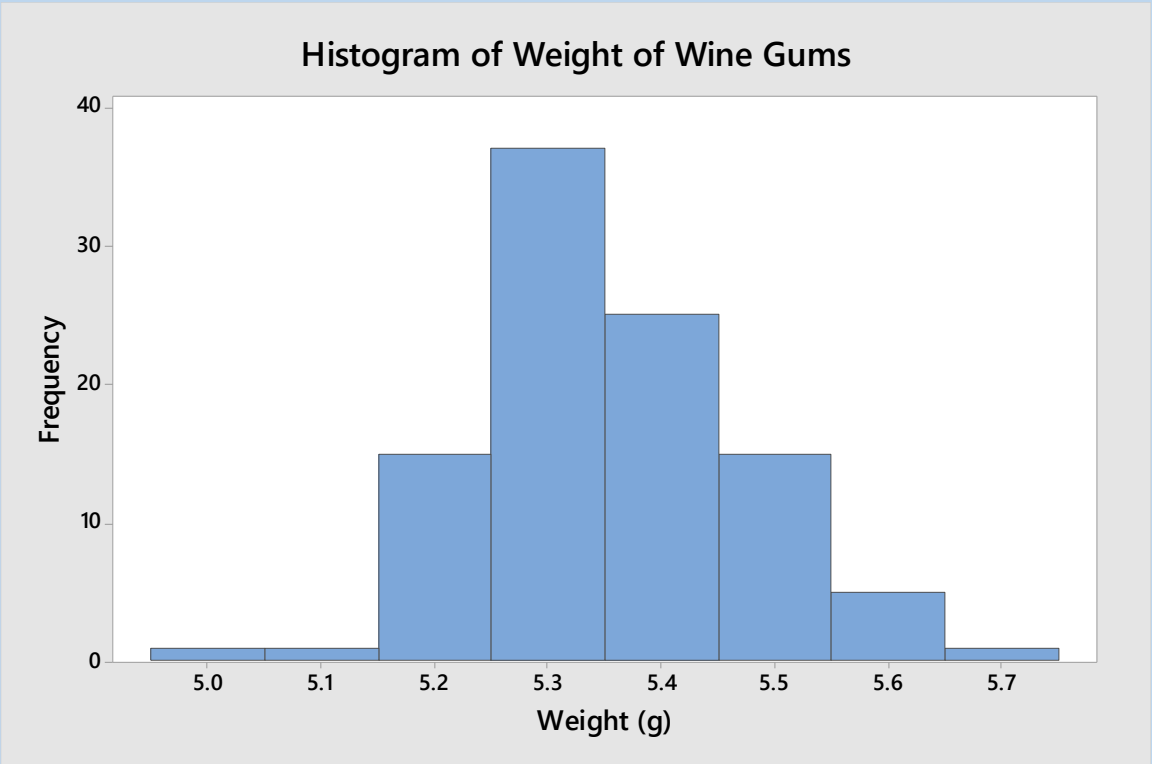


The final plot shows a cyclical pattern. There is a weekly cycle with weekends always busier than midweek. The word **seasonal** is used instead of the word **cyclical** when the cycle is one year long.

Problems 1A

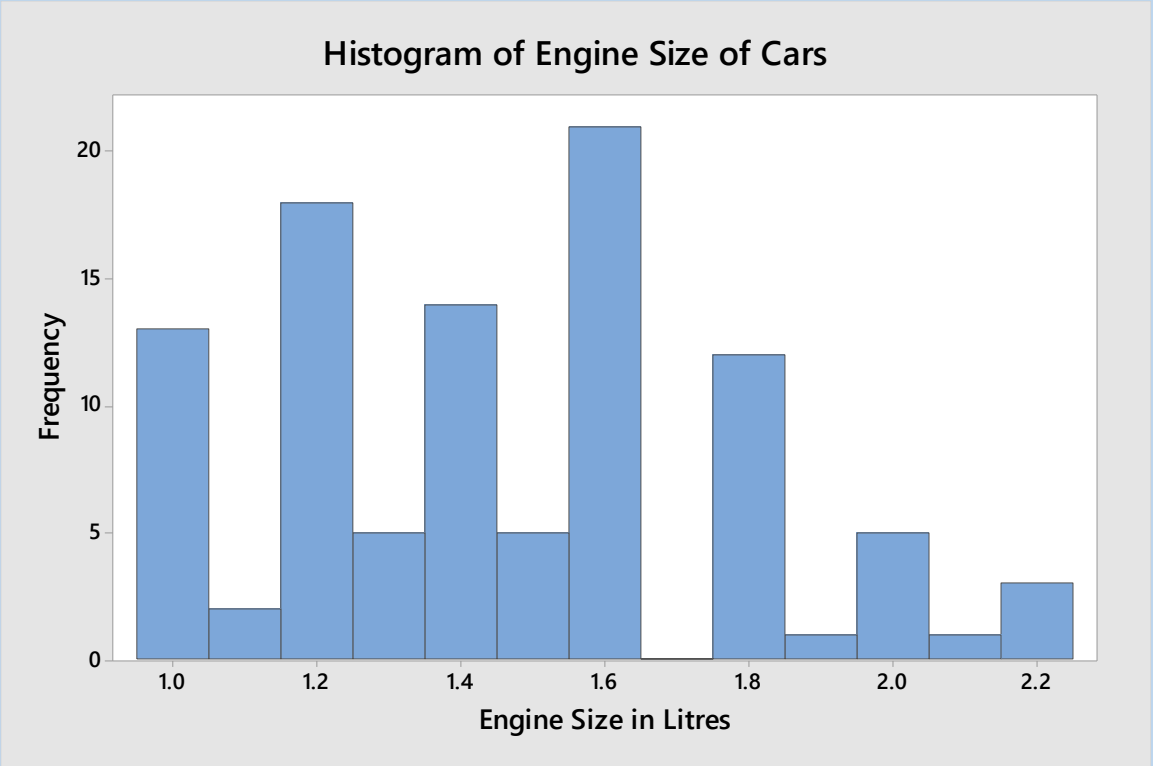
#1. The histogram below shows the weight of wine gums. What does the histogram reveal?

Fig 1.16



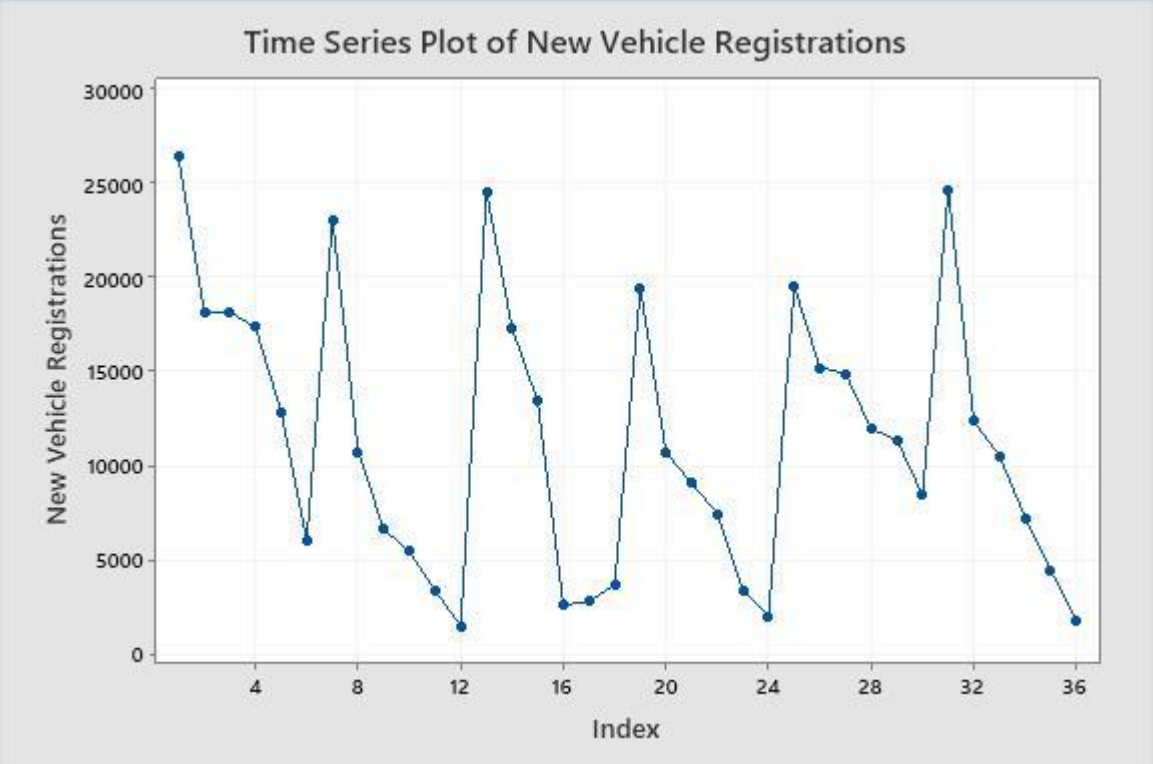
#2. The engine sizes of cars are shown in the histogram below. What does the shape of the histogram tell you about the engine sizes? Also, can you think of any other data set that might display the same peculiar pattern?

Fig 1.17



#3. Here is a time series plot of Irish monthly new vehicle registrations for the 36-month period from January 2019 until December 2021. What story do the figures tell?

Fig 1.18



Project 1A

Time Series Project

Review the last 30 days, finishing with yesterday, and record the daily value of a variable such as:
- the number of messages you received that day in a messaging app, or
- your daily spending on a debit card.

Explore these data as a time series and write a report consisting of the following sections.
(a) Identify the variable of interest.
(b) Show the data.
(c) Draw the time series plot.
(d) Identify any notable features in the shape of the plot.
(e) Express in simple language what these features are telling you about this variable.

1B. Sampling

Video Lecture https://youtu.be/5BbZgctXkZQ

We now come to a very important and surprising concept. The data that we collect and analyse are not the data that matter to us! Usually, we are interested in a much larger data set, one which is simply too big to collect. The **population** is the name given to the complete data set of interest. The population must be clearly defined at the beginning of an investigation, e.g. the shoe-sizes of all the people who live in Galway. Note that a population is not a set of people but a set of numbers, and we will never see most of these numbers.

Sometimes the measurements of interest arise in connection with some **process**, e.g. the dissolution times in water of a new type of tablet. In this case there is no population out there to begin with: we must make some of the tablets and dissolve them in water, or else there is nothing to measure. Also, the process could be repeated any number of times in the future, so the population could be infinitely large.

Once a tablet has been dissolved, it cannot be sold or used again. This is an example of a **destructive measurement**. This is another reason why we cannot measure everything: we would have plenty of data, but no merchandise left to sell.

Usually we are not able take a **census** of every unit in the population. Instead we collect a **sample** from the population and we assume that the other data in the population are similar to the data in the sample. We see only a part, but we imagine what the complete picture looks like. This kind of reasoning is called **inference**, and statistics carried out using this approach is called **inferential statistics**. If we simply describe the observed data, and make no attempt to make statements about a larger population, this is called **descriptive statistics**.

Because we use a sample to provide information about a population, it is essential that the sample is **representative**. The sample may not give perfect information, but it should not be misleading. **Bias** occurs if the sample is selected or used in such a way that it gives a one-sided view of the population.

Sampling Techniques

Random Sampling

A simple random sample is a sample selected in such a way that every unit in the population has an equal chance of being selected.

If the population is finite, this can be done by assigning identification numbers, from 1 to *N*, to all the units in the population, and then using the random number generator on your calculator to select units for the sample.

EXAMPLE Select two letters at random from the alphabet.

Assign 1 to A, 2 to B, and so on, all the way up to Z. The identification numbers can often be assigned by using the order of arrival, or the positions, of the units. Next, generate two random numbers between 1 and 27, i.e. between 1 and *N* + 1. We use *N* + 1 so that the last item has a chance to be selected. To do this, enter 27 on your calculator and press the random number key. Suppose you get 25.258, then simply ignore what comes after the decimal point and select letter number 25, the letter Y. Now press the random number key again. Suppose you get 18.862 this time. Select letter number 18, the letter R.

The same letter could be chosen twice. This is called **sampling with replacement** and it does not cause bias. Repeats can be disallowed if sampling without replacement is preferred.

If a population is infinite, or diffuse, it can be impossible to assign identification numbers. Suppose we wish to interview a random sample of 100 shoppers at a mall. It is not feasible to assign a number to each shopper, and we cannot compel any shopper to be interviewed. We have no choice but to interview willing shoppers 'here and there' in the mall, and hope that the sample is representative. However, when such informal sampling techniques are used, there is always the possibility of sampling bias. For example, shoppers who are in a hurry will be unwilling to be interviewed, but these shoppers would be more likely to answer 'yes' to some questions such as, 'Have you purchased convenience foods today?'.

Multi-Stage Sampling

Suppose you want to draw a sample of thirty players from the premiership. Rather than assigning IDs to all the players in the premiership, it is much simpler to begin by assigning IDs to all the teams in the premiership. The current league standings can be used to assign these IDs. Select a team at random using a random number, say Liverpool. This is the first stage. Now assign IDs to all the Liverpool players. The squad numbers can be used to assign these IDs. Now select a Liverpool player at random. This is the second stage.

Repeat these two stages until the required sample size is achieved. The sample is random, but using multi-stage sampling rather than simple random sampling makes the task easier when the population has a hierarchic structure. The same approach can be used to select items from an online store, by randomly selecting a page and then randomly selecting an item on that page.

Stratified Sampling

Let us suppose that we plan to open a hairdressing business in Kilkenny. We might decide to interview a random sample of adults in the city to investigate their frequency of visits to the hairdresser. If our sample just happened to include a large majority of women, then the results could be misleading. It makes better sense to firstly divide the population of adults into two strata, men and women, and to draw a random sample from each stratum.

Quota Sampling

Quota sampling is similar to stratified sampling but the units are not selected at random. Instead, either **convenience sampling** or **judgement sampling** is used to select units within the different strata. With convenience sampling, units are selected that are easy to reach, such as people who are nearby, but a convenience sample might not be representative of the population. With judgement sampling, units are selected that are considered to be representative of the population, and so the quality of the sample depends on the quality of the researcher's judgment.

Cluster Sampling

It is tempting to select naturally occurring clusters of units, rather than the units themselves. For example, we might wish to study the tread on the tyres of cars passing along a certain road. If we randomly select 25 cars and check all four tyres on each car, this is not equivalent to a random sample of 100 tyres. All the tyres on the same car may have similar tread. The selected units are random, but not independently random. If the units of interest are tyres, then select tyres, not cars.

Systematic Sampling

With this approach, units are selected at regular intervals. For example, every tenth box from a packing operation is inspected for packing defects. But suppose there are two packers, Emie and Grace, who take turns packing the boxes. Then, if the first box inspected (box 1) is Emie's, the next box inspected (box 11) will be Emie's, and so on. Grace's boxes are never inspected! All of Grace's boxes might be incorrectly packed, but the sample indicates that everything is fine. The problem here is that the sampling interval can easily correspond to some cycle of variation in the population itself.

In summary, only random samples can be relied upon to be free from bias. Random samples avoid biases due to progressive or cyclical variation in the population, or due to personal biases, whether conscious or unconscious. Data collection is crucial, and it must be done carefully and honestly. All the formulae in this book are based on the assumption that random sampling has been used. If samples are drawn by any technique that is not strictly random, then keep your eyes open for any potential sources of bias.

Types of Data

Data come in different forms but the two most important forms are **measurements** and **attributes**.

A measurement is a number. An example of a measurement is the duration of a movie. If you ask your friends "How many minutes long was the last movie that you watched?"

every one of the answers will be a number e.g. 95. We use a **mean** to give an idea of the size of the numbers.

An attribute is something that is either present or absent, yes or no. An example of an attribute is whether someone ate popcorn while watching a movie. The answer to the question "Did you eat popcorn while watching the movie?" must be either yes or no, e.g. "yes". We use a **proportion** to give an idea of how frequently an attribute occurs.

A measurement or an attribute can be referred to as a **response**, and each value of the response that is sampled is called an **observation**.

Big Data

Very large data-sets such as those that arise in connection with social media apps or customer loyalty cards are referred to as **big data**. Big data are characterised by their volume (there are a lot of data), variety (there are different types of data such as measurements, attributes, dates and times, pictures and video) and velocity (the data change quickly).

Structuring Data

When entering data into a table or a spreadsheet, enter the values of a variable in a vertical column like this, with a title in the first cell.

Table 1.1

Height
162
179
176
170
173
176

If you have a number of variables, use a column for every variable and a row for every case, like this.

Table 1.2

Height	Red Cards	Position	Outfield
162	1	Defender	1
179	0	Goalkeeper	0
176	0	Attacker	1
170	0	Goalkeeper	0
173	2	Defender	1
176	0	Attacker	1

It can be convenient to convert attribute data into numbers by means of an **indicator variable** that designates 'yes' as 1 and 'no' as 0, as in the final column.

Problems 1B

#1. Can you identify one sampling bias in each of the following sampling schemes?

(a) In order to estimate the mean number of days that outpatients wait for a hospital appointment, a radio chat show host invited listeners who are outpatients to contact the show and indicate how long they had been waiting.

(b) In order to estimate the mean number of children per family in a particular town, a researcher called to a school in that town and asked a number of the students how many children are in their family.

(c) In order to estimate the proportion of all train passengers who are satisfied with the current train timetable, a railway employee asked the first 50 passengers who boarded a train to indicate whether or not they are satisfied with the current timetable.

(d) In order to estimate the proportion of voters who would support a particular candidate, a journalist called to a number of homes one afternoon and asked the homeowners whether they intended to vote for that particular candidate.

Project 1B

Sampling Project

Select a random sample of 100 measurements from any large population of your choice, and write a report consisting of the following sections.

(a) Identify the population of interest and the measurement of interest.

(b) Describe in detail how the sample was selected from the population.

(c) Draw a histogram.

(d) Identify the important features of the shape of the histogram, and explain what these features tell you about the population.

1C. Summary Statistics

Measures of Location

Video Lecture https://youtu.be/OoLPzGBmDsY

In a conversation or in a written message, long lists of numbers are not helpful. It is better to mention a single number. A single number is easier to remember, easier to base a decision upon, and easier to compare with some other situation. For these reasons, we now consider summary statistics. A **statistic** is any quantity which is calculated from sample data, such as the minimum, the mean, etc. A statistic that summarises the information contained in the sample is called a summary statistic.

EXAMPLE The heights in metres of a random sample of trees in a forest were:

22, 28, 25, 22, 23.

The **mean** of this sample is 24. This is an **average** or a **measure of location**, because it tells where the numbers are located on the number line. It tells us typically how big the trees are. The mean is the sum of the numbers divided by the number of numbers. It is the most popular measure of location because it takes all the sample data into account, and it is well suited to data that are roughly equal. The sample mean is called ***X*bar** and its formula is shown below.

Formula 1.1

Sample Mean

$$\overline{X} = \frac{\sum X}{n}$$

There are many other summary measures of location, and we describe two of them here: the mode and the median. The **mode** is the most commonly occurring value and it is useful if the data are nearly all the same. To the question 'How many legs has a dog?' the best answer is 'four', which is the mode. Some dogs have three legs, so the mean number of legs is probably about 3.97, but this is not a useful answer.

The **median** is the middle number when the numbers are arranged in order. It is a **robust** measure of location that is insensitive to outliers. The median of the sample of tree heights above is 23, and it would still be 23 if the largest number, 28, were changed to 48, or to any large unknown number.

Although we can calculate the value of the sample mean, what we really want to know is the value of the population mean (symbol μ, pronounced 'mu'). The sample mean is merely an **estimator**. In general, sample statistics estimate population **parameters**.

Measures of Dispersion

We have been considering the heights of trees in a forest and we have estimated that the population mean is 24 metres, based on a random sample of five trees. Does this mean that all the trees are exactly 24 metres tall? Of course not! Some are taller and some are shorter. If we built a wall 24 metres tall, in front of the forest, some of the trees would extend above the top of wall, and others would stop short of it. There is a difference (or **error**, or **deviation**) between the height of each tree and the height of the wall. The average difference (i.e. the typical error, the **standard deviation**) is a useful summary measure of **dispersion** or **spread** or **variability**. The standard deviation is the root-mean-square deviation and it is calculated as follows.

Data: 22, 28, 25, 22, 23

Deviations: -2, +4, +1, -2, -1

Squared deviations: 4, 16, 1, 4, 1

Mean: (4 + 16 + 1 + 4 + 1) ÷ 4 = 6.5

Root: √6.5 = 2.5495

This is the sample standard deviation estimate, denoted by the letter ***S***. It estimates the population standard deviation, which is denoted by ***σ***, pronounced 'sigma'.

Formula 1.2

Sample Standard Deviation

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Note that the standard deviation is the typical error, not the maximum error. Also, the word 'error' does not refer to a mistake: it just means that the trees are not all the same height.

Did you notice that when calculating the mean of the squared deviations, we divided by 4 rather than 5? This quantity, ***n*-1**, is called the **degrees of freedom**, and is defined as the number of independent comparisons available. Five trees provide four independent comparisons. One tree would tell us nothing at all about variability. Dividing by *n*-1 in this way ensures that *S* is an unbiased estimator of σ.

The standard deviation squared is called the **variance**. In the example above, the variance is 6.5 square metres. The variance is a useful measure of dispersion because, when variation arises from a number of independent sources (for example, trees of different species and trees of different ages) the total variance can be found by adding the individual variances together. The symbol for the population variance is σ^2, and S^2 denotes its sample estimate. The standard deviation is more intuitive and is easier to talk about, but the variance is easier to use when performing calculations.

The sample mean and standard deviation can be easily found using a calculator in statistical mode. Follow these steps:

1. Choose statistical mode.
2. Type the first number and then press the data entry key.
3. Repeat the above step for every number.
4. Press the relevant keys to see the mean.
5. Press the relevant keys to see the standard deviation.
6. When you are finished, clear the numbers from the memory.

In some situations it is convenient to express the standard deviation as a percentage of the mean, rather than expressing it in the original units of measurement. This allows for the variation of different measurements to be compared. This is called the **Coefficient of Variation** (*CV*) or the **Relative Standard Deviation** (*RSD*). For the trees data, *CV* = 2.5495 ÷ 24 × 100 = 10.62%.

The **range** is another measure of dispersion. The range is the difference between the largest and smallest values in a data set. Unfortunately the sample range is a biased estimator of the population range. It has other weaknesses too: it is sample size dependent, and sensitive to outliers.

Measuring variation is important because in statistics we study the differences between numbers in order to gain insight. We ask: Why are the numbers different? In the case of the weights of pets, the numbers are different because some of the pets are dogs and some are cats. This is **explained variation** because we have identified a **source of variation**. But that is only part of the story. If we separate the dogs and cats, and just observe the cats, there will still be differences among their weights. This is called **unexplained variation**. Even though we can't explain this variation, we can measure it (using the standard deviation) and we can describe its behaviour (by identifying if it is normally distributed or if it follows some other pattern).

There are many processes where the aim is to achieve a certain targeted number. For example, when filling pasta into bags for sale, the aim is to achieve the nominal weight of 500g in every bag. A process is said to be **accurate** (or **unbiased**) if the process mean is on target. A process is said to be **precise** if the standard deviation is small, i.e. the results are close together.

Problems 1C

Video Tutorial: https://youtu.be/LKdOZvEb_T0

#1. Estimate the mean and the standard deviation of each of these populations, based on the sample data provided.

(a) Points awarded to a gymnast: 7, 8, 9.

(b) Shoe sizes of tram passengers: 8, 11, 2, 9, 6, 12.

#2. First guess the mean and standard deviation of the number of letters per word in this sentence, and then use your calculator to see how you did.

#3. A student selected a random sample of their friends on a social media platform and recorded the age of each person, with the following results:

19, 19, 20, 22, 20, 19, 78

(a) Calculate the mean, the mode and the median.

(b) Which of these measures of location is not a useful summary statistic to use in this situation? Explain your answer.

(c) The person whose age is recorded as 78 may actually be older than this. How does this affect the values of each of the summary statistics that you have calculated?

#4. Three vertical lines are drawn on a frequency curve, one each at the mean, the median and the mode. Identify which of (a), (b) and (c) below corresponds to each measure of location.

(a) This line divides the area under the frequency curve in half.

(b) This line touches the tallest point on the frequency curve.

(c) This line passes through the centre of gravity of the area under the frequency curve.

Project 1C

Summary Statistics

Select a random sample of 10 measurements from any large population of your choice, and write a report consisting of the following sections.

(a) Identify the population of interest and the measurement of interest.

(b) Describe in detail how the sample was selected from the population.

(c) Use your calculator to find the sample mean.

(d) Use your calculator to find the sample standard deviation.

(e) Explain in words what the sample mean and the sample standard deviation tell you about the population.

1D. Surveys

Video Lecture https://youtu.be/6UrVPF4J1cg

First Steps

The first step in any research activity is to clarify your **research objective**. Identify the population that is of interest to you and the questions that you would like to answer. Articulate your objective clearly in words at the beginning of your research project and this will guide you at every later step. If you do not clarify your research objective in this way, then you are likely to lose focus, to gather too much data, to gather the wrong kind of data, and to be unable to provide clear conclusions and actionable insight at the end of your research project. You may encounter lots of interesting information and opportunities along the way that are not central to your research objective. These can be simply noted for possible future reference and then must be left aside in order to pursue your research objective.

The next step is to find out what information is already available that sheds light on your research objective. Much of the information that you require may have already been collected by someone else and made available online or published in a book or journal. Data that has been collected by someone else is called **secondary data**. Therefore you need to carry out a **literature review** that explores these existing sources, and then summarise the information you find that is relevant to your area of research.

You are now ready to collect **primary data** which is gathered to specifically address your research objective. Before conducting a survey to gather primary data, ask yourself if it makes better sense to hire a market research company to conduct the survey for you. As well as getting a professional service, the cost may actually reduce because the company can carry out an **omnibus survey** which serves a number of clients at once by including all of their different questions on the same questionnaire so that the costs of data collection are shared among all the clients. If you decide to collect your own primary data then you can choose either an online survey tool or a paper based survey.

Questionnaire Design

The research objective will guide the selection of topics to be included and questions to be asked in the survey. Don't ask questions if you don't need to know the answers. Sometimes an enthusiastic researcher will ask respondents lots of questions but has no plan for how to use the answers that are collected. If you're not going to use data then don't collect it. Including superfluous questions means that the questionnaire will take longer to complete and therefore respondents will be less willing to participate in the survey or may drop out before completing it. Identify the questions that are required to provide insight on the research objective and limit yourself to these questions.

Questions should have high validity and high reliability. **Validity** means that the responses provide an accurate measure of what they set out to measure. For example, if the purpose of the survey is to assess the satisfaction of guests with their visit to a hotel then all important aspects of their visit should be addressed, and a survey which does not address particular aspects such as the ease of check-in or the standard of food would not be valid. **Reliability** means that the responses are consistent over time so that, if the same respondent was asked the same questions again a short time later, the responses would be similar.

Don't ask questions if you can find the answers by observation. If you are gathering facts rather than opinions then don't ask people what they think the answer is. For example, to find out how long it takes someone to walk from a hotel reception desk to a particular bedroom, don't ask someone how long they think it takes. Instead use a stopwatch to record the actual time.

Don't ask leading questions. A leading question is one that influences the respondent to give the answer that the interviewer prefers, such as, "Wouldn't you agree that the bar menu offers a good variety of snacks?". Leading questions have their place in debates or discussions but do not belong in a research study. Respondents can be influenced by a variety of personal, social and cultural factors such as wanting to look good or wanting to please the interviewer.

Don't ask questions that are confusing or ambiguous. Avoid obscure words and complicated sentences. State the question directly without negatives or double negatives. Be aware that many different people will have to engage with the same question, so try to look at the question from every possible angle. For example, a question about guest satisfaction with a hotel breakfast buffet will be encountered by some respondents who don't eat breakfast at all, some who have breakfast in their bedroom, some who bring small children to breakfast, and so on. Ask one or more colleagues to proof-read your survey questions for ambiguity. Then test out your survey questions in a **pilot survey** to see how a number of respondents engage with the questions before launching the full survey. Avoid presenting irrelevant questions to individual respondents by using a **skip pattern**, e.g. "If you answered 'no' to the previous question then go directly to question number 12."

Don't ask questions that directly invite an opinion, such as, "Is the foyer pleasantly decorated?", or questions that emphasize personal opinion, such as, "Do you think that the foyer is pleasantly decorated?". Instead invite the respondent to express

agreement or disagreement, by asking the question like this, "Do you agree that the foyer is pleasantly decorated?".

For questions that invite agreement or disagreement, it is common to use five categories, such as: strongly disagree, disagree, neither agree nor disagree, agree, strongly agree. **Likert scales** invite respondents to indicate their level of agreement or disagreement with a number of statements and the scores are then added together to provide a measure of the respondent's attitude. For example, a series of twenty Likert items would have a maximum score of 100 if the respondent gave the most positive answer to each item. Instead of asking respondents to rate their agreement, they can be asked to rate their satisfaction (very satisfied, somewhat satisfied, neither satisfied nor dissatisfied, somewhat dissatisfied, very dissatisfied), or to rate the likelihood that they would recommend a product to a friend. Seven categories can be used instead of five categories, by including 'completely satisfied' and 'completely dissatisfied' as the extreme options: but respondents may get hung up on what 'completely satisfied' actually means. Also the neutral middle category, 'neither satisfied nor dissatisfied', can be omitted to force the respondent to express either a positive or negative view, and this improves both validity and reliability.

Ask questions about the recent past. For example, instead of asking, "How many times a week do you usually go to the gym?" ask, "How many times have you been to the gym in the last week?". This question is likely to lead to an exact answer, but the word "usually" is open to interpretation and requires long-term memory and some mental arithmetic. And, of course, answers about past actions such as, "I made use of the gym", are more reliable than answers about future intentions, such as, "I will make use of the gym".

Begin with questions that engage the interest of the respondent in order to improve completion rates. Questions should be arranged in a sequence that feels natural to the respondent, moving from general questions about a particular topic to more specific questions on that topic. This is called **funnelling**. Questions about the respondent's personal and demographic details should be left until the end, unless it is necessary to screen the respondents at the beginning to determine if they should be accepted into the survey sample. Sensitive questions should be delayed until near the end of the survey after interest has been established. Emphasising that the responses are anonymous, or explaining the reason for the question, can help improve response rates to sensitive questions. For an even greater degree of anonymity, sensitive questions can be asked in a way that makes it impossible to identify which individuals gave a "yes" answer. For example, instead of asking the question, "Did you arrive late for the business meeting?" the question can be constructed as follows. "Flip a coin once. If the coin shows heads then please answer question (a). If the coin shows tails then please answer question (b). Question (a): "Did you arrive late for the business meeting?" Question (b): "Flip the coin a second time: does the coin show heads on the second flip?". The answer to either question (a) or question (b) is recorded in the one space available on the form. The proportion of genuine yes answers to question (a) can be found by adjusting for the known proportion of yes answers to question (b).

Response Scales

Use actual numbers where possible. If you want to know how long respondents spent travelling to a hotel then record the answer as a number, e.g. for the first three respondents the answers might be 28 minutes, 21 minutes, and 14 minutes. This approach provides richer information than checking a box for a crude category such as "10 minutes and less than 30 minutes". In this example the first respondent spent twice as long travelling as did the third respondent. Numbers like these that can be divided to provide a meaningful ratio ('twice as long') are said to be on a **ratio scale**. Most quantities that we measure, such as price, duration, length, weight, and so on are on a ratio scale. But if these same numbers (28, 21 and 14) were the air temperatures in degrees centigrade in three different rooms, then it would not be true to say that the first room was twice as hot as the third room. This is because zero degrees centigrade is not a true zero since there are temperatures lower than this. However it would be true to say that the fall in temperature from the first room to the second room is the same as the fall in temperature from the second room to the third room. Numbers like these that can be subtracted to provide a meaningful difference are said to be on an **interval scale**.

Often in surveys, and especially in market research surveys, we are interested in finding out the respondent's order of preference for different items. The items can be ranked using the numbers 1, 2, 3, etc. This is quite an easy task for the respondent because they begin by choosing their favourite from the options provided and then continue by choosing their favourite from the remaining options and so on. The item ranked first might be a huge favourite with the respondent, whereas they might only have a mild preference for item two over item three. All we can be sure about is that the numbers are in the correct order and for this reason numbers like these are said to be on an **ordinal scale**.

Numbers are sometimes used to identify categories that have no numerical values, such as ice-cream flavours that might be coded 1 for vanilla, 2 for chocolate, and 3 for strawberry. These numbers are being used in place of names and so numbers like these are said to be on a **nominal scale**.

A nominal scale with only two categories (yes and no) is called a **binary scale**. It can be convenient to use an **indicator variable** to code the 'yes' and 'no' answers as 1 and 0 respectively.

An answer scale that invites respondents to 'tick all that apply' is actually a series of binary questions, since each item on the list is either ticked for 'yes' or left blank for 'no'.

Survey results can be recorded in a spreadsheet in which every column represents a question and every row represents a respondent, as shown in Table 1.2.

Words and Numbers

Open-ended questions allow the respondent to volunteer any information that may be relevant. Answers consist of words rather than numbers, and can be useful for finding out about issues that matter to the respondent but which are as yet unknown to the

survey designer. Other sources of textual data include transcripts of interviews and focus groups, social media content, documents and websites. One way to begin exploring a large quantity of textual data is by using a **tag cloud**. This consists of a graphic that displays the most frequently occurring words, with larger font sizes for the words with higher frequencies. A tag cloud based on a university website is shown in Fig 1.19.

Fig 1.19

Tag Cloud Based on a University Website

Data consisting of words are explored using **qualitative data analysis** to provide insight into the internal perceptions, attitudes and motivations of the respondents. On the other hand, data consisting of numbers are analysed using **quantitative data analysis**, to describe and predict the observable outward behaviour and characteristics of either people or inanimate processes.

Internal Consistency

A questionnaire may include a number of items which all test a single underlying concept, also called a **construct**, such as customer satisfaction. For example, questions about a stay at a hotel might ask the respondent how well they enjoyed their stay, to what extent their expectations were met, how likely it is that they would recommend the hotel to a friend, whether their hotel visit was good value for money, and whether they would stay in the hotel again. Since all these items are related, there should be good agreement between the responses to the different items. This is called the internal consistency of the test and it can be measured using **Cronbach's Alpha**. This is a coefficient that has a value of 1 when there is perfect consistency between all the items and a value of 0 when there is no consistency. It is recommended that the value of the coefficient should be at least 0.7 for exploratory research, at least 0.8 for basic research, and at least 0.9 for making a decision, such as selecting a hotel as a conference venue. The value of Cronbach's Alpha will

usually increase if more items are added to the test, although its value will decrease if unrelated items are added, such as items that test two different underlying concepts, for example, satisfaction with a stay in a hotel and satisfaction with the speakers at a conference that took place in the hotel. The internal consistency can be measured in a pilot survey before distributing the questionnaire to the larger number of respondents in the full survey.

Problems 1D

#1. A number of questions from a hotel guest survey are shown below. In each case identify the appropriate response scale.

(a) Was your most recent visit for business reasons?

(b) How many nights did you stay on your most recent visit?

(c) Please rate your satisfaction with each of the following aspects of your visit:

- Ease of check in
- Helpfulness of staff
- Comfort of bedroom
- Quality of breakfast buffet

(d) Have you any additional comments?

Project 1D

Survey Project

Identify some survey topic of interest to you. Carefully articulate the research objective and carry out a literature review. Design a questionnaire and conduct the survey with a minimum of 30 respondents. Include at least two questions that provide quantitative data, two questions that provide binary data, and two open-ended questions. Summarise these data using means, proportions and narrative summaries. Record all the survey data in a structured format in a spreadsheet. Reflect on the insight that arises from your survey findings in order to draw some conclusions, recommend some actions, and suggest some ideas for further research.

Write a report consisting of the following sections:

(a) Executive summary

(b) Literature review

(c) Survey methodology

(d) Questionnaire

(e) Data summaries and data analysis

(f) Conclusions and recommendations.

2

Calculating Probability

Having completed this chapter you will be able to:

– define probability from two different standpoints;

– calculate the probabilities of simple and compound events;

– appreciate some of the subtleties of probability;

– calculate the reliability of a system of components.

Nearly all processes involve uncertainty. This is obvious in the case of simple games of chance played with cards and dice, and more complex games such as horse-racing and team sports. But uncertainty is also present in business processes, manufacturing processes and scientific measurement processes. We cannot be certain about the future behaviour of a customer, or the exact dimensions of the next unit of product, or even the result that would arise if we repeated a measurement a second time.

2A. Calculating Simple Probabilities

Video Lecture https://youtu.be/81S5wkmlXOs

Although we cannot be absolutely certain, we can have a high degree of confidence that a particular event will occur. This confidence is based on the probability of the event. Probability is a measurement of how often the event occurs, or how likely it is to occur.

For example, when a coin is tossed, the probability of 'heads' is one-half. This means that if the coin is tossed many times, it is expected that heads will occur about half of the time. Tossing a coin is an example of a **random experiment**, i.e. an experiment whose outcome is uncertain. A single toss is called a **trial**. Heads is an example of an **outcome**. An **event** is a particular outcome, or set of outcomes, in which we are interested.

Definitions of Probability

DEFINITION #1 The probability of an event means **HOW OFTEN** the event occurs. It is the proportion of occurrences when many trials are performed.

If $p = 1$, the event always occurs.
If $p = 0$, the event never occurs.
If p is close to 1, the event usually occurs.
If p is close to 0, the event rarely occurs.
In general, $0 \leq p \leq 1$.

DEFINITION #2 The probability of an event is a measure of **HOW LIKELY** the event is to occur on a single future trial.

If $p = 1$, the event is certain to occur.
If $p = 0$, the event cannot occur.
If p is close to 1, the event is likely to occur and so we are confident that it will occur.
If p is close to 0, the event is unlikely and it would be surprising if it did occur.
Again, in general, $0 \leq p \leq 1$.

The probability of an event has the same value no matter which definition we use. Definition #1 is easier to think about. But definition #2 is useful when only one trial is to be performed, e.g. a single toss of a coin. With definition #2, when we say that the probability of 'heads' is one-half, we mean that heads is just as likely to occur as not to occur. In the case of manufactured goods, definition #1 represents the perspective of the producer, and definition #2 represents the perspective of a consumer who buys a single unit of product, in relation to the event 'a unit is defective'.

Probability values that are close to 1 are interesting because they indicate that such events are likely to occur. Therefore we have **confidence** that these event will occur. This is the basis of statistical estimation (Chapter 4). Probability values that are close to 0 are interesting because they indicate that such events are unlikely to occur. We say that such events are **significant**, which means that they may have occurred for a special reason. This is the basis of statistical hypothesis testing (Chapter 5).

Calculating Probability

Classical Probability

We have already claimed that the probability of heads, when a coin is tossed, is one-half. We use a capital letter to denote an event, so we write H for heads.

$$P(\text{H}) = 1/2$$

Why is the probability of heads calculated by dividing one by two? Because there are two sides on the coin (two possible outcomes) and only one of these is heads. This approach is valid only because the different outcomes (heads and tails) are equally likely to occur. It would not be true to say that the probability that it will snow in Dublin tomorrow is one-half, because the two outcomes ('snow' and 'no snow') are not equally likely.

Formula 2.1

Classical Probability

If all the outcomes of a trial are equally likely, then the probability of an event can be calculated by the formula

$$P(E) = \frac{The\ number\ of\ ways\ the\ event\ can\ occur}{The\ number\ of\ possible\ outcomes}$$

Note that this formula, like most formulae related to probability, has a condition associated with it. The formula must not be used unless the condition is satisfied.

Empirical Probability

We now consider how to calculate the probabilities of events which do not lend themselves to the classical approach. For example, what is the probability that when a thumb-tack is tossed in the air, it lands on its back, pointing upwards? Recall our first

definition of probability: the probability of an event means the proportion of occurrences of the event, when many trials are performed. We can perform a large number of trials, and count the number of times it lands on its back. If it lands on its back *r* times, out of *n* trials, then *p* = *r*/*n* estimates the probability of the event. Of course, the result will be only an estimate of the true probability: to get a perfect answer, we would have to repeat the trial an infinite number of times.

Formula 2.2

Empirical Probability

When a large number of trials are performed, the probability of an event can be estimated by the formula

$$P(E) = \frac{The\ number\ of\ times\ the\ event\ occurred}{The\ total\ number\ of\ trials\ performed}$$

Subjective Probability

There are some events whose probabilities cannot be calculated, or estimated, by either of the two formulae presented above: for example, the probability that a particular horse will win a race, or the probability that a particular company will be bankrupt within a year. These events do not satisfy the conditions associated with either classical probability or empirical probability. To estimate the probabilities of such events, we simply make a guess. The more we know about horse-racing, or business, the more reliable our guess will be. This approach is called subjective probability. In such cases it is common to refer to the **odds** of an event rather than its probability.

Formula 2.3

Odds

$$Odds = \frac{p}{1-p}$$

For example, if the probability that your team will win this weekend is 0.8, then the odds are 4 to 1. Your team is four times more likely to win than not to win. A bookmaker might offer a price of 4 to 1 to someone who wants to bet against your team.

The Simple Rules of Probability

So far we have considered single events. We now consider how probability is calculated for combinations of events. We can use the words ‘not’, ‘and’ and ‘or’ to describe virtually any combination of events.

The ‘NOT’ Rule

This rule can be applied to any event. It has no special condition associated with it.

Formula 2.4

The ‘NOT’ Rule

For any event A

$$P(not\ A) = 1 - P(A)$$

EXAMPLE S denotes a 'Six' when a die is rolled

P(S) = 1/6

P(not S) = 1 - 1/6 = 5/6

This rule is very useful, especially if you want to calculate the probability of a complex event, A, which has a simple **complement**, not A.

The Simple Multiplication Rule (AND)

This rule can only be applied to events that are **independent**. Independence means that the probability of occurrence of one event remains the same, regardless of whether or not the other event occurs.

Formula 2.5

The Simple Multiplication Rule

For independent events A, B

$$P(A \text{ and } B) = P(A) \times P(B)$$

EXAMPLE S denotes a 'Six' when a die is rolled

P(S) = 1/6

H denotes 'Heads', when a coin is tossed

P(H) = 1/2

S and H are independent because *P*(H) remains the same whether or not S occurs

P(S and H) = 1/6 × 1/2 = 1/12

Note: This rule can be extended to deal with any number of independent events. The probability that all the events occur is the product of their probabilities.

The Simple Addition Rule (OR)

This rule can only be applied to events that are **mutually exclusive**. This means that if one event occurs, the other event cannot occur at the same time.

Formula 2.6

The Simple Addition Rule

For mutually exclusive events A, B

$$P(A \text{ or } B) = P(A) + P(B)$$

EXAMPLE F denotes a 'Five' when a die is rolled

P(F) = 1/6

E denotes an 'even number' when a die is rolled

P(E) = 3/6

F and E are mutually exclusive because they cannot both occur at once

P(F or E) = 1/6 + 3/6 = 4/6

Note: This rule can be extended to deal with any number of mutually exclusive events. The probability that any of the events occur is the sum of their probabilities.

Solving Problems

When you are presented with a problem in probability, the first step is to identify the separate events involved. Next, rephrase the problem, using only the words 'and', 'or' and 'not' to connect the events. Next, use multiplication and addition to calculate the probability of this combination of events, paying attention to the relevant assumptions in each case. Also, look out for short-cuts, e.g. if the event is complex, it may be simpler to calculate the probability that it does not occur, and subtract the result from one.

EXAMPLE

A husband and wife take out a life insurance policy for a twenty-year term. It is estimated that the probability that each of them will be alive in twenty years is 0.8 for the husband and 0.9 for the wife. Calculate the probability that, in twenty years, just one of them will be alive.

SOLUTION

Identify the separate events:

H: The husband will be alive in twenty years. $P(\text{H}) = 0.8$

W: The wife will be alive in twenty years. $P(\text{W}) = 0.9$

Rephrase the problem using the words 'AND' 'OR' and 'NOT' to connect the events.

'just one of them' means:

H and not W or not H and W

'or' becomes '+' because the two events are mutually exclusive

'and' becomes 'multiply' assuming independence

$$\begin{aligned} P(\text{just one of them}) &= P(\text{H}) \times P(\text{not W}) + P(\text{not H}) \times P(\text{W}) \\ &= 0.8 \times 0.1 + 0.2 \times 0.9 \\ &= 0.08 + 0.18 \\ &= 0.26 \end{aligned}$$

The probability that just one of them will be alive in 20 years is 26%.

Problems 2A **Video Tutorial**: https://youtu.be/U5N8JmZIyYo

#1. Explain each of the statements below twice, first using definition #1 and then using definition #2 of probability.
(a) The probability of a 'six' on a roll of a die is one-sixth.
(b) The probability that an invoice is paid within 30 days is 90%.
(c) The probability that a pig flies is 0.

#2. Calculate, or estimate, the probabilities of each of the following events:
(a) A number greater than four occurs when a die is rolled.

(b) A car is red.
(c) The current Taoiseach will be Taoiseach on the first of January next year.

#3. A single letter is drawn at random from the alphabet. The following events are defined:
W: the letter W is obtained.
V: a vowel is obtained.
Calculate the probability of each of the following events:
(a) W
(b) V
(c) W or V.

#4. A coin is tossed and a die is rolled. The following events are defined:
H: the coin shows 'heads'.
S: the die shows a 'six'.
Calculate the probability of each of the following events:
(a) H
(b) S
(c) H and S
(d) H or S.

#5. A die is rolled twice. Calculate the probability of each of the following events:
(a) two sixes
(b) no sixes
(c) a six on the first roll but not on the second roll
(d) at least one six
(e) exactly one six.

#6. Three coins are tossed simultaneously. Calculate the probability of obtaining:
(a) three heads
(b) at least one head.

#7. A firm submits tenders for two different contracts. The tenders will be assessed independently. The probability that the first tender will be successful is 70%, and the probability that the second tender will be successful is 40%.
Calculate the probability that:
(a) both will be successful
(b) neither will be successful
(c) only the first will be successful
(d) only the second will be successful
(e) at least one will be successful.

Project 2A

Estimation of Odds

(a) Identify an upcoming sporting event. (b) Estimate the probability of a particular outcome. (c) Explain how you arrived at your estimate. (d) Calculate the corresponding odds. (e) Find out the odds offered by a bookmaker for the same outcome. (f) Suggest some reasons why the bookmaker does not seem to agree with your estimate.

2B. The General Rules of Probability

The General Multiplication Rule

Video Lecture https://youtu.be/zlq-gPdEdzA

This rule can be applied to any events. The events do not need to be independent.

Formula 2.7

The General Multiplication Rule

For any events A, B

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

P(B|A) is the **conditional probability** of B given that A has occurred.

EXAMPLE A letter is drawn at random from among the 26 letters of the alphabet.
L denotes the event 'the letter drawn is from the latter half of the alphabet (N to Z)'.
V denotes the event 'the letter drawn is a vowel'.

P(L) = 13/26
P(V) = 5/26

These events are not independent. If L occurs, then the probability of V is not 5/26, but rather 2/13, because there are only two vowels in the latter half of the alphabet.

P(V|L) = 2/13

P(L and V) = *P*(L) × *P*(V|L)

P(L and V) = 13/26 × 2/13 = 1/13

The General Addition Rule

This rule can be applied to any events. They do not need to be mutually exclusive.

Formula 2.8

The General Addition Rule

For any events A, B

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

EXAMPLE F denotes 'a number smaller than five' when a die is rolled

P(F) = 4/6

E denotes an 'even number' when a die is rolled
P(E) = 3/6

F and E are not mutually exclusive because they could both occur at once, specifically when a 'one' or a 'three' occurs.

P(F or E) = *P*(F) + *P*(E) - *P*(F and E)

= 4/6 + 3/6 - 2/6

= 5/6

De Morgan's Rule

De Morgan's rule provides a simple way to replace an expression involving 'or' with an equivalent expression involving 'and', and vice versa. This can be convenient, for example, if the word 'or' arises with independent events.

Formula 2.9

De Morgan's Rule

To replace an expression with an alternative, equivalent, expression, follow these three steps:

1. Write 'not' before the entire expression.

2. Write 'not' before each event in the expression.

3. Replace every 'and' with 'or', and every 'or' with 'and'.

Example: A or B ≡ not (not A and not B)

EXAMPLE A game involves tossing a coin and rolling a die, and the player 'wins' if either 'heads' or a 'six' is obtained. Calculate the probability of a win.

H denotes 'heads', $P(H) = 1/2$

S denotes a 'Six', $P(S) = 1/6$

$P(\text{win}) = P(\text{H or S})$

H and S are not mutually exclusive events.

But H and S are independent events, so let us write:

$P(\text{win}) = P(\text{not (not H and not S)})$

$P(H) = 1/2$, $P(\text{not H}) = 1/2$

$P(S) = 1/6$, $P(\text{not S}) = 5/6$

$P(\text{win}) = 1 - (1/2 \times 5/6)$

$P(\text{win}) = 1 - 5/12$

$P(\text{win}) = 7/12$

Permutations and Combinations

The formula for classical probability (formula 2.1) can be used to calculate the probability that things are selected or arranged in a particular way. In cases like this, where things are being selected or arranged, formulae are required for factorials, permutations and combinations.

Formula 2.10

Factorial

The number of ways of arranging n things in order is called ***n factorial***, written $n!$

$$n! = n.(n-1).(n-2)...3.2.1$$

EXAMPLE In how many different ways can the letters A, B and C be arranged?

ANSWER 3! = 3.2.1 = 6

There are three ways to choose a letter for first position. There are then two ways to choose a letter for second position. That leaves one letter which must go in third position. The arrangements are:

ABC, ACB, BAC, BCA, CAB, CBA.

Formula 2.11

Permutations

The number of ways of arranging *r* things, taken from among *n* things, is written ${}^nP_r$ and is called ***n permutation r***.

$$ {}^nP_r = \frac{n!}{(n-r)!} $$

EXAMPLE The 'result' of a horse race consists of the names, in order, of the first three horses past the finishing post. How many different results are possible in a seven-horse race?

ANSWER ${}^7P_3 = 7.6.5.4.3.2.1 \div (4.3.2.1) = 7.6.5 = 210$

In this situation, seven different horses can finish in first place. After the first place has been allocated, there are six possibilities for second position. After the first and second places have been allocated, there are five possibilities for third place.

Formula 2.12

Combinations

The number of sets of *r* things that can be taken from among *n* things is written ${}^nC_r$ and is called ***n combination r***. The order of taking the things does not matter.

$$ {}^nC_r = \frac{n!}{r!(n-r)!} $$

EXAMPLE In how many different ways can 6 numbers be chosen from among 47 numbers in a lottery game?

ANSWER ${}^{47}C_6 = 47! \div 6! \div 41! = 47.46.45.44.43.42 \div (6.5.4.3.2.1) = 10{,}737{,}573$

If the six numbers had to be arranged in the correct order then the answer would be ${}^{47}P_6$. But because the order does not matter, we divide this by 6! which is the number of arrangements possible for the 6 numbers.

Problems 2B

#1. In how many different orders can 8 different numbers be arranged?

#2. Leah, Josiah and Johnny are three members of a class of 20 students. A raffle is held among the members of the class. Each member has one ticket, and the three winning tickets are drawn from a hat.

(a) What is the probability that Leah wins first prize, Josiah wins second prize, and Johnny wins third prize?

(b) What is the probability that Leah, Josiah and Johnny are the three prize-winners?

#3. It has already been calculated that there are 10,737,573 ways in which 6 numbers can be chosen from among 47 numbers in a lottery game. In how many different ways can 41 numbers be chosen from among 47 numbers? Explain your answer.

#4. In a lottery game a player selects 6 numbers, from a pool of 47 numbers. Later on six winning numbers are identified randomly from the pool. Calculate the probability that, on a single play, a player achieves:

(a) the jackpot, i.e. the player's selection matches all six winning numbers

(b) a 'match 5', i.e. the player's selection matches any 5 winning numbers

(c) a 'match 4', i.e. the player's selection matches any 4 winning numbers.

#5. Explain in words why ${}^nP_n = n!$

#6. Explain in words why ${}^nC_n = 1$.

2C. Subtleties and Fallacies

Video Lecture https://youtu.be/dNS2HMbNZC4

The concepts in probability can be subtle. And numerical answers are often counter-intuitive. For example, suppose that in a certain population, the **base rate** of people who have a latent liver disease is 1%. A medical screening programme is undertaken to identify those who have the disease. A suitable screening test is available. The **sensitivity** of the test is 95%, that is to say that 95% of the time it correctly classifies a person with the disease as having the disease. The **specificity** of the test is 98%, that is to say that 98% of the time it correctly classifies a healthy person as healthy. The people who are classified as having the disease are all sent a letter informing them that they may have the disease. Now what percentage of the people who receive these letters actually have the disease? What do you think? This problem will now be solved using Bayes' Theorem, and the answer may surprise you.

Bayes' Theorem

Let A denote the event 'the person has the disease'.

Let B denote the event 'the person is classified as having the disease'.

We are given that

$P(A) = 0.01$

$P(B|A) = 0.95$

$P(B|\text{not } A) = 0.02$

We now proceed to calculate $P(A|B)$ using the formula for Bayes' Theorem.

Formula 2.13

Bayes' Theorem

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Numerator: P(A and B) = P(A) × P(B|A) = 0.01 × 0.95 = 0.0095

Denominator: P(B) = P(A) × P(B|A) + P(not A) × P(B|not A)

= 0.01 × 0.95 + 0.99 × 0.02

= 0.0293

P(A|B) = 0.0095 ÷ 0.0293 = 0.3242 = 32.42%

The answer is not intuitive. It turns out that only a minority of the people who receive the letters have the disease. This happens because the number of incorrect classifications that occur with the large number of healthy people exceeds the number of correct classifications that occur with the small number of people who have the disease. For this reason, many experts have advised that certain medical screening programmes should be discontinued, because unnecessary anxiety is caused to a large number of people.

The Prosecutor's Fallacy

Suppose that a defendant is charged with burglary and the prosecutor presents this argument. 'It is unlikely that an innocent person would have glass fragments under their shirt cuffs, and yet the defendant has glass fragments under their shirt cuffs.' This argument seems to imply that the defendant is guilty, but let us unpack the argument.

Let A denote 'a person matches a certain description', in this case the person has glass fragments under their shirt cuffs.

Let B denote 'a person is innocent'.

The prosecutor has stated that P(A|B) is small, by using the word 'unlikely'. But this does not mean that P(B|A) is small.

There are many innocent people, such as those who install and repair windows, who may have glass fragments under their shirt cuffs. So it may be quite likely that a person with glass fragments under their shirt cuffs is innocent.

This mistake of 'reversing the conditional' is referred to as the prosecutor's fallacy. Here is another example. Saying that most shoplifters wear hoodies is not the same as saying that most people who wear hoodies are shoplifters.

False Assumption of Independence

Witnesses to a robbery in Los Angeles testified that one of the perpetrators had been a man with a beard and a moustache. Later that day a police officer arrested a suspect who answered this description. When the case was brought to trial, part of the prosecutor's presentation to the jury went as follows:

P(man with beard) = 0.1

P(man with moustache) = 0.25

The prosecutor then proceeded to argue that:

P(man with beard and moustache) = 0.025

What is wrong with this argument?

Confusion about random match probability (RMP)

If there is a probability of 'one in a million' that a person possesses a certain attribute, then it should not be surprising to find that many people possess that attribute. Based on a **frequentist approach**, in a population of five million people we would expect the attribute to arise about five times. So 'one in a million' does not describe the probability that the attribute will ever arise, but rather the probability that it will arise in the case of a particular individual selected at random. Random match probability can be used with DNA evidence to build a case against a person who is *already* a suspect.

An Illustration

The table below shows how data can be **cross-classified** according to two sets of criteria (gender and smartphone operating system). This is a 2×2 table because there are only two rows and two columns of data in the table proper. The totals are displayed in the **margins** and do not provide any additional information. A number of different probabilities can be calculated in relation to the events: F (female) and I (iPhone).

Table 2.1

Gender	**iPhone**	**Android**	*Totals*
Male	20	30	50
Female	26	24	50
Totals	46	54	100

P(F) = 50/100 = 0.50	**marginal probability** of F
P(I) = 46/100 = 0.46	marginal probability of I
P(F and I) = 26/100 = 0.26	**joint probability** of F and I
P(F|I) = 26/46 = 0.5652	conditional probability of F given I
P(I|F) = 26/50 = 0.52	conditional probability of I given F
P(F|I) ≠ *P*(I|F)	prosecutor's fallacy
P(F|I) ≠ P(F)	I,F are not independent
P(F or I) = 1 - 0.3 = 0.7	De Morgan's rule
P(F or I) = *P*(F) + *P*(I) - *P*(F and I)	General addition rule
= 0.5 + 0.46 - 0.26 = 0.7	
P(F and I) = *P*(F) x *P*(I|F)	General multiplication rule

= 0.50 x 0.52 = 0.26

P(F and I) = P(I) x P(F|I) General multiplication rule

= 0.46 x 0.5652 = 0.26

P(F|I) = P(F and I) ÷ P(I) Bayes' theorem

= 0.26 ÷ 0.46 = 0.5652

Likelihood Ratio

The likelihood ratio is the ratio of two probabilities of the same event under different hypotheses. The higher the ratio, the more likely it is that the first hypothesis is true. For example, suppose that an iPhone is left behind when these people leave the room. How much more likely is it that the iPhone belongs to a female rather than a male?

The Likelihood Ratio, $LR = P(I|F) \div P(I|M)$ = 0.52 ÷ 0.4 = 1.3

It is 1.3 times more likely that the iPhone belongs to a female rather than a male.

Bayesian Approach

Suppose that the iPhone which was left behind has a pink cover. Because of the pink cover, we might immediately consider that it is ten times more likely to belong to a female. This number, 10, is called the **prior odds**. Next, we examine the scientific and statistical evidence as shown above which informs us that the likelihood ratio is 1.3. Finally we multiply the prior odds by the likelihood ratio to obtain the **posterior odds** which informs us that the iPhone is thirteen times more likely to belong to a female.

Statistical Approaches to DNA evidence

Using the frequentist approach to DNA evidence, we could say, 'The chance of obtaining these matching profiles if the blood came from a random person unrelated to the suspect is one in a million.'

Using the likelihood ratio approach to DNA evidence, we could say, 'The results of the DNA analysis are one million times more likely if the DNA came from the suspect than if the DNA came from a random unrelated person in the population.'

Using the Bayesian approach to DNA evidence, the prior odds is multiplied by the likelihood ratio to obtain the posterior odds in favour of the prosecution hypothesis.

Problems 2C

#1. There are many reasons why an unsound conclusion is sometimes drawn when reasoning with probability. Here is a list of some of these reasons, labelled i, ii, etc.

i. Two events are assumed to be mutually exclusive for no good reason.

ii. A sample proportion is assumed to be the same as a population proportion.

iii. Bayes' theorem is incorrectly applied.

iv. Prior and posterior probabilities are not well understood.

v. The likelihood ratio is incorrectly calculated.

vi. The conditional probability of A given B is confused with the conditional probability of B given A.

vii. A combination is confused with a permutation.

viii. The random match probability is not well understood.

ix. A joint probability is confused with a marginal probability.

x. Two events are assumed to be independent for no good reason.

In each of the following five different scenarios, an unsound conclusion is drawn in the final sentence. In each case, identify which one of the reasons listed above best describes the mistake in the reasoning. For example, if you think that the mistake in scenario (a) corresponds to i above, then just write "(a) i" for your answer at part (a).

(a) Shoplifters usually wear hoodies. George wears a hoodie. Therefore it's likely that George is a shoplifter.

(b) The probability that a man has a beard is 0.1 and the probability that a man has a moustache is 0.25. Therefore the probability that a man has a beard and a moustache is $0.1 \times 0.25 = 0.025$.

(c) The getaway car was a yellow hatchback with black doors. Only one car in a million has these features. Police have found a car which matches this description. Therefore it is almost certain that this is the getaway car.

(d) 20% of students own a car and 30% of students own a bicycle. Therefore 50% of students own a car or a bicycle.

(e) On 50% of Tom's shopping trips he travels by bus and on the other occasions he walks. On 10% of Tom's shopping trips he buys chocolate. Therefore the probability is 5% that on one of Tom's shopping trips he walks and buys chocolate.

#2. 5% of cattle in a national herd have a particular disease. A test is used to identify those animals who have the disease. The test is successful in 70% of genuine cases, but also gives a 'false positive' result for healthy cattle, with probability 0.002. What proportion of the animals, identified by the test, actually have the disease?

#3. After a busy weekend at a campsite, a pair of shoes was left behind by one of the campers. The shoes were size 11, which indicates that the owner is 350 times more likely to be male than female. The weekend records show that 571 campers were at the campsite and 390 of these were male. Calculate the likelihood ratio and combine this with the prior odds to find the posterior odds that the owner of the shoes is male.

2D. Reliability

Video Lecture https://youtu.be/0hK7qmCPzaQ

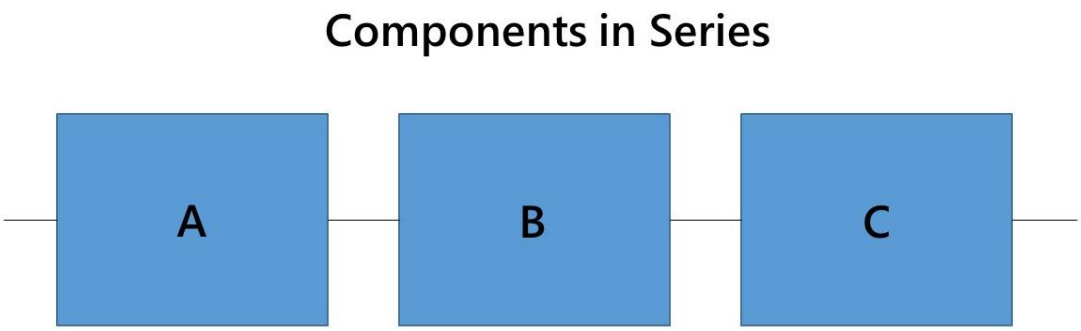
Reliability is defined as the probability that a product will function as required for a specified period. Where a product consists of a system of components whose individual reliabilities are known, the reliability of the system can be calculated.

Components in Series

When components are in series it means that all the components must function in order for the system to function. Most systems are like this.

EXAMPLE A torch consists of a battery, a switch and a bulb, assembled in series. The reliabilities of the battery, switch and bulb are 0.9, 0.8 and 0.7 respectively. Calculate the system reliability.

Fig 2.1



A: battery functions, $P(A) = 0.9$

B: switch functions, $P(B) = 0.8$

C: bulb functions, $P(C) = 0.7$

P(torch functions) = P(A and B and C)

Assuming independence

P(torch functions) = $P(A) \times P(B) \times P(C)$

P(torch functions) = $0.9 \times 0.8 \times 0.7 = 0.504$

The independence condition requires that a good battery has the same chance of being combined with a good switch, as does a bad battery, etc. This condition tends to be satisfied by random assembly.

Formula 2.14

Reliability of a System of Components in Series

The reliability of a system of components in series, assuming independence, is the product of the reliabilities of the individual components.

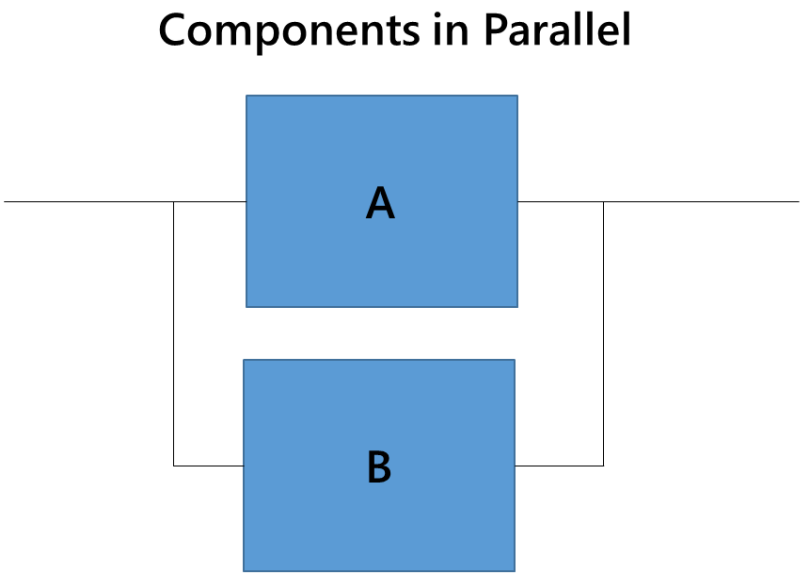
Components in Parallel

Parallel systems include components in back-up mode. If one component fails, there is another component to take its place. The back-up components can be either in standby mode, such as a standby generator to provide power during a mains failure, or in active mode, such as stop-lamps on a vehicle, where all of the components are active even before one fails, but the calculations are the same either way.

EXAMPLE The reliability of the mains power supply is 0.99 and the reliability of a standby generator is 0.95.

Calculate the system reliability.

Fig 2.2



A: mains functions, $P(A) = 0.99$, $P(\text{not A}) = 0.01$

B: generator functions, $P(B) = 0.95$, $P(\text{not B}) = 0.05$

$P(\text{system functions}) = P(\text{A or B})$

$P(\text{system functions}) = 1 - P(\text{not A and not B})$

Assuming independence

$P(\text{system functions}) = 1 - 0.01 \times 0.05$

$P(\text{system functions}) = 1 - 0.0005 \quad = 0.9995$

Formula 2.15

Reliability of a System of Components in Parallel

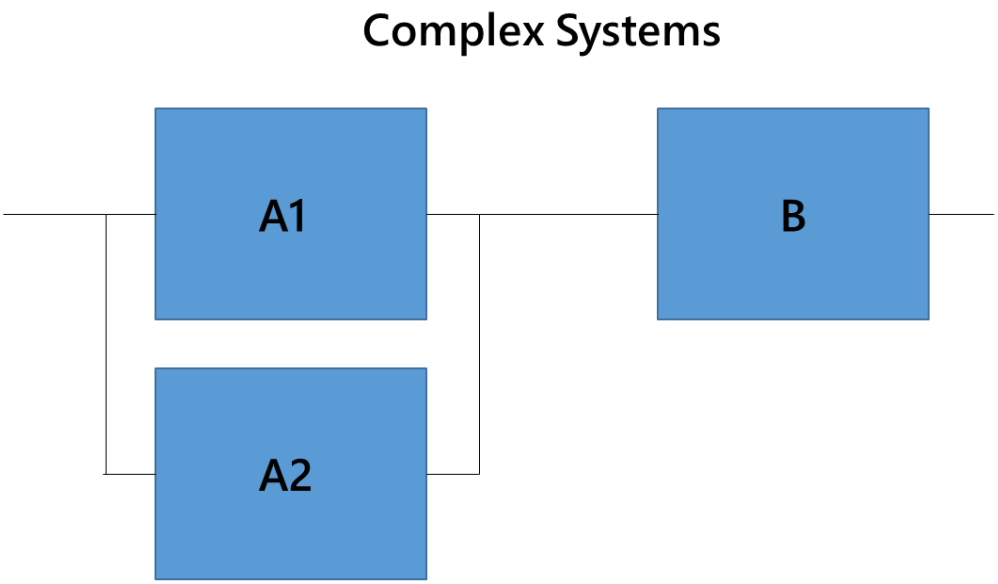
The unreliability of a system of components in parallel, assuming independence, is the product of the unreliability of each of the individual components, where unreliability is defined as 1 – reliability.

Complex Systems

Complex systems contain subsystems. The reliability of each subsystem should be calculated first. The subsystem can then be regarded as a single component.

EXAMPLE A projector has two identical bulbs and one fuse. The reliability of a bulb is 0.9 and the reliability of a fuse is 0.95. All other components are assumed to be 100% reliable. Only one bulb is required for successful operation. Calculate the reliability of the projector.

Fig 2.3



A1: bulb 1 functions, $P(A1) = 0.9$, $P(\text{not } A1) = 0.1$

A2: bulb 2 functions, $P(A2) = 0.9$, $P(\text{not } A2) = 0.1$

A: bulb sub-system functions

$P(A) = 1 - 0.1 \times 0.1$

$P(A) = 1 - 0.01 = 0.99$

$P(\text{projector functions}) = P(\text{A and B})$

$P(\text{projector functions}) = 0.99 \times 0.95 \quad = 0.9405$

Problems 2D

#1. A vacuum cleaner consists of a power supply, a switch and a motor assembled in series. The reliabilities of these components are 0.95, 0.90 and 0.80, respectively.

(a) Calculate the reliability of a vacuum cleaner.

(b) If a customer buys two vacuum cleaners, in order to have a second vacuum cleaner available as a back-up, calculate the reliability of the system.

(c) If a customer buys two vacuum cleaners, in order to have a second vacuum cleaner available to provide spare parts as required, calculate the reliability of the system.

#2. A commuter can drive from home to work door-to-door by car, or else take a bus from home to the train station and then take a train to work. The reliabilities of the car, bus and train are 0.80, 0.90 and 0.95 respectively, and these three are independent. Calculate the system reliability for the journey to work.

#3. A stirrer consists of three independent components: a battery, a motor, and a paddle. The reliabilities of each of these components are 0.90, 0.95 and 0.85, for the battery, motor, and paddle, respectively. All three components are required for successful operation of the stirrer.

(a) Calculate the reliability of the stirrer.

(b) Damaged paddles are easily replaced. If a second paddle is included, calculate the reliability of the stirrer.

(c) Batteries are also easily replaced. If a second paddle and a second battery are included, calculate the reliability of the stirrer.

(d) Assume that a second paddle and a second battery are already included. Which of the following two options leads to a higher system reliability: adding a third battery or adding a third paddle?

(e) How many paddles are required to bring the reliability of the paddle sub-system to over 99.9%?

#4. The components of a wireless network are: a socket, a router and an adaptor. The three components are independent, and all three components must operate for successful operation of the network. The reliabilities of these three components are 0.99, 0.98 and 0.60 for the socket, the router and the adaptor, in that order.

(a) Calculate the reliability of the network.

(b) If a second adaptor is available as a back-up, calculate the reliability of the network.

(c) If a second and a third adaptor are available for back-up as required, calculate the reliability of the network.

(d) If an unlimited number of adaptors are available for back-up as required, calculate the reliability of the network.

(e) What is the minimum number of adaptors that are required so that the reliability of the network will exceed 95%?

3

Using Distributions

Having completed this chapter you will be able to:

– recognise some common statistical distributions;

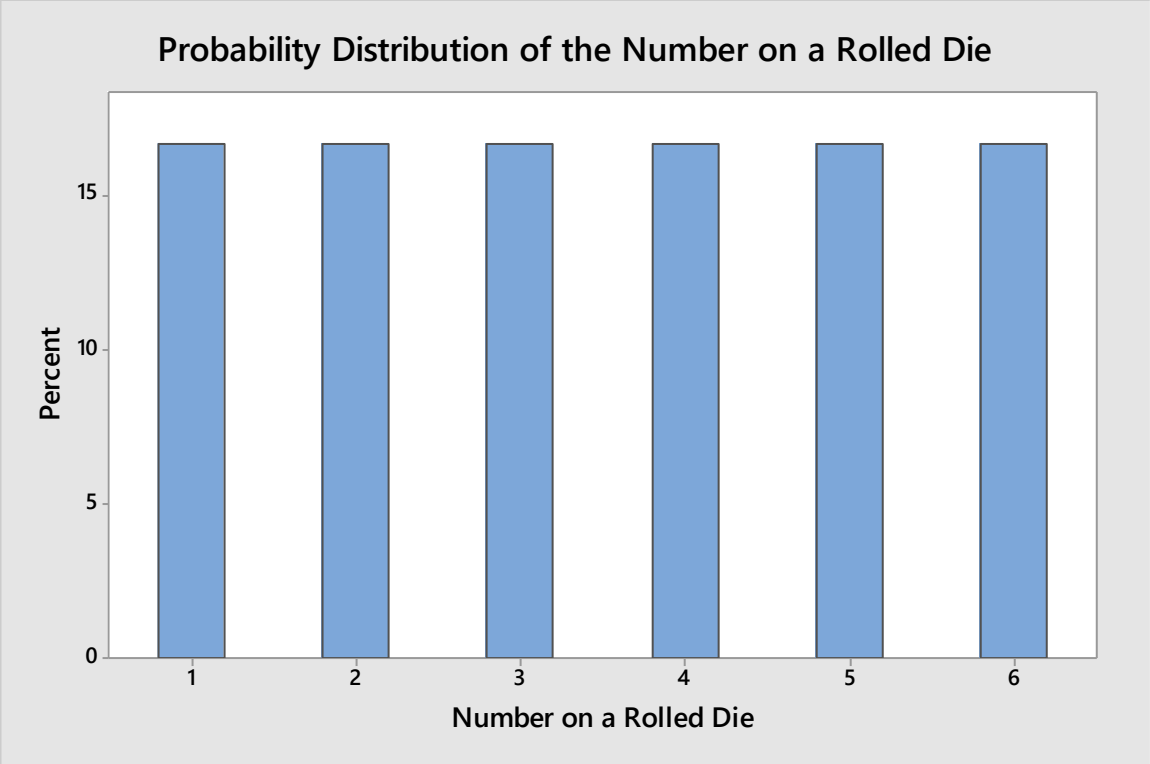
– calculate probabilities associated with these distributions.

3A. Random Variables

Video Lecture https://youtu.be/JyAw7pOeAfQ

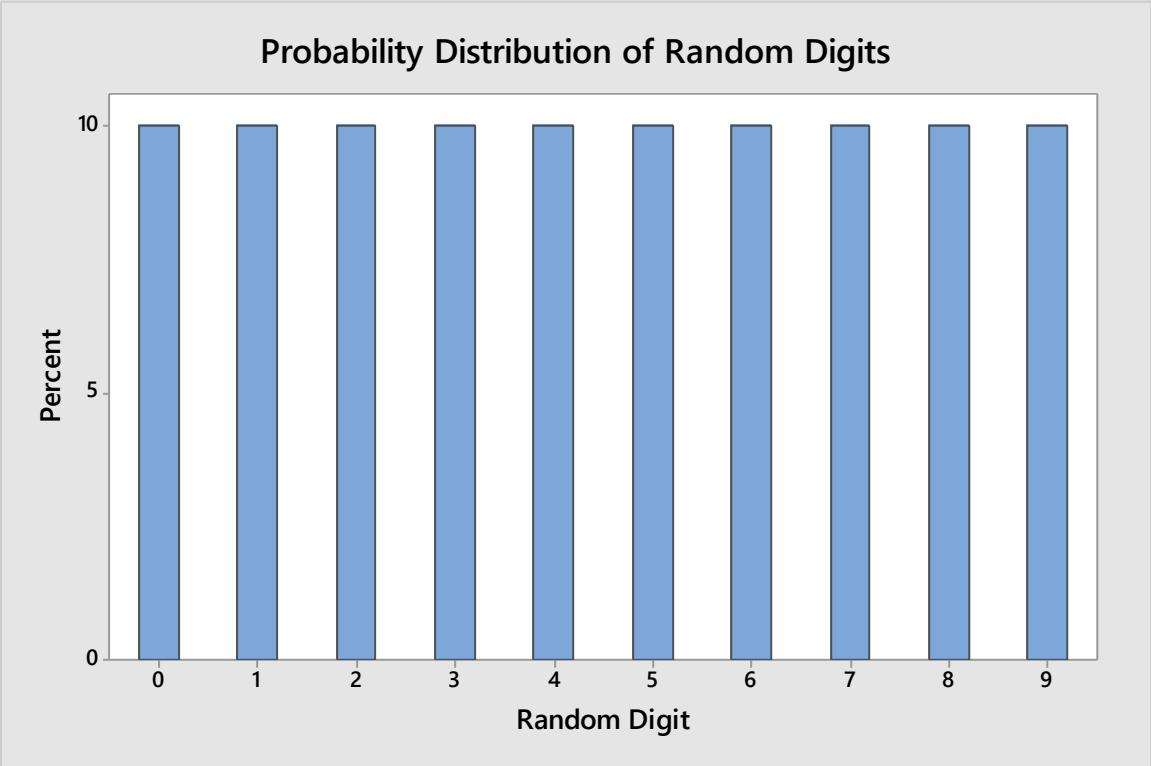
An experiment such as rolling a die is called a **random experiment**, because the outcome is uncertain. And because the outcome is a number, that number is called a **random variable**. All we can say for sure is that the outcome will come from the set of all possible outcomes, {1, 2, 3, 4, 5, 6}, called the **sample space**. In this case, all the outcomes in the sample space are equally likely. The **probability distribution** is shown below, and the probabilities add up to 1 because one or other of these numbers is certain to occur.

Fig 3.1



Selecting a random digit gives rise to a similar situation. The same rule applies about the outcomes being equally likely. The probability distributions even look alike, having bars that are equally tall.

Fig 3.2



These random variables are alike in that they are from the same family, called the **uniform distribution**. They are different members of the family as can be seen from their different **parameter** values, $k = 6$ and $k = 10$, where a parameter means a number that is constant within a given situation.

In general, if we can identify the distribution of a random variable, then we have a model that describes its behaviour. This can be useful in two different ways. On the one hand, if the data fit the model well, we can predict the behaviour of the variable by calculating the probability of occurrence of the different values of the variable. And on the other hand, if the data do not fit the model well, this alerts us that there is some other factor affecting the process, and by searching for this factor we may find some previously unknown problem or opportunity.

Problems 3A

#1. An experiment consists of tossing a coin. Instead of expressing the outcome as heads or tails, the outcome is expressed as the number of heads that appear on that toss.
(a) What is the sample space for this experiment?
(b) What is the name of the distribution that describes the behaviour of this variable?
(c) What is the value of the parameter?

3B. The Normal Distribution

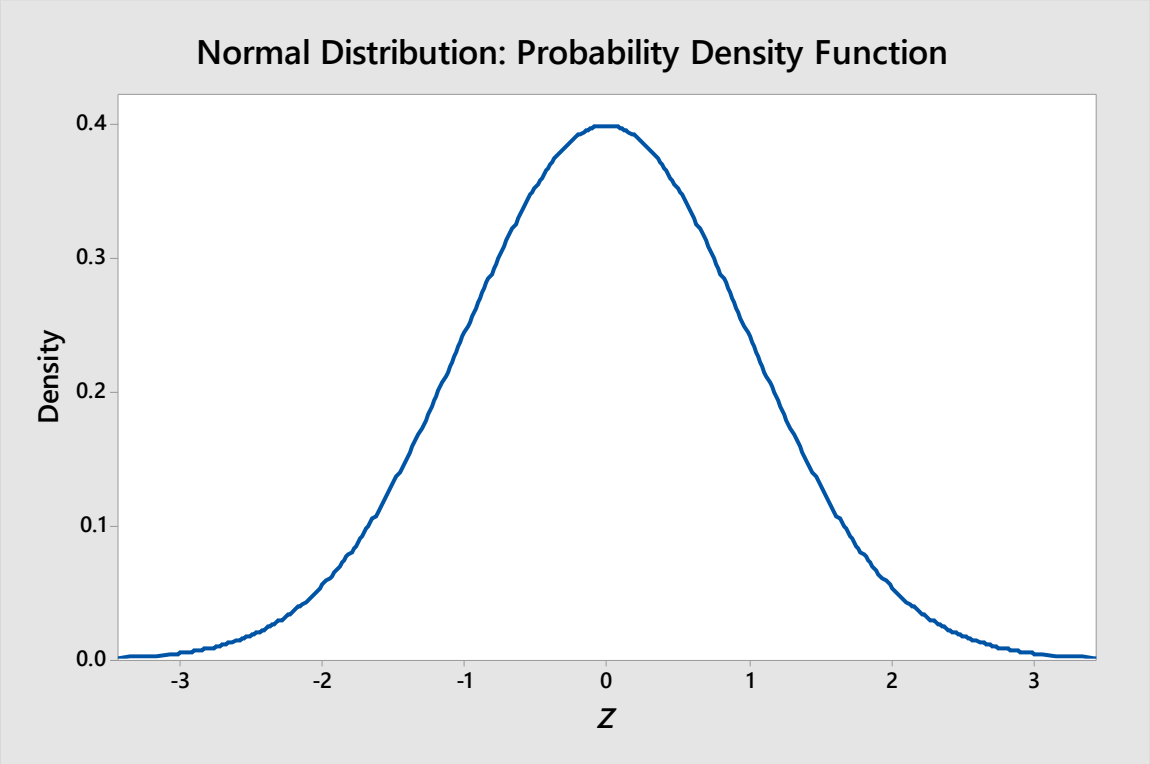
Video Lecture https://youtu.be/KkU5p_HiA78

The normal distribution describes variables from a wide variety of situations, including measurements that occur in nature, manufacturing dimensions, and errors of

measurement. It always has the same characteristic bell shape. The normal distribution has two parameters: the mean, μ, and the standard deviation, σ.

The normal distribution is a **continuous distribution.** The **probability density** is represented by the height of the curve and the total area under the curve is 1, unlike **discrete distributions** where the heights of the bars add up to one. The probability of occurrence of a value in any particular range is the area under the curve in that range.

Fig 3.3



To calculate normal probabilities, we first compute the **standard normal score**, z.

Formula 3.1

Normal Distribution

$$z = \frac{X - \mu}{\sigma}$$

z is the 'standard normal score'

The z-score represents the number of standard deviations that a value is above the mean. It follows that $z = 0$ represents the mean, positive z-scores represent values above the mean, and negative z-scores represent values below the mean. The normal distribution table (see appendix) gives the cumulative probability for any z-score.

EXAMPLE 1

The heights of men are normally distributed with mean, $\mu = 175$ cm, and standard deviation, $\sigma = 10$ cm. What is the proportion of men that are taller than 182.5 cm?

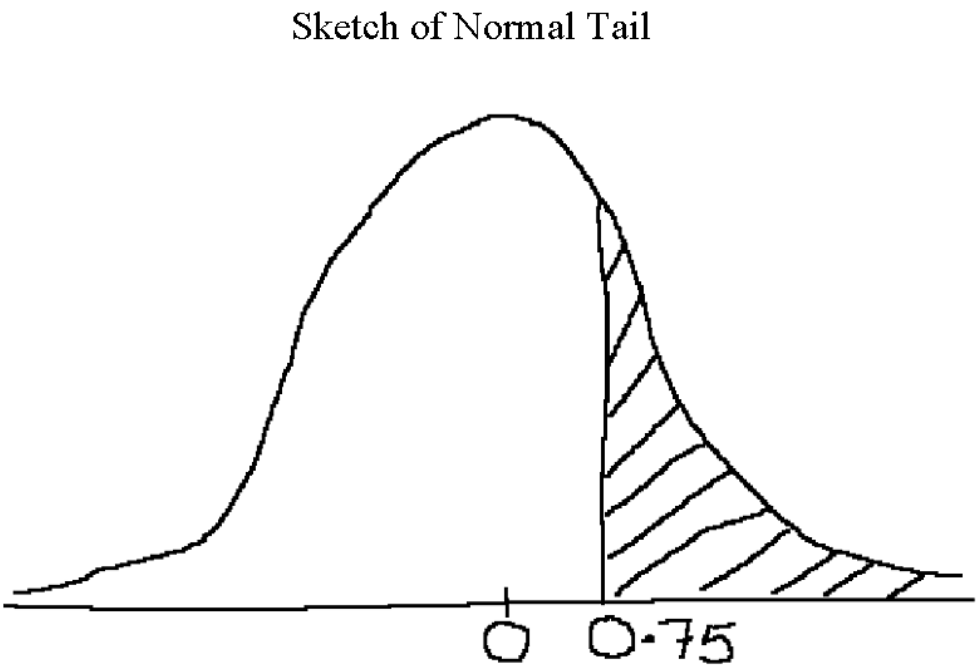
SOLUTION 1

$z = (X - \mu) / \sigma$

z = (182.5 - 175) / 10

z = 0.75

Fig 3.4



Having calculated z, we sketch a bell-curve and put a tick-mark at zero in the middle. We mark our z-score on the left or right as appropriate (on the right this time, because we have +0.75), and shade the area to the left or right of z as appropriate (to the right this time, because we have 'taller than'). We now ask ourselves whether the shaded area is a minority or a majority of the total area under the curve (a minority this time, because we can see that the shaded area is less than 50% of the total area under the curve).

The normal table gives p = 0.7734, a majority.

We require 1 - p = 1 - 0.7734 = 0.2266.

22.66% of men are taller than 182.5 cm.

EXAMPLE 2

The heights of men are normally distributed with mean, μ = 175 cm, and standard deviation, σ = 10 cm. What is the proportion of men that are between 165 cm and 170 cm in height?

SOLUTION 2

These two X values will correspond to two z-scores.

X_1 = 165

X_2 = 170

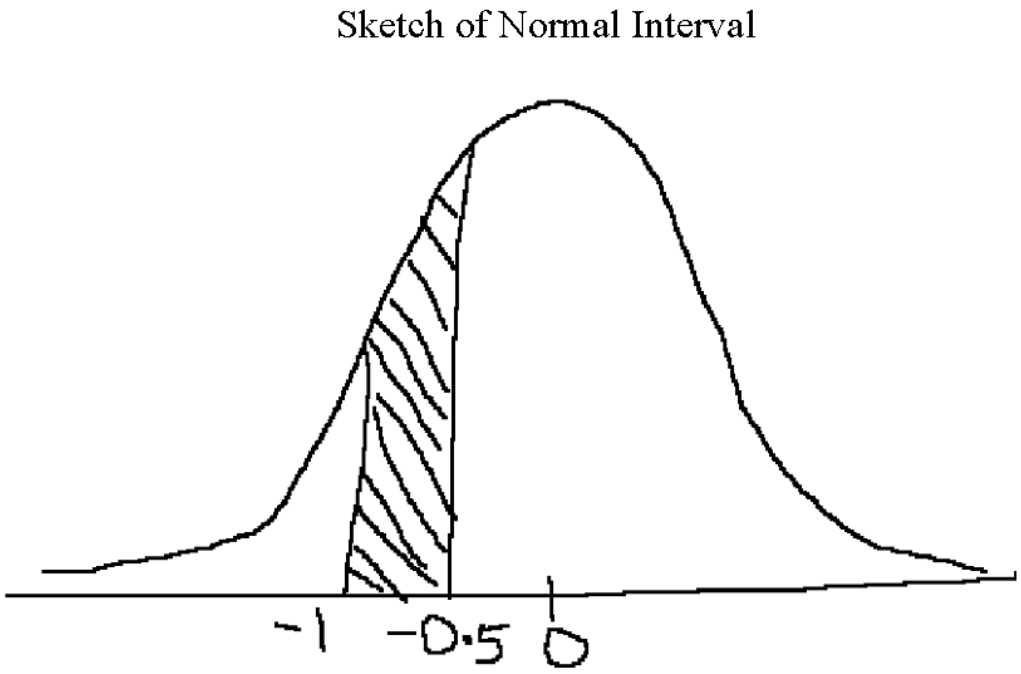
$z = (X - \mu) / \sigma$

z_1 = -1

$z_2 = -0.5$

Now we sketch a bell-curve and mark the two *z*-scores where they belong.

Fig 3.5



The following four steps can be used to find the probability corresponding to an interval.

Left tail area	= 1 - 0.8413 = 0.1587	
Right tail area	= 0.6915	
Total tail area	= 0.8502	
Interval area	= 1 - 0.8502	= 0.1498.

14.98% of men are between 165 cm and 170 cm in height.

Problems 3B **Video Tutorial:** https://youtu.be/GXX2A6HwUoo

#1. The heights of corn stalks are normally distributed with mean μ = 16 cm, and standard deviation σ = 2 cm. Calculate the proportion of stalks with heights that are:
(a) shorter than 14 cm
(b) between 13 cm and 17 cm
(c) shorter than 17.5 cm
(d) taller than 18.25 cm
(e) taller than 13.8 cm
(f) between 11 cm and 13.5 cm
(g) between 17.625 cm and 18.475 cm
(h) taller than 16 cm
(i) taller than 30 cm
(j) exactly 17 cm
(k) between 12.08 cm and 19.92 cm.

#2. State your answer to problem #1 (k) as a 'prediction interval' by completing the following sentence: 'It can be stated with... confidence that if a corn stalk is selected at random, its length will be between... cm and... cm.'

#3. The heights of women are normally distributed with mean 167.5 cm and standard deviation 7.5 cm. A range of T-shirts are made to fit women of different heights as follows:

Small T-shirt: 155 to 165 cm
Medium T-shirt: 165 to 175 cm
Large T-shirt: 175 to 185 cm.

What percentage of the population is in each category, and what percentage is not catered for?

#4. The times that different people spend reading a blog are normally distributed with mean 70 seconds and standard deviation 5 seconds.

(a) What proportion of the reading times are greater than 82 seconds?

(b) What proportion of the reading times are between 65 and 80 seconds?

(c) What proportion of the reading times are between 55 and 65 seconds?

(d) It is thought that a reading time greater than 120 seconds will cause the reader to lose interest. How often will such reading times occur?

(e) Calculate a 95% prediction interval for the reading time of a randomly selected person.

#5. The fill-volumes of bottles are normally distributed with mean 169 ml and standard deviation 2 ml.

(a) What proportion of the bottles have a fill-volume less than 168 ml?

(b) What proportion of the bottles have a fill-volume between 168 ml and 170 ml?

(c) What proportion of the bottles have a fill-volume between 165 ml and 167 ml?

(d) What proportion of the bottles have a fill-volume of exactly 170 ml?

(e) It is required to adjust the mean fill-volume to ensure that no more than 1% of bottles have a fill-volume below 168 ml. What should the new fill-volume be? (Assume that the standard deviation remains unchanged.)

3C. Discrete Distributions

Video Lecture https://youtu.be/EvpEUvy102I

Discrete random variables are limited to a particular set of values, as opposed to continuous random variables that can take on any value in a particular range. Counts are discrete random variables because the value is always an integer. There are two common distributions that describe data that consist of counts, namely the **binomial** distribution and the **Poisson** distribution.

The Binomial Distribution

The binomial distribution describes the number of occurrences of an event, on a fixed number of similar trials. On each trial the event may either occur or not occur.

Examples of binomial variables include: the number of bull's-eyes achieved when a player throws three darts; the number of 'heads' that occur when a coin is tossed ten times; the number of 'yes' answers given to a question posed to 1000 respondents in an opinion poll; the number of defective handsets in a sample of twenty handsets. These events could be called 'successes', but that name implies that all such events are desirable, so 'occurrences' is a better name. It is assumed that the trials are independent, i.e. if the first dart hits the bull's-eye, this does not affect the second throw. The probability of occurrence is the same on each trial, and does not change due to confidence, tiredness, etc.

For a binomial distribution with n trials, the sample space is {0, 1, 2, 3, … n}

The Poisson Distribution

The Poisson distribution describes the number of occurrences of an event, within a fixed interval of opportunity.

Examples of Poisson variables include: the number of fish caught in a day; the number of potholes on a 1 km stretch of road; the number of scratches on a smartphone screen; the number of cherries in a pot of cherry yogurt. These examples show that the interval can be an interval of time, or of length, or of area, or of volume. There is no limit, in theory, to the number of times the event could occur. These events could be called 'accidents', but that name implies that all such events are undesirable, so 'random events' is a better name. It is assumed that the events are independent, e.g. if a fish has been caught in the past 5 minutes, that does not make it any more or less likely that a fish will be caught in the next five minutes.

For a Poisson distribution, the sample space is {0, 1, 2, 3, …}

Problems 3C

#1. Which of the following variables are binomially distributed?
Q: the number of left-handed people sitting at a table of four people
R: the number of red cars in a random sample of 20 cars
S: the time taken by a show-jumping contestant to complete the course
T: the outdoor temperature on a summer's day
U: the number of months in a year in which there are more male than female births in a maternity unit
V: the number of weeds in a flower-bed
W: the weight of a banana
Y: the duration of a telephone call
Z: the number of players on a football team who receive an injury during a game.

#2. Which of the following variables are Poisson distributed?
N: the number of goals scored in a football match
O: the number of students per year that are kicked by horses in a riding school

P: the number of days in a week on which the Irish Independent and the Irish Times carry the same lead story
Q: the weight of a cocker spaniel puppy
R: the number of crimes committed in Dublin in a weekend
S: the number of telephone calls received per hour at an enquiry desk
T: the number of weeds in a flower-bed
U: the length of an adult's step
V: the number of bulls-eye's a player achieves with three darts.

3D. Binomial and Poisson Calculations

Video Lecture https://youtu.be/ylmxmzBKyTo

The Binomial Formula

Binomial probabilities can be calculated if the parameters, *n* and *p*, are known.

Formula 3.2

Binomial Distribution

$$P(r) = {}^{n}C_{r} \times p^{r} \times (1-p)^{n-r}$$

n = the number of trials

p = the probability of occurrence on a single trial

r = the exact number of occurrences

P(r) is the probability of exactly *r* occurrences.

EXAMPLE 10% of eggs are brown. If 6 eggs are selected at random and packed into a carton, what is the probability that exactly two are brown?

SOLUTION

n = 6

p = 0.1

r = 2

$P(r) = {}^{n}C_{r} \times p^{r} \times (1\text{-}p)^{n\text{-}r}$

$P(2) = {}^{6}C_{2} \times (0.1)^{2} \times (1 - 0.1)^{6\text{-}2}$

$P(2) = {}^{6}C_{2} \times (0.1)^{2} \times (0.9)^{4}$

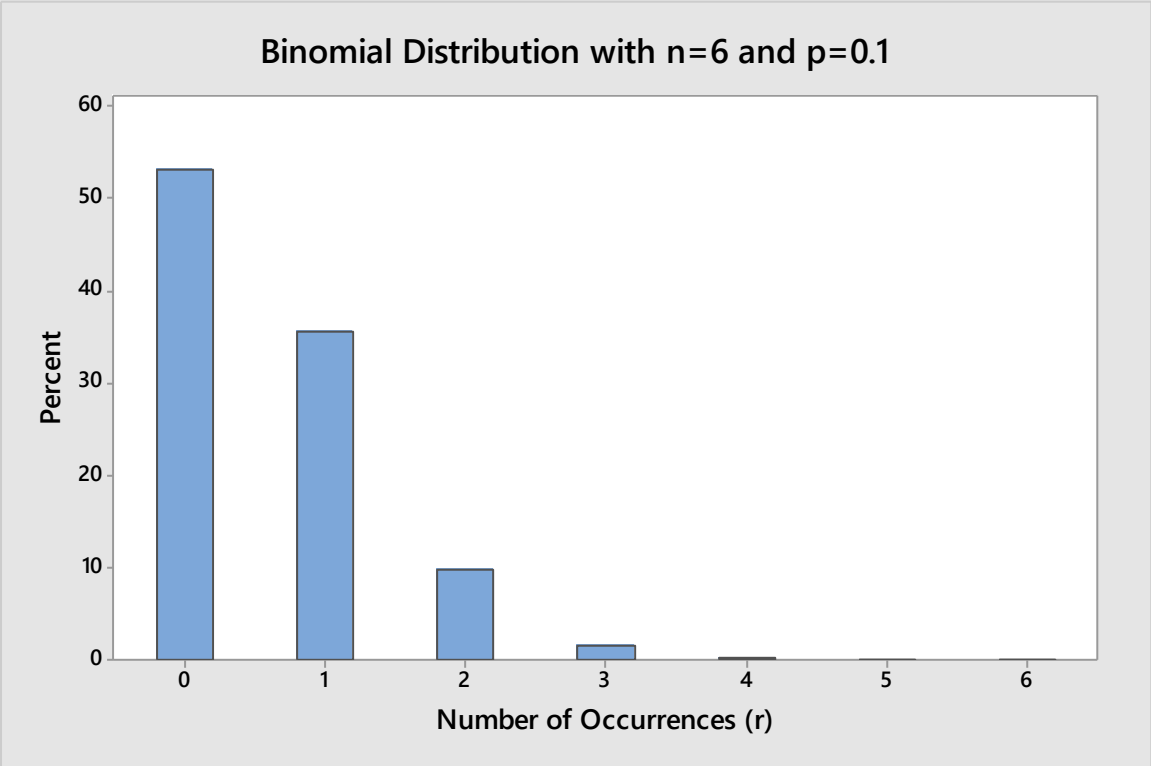
$P(2) = 15 \times 0.01 \times 0.6561$

$P(2) = 0.098415$

The probability that a carton contains exactly two brown eggs is 0.098415, or, to put it another way, 9.84% of all such cartons contain exactly two brown eggs.

The complete probability distribution is shown in Figure 3.6.

Fig 3.6



Sometimes we are not interested in an exact number of occurrences, but in any number of occurrences more than, or less than, some threshold value.

EXAMPLE

A jury consists of 12 citizens, selected at random, from a population which is 55% female. What is the probability that the jury will have at least three female members?

SOLUTION

$n = 12$

$p = 0.55$

'at least three' means that $r = 3$ or 4 or... or 12

P('at least three') = $P(3) + P(4) + ... + P(12)$

(The probabilities can be added because the exact outcomes are mutually exclusive.)

P('at least three') = $1 - [P(0) + P(1) + P(2)]$

(It is quicker to use the r values not on the list instead, and subtract their sum from 1.)

$P(0) = {}^{12}C_0 \times 0.55^0 \times (0.45)^{12} = 0.000069$

$P(1) = {}^{12}C_1 \times 0.55^1 \times (0.45)^{11} = 0.001011$

$P(2) = {}^{12}C_2 \times 0.55^2 \times (0.45)^{10} = 0.006798$

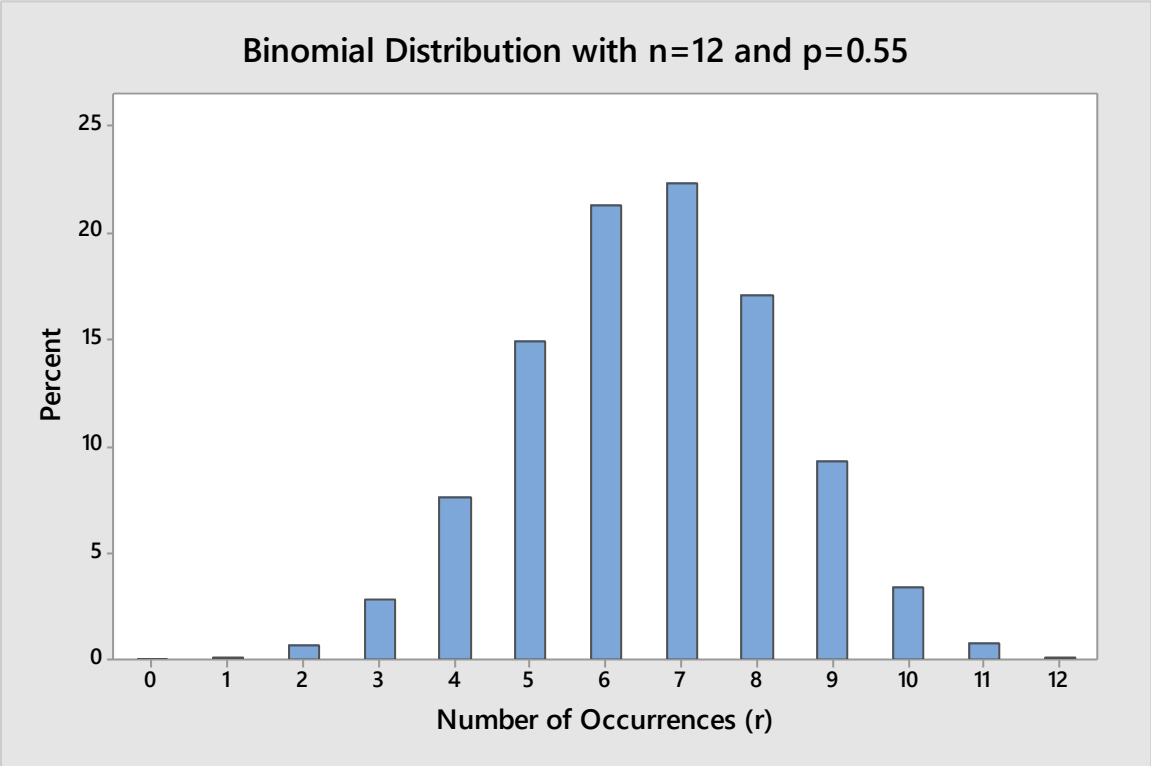
$P(0) + P(1) + P(2) = 0.007878$

$1 - 0.007878 = 0.992122$

99.21% of juries will have at least three female members.

The complete probability distribution is shown.

Fig 3.7



The Normal Approximation to the Binomial Distribution

We now consider an alternative solution to the jury problem.

Formula 3.3

Normal Approximation to the Binomial Distribution

$\mu = n.p$ and $\sigma = \sqrt{n.p(1-p)}$ are the parameters of the normal approximation to the binomial distribution, subject to the conditions $n.p > 5$ and $n.(1\text{-}p) > 5$.

$\mu = 12 \times 0.55 = 6.6$

$\sigma = \sqrt{(12 \times 0.55 \times 0.45)} = 1.723$

'at least three' means that $X > 2.5$

(Note the correction for continuity. Because the normal distribution is continuous and the binomial distribution is discrete, all the values between 2.5 and 3.5 are allocated to $r = 3$.)

$z = (2.5 - 6.6) / 1.723 = -2.38$

The normal table gives $p = 0.9913$.
99.13% of juries will have at least two female members.
This answer approximates closely to the exact answer obtained using the binomial distribution.

The Poisson Formula

Poisson probabilities can be calculated if the parameter, λ, is known.

Formula 3.4

Poisson Distribution

$$P(r) = \frac{e^{-\lambda} \times \lambda^r}{r!}$$

λ = the mean number of occurrences per interval

P(r) is the probability of exactly *r* occurrences.

EXAMPLE A company receives three complaints per day on average. What is the probability of receiving more than two complaints on a particular day?

SOLUTION $\lambda = 3$ and 'more than two' means that $r = 3$ or 4 or 5 or …

P('more than two') = $P(3) + P(4) + P(5) + \ldots$

P('more than two') = $1 - \{ P(0) + P(1) + P(2)\}$

$P(0) = e^{-3} \times 3^0 / 0! = 0.0498$

$P(1) = e^{-3} \times 3^1 / 1! = 0.1494$

$P(2) = e^{-3} \times 3^2 / 2! = 0.2240$

$P(0) + P(1) + P(2) = 0.4232$

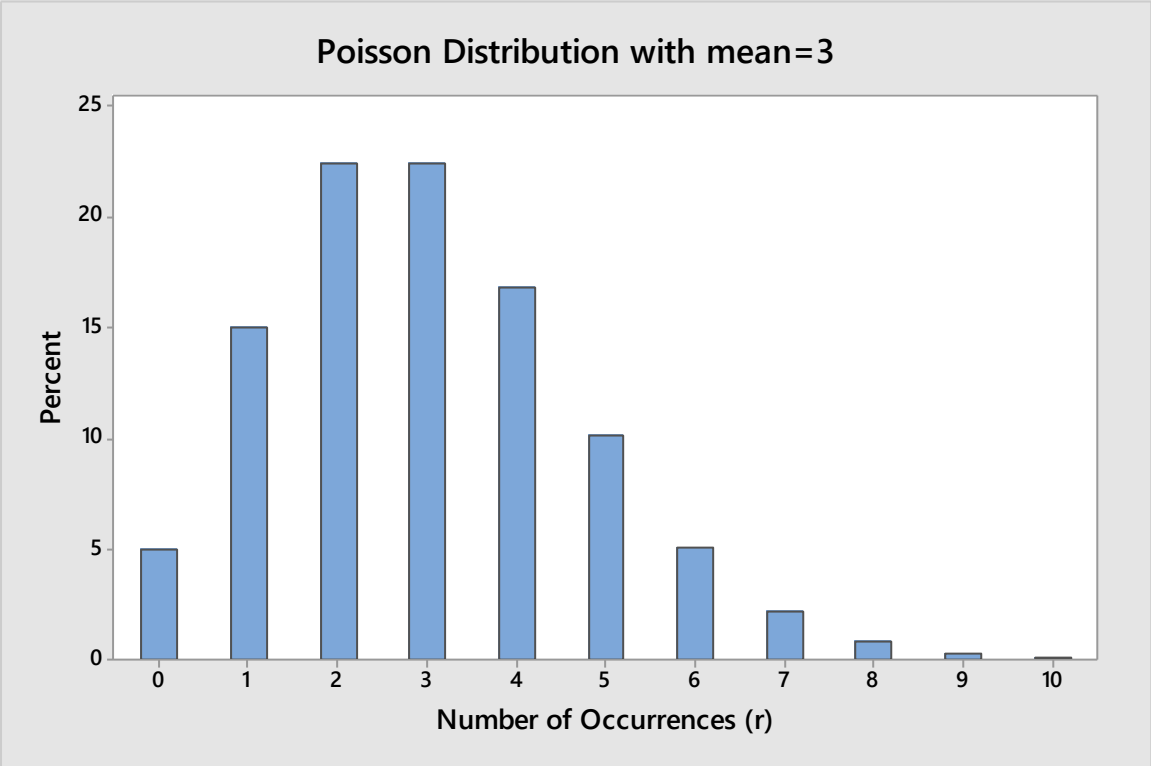
$1 - 0.4232 = 0.5768$

57.68% is the probability of receiving more than two complaints on a particular day.

The complete probability distribution is shown in the following graph.

Note that for large values of lambda, the Poisson distribution can be approximated by a normal distribution with $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$.

Fig 3.8



Problems 3D

#1. Construct the complete probability distribution for the number of heads obtained when a coin is tossed four times, i.e. calculate $P(0)$, $P(1)$, $P(2)$, $P(3)$, and $P(4)$. Verify that the sum of these probabilities is one, and explain in words why this is so.

#2. A certain machine produces bandages, 5% of which are defective. Every hour a sample of 6 bandages are selected at random and inspected. What proportion of these samples contain:
(a) exactly two defectives
(b) more than two defectives
(c) two defectives, at most
(d) fewer than two defectives
(e) at least two defectives?

#3. 10% of customers pay by direct debit. In a random sample of 300 customers, what is the probability that fewer than 40 pay by direct debit?

#4. 5% of airline passengers fail to show up for their flights. Out of a random sample of 150 passengers, what is the probability that more than 10 fail to show up?

#5. The number of blemishes that occur on a roll of carpet during manufacture is Poisson distributed, with mean 0.4 blemishes per roll. What percentage of the rolls are classified as:
(a) 'perfect', having no blemishes
(b) 'imperfect', having one or two blemishes
(c) 'scrap', i.e. all the rest?

#6. The number of customers who enter a shop per hour is Poisson distributed with mean 8. Calculate the percentage of hours in which:
(a) fewer than two customers enter
(b) more than two customers enter.

3E. Other Distributions

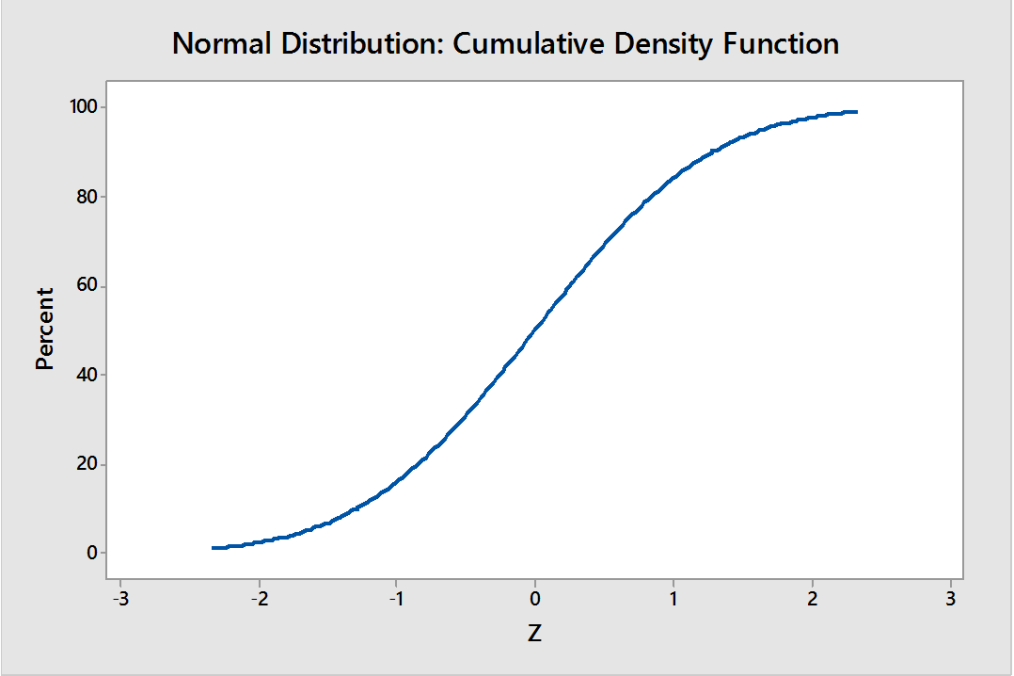
Video Lecture https://youtu.be/LpNtMOvGI8w

The following is a brief introduction to some other common distributions. The **exponential** distribution describes the intervals between random (i.e. Poisson) events. As such, it provides a simple model for product lifetimes, provided that failures occur at random. The **Weibull** distribution is a more flexible lifetime model that can take account of burn-in and wear-out fail modes. The **lognormal** distribution is a positively skewed distribution that is useful for describing populations where the larger values can be very large, such as total bacteria counts.

How can we know what distribution to fit to a set of data? Ideally, we may know that the process operates in a way that satisfies certain assumptions and so we choose the corresponding distribution. If we have no such prior knowledge, we can look at a histogram of the data to see whether its shape reminds us of some distribution. We can follow this up by checking the fit of one or more distributions using a **probability plot**. This procedure is explained here using the normal distribution.

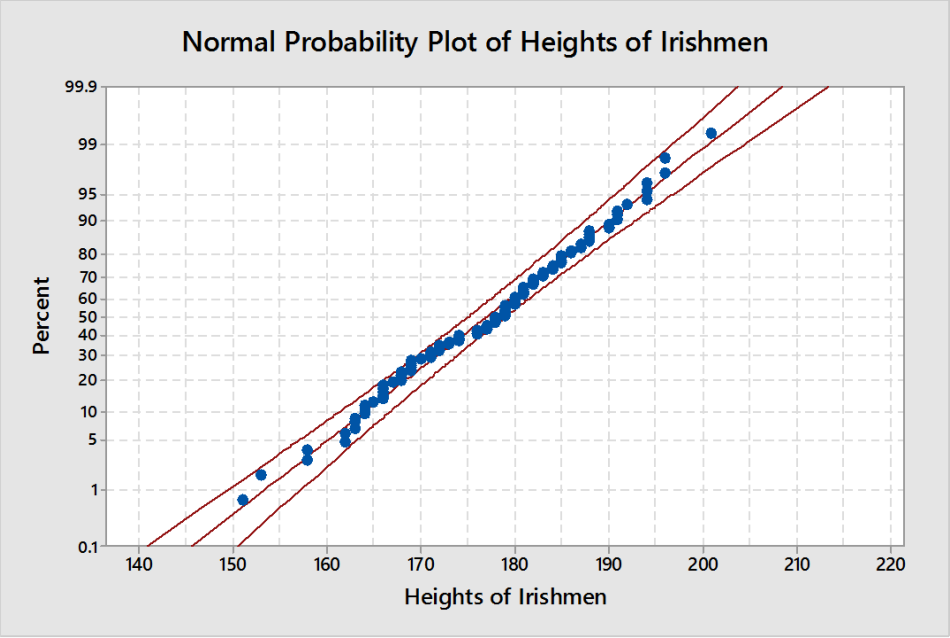
The normal probability density function was illustrated in Fig 3.3. The normal cumulative density function is shown below.

Fig 3.9



This curve is sigmoidal in shape, but if the two ends of the scale on the vertical axis were stretched, it could be made linear. This is what a probability plot does. It plots the cumulative probabilities using a suitably transformed vertical axis, so that if the data are normal, the graph will be linear.

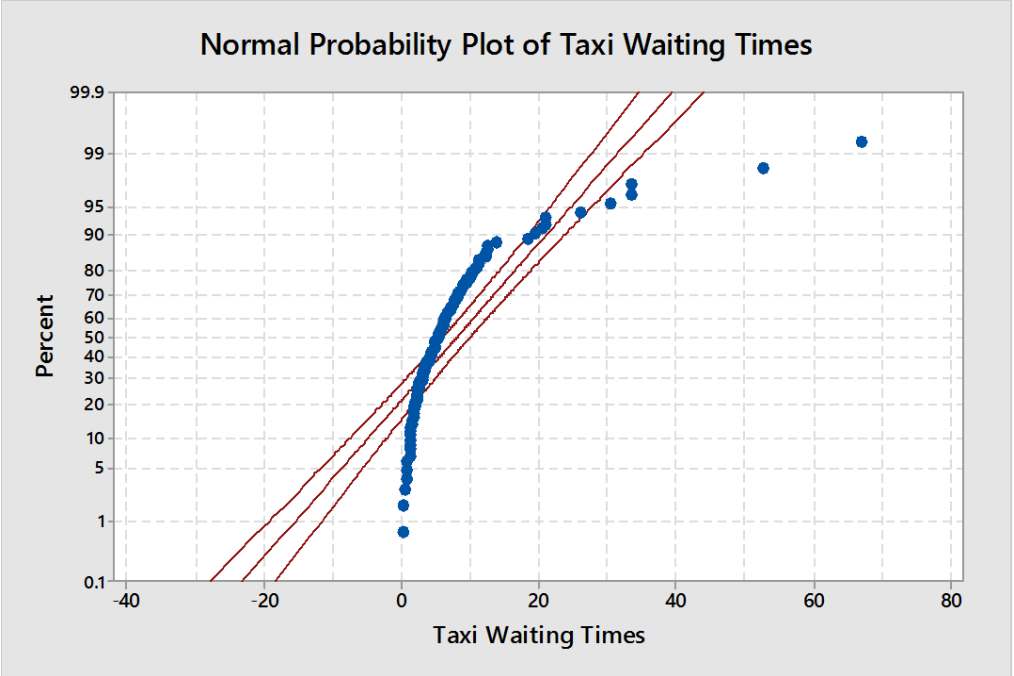
Fig 3.10



The plot above indicates that the heights of Irishmen are quite normal. There is some scatter, which we expect anyway, but the points are approximately linear.

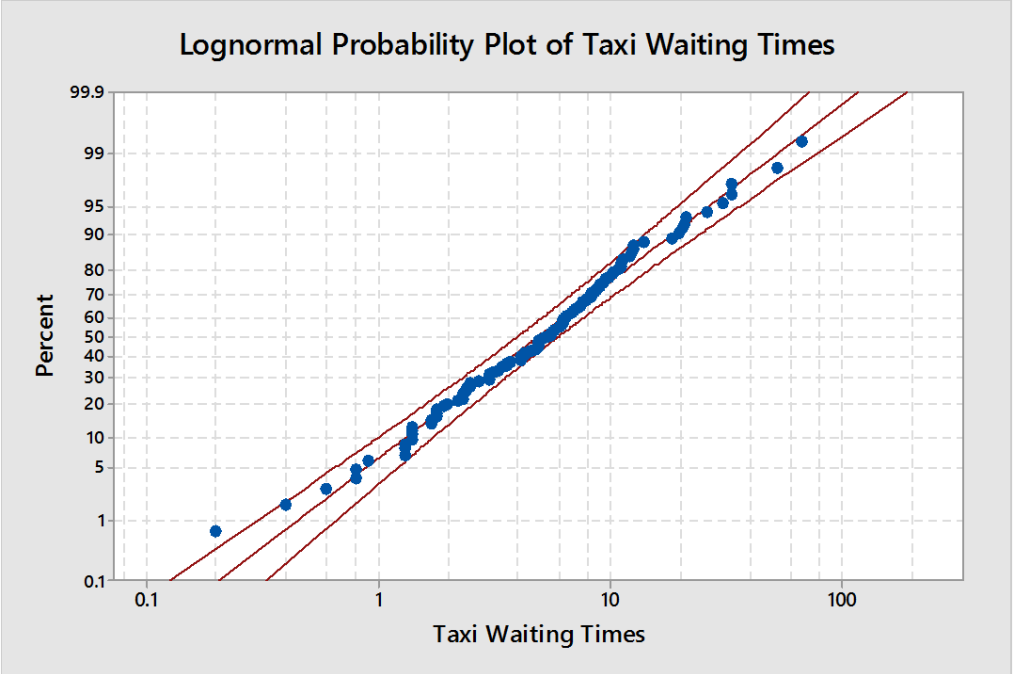
The graph below is a normal probability plot of the times spent waiting for a taxi. The graph shows clearly the data are not normal. This would also be obvious from a histogram.

Fig 3.11



A lognormal distribution is a good fit to the taxi data, as the following graph shows.

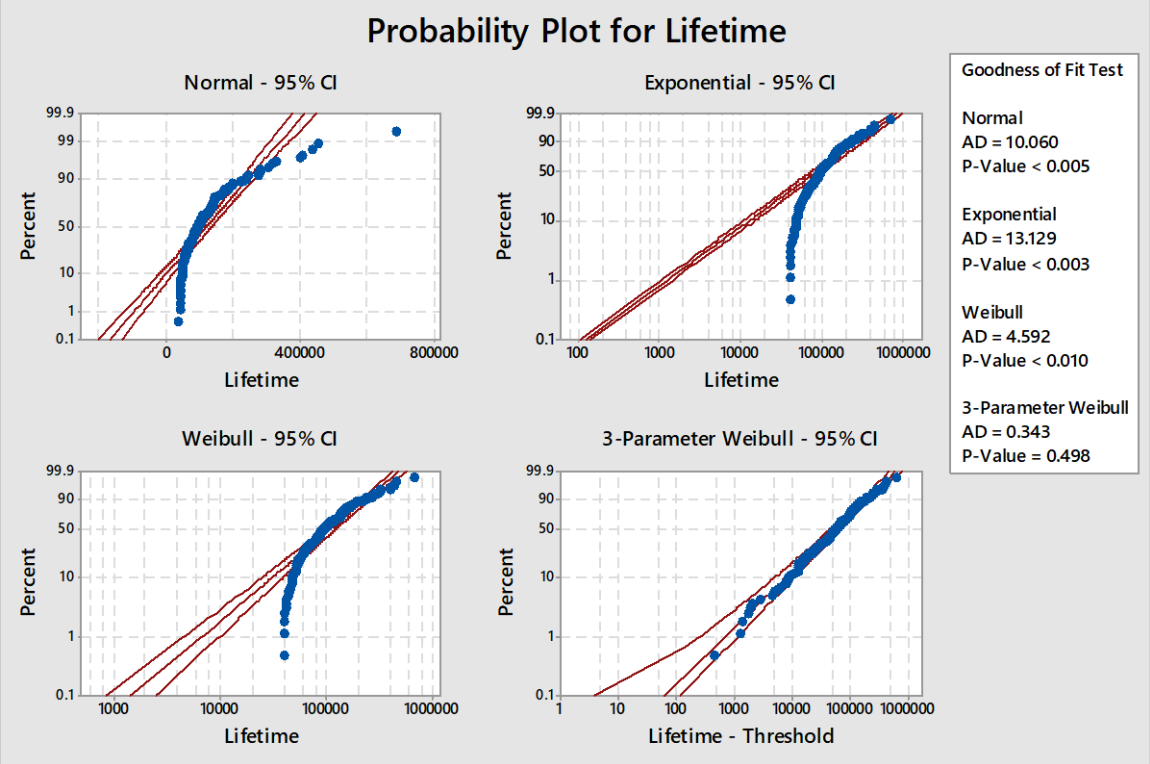
Fig 3.12



Problems 3E

#1. The lifetimes of vehicle bearings, in kilometres, are explored below using a number of alternative distributions in order to identify a model that would fit the data reasonably well. Can you say which of these distributions seems to fit best and why?

Fig 3.13



4

Making Estimates

Having completed this chapter you will be able to:

– estimate means, proportions and standard deviations;

– calculate confidence intervals;

– determine the sample size required to make an estimate;

– estimate the difference between means or proportions in two populations.

4A. How Samples Behave

Video Lecture https://youtu.be/ZgQ8I4Trhu8

We begin this chapter by exploring a number of ideas about sampling.

Random samples are better

Random samples are better than non-random samples because they tend to be unbiased. It often happens that we are interested in some large population and we wish to estimate the mean of that population. We can draw a random sample and use the sample mean as an estimate of the population mean. The estimate will not be perfect but it will be useful.

If we do this over and over again with the same population, then the sample means will be different each time because each sample will be different. Therefore, some of the sample means will overestimate the population mean and some will underestimate it. But the estimates won't all tend to be too high. And they won't all tend to be too low either. Overall, the estimates will be centred on the actual population mean. This is what **unbiased** means.

We say that the **expected value** of the sample mean is μ.

Formula 4.1

The Expected Value of the Sample Mean

$$E(\overline{X}) = \mu$$

To summarise, a sample mean is an unbiased estimator of the population mean, provided that the sample is random.

This is also true for proportions. If we take a random sample from a large population in order to estimate the proportion of the units in the population that have some particular attribute, then the sample proportion provides an unbiased estimate of the population proportion.

Bigger samples are better

When a random sample is drawn in order to estimate the population mean, it is better to use a larger sample rather than a smaller one, because the mean of a larger sample will usually be closer to the actual population mean. The typical difference between a sample mean and the population mean is called the standard error of the sample mean. In general, the **standard error** of an estimator is a measure of how close it is, typically, to the parameter that it estimates.

Formula 4.2

The Standard Error of the Sample Mean

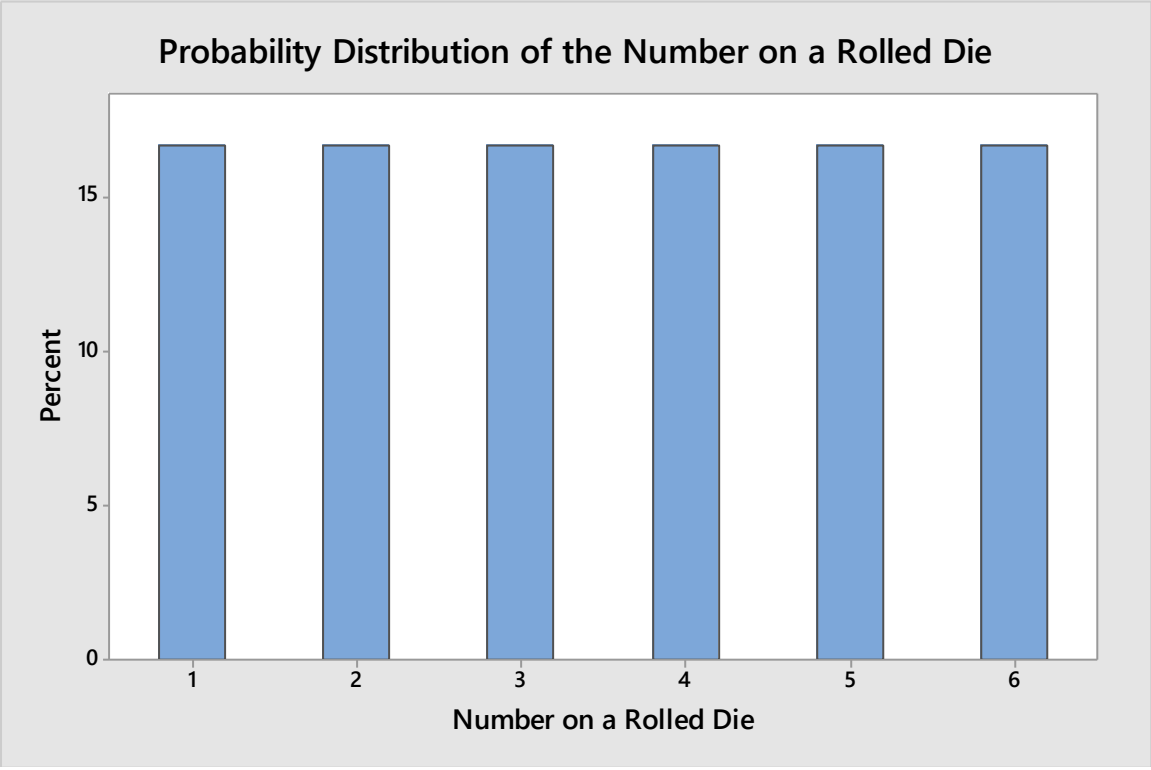
$$SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Notice that the standard error reduces as the sample size, *n*, increases. This confirms our intuition that a larger sample provides a more precise estimate of μ than a smaller sample.

Sample means follow a predictable pattern

This next discovery is very surprising. We already know that statistical distributions can have many different shapes. We might expect that the distribution of sample means from different populations would also have different shapes. But this is not so! Sample means all have the same distribution, no matter what population the samples are taken from. To illustrate this, we will consider the population of individual numbers on a rolled die. This population has a uniform shape.

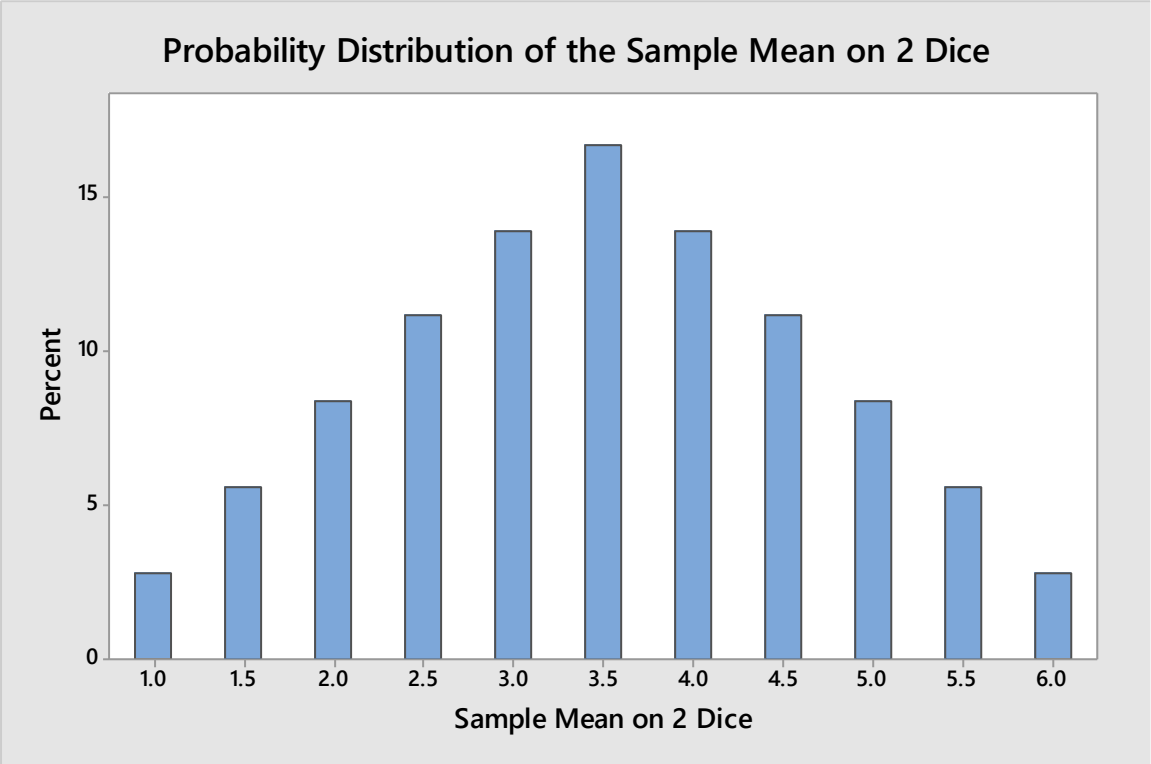
Fig 4.1



Now consider sample means for *n* = 2 dice.

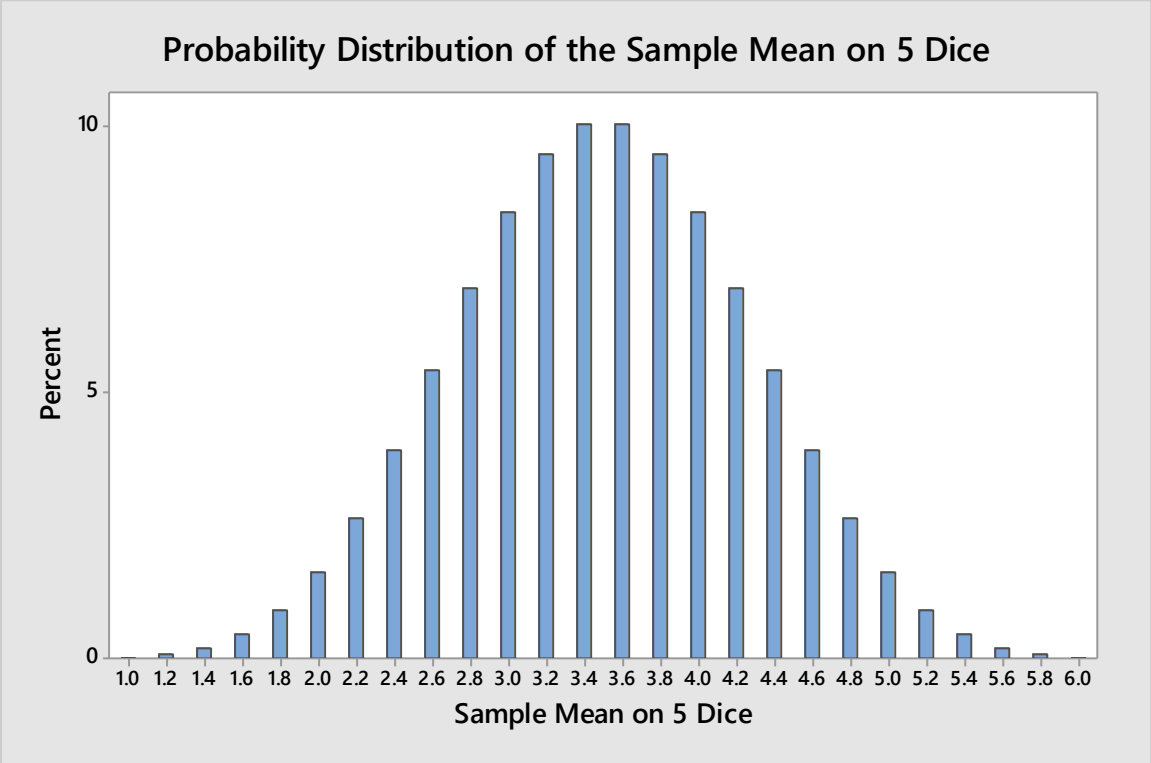
The sample mean is more likely to be in the middle (e.g. 3) than at one of the extremes (e.g. 6). This is because there are many ways to get a mean of 3 (e.g. 3&3 or 2&4 or 4&2 etc.) but there is only one way to get a mean of 6 (i.e. 6&6).

Fig 4.2



Now consider sample means for n = 5 dice.

Fig 4.3



Notice that the distribution is normal. It turns out that sample means taken from any population are normally distributed if the samples are big enough. And how big is 'big enough'? If the parent population is normal, then $n = 1$ is big enough. If it is symmetric, then $n = 5$ is usually big enough. Very skewed populations require larger n. The really surprising thing is that any population, even if it's not a normal population, will produce a normal distribution of sample means.

Our series of discoveries since the beginning of this chapter can be summarised by a particular formulation of the **Central Limit Theorem**.

Formula 4.3

The Central Limit Theorem

When random samples of size n are drawn from **any** population with mean μ and standard deviation σ, the sample means tend to form a normal distribution, with expected value μ, and standard error $\sigma/\sqrt{n}$.

Problems 4A

#1. The standard deviation of the weights of food portions is 20 grams. The population mean is estimated by taking a random sample. What is the standard error of this estimate if the sample size is:
(a) 4 portions?
(b) 100 portions?

#2. The population of taxi waiting times has a skewed distribution with mean 8 minutes and standard deviation 10 minutes. Random samples of size 30 are drawn from this population.
(a) What is the expected value of the sample mean?
(b) What is the standard error of the sample mean?
(c) What is the distribution of the sample mean?

4B. Confidence Interval for a Mean or Proportion

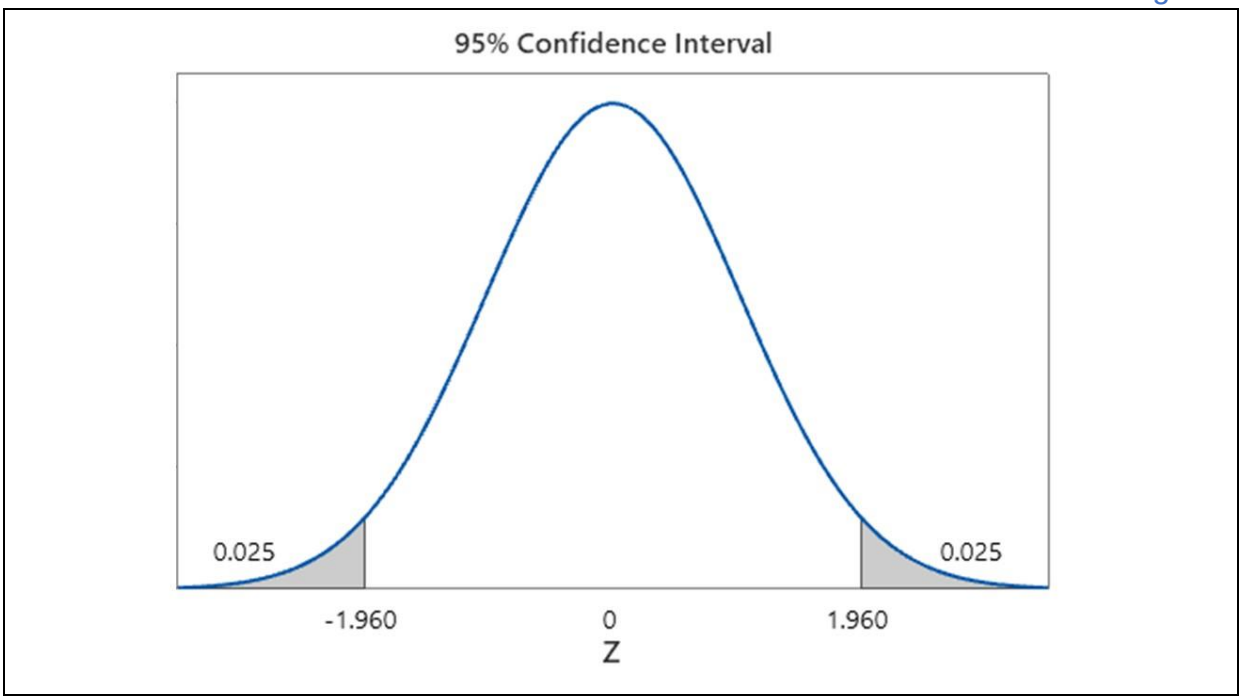
Video Lecture https://youtu.be/5aZ0myUR3k0

Confidence Interval for a Population Mean (when sigma is known)

We now know the distribution of the sample mean. If we take many random samples, each of size n, from any population, we know that the sample means will tend to form a normal distribution, with expected value μ, and standard error $\sigma/\sqrt{n}$. But we also know that there is a probability of 95% that a value from a normal distribution has a z-score lying inside the interval between -1.96 and +1.96. Therefore, if we take many random samples, 95% of the sample means will lie within 1.96 standard errors of the population mean.

It follows that if we draw just one random sample, there is a 95% probability that the sample mean will lie within 1.96 standard errors of the population mean. On this basis, we can use the sample mean to construct a **confidence interval** that contains the unknown population mean with 95% confidence.

Fig 4.4



Formula 4.4

95% Confidence Interval for a Population Mean (if σ is known)

$$\mu = \overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

This formula can be used only if σ is known. There are situations where this is the case. Suppose I pick up a stone and weigh it 1000 times on a kitchen scale. The results will differ slightly from each other, because the scale is not perfect. I can calculate the mean of the results, and this tells me something about the stone; it tells me how heavy it is. I can also calculate the standard deviation of the results, but this tells me nothing about the stone. It tells me something about the scale; it tells me how precise it is. Now if you pick up a different stone and begin to weigh it on the same scale, we know that the standard deviation for your results will be the same as for mine, because you are using the same weighing process. So we can use the formula above to estimate your process mean, because we already know the standard deviation from previous experience. In summary, if we have prior knowledge of the standard deviation of a process, we can use this knowledge to estimate the current process mean.

EXAMPLE The cycle times of a washing machine are known to be normally distributed with a standard deviation of 2 seconds. Construct a 95% confidence interval for the population mean cycle time, if a random sample of three cycles, measured in seconds, were as follows: 2521, 2526, 2522.

SOLUTION

$\bar{X}$ = 2523

σ = 2

$n = 3$

$\mu = 2523 \pm 1.96 \times 2 / \sqrt{3}$

$\mu = 2523 \pm 2.26$

$2520.74 < \mu < 2525.26$

We can state with 95% confidence that the mean cycle time, of all the cycles of this washing machine, lies between 2520.74 and 2525.26 seconds. On 95% of occasions when we construct such a confidence interval, the interval will include the population mean. There is a 5% probability that we selected an unfortunate sample, leaving the population mean lying outside the interval. Other confidence levels, such as 99%, could be used instead, but this book uses 95% throughout. The simple estimate, 2523, is called a **point estimate**.

Confidence Interval for a Population Mean (when sigma is unknown)

It would be great if the previous formula could be used with a population having an unknown standard deviation. Perhaps we could replace σ in the formula with its sample estimate, S. However, because S is only an estimate of σ, we would lose some of our confidence, and we would have less than 95% confidence in the result. One way to restore our confidence would be to widen the interval by using a number bigger than 1.96. But the problem is, what number should we use? This problem was solved in 1908 by William Sealy Gossett, a statistician at Guinness's brewery in Dublin. Writing under the pen-name 'Student', Gossett gave us the *t*-distribution (see Table 2 in the appendix of Statistical Tables). The *t*-distribution is closely related to the normal distribution, but while the normal *z*-score measures in units of σ, the *t*-score measures in units of S. The *z*-score can be defined as 'the number of standard deviations that a value is above the mean', but the *t*-score is defined as 'the number of estimated standard deviations that a value is above the mean'. There is a different *t*-distribution for every different number of degrees of freedom, n-1, on which the standard deviation estimate is based. The great thing is that we now have a confidence interval formula that requires no prior knowledge of the population; all we need is a random sample. Formula 4.4 is a special case of Formula 4.5, because $t_\infty = z$. This formula represents the culmination of all the preceding theory in this chapter, and it is immensely useful.

Formula 4.5

95% Confidence Interval for a Population Mean (if σ is unknown)

$$\mu = \overline{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$$

EXAMPLE A random sample of flight times last year between Dublin and Edinburgh, gave the following results, in minutes: 45, 49, 43, 51, 48.
Calculate a 95% confidence interval for the population mean.

SOLUTION

$\bar{X} = 47.2$

$S = 3.19$

$n = 5$

To find the value of t, we refer to the table of the t-distribution.
The degrees of freedom is n-1 and in this case $n = 5$ so $df = 4$.
And because the confidence interval is two-sided, we read from the 2.5% column.
This assigns 2.5% to both the upper and lower tails and leaves 95% in the interval.

$t_{n-1} = 2.776$

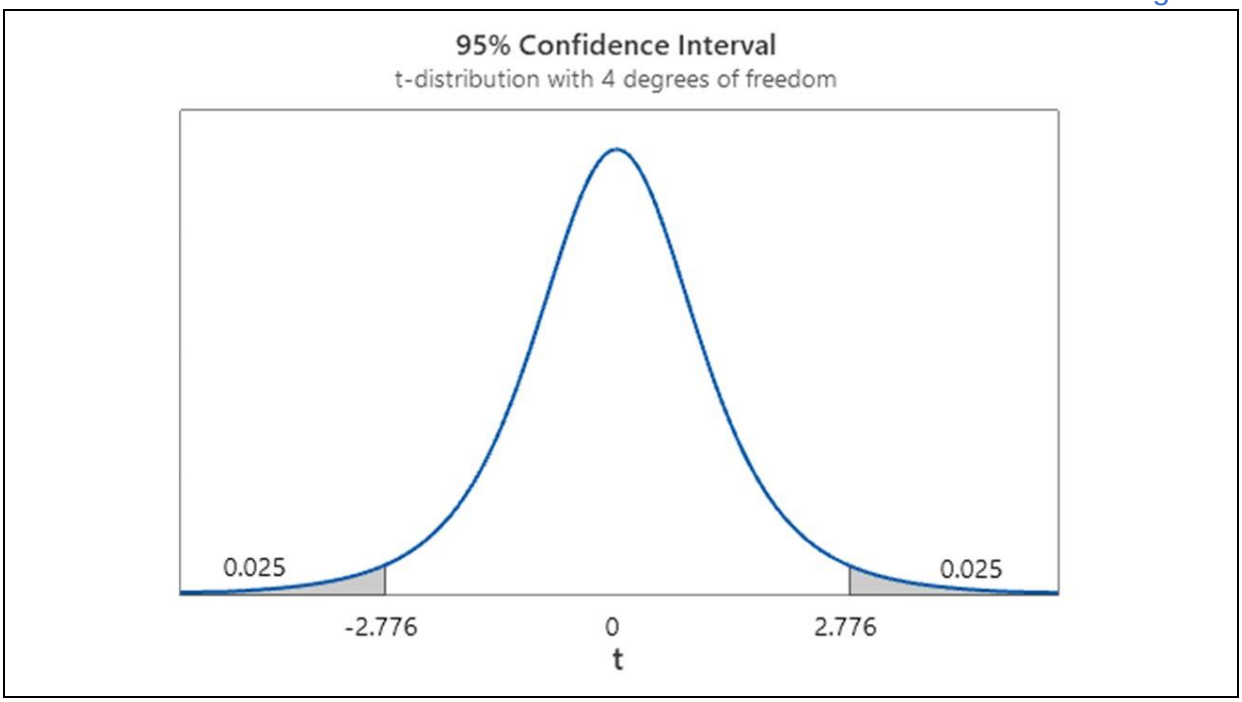
$\mu = 47.2 \pm 2.776 \times 3.19 / \sqrt{5}$

$\mu = 47.2 \pm 3.96$

$43.24 < \mu < 51.16$

We can state with 95% confidence that the mean flight time, for all flights between Dublin and Edinburgh last year, lies between 43.24 and 51.16 minutes.

Fig 4.5



Confidence Interval for a Population Proportion

A population proportion means the proportion of units in a population that possess some attribute of interest, e.g. a population could consist of customers, and the attribute could be that the customer pays with a debit card, or a population could consist of products, and the attribute could be that the product is defective. We use the Greek letter π to denote the population proportion. If a random sample of n units includes r units that possess the attribute, then the sample proportion, $P = r/n$, is an unbiased estimate of the population proportion, π, with standard error $\sqrt{[\pi(1-\pi)/n]}$. If π is unknown, the standard error is also unknown, but it can be estimated by $\sqrt{[P(1-P)/n]}$. Hence the following formula provides an approximate confidence interval for π. An

exact confidence interval, based on the binomial distribution, can be obtained by an iterative procedure, but it is too tedious for calculation by hand.

Formula 4.6

Approximate 95% Confidence Interval for a Population Proportion

$$\pi = P \pm 1.96\sqrt{\frac{P(1-P)}{n}}$$

EXAMPLE Out of 50 students randomly selected from a campus, 16 had been bitten by a dog at some time. Calculate an approximate 95% confidence interval for the population proportion.

SOLUTION

P = 16 / 50 = 0.32

π = 0.32 ± 1.96 √ [(0.32 × (1 - 0.32) / 50]

π = 0.32 ± 0.13

0.19 < π < 0.45

We can state with approximately 95% confidence that between 19% and 45% of all students at the campus had been bitten by a dog at some time.

Problems 4B

#1. A process that fills bags with rice is known to have a standard deviation of 0.5 grams. A random sample of bags filled for a certain customer had weights as follows, in grams.

497.15, 498.21, 497.93, 497.46, 498.91, 497.61

Calculate a 95% confidence interval for the population mean. Express your answer in words.

#2. It is known that the standard deviation of the diameters of plastic tubes, made by a certain process, is 0.05 mm. Five tubes were randomly selected from today's production and their diameters were measured, with the following results in mm. Calculate a 95% confidence interval for the population mean. Express your answer in words.

12.01, 12.05, 12.08, 12.02, 12.11

#3. A random sample of service calls by a certain engineer involved journeys of the following distances, in km. Calculate a 95% confidence interval for the population mean. Express your answer in words.

17.1, 9.2, 4.0, 3.1, 20.7, 16.1, 11.0, 14.9

#4. A large number of delegates attended a week-long conference. A random sample of five delegates was selected, and the expenses that they each had incurred on food and accommodation was observed, in euro, as follows.

848, 884, 902, 721, 812

Calculate a 95% confidence interval for the population mean. Express your answer in words.

#5. A random sample of 200 booklets was selected from a large consignment, and 60 of these were found to have defective binding. Calculate an approximate 95% confidence interval for the population proportion. Express your answer in words.

#6. Out of 633 students randomly sampled at a university campus, 105 walk from home to their classes. Calculate an approximate 95% confidence interval for the population proportion. Express your answer in words.

Project 4B

Estimation of a Mean

Select a random sample of ten measurements from any large population of your choice and write a report consisting of the following sections.

(a) Identify the population of interest and the measurement of interest.
(b) Describe in detail how the sample was selected from the population.
(c) Show the data, and calculate the sample mean and the sample standard deviation.
(d) Calculate a 95% confidence interval for the population mean.
(e) State clearly in words what the confidence interval tells us.

4C. Sample Size for Estimation

Video Lecture https://youtu.be/MrVK7fQWPbk

Sample Size for Estimating a Population Mean

Larger samples provide more precise estimates. To determine the sample size required in any situation, we first need to identify some practical size, $\pm\delta$ (delta), for the **margin of error** that can be tolerated. Secondly, we require an estimate of the standard deviation to use as a **planning value**. If no such estimate is available, we simply proceed with a **pilot sample** of any convenient size. The standard deviation of the pilot sample is used as a planning value to calculate the full sample size required, and then more data can be collected as necessary. The value for the sample size suggested by the formula must always be rounded up to the next integer, because a sample size must be an integer, and any sample smaller than what the formula suggests is too small.

Formula 4.7

Sample Size for Estimating a Population Mean

$$n = \left(\frac{1.96 \times \sigma}{\delta} \right)^2$$

EXAMPLE A random sample of four fish was taken at a fish-farm. The weights of the fish, in grams, were 317, 340, 363 and 332. How many fish must be weighed in order to estimate the population mean to within ± 10 grams, with 95% confidence?

SOLUTION

S = 19.2, the sample standard deviation is the planning value

δ = 10, the margin of error

$n = (1.96 \times 19.2 / 10)^2$

n = 14.16, rounded up becomes 15.

We require 15 fish. Since we already have 4 fish in the pilot sample, 11 more fish must be sampled to bring the **cumulative sample** size up to 15.

Sample Size for Estimating a Population Proportion

To determine the sample size required to estimate a population proportion, we require an initial estimate of the population proportion to use as a **planning value**, P. This can be based on a pilot sample or on our prior knowledge of similar populations. Alternatively, we can use P = 0.5, which is the most conservative planning value, and will lead to a sample size that is certainly big enough. Sample sizes required for estimating proportions are typically much larger than those required for estimating means, because each observation provides only a 'yes' or 'no' response. Also, larger samples are required to estimate proportions that are close to 0.5, compared to proportions that are close to either 0 or 1.

If the population is finite, the calculated sample size for estimating a mean or a proportion can be reduced by multiplying it by 1-f where f is the sampling fraction, $f = n/N$, but since f is usually very small it is common practice not to bother about this finite population correction.

Formula 4.8

Sample Size for Estimating a Population Proportion

$$n = \frac{1.96^2 \times P(1-P)}{\delta^2}$$

EXAMPLE How large a sample of voters is required to estimate, to within 1%, the level of support among the electorate for a new proposal to amend the constitution?

SOLUTION

P = 0.5, conservative planning value

δ = 0.01, the margin of error is 1%

$n = 1.96^2 \times 0.5 \times (1 - 0.5) / 0.01^2$

n = 9604. We require a sample of 9604 voters.

Problems 4C

#1. The lifetimes of three rechargeable batteries randomly selected from a batch were measured, with the following results, in hours: 168, 172, 163. How large a sample is required to estimate the population mean to within ± 1 hour, with 95% confidence?

#2. The standard deviation of the lengths of roof-tiles from a certain supplier is known to be 3 mm. How large a sample is required to estimate the population mean to within ± 0.5 mm, with 95% confidence?

#3. How large a sample is required to estimate, to within 10%, the proportion of pizza orders from a certain outlet that require delivery?

#4. How large a sample is required to estimate, to within 3%, the level of support for a political party which traditionally receives 22% support in opinion polls?

Project 4C

Estimation of a Proportion

This is a group project. Students arrange themselves into groups of two, three or four students. The group then identifies a large population and draws a random sample of fifty units from that population. Each student selects a different attribute and observes the number of occurrences of their particular attribute in the sample. Each student analyses the data for their own attribute, and writes an individual report on the entire project in relation to that attribute. The report must consist of the following sections.

(a) Identify the population of interest and the attribute of interest.
(b) Describe in detail how the sample was selected from the population.
(c) Calculate an approximate 95% confidence interval for the population proportion.
(d) State clearly in words what the confidence interval tells us.
(e) Calculate the sample size required to estimate the population proportion to within ±1% with 95% confidence.

4D. Estimating a Standard Deviation

Video Lecture https://youtu.be/urUAlOe8Jsw

Estimating σ Using a Single Sample

The sample standard deviation estimate, S, is based on n-1 degrees of freedom. It is an unbiased estimator of the population standard deviation, σ. The divisor is n-1 rather than n, to compensate for the fact that the estimated deviations are 'too small' as a result of being measured from the sample mean rather than the population mean.

Estimating σ Using Two or More Samples

If samples are available from two or more populations, that are known or assumed to have equal variance, the information in the samples can be combined to provide a **pooled variance estimate**. In practice, the equal variance assumption is sometimes made to simplify the analysis. If the samples are of equal size, the pooled variance estimate is simply the mean of the sample variances. If the samples are of unequal size, a **weighted mean** must be calculated, giving weight to each sample variance in proportion to the number of degrees of freedom on which it is based. This formula for two samples can easily be extended to deal with more than two samples.

Formula 4.9

Pooled Variance Estimate

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 - 1 + n_2 - 1}$$

Degrees of freedom = n_1-1 + n_2-1 = n_1 + n_2 - 2

EXAMPLE

A depth sounder was used to take three repeat measurements at a fixed point near the coast, and another six measurements at a fixed point further out to sea. The results, in metres, were as follows: 28, 29, 26 (coast) and 36, 38, 37, 35, 35, 36 (sea). Estimate the standard deviation of repeat measurements for this device.

SOLUTION

$n_1 = 3$

$S_1 = 1.528$

$n_2 = 6$

$S_2 = 1.169$

$S^2 = [(3\text{-}1) \times 1.528^2 + (6\text{-}1) \times 1.169^2] \div (3\text{-}1 + 6\text{-}1)$

$S^2 = 1.643$ is the pooled variance estimate, and

$S = 1.282$ is the pooled standard deviation estimate.

Confidence Interval for σ

When random samples are drawn repeatedly from a normal population, the distribution of the sample standard deviations is related to a **chi-square distribution**. This is a skewed distribution that is bounded below by zero. The exact shape of a chi-square distribution depends on its degrees of freedom, n-1. The typical shape of a chi-square distribution is illustrated in *Fig 5.9* in chapter 5. We can use the upper and lower 2.5% points of the distribution to construct a 95% confidence interval for σ. Notice that the interval is not symmetric about the point estimate.

Formula 4.10

Confidence Interval for σ

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{upper}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{lower}}}$$

This formula applies to a normally distributed population.

EXAMPLE

A random sample of five loaves had the following lengths, in cm:

26.9, 28.4, 31.6, 27.8, 30.2

Calculate a 95% confidence interval for sigma, the population standard deviation.

SOLUTION
$n = 5$

$S = 1.898$

degrees of freedom, $df = 4$

From the chi-square table, we find

upper 2.5% point = 11.143

lower 2.5% point = 0.4844 (the upper 97.5% point)

$\sqrt{[4\ (1.898)^2 \div 11.143)]} = 1.137$

$\sqrt{[4\ (1.898)^2 \div 0.4844)]} = 5.454$

$1.137 < \sigma < 5.454$

We can state with 95% confidence, assuming normality, that the standard deviation of the lengths of all the loaves lies between 1.137 and 5.454 cm.

Problems 4D

#1. Kevin weighed himself three times in succession on a scale and the readings, in kg, were:

25.4, 25.6, 24.9

Rowena weighed herself twice on the same scale and her results were:

31.2, 31.9

Use all the data to estimate the standard deviation of repeat measurements for this scale.

#2. A 100 m hurdler was timed on five practice events, with the following results, in seconds.

13.01, 13.05, 12.99, 13.06, 13.03

Calculate a 95% confidence interval for σ.

4E. Estimating a Difference between Means or Proportions

Video Lecture https://youtu.be/PqKgcgtqfhA

Comparing Means Using Paired Samples

Often we are interested in the difference between two population means, rather than the means themselves, for example, how much more expensive one retail outlet is compared to another. If possible, it is best to use **paired data**. This means that we sample the same items in the two retail outlets. This gives rise to pairs of values, one from each population. The advantage of pairing is that we are comparing like with like, and so neither population is favoured by the list of items that are sampled.

EXAMPLE The prices of a number of randomly selected greengrocery items were compared in two outlets, Max's and Ben's. The following data show the price in cent, for each item, in each outlet.

Table 4.1

Item	Max	Ben
Pears	119	135
Lettuce	109	125
Bananas	123	115
Mango	89	123
Avocado	75	105
Lemon	32	48
Celery	99	115

SOLUTION

First, we calculate the difference between each pair of values.

The differences (Max's minus Ben's) are:

-16, -16, +8, -34, -30, -16, -16

Then we use Formula 4.5 to estimate the mean difference.

$\bar{X}$ = -17.14, the mean difference

S = 13.46

n = 7

t_{n-1} = 2.447

μ = -17.14 ± 2.447 × 13.46 ÷ √7

μ = -17.14 ± 12.45

-29.59 < μ < -4.69

We can state with 95% confidence that Max's is between 4.69 and 29.59 cent less expensive than Ben's, per item, on average.

Comparing Means Using Independent Samples

We may wish to compare the means of two populations in a situation where it is not feasible to obtain paired data. In such cases, we draw two independent random samples, one from each population. Assuming that the population variances are equal, we first compute the pooled variance estimate using Formula 4.9, and then use the Formula 4.11 to construct the confidence interval.

Formula 4.11

Confidence Interval for Difference Between Means

$$\mu_1 - \mu_2 = \overline{X}_1 - \overline{X}_2 \pm t\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}$$

S^2 is the pooled variance estimate, assuming equal variances.

EXAMPLE Two independent random samples were selected, of apartments in Lucan and Naas. All apartments were of a similar size and standard. The monthly rents, in euro, are shown below. Calculate a 95% confidence interval for the population mean difference in monthly rents.

SOLUTION

Lucan: 950, 875, 850, 900.
Naas: 825, 850, 875, 880, 860.
First use Formula 4.9 to find S^2
$n_1 = 4$
$S_1 = 42.70$
$n_2 = 5$
$S_2 = 21.97$
$S^2 = [(4\text{-}1) \times 42.70^2 + (5\text{-}1) \times 21.97^2] \div (4\text{-}1 + 5\text{-}1)$
$S^2 = 1057.23$

Now use Formula 4.11 to construct the confidence interval.

$\bar{X}_1 = 893.75$
$\bar{X}_2 = 858.00$
$df = 7$
$t = 2.365$, from the t tables
$\mu_1 - \mu_2 = 893.75 - 858.00 \pm 2.365 \sqrt{} (1057.23 \div 4 + 1057.23 \div 5)$
$\mu_1 - \mu_2 = 35.75 \pm 51.58$
$-15.83 < \mu_1 - \mu_2 < 87.33$

We can state with 95% confidence that Lucan is between €15.83 less expensive and €87.33 more expensive than Naas for apartment rental, on average. Because the confidence interval includes zero, it may be that the mean difference is zero.

Difference between Proportions

When we estimate a difference between proportions in this context, we are referring to proportions in two distinct populations. For example, we might estimate the difference between the proportion of Wicklow voters who support the Labour Party and the proportion of Mayo voters who support the Labour Party. We would not use this formula to estimate the difference between the proportion of Wicklow voters who support the Labour Party and the proportion of Wicklow voters who support the Green Party.

Formula 4.12

Approximate 95% Confidence Interval for Difference Between Proportions

$$\pi_1 - \pi_2 = P_1 - P_2 \pm 1.96\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

EXAMPLE In a sample of 200 shoppers in Liffey Valley Shopping Centre, 98 purchased clothing. In a sample of 300 shoppers in Blanchardstown Shopping Centre only 34 purchased clothing. Calculate a 95% confidence interval for the difference between the population proportions.

SOLUTION

$n_1 = 200$ $P_1 = 98 / 200 = 0.49$

$n_2 = 300$ $P_2 = 34 / 300 = 0.1133$

$\pi_1 - \pi_2 = (0.49 - 0.1133) \pm 1.96 \sqrt{} [0.49 (1 - 0.49) \div 200 + 0.1133 (1 - 0.1133) \div 300]$

$= 0.3767 \pm 0.0780$

$0.2987 < \pi_1 - \pi_2 < 0.4547$

We are 95% confident that between 29.87% and 45.47% more shoppers purchased clothing in Liffey Valley Shopping Centre, compared to Blanchardstown Shopping Centre.

Problems 4E

#1. Six internal wall-to-wall measurements were taken with a tape measure, and the same measurements were also taken with a handheld laser device. The results, in cm, are shown below, in the same order each time. Calculate a 95% confidence interval for the mean difference.

Tape Measure:	602,	406,	478,	379,	415,	477
Laser Device:	619,	418,	483,	386,	413,	489

#2. Calculate a 95% confidence interval for the difference in weight between Kevin and Rowena. The data are provided in problems 4D #1.

#3. An industrial process fills cartridges with ink. At 10 a.m. a sample of fills were measured, with the following results, in ml:

55, 49, 57, 48, 52

At 11 a.m. another sample was taken, with the following results, in ml:

53, 59, 50, 49, 52

Calculate a 95% confidence interval for the difference between the population means.

#4. Out of a random sample of 150 visitors at a theme park in summer, 15 had prepaid tickets, but in a random sample of 800 visitors at the theme park in winter, only 9 had prepaid tickets. Calculate a 95% confidence interval for the difference between the population proportions.

Project 4E #1

Estimation of a Mean Difference using Paired Samples

Calculate a confidence interval for the mean price difference per item between two online retailers that sell some identical products. Select a sample of eight prices from each retailer. Write a report consisting of the following sections.

(a) Provide the names of the two retailers.

(b) Describe in detail how you selected the samples and show the sample data.

(c) Calculate the confidence limits.

(d) Clearly state your conclusion in words.

Project 4E #2

Estimation of a Difference between Means using Independent Samples

Calculate a confidence interval for the difference between the means of two populations. There must be a reason why this difference is important to you. Select two samples, one from each population, and each sample should include at least six observations. Write a report consisting of the following sections.

(a) Identify the two populations and the measurement of interest. Explain why the difference between the means of these two populations is important to you.

(b) Describe in detail how you selected the samples and show the sample data.

(c) Calculate the confidence limits.

(d) Clearly state your conclusion in words.

Project 4E #3

Estimation of a Difference between Proportions

Calculate a confidence interval for the difference between two population proportions. There must be a reason why this difference is important to you. Select two samples, one from each population, and each sample should include at least one hundred observations. Write a report consisting of the following sections.

(a) Identify the two populations and the attribute of interest. Explain why the difference is important to you.

(b) Describe in detail how you selected the samples and show the summarised sample data.

(c) Calculate the confidence limits.

(d) Clearly state your conclusion in words.

5

Testing Theories

Having completed this chapter you will be able to:

– formulate and carry out various hypothesis tests;

– analyse data from a clinical trial.

5A. Introduction to Hypothesis Testing

Video Lecture https://youtu.be/FJYjD1L74vM

The lady tasting tea

It was a warm July afternoon. Muriel and some friends were having tea on the lawn. Ronald was there too. Muriel remarked that tea tastes better when the milk is added to the cup first. Ronald didn't believe that she could tell the difference. Here we have two opinions. Is there any way to prove who is right? Ronald proposed an experiment. He would pour eight cups of tea, four with the milk added first and four with the tea added first. He would then present the eight cups to Muriel in random order and see if she could identify which cups had the milk added first. Muriel agreed to the challenge.

When Muriel tasted the cups of tea, she correctly identified every one of the eight cups. This proved that Muriel was able to tell the difference, because if she was simply guessing, it is unlikely that she would have got all the answers correct.

Statistical hypothesis testing

The story about the lady tasting tea illustrates how theories are tested using statistics. In fact, Ronald later became a famous statistician, Sir Ronald Fisher, who developed many of the ideas that we use in statistics today.

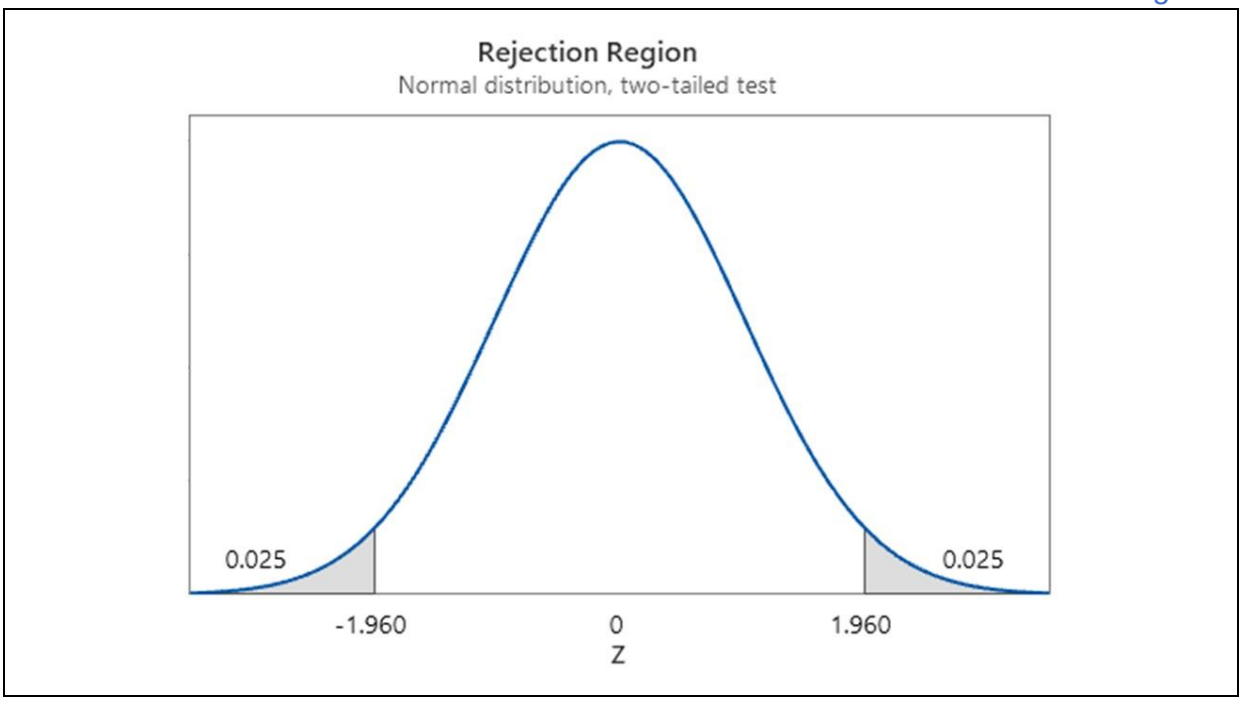
The first step in testing a theory is to clearly state the theory that is to be tested. The theory to be tested is called the **null hypothesis**. Hypothesis is another word for theory. It is 'null' because at the beginning it is neither proven nor disproven. The null hypothesis is represented by the symbol $\mathbf{H_0}$. The null hypothesis in the tea-tasting story is Ronald's belief that Muriel cannot tell the difference. In statistical hypothesis testing, the null hypothesis might say that the population mean is equal to some specified value, and then a sample is taken in order to test this hypothesis.

If the null hypothesis is rejected, it is rejected in favour of the **alternative hypothesis**, $\mathbf{H_1}$. The alternative hypothesis can be either vague or specific. A vague alternative hypothesis represents any departure from the null hypothesis, while a specific alternative hypothesis specifies a particular direction in which the data will disagree with the null hypothesis. In the tea-tasting story, the alternative hypothesis is specific: we want to find out if Muriel's answers are better than random guesses. We are not simply asking if her answers are different from random guesses, either better or worse. In a statistical hypothesis test, where the null hypothesis states that the population

mean is equal to some specified value, a vague alternative hypothesis would state that the population mean is not equal to the specified value. But the specific alternative hypothesis would state that the population mean is less than the specified value, or else the specific alternative hypothesis would state that the population mean is greater than the specified value. We decide in each context whether the alternative should be vague or specific, by asking ourselves if the null hypothesis would be proven wrong by any difference at all, or only by a difference in a particular direction. A vague alternative hypothesis will lead to a two-sided test, also called a **two-tailed test**, while a specific alternative hypothesis will lead to a one-sided test, also called a **one-tailed test**.

Before we perform the test, it is assumed that the null hypothesis is true. Then the sample is drawn and we observe the data. Then we ask the important question, which is, 'What is the probability that this kind of data would arise, if the null hypothesis is true?'. This probability is called the ***p*-value**, or simply ***p***. A small *p*-value provides evidence against the null hypothesis. If *p* is less than 5% then the null hypothesis is rejected at the 5% **significance level**.

Fig 5.1



Now, in the tea-tasting story, there are 70 different ways in which Muriel could have chosen a set of 4 cups out of the 8 cups that were presented to her. Only one of these is the correct choice, so the chance of Muriel guessing correctly is 1 in 70 which is 1.4%. Muriel did guess correctly so the *p*-value is 1.4%. This is less than 5% so the null hypothesis is rejected in this case.

It is possible for a null hypothesis to be rejected even though it is true, simply because an unfortunate sample has been drawn. This is called a **type 1 error**. The probability of occurrence of a type 1 error is the level of significance (**alpha**) which is usually 5%. It is also possible for a null hypothesis to be accepted even though it is false and this is called a **type 2 error**. Type 2 errors can occur for a number of reasons: perhaps the

sample is unfortunate, or the sample may be just too small, or perhaps the null hypothesis is not exactly true but is close to the truth. If a type 1 error has serious consequences, we can choose a significance level smaller than 5% ('significant'), such as 1% ('very significant') or 0.1% ('highly significant'). Also, if multiple hypotheses are to be tested, the probability of one or more type 1 errors occurring is increased, i.e. the **family error rate** is higher than the **individual error rate**. This is another reason for choosing a smaller significance level in certain cases.

The court analogy

The way that statistics is used to test theories is similar to the way that cases are tried in court. In court, a defendant is charged with an offence, and initially it is assumed that 'the defendant is innocent'. This is the null hypothesis. But if the evidence reveals facts that are unlikely to arise in the case of an innocent defendant, then the null hypothesis is rejected and the defendant is asserted to be guilty.

Notice that the null hypothesis is an assumption, and if the null hypothesis is accepted it has not been proven to be true, it has simply not been proven to be false. On the other hand, if the null hypothesis is rejected, then the alternative hypothesis is asserted to be true. This corresponds to the situation in a court case, where the defendant enjoys the presumption of innocence, and it is the responsibility of the prosecution to provide evidence that proves beyond reasonable doubt that the defendant is guilty.

Hypothesis Testing Procedure

There are many different kinds of statistical hypothesis tests but these steps are followed in every case.

1. State H_0 and H_1.
2. Draw a random sample.
3. Calculate the **test statistic** using an appropriate formula.
4. Look up statistical tables to find the **critical value** that identifies an area of 5% called the **rejection region**.
5. If the test statistic is in the rejection region then $p < 5\%$, so reject H_0. Otherwise, accept H_0.

If software is used to do the calculations at step 3, then the software will provide an exact *p*-value and so there is no need to look up statistical tables. Instead we just read the *p*-value from the software output to see if it falls below 5%, and then we can decide whether to accept or reject the null hypothesis.

Problems 5A

#1. A shoe manufacturer places an order with a supplier for a batch of shoelaces made to a specified target length. Before the batch is delivered, a random sample is drawn from the batch in order to test the null hypothesis that the batch mean is equal to the specified target length. In this context, what is meant by a type 1 error, what is meant by a type 2 error, and which error would be more costly?

#2. A manager is concerned about the delays experienced by customers while waiting to be served at a local outlet of a restaurant chain. One day, the manager notices that the customer waiting times seem to be longer than usual. The manager records all the

customer waiting times for the next hour, and then refers to the restaurant chain's website which gives a commitment to achieve a particular mean waiting time for customers. The manager decides to carry out a statistical hypothesis test of this hypothesised mean waiting time, using the recorded customer waiting times from that hour as the sample data. Is there anything wrong with this approach?

5B. Testing a Mean or Proportion

Video Lecture https://youtu.be/5jiFQC0qjPM

Testing a Mean When Sigma is Known

This approach can be used to test whether the mean of a population is equal to some particular value. This may be a value that the mean is supposed to be, or believed to be, or claimed to be. It is a theoretical or hypothesised value which can then be tested using sample data.

Formula 5.1

One-sample *z*-test

H_0: μ = some value

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Use z if σ is known.

Notice that the numerator in the **one-sample *z*-test** measures the difference between the observed sample mean and the hypothesised population mean. The denominator is the standard error, i.e. how big we would expect this difference to be. The test statistic therefore shows the scale of the disagreement between the evidence and the theory. The *p*-value measures how likely it is that so much disagreement would arise by chance.

EXAMPLE Bags of peanuts are claimed to have weights that conform to 'average is not less than 25 g'. It is known that the standard deviation of the bag weights is σ = 3 g. Test the claim, if a random sample of bag weights were as follows:

21, 22, 22, 26, 19, 22

SOLUTION

H_0: μ = 25

Note that H_0 must state a single value for μ.

H_1: μ < 25

This specific alternative hypothesis will lead to a one-tailed test, using the left tail, because only a 'less than' value can lead to rejection of the null hypothesis.

$\bar{X}$ = 22

n = 6

z = (22 - 25) ÷ (3 / √6) = -2.449

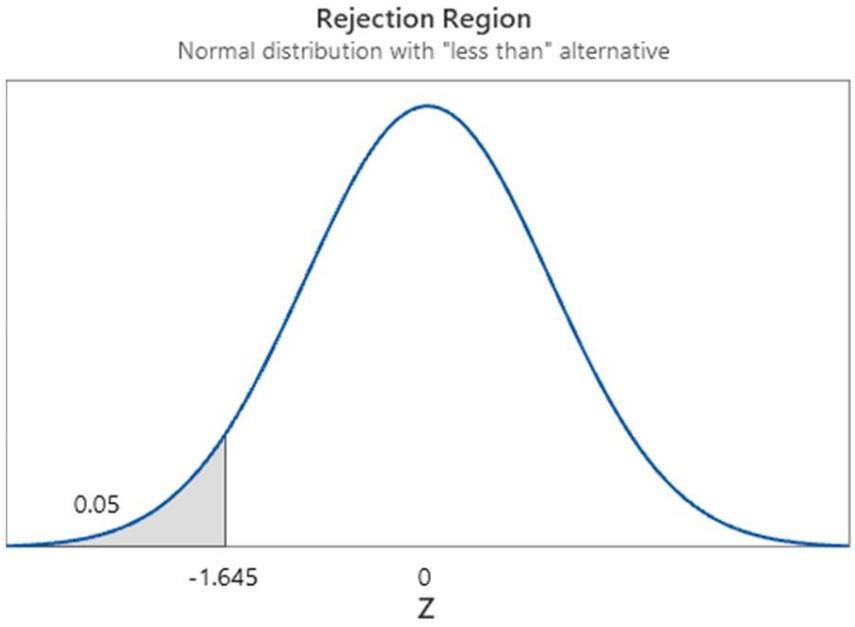
The tables suggest that critical z = 1.645

Look in the *t*-distribution tables, in the final row (z), in the 5% column (one-tailed test).

Since we are dealing with the left tail, z must be negative, so critical z = -1.645.

Any value more extreme than -1.645 is in the rejection region.

Fig 5.2



-2.449 is in the rejection region (p < 5%).
Reject H_0: $\mu = 25$, in favour of H_1: $\mu < 25$
The claim that the 'average is not less than 25 g' is rejected at the 5% level.

Testing a Mean When Sigma is Unknown

Usually the population standard deviation is unknown, and a **one-sample t-test** must be used. We repeat the previous example, without assuming prior knowledge of σ.

Formula 5.2

One-sample *t*-test

H_0: μ = some value

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Use t if σ is unknown.

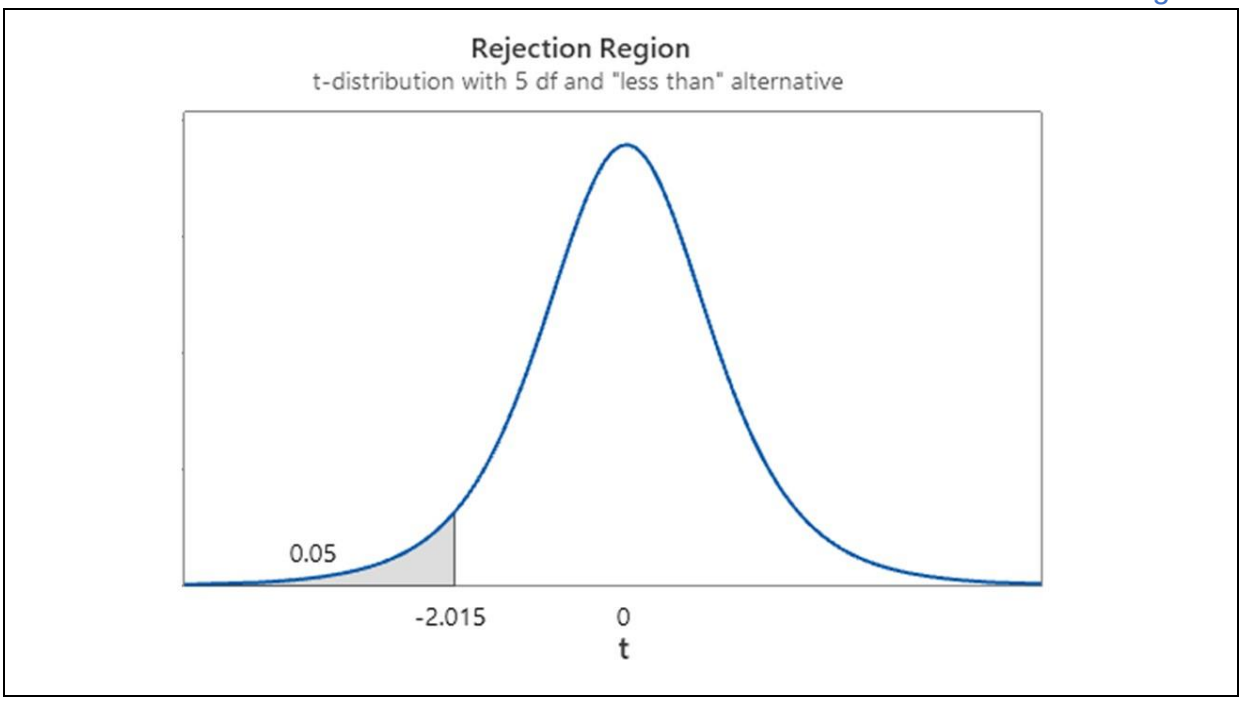
EXAMPLE Bags of peanuts are claimed to have weights that conform to 'average is not less than 25 g'. Test this claim, if a random sample of bag weights were as follows:

21, 22, 22, 26, 19, 22

SOLUTION

H_0: $\mu = 25$
H_1: $\mu < 25$
$\bar{X} = 22$
$S = 2.28$
$n = 6$
$df = 5$
$t = (22 - 25) \div (2.28 / \sqrt{6}) = -3.22$
The tables identify critical $t = -2.015$

Fig 5.3



-3.22 is in the rejection region ($p < 5\%$)
Reject H_0: $\mu = 25$, in favour of H_1: $\mu < 25$

We reject, at the 5% level, the claim that the 'average is not less than 25 g'.

Testing a Proportion

We may have a theory about a population proportion, for example, the proportion of voters who favour a certain candidate, or the proportion of manufactured units that are defective. If we draw a random sample from the population, we can use the observed sample proportion to test the theory about the population proportion, using a one-sample *P*-test.

Formula 5.3

One-sample *P*-test

H_0: π = some value

$$z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Conditions: $n.\pi \geq 5$ and $n.(1-\pi) \geq 5$

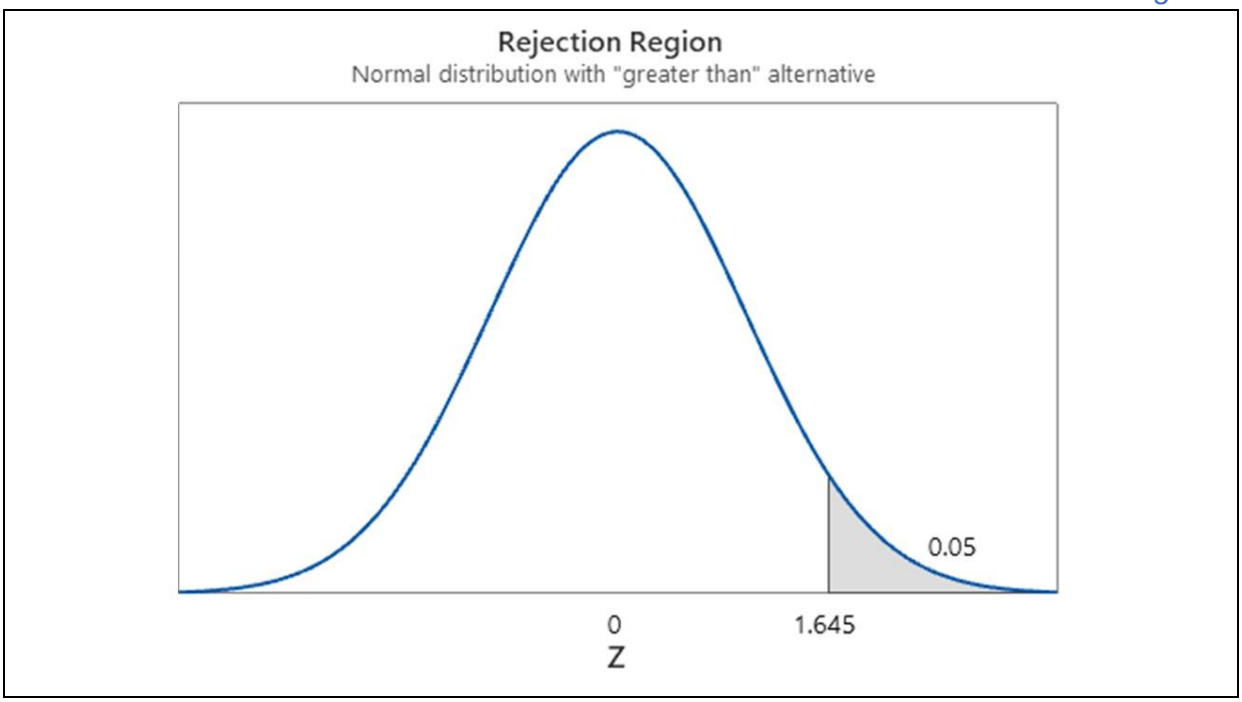
EXAMPLE A courier made a commitment that 'not more than 8% of all deliveries will be late'. A random sample of 315 deliveries included 30 late deliveries. Can it be asserted that the commitment has been violated?

SOLUTION

H_0: $\pi = 0.08$
H_1: $\pi > 0.08$
$n = 315$
$P = 30/315 = 0.095$
$z = (0.095 - 0.08) \div \sqrt{(0.08 \times 0.92 / 315)} = 0.98$

Critical $z = 1.645$

Fig 5.4



0.98 is not in the rejection region ($p > 5\%$).
Accept H_0. The data do not establish, at the 5% level, that the proportion of all deliveries that are late exceeds 8%. Therefore it cannot be asserted that the commitment has been violated.

Problems 5B

Video Tutorial: https://youtu.be/3KJcgl33qvg

#1. Beads of glue are applied to components during product assembly. The mean weight of these beads is supposed to be 7 grams and the standard deviation is known to be 0.15 grams. A random sample of five beads was selected in order to check if the correct mean weight is being achieved. The data are shown below, in grams. Use these data to test the hypothesis that the population mean is 7 grams.

6.92, 7.06, 6.88, 6.79, 6.93

#2. The weights of chocolate bars are described on the wrappers with a message claiming that the mean is not less than 50 grams. The data below show the weights in grams of a random sample of six of these bars. Use these data to test the claim about the mean weight.

49.76, 49.72, 50.13, 50.04, 49.54, 49.72

#3. A tutor prepared a laboratory assignment for students in a large class. The tutor believes that the mean time that students will need to complete this assignment is 80 minutes. In order to test this belief, the tutor randomly selected eight students from the class, gave them the laboratory assignment to complete, and recorded the time spent by each of the eight students. The data are shown below, in minutes. Perform a hypothesis test to investigate if the population mean is 80 minutes.

75, 81, 72, 75, 69, 63, 70, 72

#4. It is claimed that 'at least half' of the passengers who use the Galway to Dublin train are business travellers. A random sample of 220 passengers was taken on this route, and just 92 of the sampled passengers were found to be business travellers. Use a hypothesis test to investigate if the claim can be accepted.

#5. A botanist predicted that one in four of a large consignment of laburnum trees would have pink flowers. Out of a random sample of 96 of these trees, only 15 had pink flowers. Perform a hypothesis test to assess the botanist's prediction.

Project 5B

Hypothesis Test Project

Identify some inexpensive consumer product which has a clearly stated label claim regarding the mean contents. Carry out a statistical hypothesis test of this claim by selecting and measuring a random sample of five units of product. You will need to have a measuring instrument which is capable of distinguishing between the different units of product. Write a report consisting of the following sections.
(a) Identify the product by name, and state the label claim.
(b) Describe in detail how you selected the sample.
(c) Describe how you performed the measurements and display the data.
(d) Select an appropriate test, calculate the value of the test statistic, and compare the calculated value with the critical value from the tables.
(e) State your conclusion in simple language that can be understood by someone with no knowledge of statistics.

5C. Difference between Means

Video Lecture https://youtu.be/gmmh3Mmn9Uc

Often we want to compare two populations to see if there is a difference between them. There are two ways to do this. The preferred approach is to use paired samples. If this is not possible, independent samples can be used.

Testing Means Using Paired Samples

The one-sample *t*-test can be used with paired samples to test the hypothesis that the mean difference is zero. The data used in the test are the differences between each pair of observations, and the test is called a **paired *t*-test**.

EXAMPLE

Grace used a titration method to measure the vitamin C concentration of peppers. The data show the results for peppers stored in the fridge, and for similar peppers stored at room temperature. Do the data assert that storage condition makes any difference to vitamin C concentration?

Table 5.1

Item	Fridge	Room Temperature
Green Pepper	40.2	35.5
Yellow Pepper	34.7	31.6
Orange Pepper	31.5	27.3
Red Pepper	30.2	21.8

SOLUTION

H_0: $\mu = 0$

The mean difference is assumed to be zero.

H_1: $\mu \neq 0$

This is a vague alternative hypothesis because we are looking for 'any difference'.

Differences (fridge minus room): 4.7, 3.1, 4.2, 8.4

$\bar{X} = 5.1$

$S = 2.30$

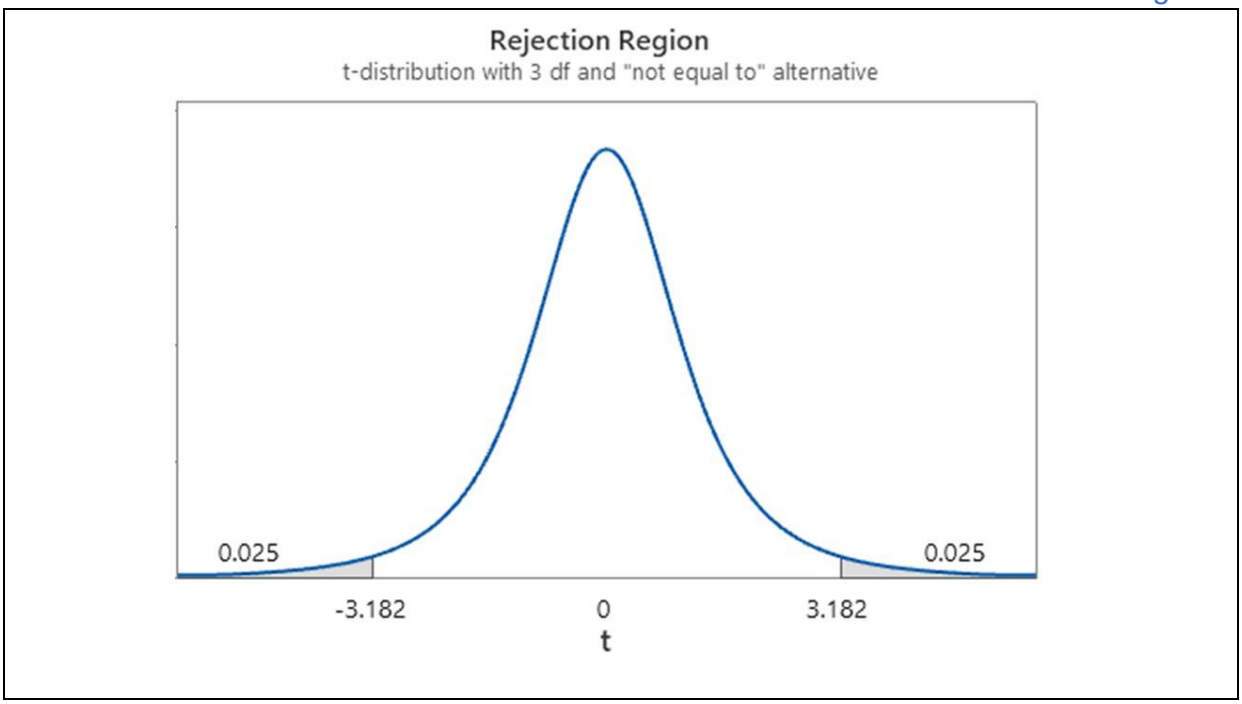
$n = 4$

$df = 3$

$t = (5.1 - 0) \div (2.30 / \sqrt{4}) = 4.43$

The tables identify critical $t = 3.182$

Fig 5.5



4.43 is in the rejection region ($p < 5\%$).

Reject H_0: $\mu = 0$, in favour of H_1: $\mu \neq 0$

The null hypothesis, that the mean difference is zero, is rejected at the 5% level. Storage conditions do make a difference. The mean Vitamin C content is higher for peppers stored in the fridge.

Testing Means Using Independent Samples

Formula 5.4

Two-sample *t*-test

H_0: $\mu_1 = \mu_2$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S^2/n_1 + S^2/n_2}}$$

S^2 is the pooled variance estimate, assuming equal variances.

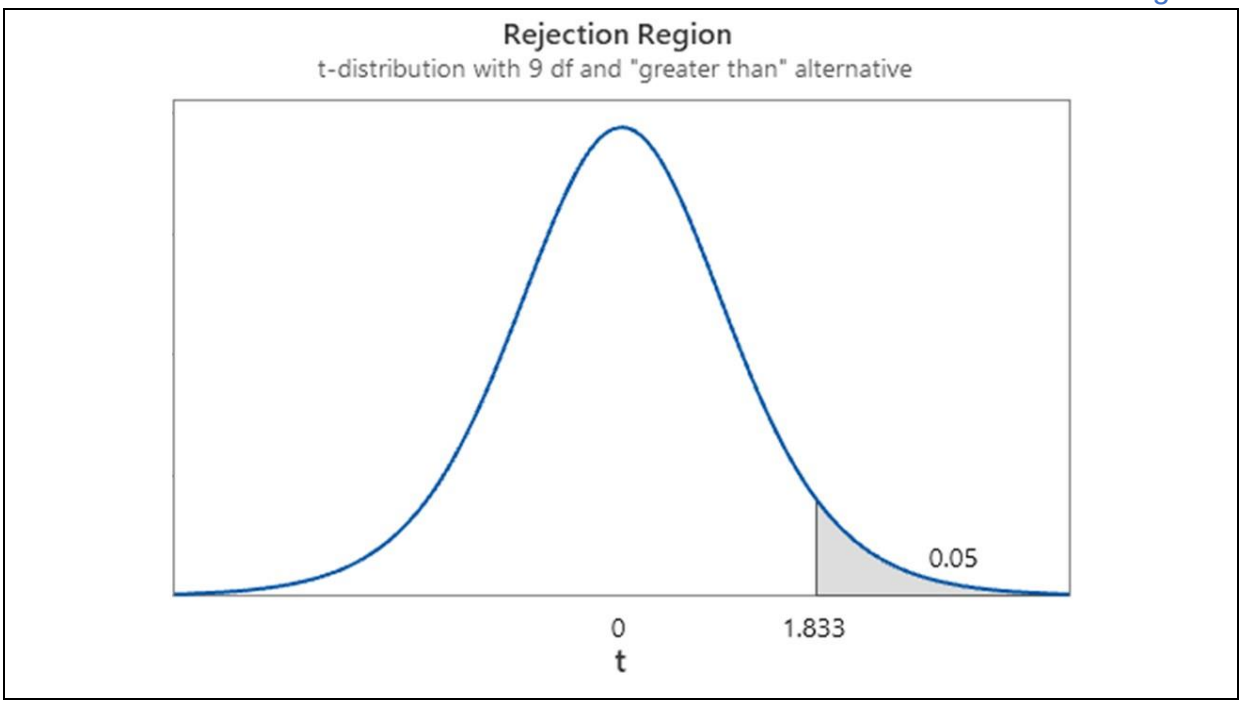
EXAMPLE The following measurements of 'biological oxygen demand' were taken from the influent and effluent of a groundwater treatment tank. A two-sample *t*-test can be used to see if the effluent measurements are lower, on average.

Influent: 220, 198, 198
Effluent: 186, 180, 180, 174, 177, 165, 174, 171

SOLUTION

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$
$n_1 = 3$ $n_2 = 8$
$S_1 = 12.70$ $S_2 = 6.40$
$S^2 = 67.73$ This is the pooled variance estimate using Formula 4.9.
$df = 9$
$\bar{X}_1 = 205.33$
$\bar{X}_2 = 175.88$
$t = (205.33 - 175.88) \div \sqrt{(67.73 / 3 + 67.73 / 8)} = 5.29$
The tables identify critical $t = 1.833$

Fig 5.6



5.29 is in the rejection region ($p < 5\%$).

Reject H_0: $\mu_1 = \mu_2$, in favour of H_1: $\mu_1 > \mu_2$

The data assert, at the 5% level, that the average effluent measurements are lower than the average influent measurements.

To test whether the means of more than two populations are all equal, use one-way ANOVA, presented in chapter 7.

Problems 5C

#1. The fuel consumption of a random sample of five vehicles, in litres per 100 km, was measured before servicing, and again afterwards. The data are shown in Table 5.2. Can it be asserted at the 5% level that fuel consumption tends to be lower after servicing?

Table 5.2

Driver	Before	After
Jake	13	12
Rosa	15	13
Charles	16	14
Terry	15	15
Amy	14	13

#2. A random sample of seven workers who attended a course to improve their typing speeds, and their speeds in words per minute before and after the course, are listed below. Do these data establish at the 5% level that the course is generally effective?

Table 5.3

Worker	Before	After
Britta	23	25
Jeff	37	37
Annie	39	43
Abed	45	50
Pierce	46	45
Shirley	39	46
Troy	51	54

#3. Kevin weighed himself three times in succession on a scale and the readings, in kg, were: 25.4, 25.6, 24.9. Rowena weighed herself twice on the same scale and her results were: 31.2, 31.9. Do these data assert, at the 5% level, that Rowena is heavier?

Project 5C

Hypothesis Test with Two Populations

Perform a hypothesis test that will allow you to investigate if there is a difference between the mean monthly rental for apartments in different geographical areas. Sample six apartments from each area, and write a report consisting of the following sections.

(a) Identify the two areas by name.

(b) Describe in detail how you performed the sampling and display the sample data.

(c) State the null hypothesis and the alternative hypothesis.

(d) Select an appropriate test, calculate the value of the test statistic, and compare the calculated value with the critical value from the tables.

(e) State your conclusion in simple language that can be understood by someone with no knowledge of statistics.

5D. Contingency Tables

Video Lecture https://youtu.be/kT4vHh_gNxI

Sometimes we want to compare two population proportions, e.g. we may want to prove that a higher proportion of successful outcomes occur when a particular strategy is applied in medicine or business. The following example tests whether a higher proportion of women, compared to men, choose an iPhone rather than an Android as their smartphone. Two independent random samples are taken, one from each gender. The data are presented in a 2×2 table, with two rows for the two genders, and two columns for the two types of smartphone.

Table 5.4

Gender	iPhone	Android	*Totals*
Male	20	30	50
Female	26	24	50
Totals	46	54	100

A table like this is called a **contingency table**, because it is used to test whether the proportion of smartphone types is contingent (i.e. dependent) on gender.

Formula 5.5

Contingency Tables

H_0: No association exists between the row and column categories.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O is the observed frequency and E is the expected frequency.

E = row total × column total ÷ grand total

Condition: Every $E \geq 5$

df = (number of rows - 1) × (number of columns - 1)

This test is called a **chi-square test of association** because we are investigating whether there is an association between gender and smartphone type. It can also be called a **test of independence** because we are investigating whether the probability of someone choosing a particular smartphone type is independent of their gender.

H_0: No association exists between gender and smartphone type.

First, the marginal totals are calculated, if they are not already displayed. Notice that the rows and columns containing the marginal totals are not counted as part of the table: this is a 2×2 table.

Next, the expected frequencies are calculated, using the formula for *E* above, and these are displayed below the observed frequencies. The expected frequencies represent the numbers that would be expected in the cells if the rows and columns are independent, i.e. if the null hypothesis is true.

E is calculated for every cell. In the first cell, $E = 50 \times 46 \div 100 = 23$
Of course, the expected frequencies must also add up to satisfy the marginal totals.

Table 5.5

Gender	iPhone	Android	*Totals*
Male	20 23	30 27	50
Female	26 23	24 27	50
Totals	46	54	100

The next step is to calculate the chi-square contribution in every cell, using the formula. These contributions can be displayed below the expected frequencies. The chi-square contributions measure the degree of disagreement between the observed and expected frequencies, which is a measure of the degree of disagreement between the data and the null hypothesis.

Table 5.6

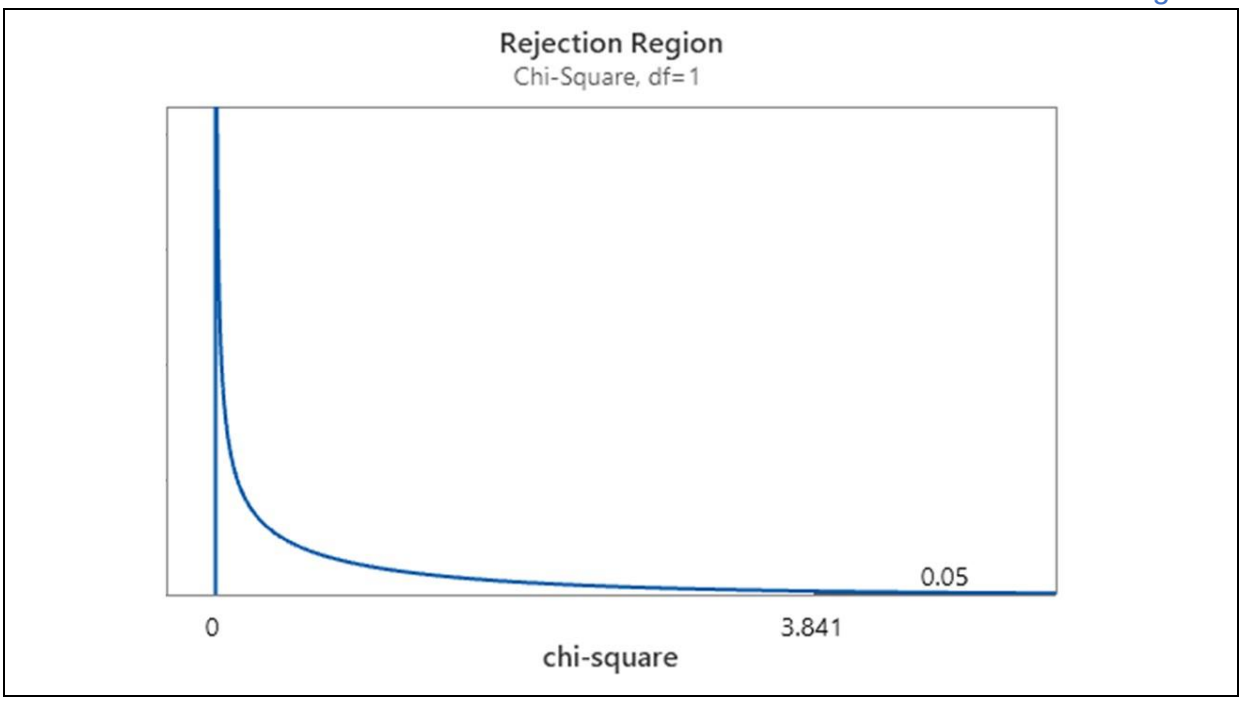
Gender	iPhone	Android	*Totals*
Male	20 23 0.3913	30 27 0.3333	50
Female	26 23 0.3913	24 27 0.3333	50
Totals	46	54	100

Finally, the chi-square contributions from all the cells are added together to provide the value of the test statistic, called the Pearson chi-square statistic.

$\chi^2 = 1.4492$

The chi-square distribution table indicates that the critical value of chi-square with 1 degree of freedom is 3.841. We use the upper tail of the distribution, because we wish to know whether the observed and expected frequencies are significantly different, not whether they are significantly similar. There is one degree of freedom because the degrees of freedom in a 2×2 table is $df = (2-1) \times (2-1) = 1$.

Fig 5.7



1.4492 is not in the rejection region.
$p > 5\%$

Therefore the null hypothesis, which states that no association exists between gender and smartphone, is accepted at the 5% level. The data do not establish, at the 5% level, that women are any more likely than men, or any less likely than men, to choose a particular smartphone type.

Contingency Tables larger than 2 x 2

Contingency tables with two rows and two columns are very common in medical research and elsewhere, where an intervention is either applied or not (two rows) and the outcome is either successful or not (two columns). However, a test of association can be used to compare proportions in any number of populations, not just two. In fact, a contingency table with any number of rows and any number of columns can be used to test for association between the row categories and the column categories.

EXAMPLE The numbers of people whose lives were saved, and lost, when the Titanic sank, are shown below. The figures are broken down by class. Is there an association between class and survival?

Table 5.7

Class	Saved	Lost
First class	202	123
Second class	118	167
Third class	178	528
Crew	192	670

SOLUTION

H_0: No association exists between class and survival.

Table 5.8

Class	Saved	Lost	*Totals*
First class	202 103.0 95.265	123 222.0 44.175	*325*
Second class	118 90.3 8.505	167 194.7 3.944	*285*
Third class	178 223.7 9.323	528 482.3 4.323	*706*
Crew	192 273.1 24.076	670 588.9 11.164	*862*
Totals	*690*	*1488*	*2178*

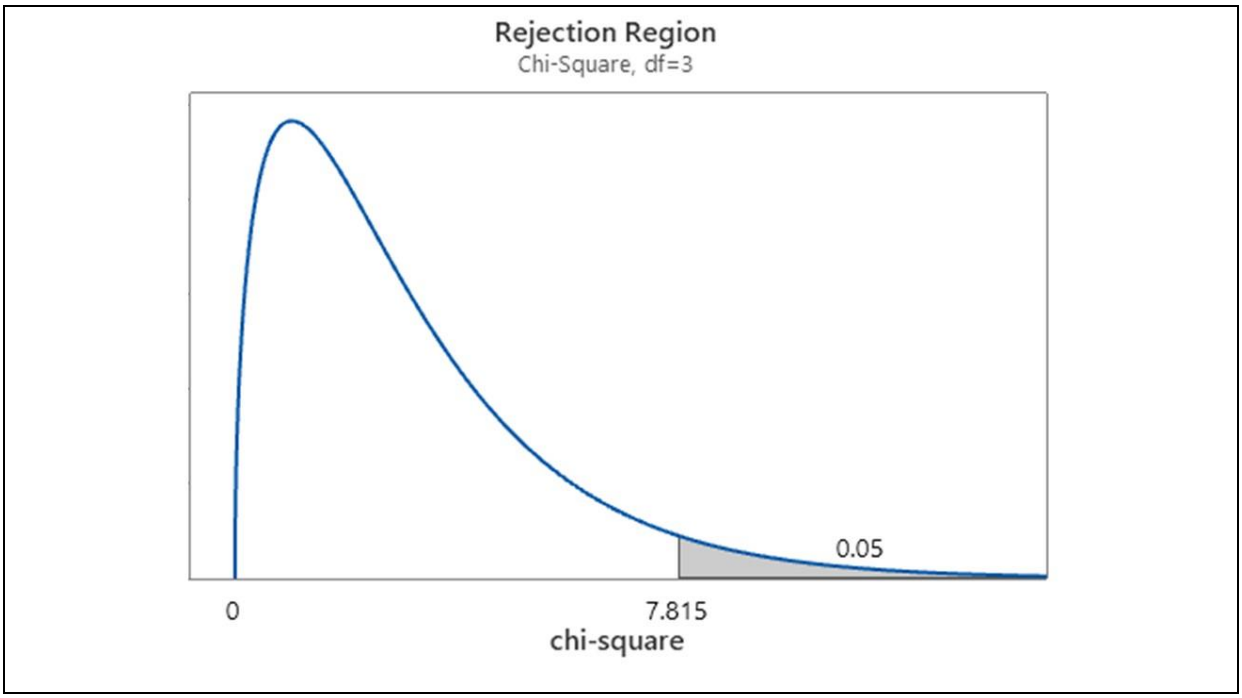
Table 5.8 shows the marginal totals. The observed frequencies, the expected frequencies and finally the chi-square contributions are shown in the cells.

$\chi^2 = 200.775$

In this case, $df = 3$

The critical value of chi-square with 3 degrees of freedom is 7.815.

Fig 5.8



200.775 is in the rejection region, so $p < 5\%$.
H_0 is rejected at the 5% level.

Looking at the different cells we notice that the largest chi-square contribution corresponds to first class passengers who were saved, so it makes sense to mention that first in our conclusions:
The data establish, at the 5% level, that an association does exist between class and survival. First class passengers were more likely than others to be saved, and crew members were less likely than others to be saved.

Sampling for a Test of Association

When collecting data for a test of association, it is OK to use the entire population of data if it is available, as in the case of the Titanic data. If sampling is to be used, sufficient data must be collected so that the expected count in every cell is at least five.

There are three equally valid approaches to sampling for a test of association, and usually whichever approach is most convenient is the one that is used.

1. Collect a single random sample from the entire population, and classify every observation into the appropriate row and column. For example, a single random sample of students could be selected from among the students at a college, and each selected student could then be classified by their gender, either male or female, and also by their smartphone type, either iPhone or Android. An advantage of collecting the data in this way is that it can be used to not only perform a contingency table analysis, but also to estimate the proportion of the population that belongs in any row or column category, e.g. the proportion of students who are male, or the proportion of students who have an iPhone.

2. Collect a random sample from each row, and classify every observation into the appropriate column. For example, collect a random sample of male students and classify each selected student by their smartphone type, either iPhone or Android. Then collect another random sample of female students and again classify each selected student by their smartphone type, either iPhone or Android. An advantage of collecting the data in this way is that we can ensure sufficient representation of sparse categories, e.g. if there are only a small number of male students in the college, we can ensure that we have sampled sufficient numbers of male students by targeting them in this way.

3. Collect a random sample from each column, and classify every observation into the appropriate row. For example, collect a random sample of students with iPhones, and classify each selected student by their gender, either male or female. Then collect another random sample of students with Android smartphones and again classify each selected student by their gender, either male or female. An advantage of collecting the data in this way is that we can ensure sufficient representation of a sparse smartphone type, either iPhone or Android.
Now suppose that there are more than two categories of smartphone, e.g. iPhone, Android and Windows. And suppose that Windows smartphones are so sparse that after collecting the data the number of Windows smartphones is insufficient to carry out a test of association, because some of the expected frequencies are less than five. We could combine two categories together, e.g. combine the Android and Windows categories, thus reducing the number of categories to two: 'iPhone' and 'Other'. If it doesn't make sense to combine categories, the sparse category could be simply omitted from the analysis.

Problems 5D

Video Tutorial: https://youtu.be/hhChzSuxFZg

#1. A study on the effectiveness of ADT (androgen-deprivation therapy) as a protective treatment against COVID-19 observed the rates of infection in a population of men, some of whom were on ADT and some of whom were not. Does the evidence assert that the infection rates are different for the men who are on ADT?

Table 5.9

COVID-19	Men on ADT	Men not on ADT
Infected	4	114
Not infected	5269	37047

#2. A number of students were randomly sampled from courses in Computing, Science and Engineering, and then classified by gender. Is there an association between gender and course?

Table 5.10

Gender	Computing	Science	Engineering
Male	20	17	44
Female	19	45	4

#3. The numbers of people whose lives were saved, and lost, when the Lusitania sank, are shown below. The figures are broken down by class. Is there an association between class and survival? Can you explain the findings?

Table 5.11

Class	Saved	Lost
First class	113	177
Second class	229	372
Third class	134	236
Crew	292	401

#4. Out of a random sample of 150 visitors at a theme park in summer, 15 had prepaid tickets, but in a random sample of 800 visitors at the theme park in winter, only 9 had prepaid tickets. Do these data assert, at the 5% level, that the population proportions are different?

#5. Broken biscuits are occasionally noticed in boxes awaiting shipment. A random sample of 100 boxes packed by hand included 3 boxes with some broken biscuits. A random sample of 100 boxes packed by machine included 17 boxes with some broken biscuits. Do these data indicate that the problem is related to the packing method?

Project 5D

Test of Association

Design and carry out a test of association on any population of your choice. Choose a population of items such that every item in the population can be classified according to two sets of categories. These two sets of categories correspond to the rows and columns of your contingency table. Write a report consisting of the following sections.
(a) Describe the population of interest and the purpose of your study.
(b) Explain which one of the three different sampling approaches suits your study best.
(c) Show the table of observed and expected frequencies, and present the analysis.
(d) State your conclusions correctly and completely, using only simple language.
(e) Suggest two possible causes of association between the sets of categories you have studied, even if no association has been proven.

5E. Sample Size and Power

Video Lecture https://youtu.be/3rWGUWyv0dQ

Practical Significance and Statistical Significance

In everyday life, when someone refers to a 'significant difference' they mean a difference that is big enough to be important. For example, there may be a difference between the nominal mean weight of bags of peanuts, and the actual population mean. But some differences are too small to be important. Suppose the difference is 2 grams. Is this important? To answer that question, we need to ask the customer, or an expert in nutrition acting on behalf of the customer.

However, in statistics, when we refer to a 'statistically significant difference' we mean that we are sure that a difference exists, i.e. on the basis of the sample data, the null

hypothesis is rejected at the 5% level. The magnitude of the difference may or may not be of practical importance. In order to find out if a certain difference is statistically significant, we would ask a statistician, or someone who is competent in statistics.

We need to bring these two ideas together. There is no point in using statistics to identify differences that are so small as to be of no interest to anyone. On the other hand, we want to make sure not to overlook a difference that is large enough to be important. In order to identify differences that are important, and ignore differences that are trivial, we need to choose an appropriate sample size.

Choosing the Sample Size

Suppose that you want to look for a needle in a haystack. How much time would you set aside for the task? The answer depends on three things:
1. How small is the needle? A smaller needle requires more time.
2. How big is the haystack? A larger haystack requires more time.
3. How badly do you want to find that needle? Increasing your chance of finding the needle requires more time.

In statistical hypothesis testing, the situation is very similar.

1. Delta

Firstly, we are looking for something. We are looking for a difference, which is like looking for a needle. We are looking for a difference, say, between the actual mean weight of bags of peanuts and the hypothesised mean. And we are looking for a difference that is important in size. So we begin by asking the customer to identify the size of this difference, which we call delta. The smaller the value of delta, the larger the sample size that will be needed.

2. Sigma

Secondly, it is difficult to find what we are looking for because it is hidden by random variation, in the same way that a haystack hides a needle. The random variation is measured by the standard deviation of the population. Its value may be known from previous experience or estimated from a pilot sample. The larger the standard deviation, the larger the sample size that will be needed.

3. Power

Power is the probability that we will find what we are looking for, like the chance of finding the needle in the haystack. In other words, it is the probability of rejecting the null hypothesis, if the size of the difference is delta. As a rule of thumb, we select a power of 80% for detecting a difference of size delta. This gives a good chance of detecting a practically significant difference if one exists. If the difference is greater than delta, then the probability of rejecting the null hypothesis will be even greater than 80%. And if the difference is less than delta, then the probability of rejecting the null hypothesis will be less than 80%.

Having identified the values of delta, sigma and power, statistical software can be used to compute the required sample size. It is also necessary to specify whether a one-tailed or two-tailed test is envisaged. A one-tailed test requires a smaller sample size, because we know in advance which tail to look in, to see if the test statistic is in the rejection region.

If a sample size was decided arbitrarily in advance, and then a hypothesis test leads to the null hypothesis being accepted, it is important to find out if the test was sufficiently powerful. Statistical software can be used in this case to compute the power of the test. If the power was low, then the hypothesis test does not tell us much.

Research versus Validation

We may use the word **research** to refer to the activity of looking for a difference, as described above. However, on other occasions, our objective is to prove that there is no difference. We call this **validation**. This is like trying to prove that there is no needle in the haystack. This requires a larger sample size, because we will not be willing to assume that there is no needle (based on absence of evidence), but rather we want to assert that there is no needle (based on evidence of absence). We must make sure to take a sample so big that if there is a difference, we will find it. We achieve this by selecting a power of 95% for detecting a difference of size delta. This will call for a larger sample, and if we are still unable to reject the null hypothesis, then we are no longer merely assuming it to be true, we are asserting it to be true. In summary, a power of 80% is recommended for research, and 95% for validation.

Problems 5E

#1. If a statistical hypothesis test is compared to a court case, where the null hypothesis corresponds to the assumption of innocence, then what does validation correspond to in this situation?

#2. It is believed that the regular consumption of liquorice may cause an increase in blood pressure. A hypothesis test is planned to test the difference in blood pressure of a sample of volunteers before and after a prescribed period of liquorice consumption. When considering how many volunteers are needed for the study, a value for delta will be required. How would you identify this value for delta?

5F. Tests of Variances and Goodness-of-fit

Video Lecture https://youtu.be/8eAeOL1sfV0

One-sample Variance Test

The **chi-square** statistic can be used to test a hypothesis concerning a population variance. The upper 5% point of the distribution identifies the rejection region for a 'greater than' alternative, and the lower 5% point (the upper 95% point) for a 'less than' alternative. For a 'not equal to' alternative, the upper 2.5% and 97.5% points are used.

Formula 5.6

One-sample Variance Test

H_0: σ^2 = some value

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

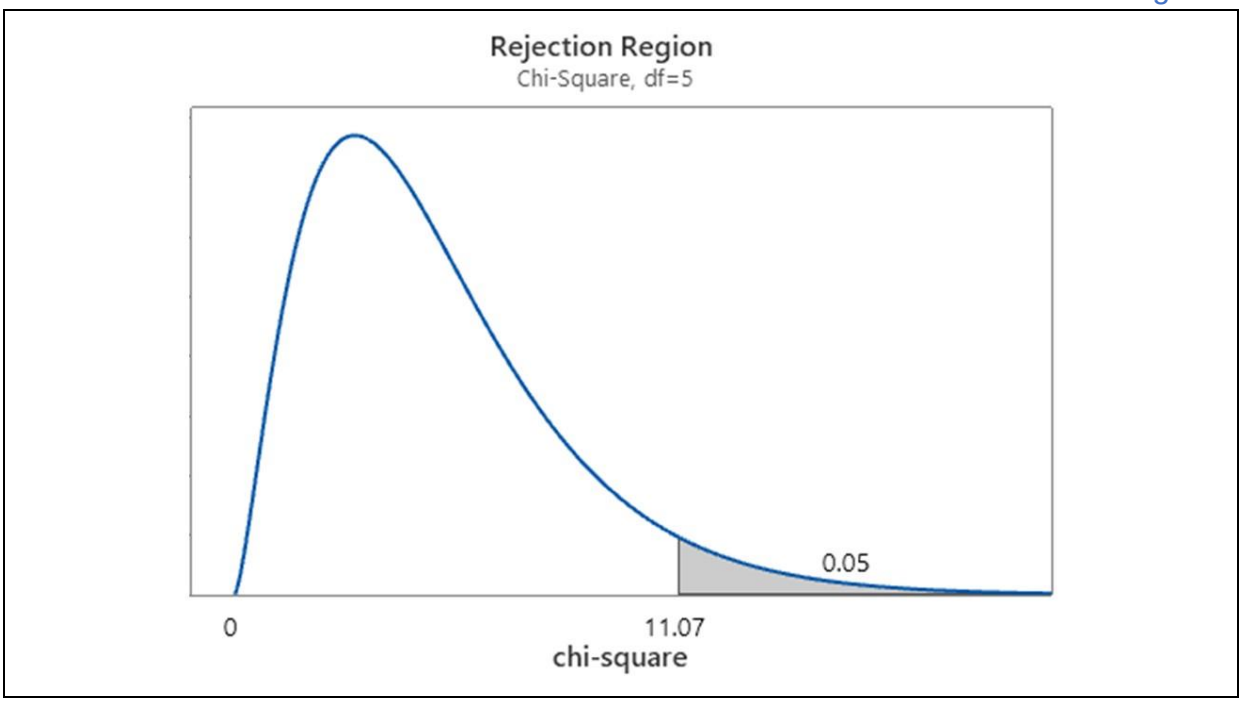
Condition: A normal population.

EXAMPLE The diameters of bagels are required to have a standard deviation of 2 mm or less. It is assumed that the bagels currently being produced conform to this requirement. Do these randomly sampled data assert otherwise at the 5% level?
Diameters in mm: 97, 94, 96, 95, 99, 93

SOLUTION

H_0: $\sigma^2 = 4$, the test uses the variance, not the standard deviation
H_1: $\sigma^2 > 4$
$n = 6$
$S = 2.16$
$\chi^2 = (6-1) \times 2.16^2 / 4$
$\chi^2 = 5.832$
$df = 5$
Critical $\chi^2 = 11.070$

Fig 5.9



5.832 is not in the rejection region.
Accept H_0.
These data do not assert, at the 5% level, that the standard deviation exceeds 2 mm.

Two-sample Variance Test

If two samples are drawn from normal populations, the ratio of the sample variances follows an *F* distribution. We define the ***F*-statistic** as the ratio of the larger variance to the smaller variance, and this always leads to a one-tailed test, using the upper 5% point of the *F* distribution. The *F* distribution has degrees of freedom for the numerator and denominator, in that order.

Formula 5.7

Two-sample Variance Test

H_0: $\sigma_1^2 = \sigma_2^2$

$$F = \frac{S_1^2}{S_2^2} \text{ where } S_1^2 \geq S_2^2$$

Condition: Normal populations.
The degrees of freedom are n_1-1 and n_2-1 for the numerator and denominator.

EXAMPLE

Paint can be applied by two modes: brush or sprayer. We are interested to know if the thickness is equally consistent for the two modes of application. Two random samples of paint specimens were selected, one from each mode. The thickness of each paint specimen was measured in microns, with the following results.

Brush: 270, 295, 315, 249, 296, 340
Sprayer: 325, 333, 341, 334, 317, 342, 339, 321

Can it be asserted, at the 5% level, that the variances are unequal for the two modes of application?

SOLUTION

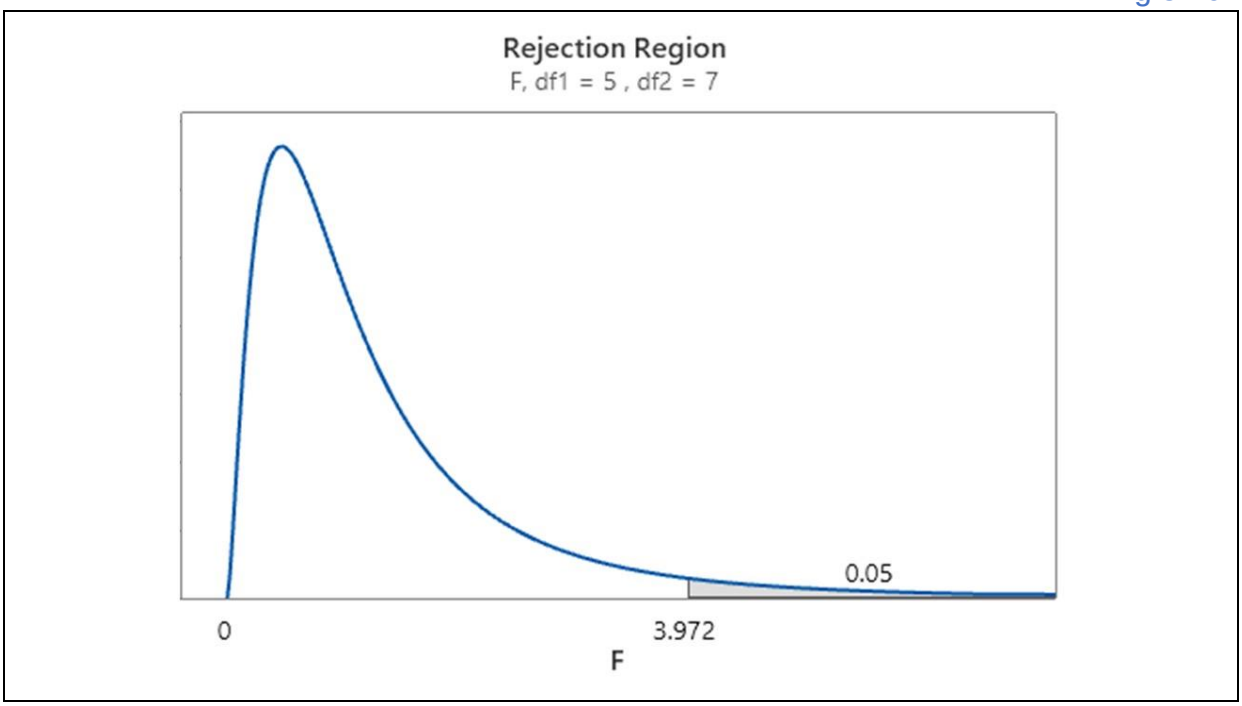
H_0: $\sigma_1^2 = \sigma_2^2$

$n_1 = 6$
$n_2 = 8$
$S_1 = 32.13$
$S_2 = 9.47$

$F = 11.51$

$df = 5, 7$
Critical $F = 3.972$

Fig 5.10



11.51 is in the rejection region ($p < 5\%$).
Reject H_0.

Yes. The data assert that sprayed paint exhibits less variation in thickness than brushed paint.

Test of Goodness-of-fit

We may wish to test the **goodness-of-fit** of some data to a particular distribution, e.g. a Poisson distribution. This may sound like a pointless exercise, but in fact it is one of the most interesting of all hypothesis tests. Suppose we count the number of breakdowns per day on a photocopier and we find that the Poisson distribution is not a good fit to the data, this indicates that the assumptions which apply to the Poisson distribution do not apply to that situation. The Poisson distribution assumes that events occur at random. If the Poisson distribution is not a good fit, then the events do not occur at random: perhaps they occur at regular intervals (i.e. uniform, like a 'dripping tap'), or perhaps they occur in clusters (i.e. contagious, 'it never rains but it pours'). This gives us useful information which may enable us to prevent future breakdowns. If the breakdowns occur at regular intervals, they may be caused by wear-out of some component: this could be addressed by scheduling regular servicing or replacement of that part. If the breakdowns occur in clusters, it may indicate that the remedy being applied is ineffective, so a new repair strategy needs to be devised. This is just one example: the interpretation will vary greatly from one case to another.

Formula 5.8

χ^2 Goodness-of-fit Test

H_0: The population has some specified distribution

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O is the observed frequency and E the expected frequency in each of the k categories.

$df = k - j - 1$, where j is the number of parameters estimated from the data.

Conditions:

(i) $k \geq 5$

(ii) Every $E \geq 1$ Adjacent categories may be combined to achieve this.

EXAMPLE

A die was rolled 30 times. The outcomes are summarised below. Is it a fair die?

Table 5.12

Outcome	1	2	3	4	5	6
Frequency	6	3	5	8	5	3

SOLUTION

H_0: The population has a uniform distribution with $k = 6$.

There were a total of 30 rolls, so the expected frequencies are: 5, 5, 5, 5, 5, 5.

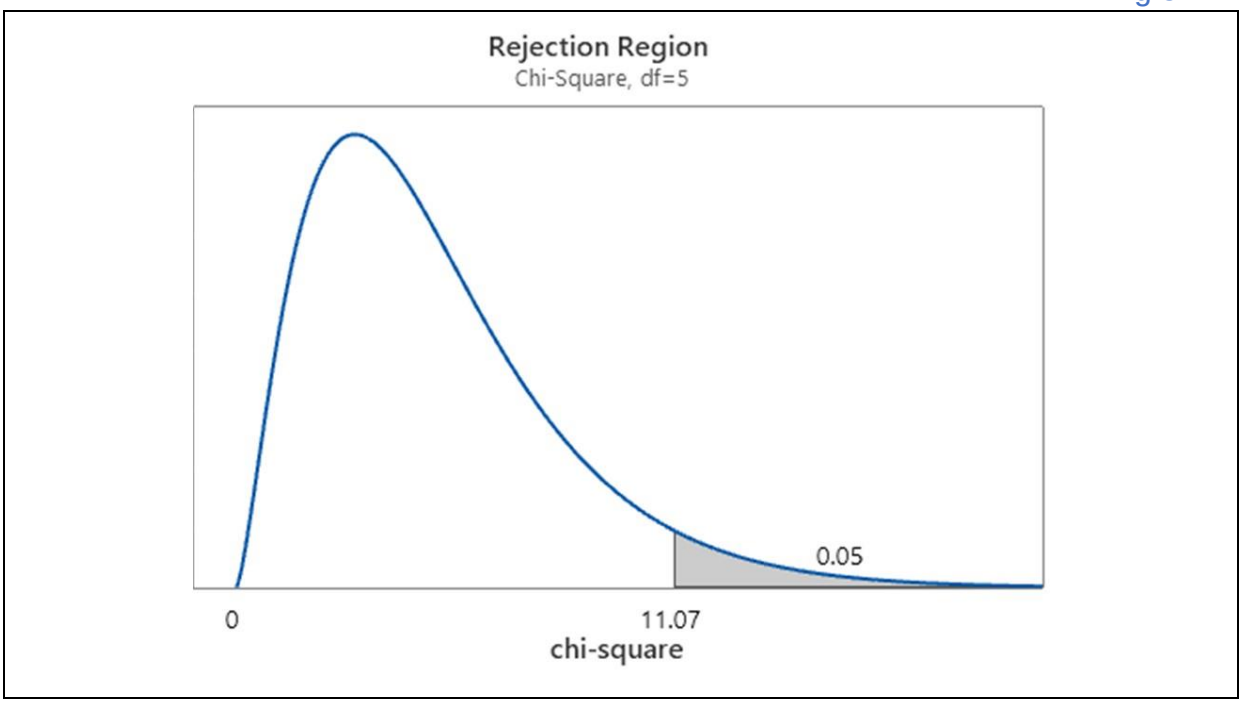
$\chi^2 = (6-5)^2/5 + (3-5)^2/5 + (5-5)^2/5 + (8-5)^2/5 + (5-5)^2/5 + (3-5)^2/5$

$\chi^2 = 1/5 + 4/5 + 0/5 + 9/5 + 0/5 + 4/5 = 3.6$

$df = 6 - 0 - 1 = 5$ No parameters were estimated from the data.

Critical value of $\chi^2 = 11.070$

Fig 5.11



3.6 is not in the rejection region.
Accept H_0.
The data are consistent with the assumption that the die is fair. We cannot assert, at the 5% level, that the die is 'weighted' in favour of some numbers at the expense of others.

Problems 5F

#1. A process that fills masses of powder into pots has a standard deviation of 5 mg. A new process has been tried out and the masses of nine quantities of powder filled by this process are shown. Do the data assert, at the 5% level, that the new process is less variable?
Mass in mg: 251, 255, 252, 250, 254, 253, 253, 251, 252

#2. A hotel provides a coach to take guests to the airport. There are two alternative routes: through the city streets, or via a ring road. A random sample of journeys was observed for each route, and the time was recorded in minutes on each occasion. Do the data below assert, at the 5% level, that that either one of the two routes has more consistent journey times than the other?
City Streets: 40, 36, 36, 40, 33
Ring Road: 35, 36, 37, 34, 37

#3. Two suppliers provided similar gearbox components to an automotive manufacturer. A random sample of components was taken from each supplier and the diameters were measured, in mm. Is there significant evidence of a difference in precision between the two suppliers?
Supplier A: 12.51, 12.48, 12.51, 12.50, 12.51, 12.52
Supplier B: 12.50, 12.50, 12.48, 12.51

#4. A group of 29 students were each asked to choose a random digit between 0 and 9 inclusive. Do their data, in Table 5.13, appear to come from a uniform distribution?

Table 5.13

Digit	0	1	2	3	4	5	6	7	8	9
Students	1	2	0	4	4	7	3	6	2	0

#5. One Saturday, the number of goals scored in each of the 56 matches played in the English and Scottish leagues was recorded. Is the Poisson distribution a good fit?

Table 5.14

Goals	0	1	2	3	4	5	6	7
Matches	6	13	15	12	5	1	1	2

#6. A consignment of eggs was found to contain a large number of cracked eggs. Before disposing of the consignment, the number of cracked eggs in each carton of 6 eggs was counted. Is the binomial distribution a good fit? Do you think the damage occurred before or after the eggs were put into the cartons?

Table 5.15

Cracked	0	1	2	3	4	5	6
Cartons	80	14	12	20	16	21	40

#7. Review section 3E in chapter 3 and attempt problems 3E. Section 3E deals with goodness-of-fit for continuous distributions.

5G. Clinical Trials

Video Lecture https://youtu.be/w47YRvQGTL0

Conduct of Clinical Trials

A clinical trial is an experiment to investigate the effectiveness of a new treatment. Clinical trials proceed in phases which test the new treatment for safety (phase 1) for efficacy (phase 2) and for superiority to the current treatment (phase 3). Typically, the experimental treatment (a new drug or intervention) is compared with a control (an existing treatment or placebo). If it is unknown which treatment will most benefit the patient then, and only then, is it ethical to carry out the trial (the principle of **equipoise**). Designed clinical trials are necessary because observational studies of clinical outcomes can be misleading. For example a certain hormone therapy was believed to be beneficial because women who took this treatment had lower rates of heart disease. However, a clinical trial revealed that the hormone therapy actually increased the risk of heart disease and other illnesses. The positive outcomes that had previously been observed were due to the better baseline health status of the women who had chosen to take the hormones: they were healthier, more active and less overweight than women who did not take the hormones. In a clinical trial, individuals do not choose their treatment: it is assigned to them.

Recruitment

The first step in a clinical trial is to draw up a list of participants who meet specified criteria, so that they are eligible to be allocated to either the experimental group or the control group. Finding a sufficient number of suitable participants can be a challenge. Some may have a preference for a particular treatment. Even those who do give their informed consent may withdraw later on. One way to increase the pool of suitable participants is to conduct a multicentre trial at a number of locations simultaneously.

Allocation

Next, the participants are randomly assigned to the different treatment groups so that there is an equal chance of each individual being allocated to each group. This random assignment is carried out so that the two groups differ only with regard to the treatment, and are as similar as possible with regard to all other characteristics both known and unknown.

If the participants are different in relation to some prognostic variable such as gender, then stratified randomization is carried out in order to achieve balance for the different treatments within these strata. Age may also be used as a prognostic variable for stratification. Adding multiple prognostic variables is not advised since it reduces the number of participants available for allocation in each stratum: if the number falls to zero or one in a stratum then randomization in that stratum is not possible.

For the avoidance of bias, clinical trials are typically carried out either **single-blind** or **double-blind**. In a single-blind trial, no participant knows whether they have been assigned to the experimental group or the control group. In a double-blind trial, the researchers do not know which participant is in which group either. This avoids a situation where the participants in the experimental group might have a greater expectation of recovery than the participants in the control group, or a situation where the researchers might behave in a way that communicates higher expectations to these participants (the **Rosenthal effect**).

Follow-up

The participants are followed up so that all the outcomes can be observed. This is not as easy as it sounds. As a clinical trial proceeds, some participants may drop out or otherwise be lost to follow-up. This raises the problem of how to handle these missing data. The easy answer is simply to ignore the missing data but this will cause bias unless the data are missing at random. Unfortunately it is possible that drop-out may in some way be related to positive or negative treatment outcomes. A preferred approach is to use some **imputation** strategy to fill in the missing values, such as replacing the missing value with a value taken from another participant with similar characteristics (**hot deck method**) or replacing the missing value with multiple values (**multiple imputation**) in order to take account of the uncertainty involved. However, nothing compensates properly for missing data.

Data Set for Analysis

After the trial, the data are analysed so that conclusions can be drawn and results published. Strange though it may seem, the complete set of data that arises for all the original participants should be used in the analysis, even if some of these participants

failed to comply with the protocol, and even if some of them never took the drug that they were allocated. There are two reasons for taking this approach, which is called **intention-to-treat** analysis. The first reason is that non-compliance may be related to some adverse side-effects of the drug. The second reason is that compliant participants, regardless of whether they are in the experimental group or the control group, tend to have different clinical outcomes than non-compliant participants. Excluding the data from non-compliant participants in the experimental group would have the effect of unbalancing the study design, i.e. the treatment group would include only compliant participants but the control group would include both compliant and noncompliant participants. After conducting the intention-to-treat analysis, secondary analyses can be conducted taking account of the treatment that the participants actually received (**as-treated** analysis) or focusing on the participants who complied strictly with the prescribed regimen (**per-protocol** analysis) but it must be borne in mind that these secondary analyses are prone to bias.

Purpose

A clinical trial typically aims to identify which treatment is superior. Two-tailed hypothesis tests should be used because we will not know in advance which treatment is superior, even if we think we do know. Sometimes the purpose of a clinical trial is to demonstrate **non-inferiority** of an alternative drug which may be less invasive or less costly. In such cases it is appropriate to use a one-tailed test. Note that after being approved by the regulatory authority, the new treatment continues to be monitored for rare and long-term side effects (phase 4).

Sample Size

In order to determine the sample size required for a clinical trial, a clinical judgement must be made about the magnitude of the therapeutic effect that is required in order to claim superiority or non-inferiority. Power and sample size calculations can then be carried out in the usual way to ensure that the study will be large enough to have a good chance of identifying an effect that is large enough to be clinically important.

Analysis of Clinical Trials: Discrete Response Variable

In some clinical trials the **end-point** is a binary outcome, e.g. it is of interest whether the patient experienced remission, and so the response is either yes or no.

EXAMPLE

A randomised double-blind clinical trial was carried out to investigate whether a certain intervention was effective in the remission of psoriasis. The numbers of patients in the treatment and control groups who did and did not experience remission after 30 days are shown in the table.

Table 5.16

	Remission	No Remission	*Totals*
Intervention	27	11	38
No Intervention	15	23	38
Totals	42	34	76

This is a contingency table. See chapter 5 for details. The null hypothesis states that there is no association between the treatment group and remission status. In other words, the proportion of patients experiencing remission is equal in the two groups.

To complete the analysis we must calculate the expected frequencies and the chi-square contributions.

Table 5.17

	Remission	No Remission	*Totals*
Intervention	27 21 1.714	11 17 2.118	38
No Intervention	15 21 1.714	23 17 2.118	38
Totals	42	34	76

The chi-square contributions are added together to provide the test statistic.

$\chi^2 = 7.664$

The critical value of chi-square with one degree of freedom is 3.841.

The null hypothesis is rejected.

There is an association between the treatment group and remission status.

Checking in the first cell in the table we notice that a greater number of patients than expected in the intervention group experienced remission: 27 compared to 21.

It can be concluded that this intervention is effective in the remission of psoriasis.

Analysis of Clinical Trials: Continuous Response Variable

A continuous response variable is a clinical measurement on some numerical scale, e.g. blood pressure.

EXAMPLE

A clinical trial was carried out to investigate whether drug A is more effective than drug B for reducing systolic blood pressure. The reduction in blood pressure for each patient in each group is shown. Each number represents the change, calculated as blood pressure before minus blood pressure after, in units of mm Hg.

Drug A: -3, 11, -3, 1, -5, 3, 15

Drug B: 19, 13, 19, 4, 23, 21, 4

The null hypothesis states that there is no difference between the mean of the two groups. In other words, the drug chosen has no effect on the mean reduction in blood pressure.

Assuming that the groups are independent, a two-sample *t*-test can be used to analyse the data. See chapter 5 for details.

The sample mean and standard deviation are 2.71 and 7.61 for drug A, and 14.71 and 7.93 for drug B.

The pooled standard deviation estimate is 7.77, with 12 degrees of freedom.

The test statistic, t = -2.89

The two-tailed critical value of t with 12 degrees of freedom is -2.179 and +2.179.

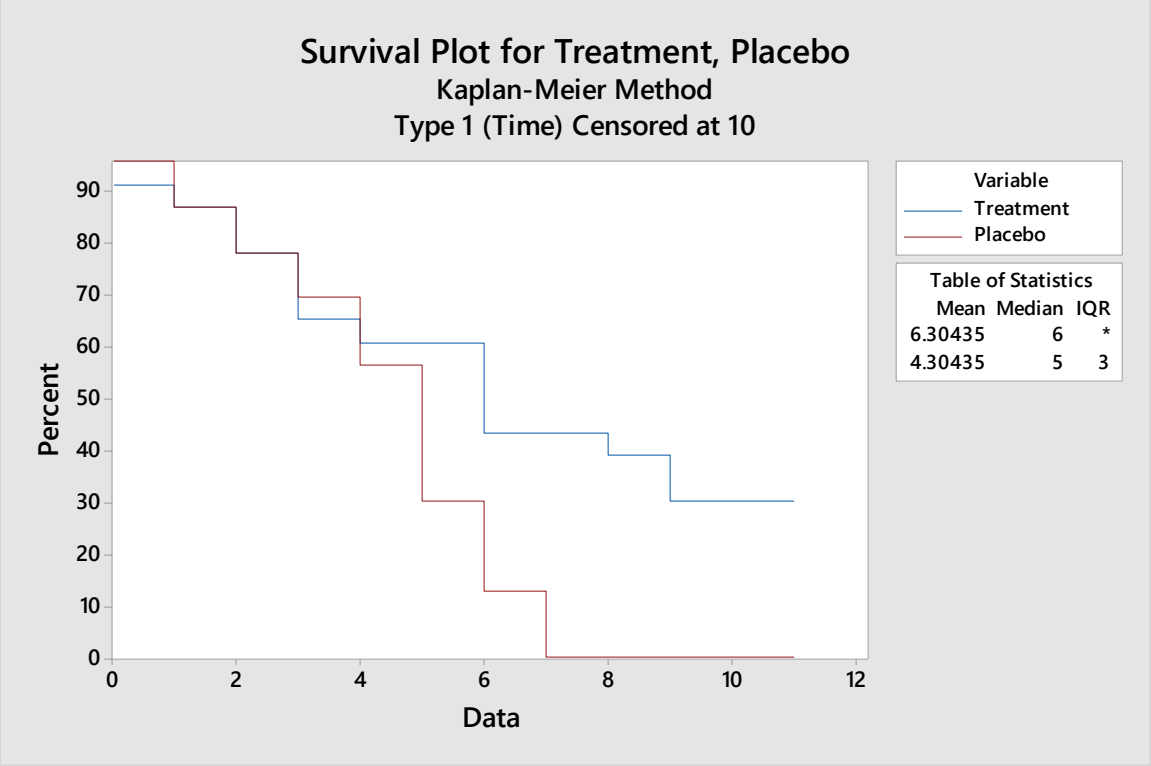
The null hypothesis is rejected at the 5% level.

It can be concluded that the mean reduction in blood pressure is greater with Drug B.

Analysis of Clinical Trials: Time-to-event response variable

Many studies measure the time to an event such as death, heart attack, etc. Such data are referred to as **survival data**. Of course, some participants may not have experienced the event at the end of the follow-up period and so their survival times are **right-censored**. Survival data are analysed by the **Kaplan-Meier** method and are illustrated in a **survival plot**.

Fig 5.12



A **log-rank test** can be used to compare the two groups. The null hypothesis states that the two groups have identical survival distributions. In this example, the *p*-value,

provided by Minitab® software, is less than 0.05 so it can be concluded that the survival distributions are different.

Method	Chi-Square	DF	P-Value
Log-Rank	7.13644	1	0.008

Problems 5G

#1. Skin lesions are sometimes removed by shave biopsy. Treating the area by keeping it covered and moist for a number of days afterwards may reduce scarring. A randomised trial was carried out with 80 patients, half of whom were assigned to the treatment and half to the control group. Every patient was assessed after five days when the scarring was classed as satisfactory or not satisfactory. Of the 40 patients in the treatment group, 33 were classed as satisfactory, and of the 40 patients in the control group, 31 were classed as satisfactory. Is the difference in proportions significant?

#2. Participants in a study were randomly assigned to two different weight-loss regimes. The weight loss, in kg, of each individual is shown below. Do these data support the assertion that one regime is more effective than the other?

Regime 1: 1.5, 0.2, 0.8, 1.9, 2.4

Regime 2: 2.0, 0.8, 2.7, 1.6, 1.2

Project 5G

Clinical Trial

Review a paper that describes the conduct of a clinical trial. Briefly discuss what the paper says, or omits to say, under each of the six headings below.
For example, under the heading "Recruitment", does the paper mention inclusion criteria and informed consent? If it does, then quote a few words from the document to show how this issue was addressed. If it does not, then make a note of the omission, and write a few lines to explain how this issue should have been addressed.
(a) Recruitment
(b) Allocation
(c) Follow-up
(d) Data set for analysis
(e) Purpose
(f) Sample size

6

Making Predictions

Having completed this chapter you will be able to:

– identify variables that are useful for making predictions;

– make predictions using regression models;

– use multivariate techniques to explore data-sets and suggest categorisations.

6A. Correlation

Video Lecture https://youtu.be/wx_hYGQYtGk

Sometimes we wish to know the value of a variable that cannot be observed directly. It could be something that is difficult to measure, like temperature. Or it could be something that will not be available until the future, like the number of pages that can be printed with a recently purchased ink cartridge. Or it could be something that is unknown, like the height of an intruder who departed without being seen.

However, there may be some other variable that can be measured instead, and perhaps this other variable may be useful for estimating the value of the variable of interest. This other variable is called a **predictor variable** and the variable of interest is called the **response variable**.

So, instead of measuring temperature, we could measure the length of a thread of liquid in a glass tube. Instead of measuring the number of pages that can be printed with a recently purchased ink cartridge, we could measure the volume of ink in the cartridge. Instead of measuring the height of an intruder, we could measure the size of a footprint left by the intruder. In each case, the value of the predictor variable could be useful for estimating the value of the response variable.

Scatterplots

The first step is to investigate whether there seems to be any relationship between the two variables. This is called **correlation**, and it can be explored with a scatterplot. The predictor goes on the horizontal axis and the response goes on the vertical axis. A sample is drawn, and the values of both variables are observed. The resulting data-pairs are plotted as points on the graph.

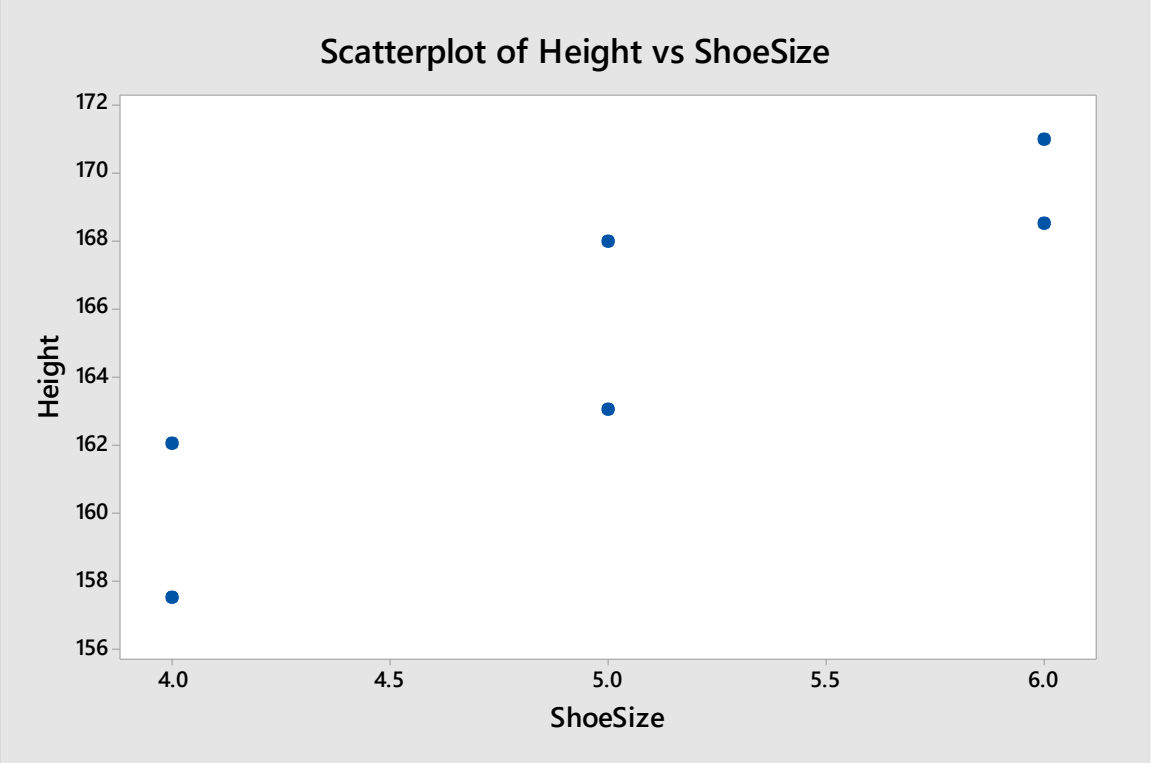
EXAMPLE 1

The shoe-sizes and heights of a number of women, randomly selected from a large class, were observed. The data are represented in the following scatterplot.

We can see that as shoe-size increases, height tends to increase also, i.e. there is some **positive correlation** between shoe-size and height. The correlation is fairly strong but not very strong. There definitely seems to be a relationship but there is a lot of random scatter as well. We could say that there is some **explained variation**

and also some **unexplained variation**. Knowing someone's shoe-size would be useful for estimating their height, but it would not provide a perfect estimate.

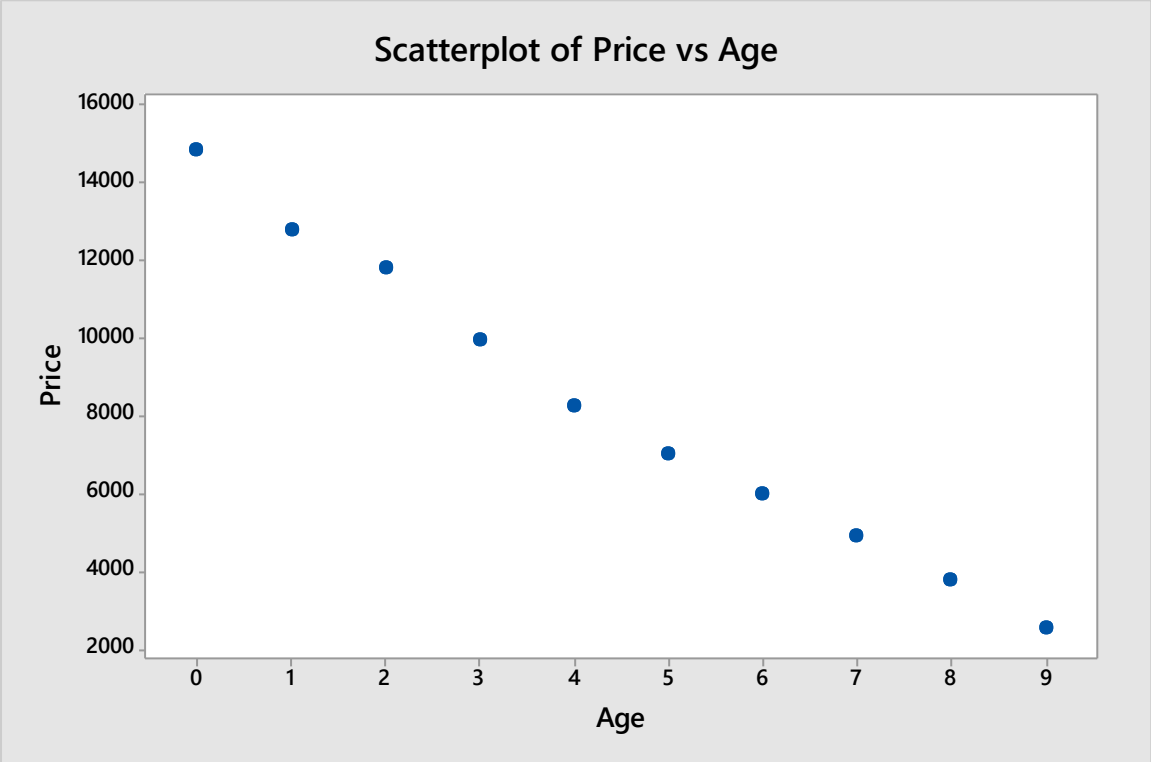
Fig 6.1



EXAMPLE 2

A number of Nissan Micra cars were sampled from a car sales website. The age and the price of each selected car is represented in the scatterplot below.

Fig 6.2

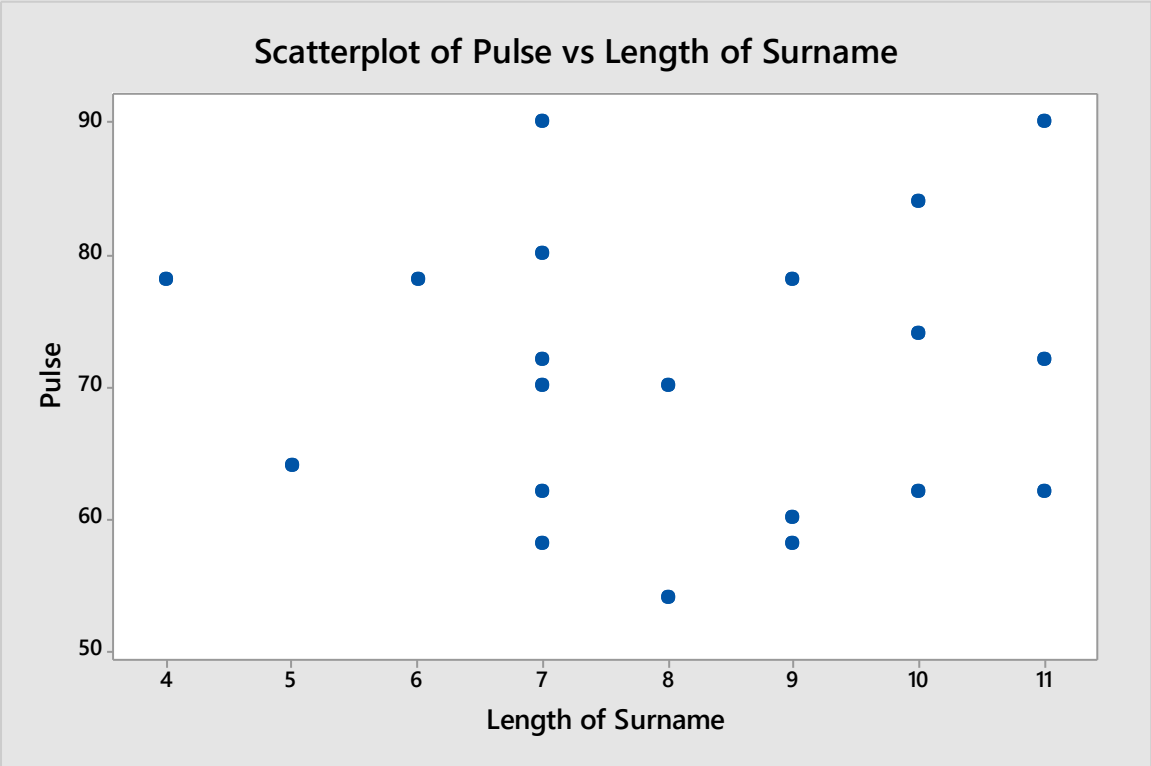


In this case, there is **negative correlation**, i.e. as age increases, price tends to decrease. And on this occasion the correlation is strong: the points are not widely scattered. Instead, they exhibit a high degree of linearity. It seems that age explains a lot of the variation in the prices of these cars.

EXAMPLE 3

A number of patients were sampled from a hospital and in each case their pulse, measured in beats per minute, and the number of letters in their surname were observed. The data are represented in the scatterplot below.

Fig 6.3



It's no surprise that these two variables are not correlated at all, either positively or negatively. A scatter diagram of two uncorrelated variables will reveal a set of randomly scattered points. As *X* increases, *Y* does not tend to either increase or decrease. This can be called a **plum-pudding model**, because the points are randomly scattered, like the currants in a plum pudding.

Correlation does not prove that *X* causes *Y*. When a scatterplot of observed values of *X* and *Y* reveals correlation between the two variables, we say that there is a predictive relationship between *X* and *Y*. We can predict the value of *Y* more precisely if the value of *X* is known. We do not assume that *X* causes *Y*. For example, if we take a sample of countries from among the countries of the world, and observe the number of TV sets per person in each country (X), and the life expectancy in each country (Y), we will find that there is positive correlation between these two variables. But this does not prove that importing more TV sets would cause people to live longer. To prove **causation** requires a designed experiment in which the values of *X* are randomly assigned. (See chapter 7 for details.)

Correlation Coefficient

The correlation coefficient (r) is a statistic that measures the degree of linearity between two variables. Its value always lies between -1 and 1, inclusive. The sign of r (+ or -) indicates the type of correlation (positive or negative). The size of r indicates the strength of the correlation. If $r = 0$ then there is no correlation, while if $r = 1$ or if $r = -1$ there is perfect correlation with all the points lying on a straight line. Values close to zero indicate that the correlation is weak, while values close to 1 or close to -1 indicate that the correlation is strong.

The value of r is tedious to calculate by hand, but you can use your calculator or statistical software to calculate it for you.

EXAMPLE 1

For the shoe-size and height data, $r = 0.889$

This value indicates positive correlation, since the value of r is positive. This tells us that as shoe-size increases, height tends to increase.

This value also indicates correlation that is fairly strong but not very strong, since the magnitude of r is well above zero but not very close to 1. This tells us that shoe-size is useful for predicting height, but the prediction will not be very precise.

EXAMPLE 2

For the age and price data, $r = -0.996$

This value indicates negative correlation, since the value of r is negative. This tells us that as the age of a car increases, the price tends to go down.

This value also indicates very strong correlation, since the magnitude of r is very close to 1. This tells us that age is useful for predicting price, and the prediction will be quite precise.

EXAMPLE 3

For the surname and pulse data, $r = 0.009$

This value indicates little or no correlation, since the value of r is close to 0. This tells us that surname length is not useful for predicting pulse.

All of these values for the correlation coefficient are based on sample data, so they are estimates of the population correlation coefficient, which is denoted by the Greek letter rho, symbol ρ.

Coefficient of Determination

The values of a response variable, Y, are not all the same. Now let us try to explain why the values of Y are different. Some of the differences in Y can be explained by differences in X, because when X changes, Y tends to change too. But some of the differences in Y cannot be explained by differences in X, because even when X stays the same, there is some variation in Y. So how much of the variation in Y can we explain?

The square of the correlation coefficient, r^2, is called the coefficient of determination. It is usually expressed as a percentage, and it measures the proportion of the variation

in Y which is explained by the variation in X. The sample value of r^2 can be **adjusted** to provide an unbiased estimate of the r^2 value in the population.

EXAMPLE 1

For the shoe-size and height data, *R-Sq(adj)* = 73.8%

This indicates that 73.8% of the variation, in the heights of all the women in that class, is 'explained' by the variation in shoe-size. If we only observed the heights of women who all have the same shoe-size, the variance in height would reduce by about 73.8%.

The **residual variation**, 100% - 73.8% = 26.2%, is unexplained, which means that it cannot be explained by shoe-size. Even for women with the same shoe-size, there is still some variation in their heights. This variation can be estimated using the sample standard deviation, S, and for the shoe-size and height data, S = 2.57 cm. This is the standard deviation estimate of the heights of women who wear the same shoe-size.

EXAMPLE 2

For the age and price data, *R-Sq(adj)* = 99%

This indicates that 99% of the variation, in the prices of all the Nissan Micra cars for sale on that website, is explained by the variation in the ages of the cars. Older cars cost less, of course! If we only look at the prices of cars that are all the same age, they would be much more alike in price; the variance in price would reduce by about 99%.

The residual variation, 100% - 99% = 1%, cannot be explained by age. It may be related to a multitude of things such as the car's location, colour, condition, and so on. When we look at cars that are the same age, we still find some variation in the prices. This variation can be estimated using the sample standard deviation, S, and for the age and price data, S = €407. This estimates the standard deviation of the prices of cars that are all the same age.

EXAMPLE 3

For the surname and pulse data, *R-Sq(adj)* = 0%

This indicates that none of the variation, in the pulse measurements of all the patients in that hospital, is explained by the variation in surname length. As the surname changes, the pulse doesn't tend to change at all. And if look at patients with surnames that are all the same, the variance in the pulse measurements does not reduce at all.

The residual variation, 100% - 0% = 100%, i.e. all the variation, is still unexplained. This variation can be estimated using the sample standard deviation, S, and for the surname and pulse data, S = 11 beats per minute. This estimates the standard deviation of the pulse measurements of all the people with the same length of surname. But in this case, because it represents 100% of the variation, it estimates the standard deviation of the pulse measurements of all the people in the hospital.

Problems 6A

#1. (Activity) Select a sample of car journeys and in each case observe the journey distance (X) and the journey time (Y). Find the value of the correlation coefficient and explain in words what it means. Also find the value of the coefficient of determination and explain in words what it means.

6B. Regression Line

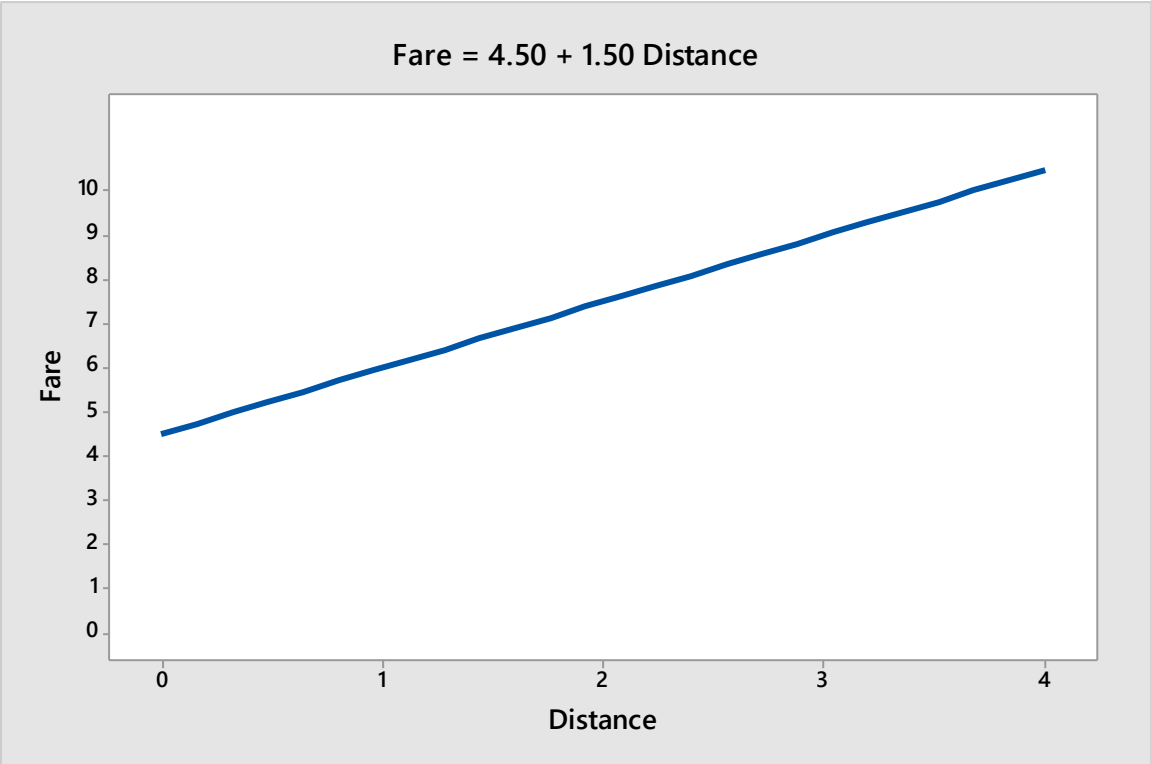
Video Lecture https://youtu.be/A2buMCctYf8

Equation of a Line

Having established that a linear relationship exists between two variables, the next step is to identify the line that best describes the relationship. The equation of a line has the form $Y = a + bX$ where X and Y are the two variables. X is the independent variable (the predictor) and Y is the dependent variable (the response). The parameter ***b*** is the key to the relationship, and we should look at its value first. It tells us how much Y increases when X increases by one unit. It is called the **slope** of the line.

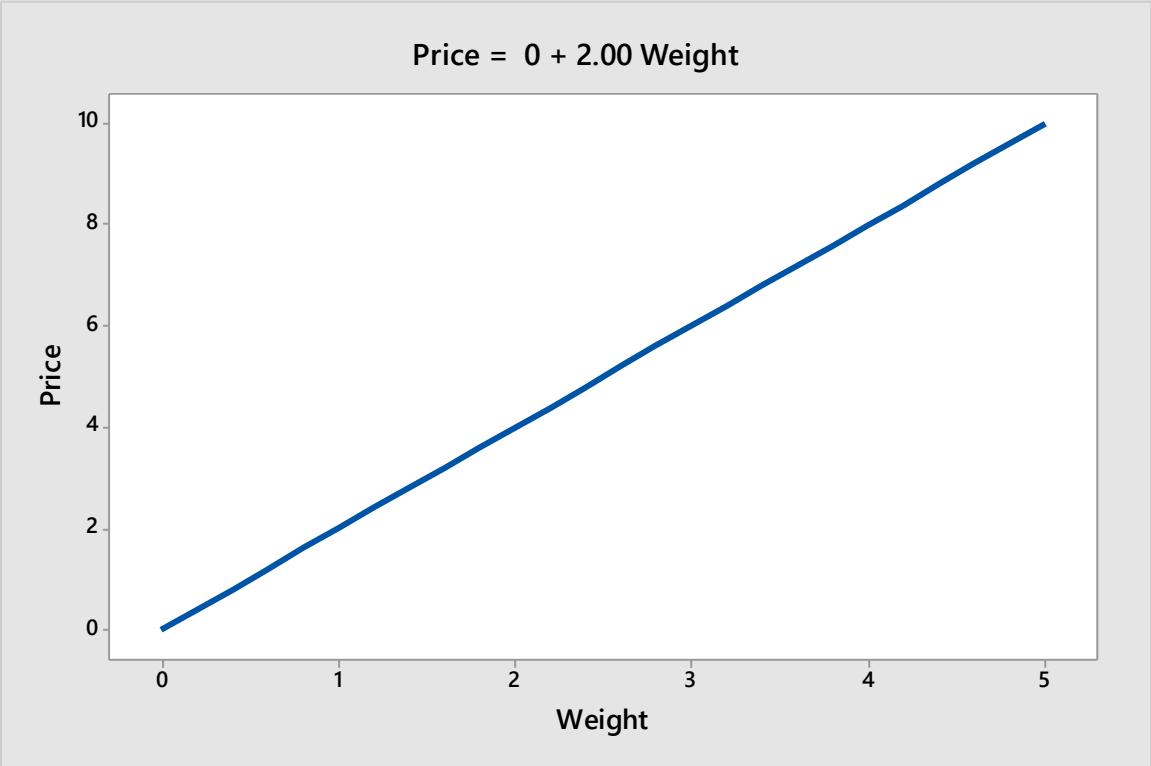
Consider this story about a taxi fare. The amount of a taxi fare, Y, in euro, depends on the length of the journey, X, in km. Suppose that the cost per km is €1.50, then $b = 1.5$ since for every one extra kilometre, the fare increases by €1.50. But this is not the full story. If you call a taxi, even before you begin the journey, there is an initial fee of about €4.50. This parameter, ***a***, is called the **intercept**. It is the value of Y when X is zero. You could say that ***a*** represents how big Y is at the start, and that ***b*** represents how fast it grows.

Fig 6.4



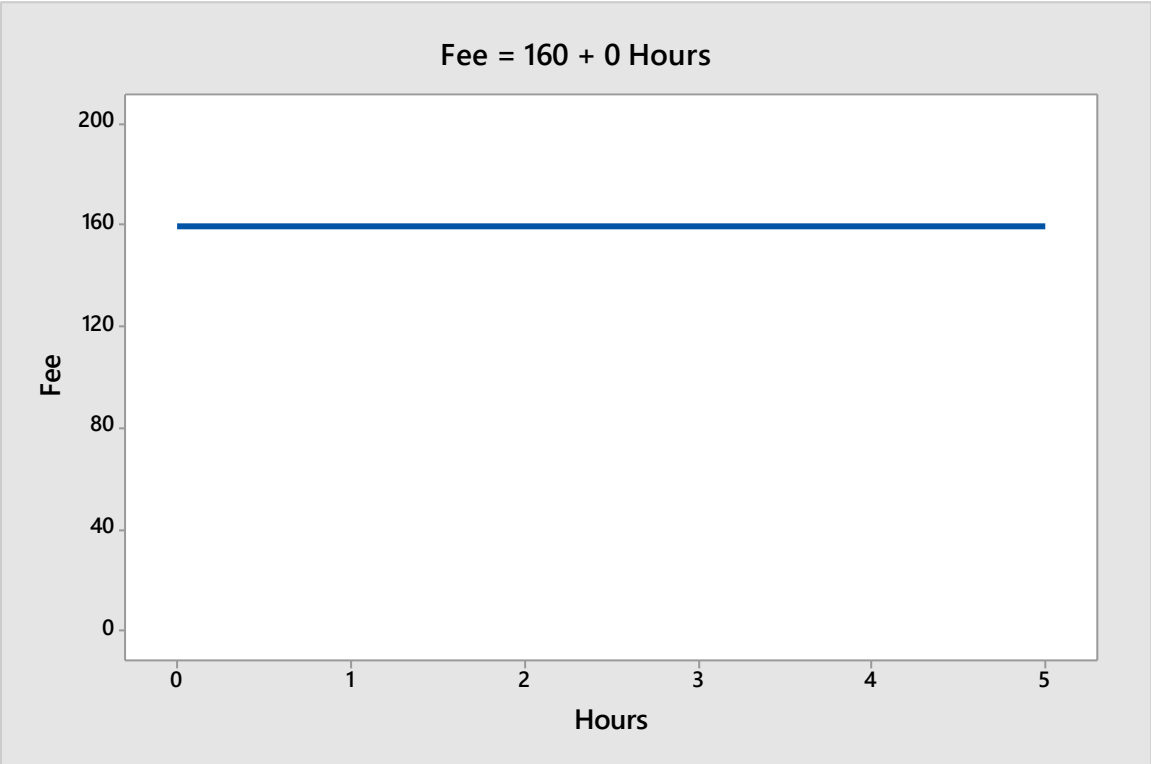
Consider another story about a visit to a shop to buy bananas. The price that you will pay, Y, in euro, depends on the weight of the bananas, X, in kg. Bananas cost €2 per kg, so for every additional kg of bananas that you take, you will pay an extra €2. But, if you decide to come away with no bananas, you don't pay anything, so $a = 0$. We say that the line passes through the origin. Not only is there a linear relationship between X and Y, but X and Y are **directly proportional**. This means that if the value of X is doubled, for example, the value of Y will double also. Any change in X is matched by an identical percentage change in Y.

Fig 6.5



The final story concerns paying a television licence fee. You might think that the television licence fee, *Y*, in euro, would depend on the number of hours per week, *X*, that you spend watching television. But this is not so. Every additional hour per week adds nothing to the licence fee, so $b = 0$. If you have a TV set and never watch it, you must pay the licence fee of €160. So in this case, $a = 160$.

Fig 6.6



Regression Equation

A regression equation is an equation of a line that represents the underlying relationship between two variables, even though the points are scattered above and below the line. The idea is that, even though the individual values of *Y* are scattered, the mean value of *Y,* for each different value of *X,* does lie on the line. The equation of the population regression line is

$$Y = \alpha + \beta X$$

where β is the increase in the mean value of *Y* for every one unit increase in *X*, and α is the mean value of *Y* when *X* is zero. Of course, to find out the values of α and β we would need to take a census of the entire population, so we will usually have to be content with sample estimates. The sample regression equation is

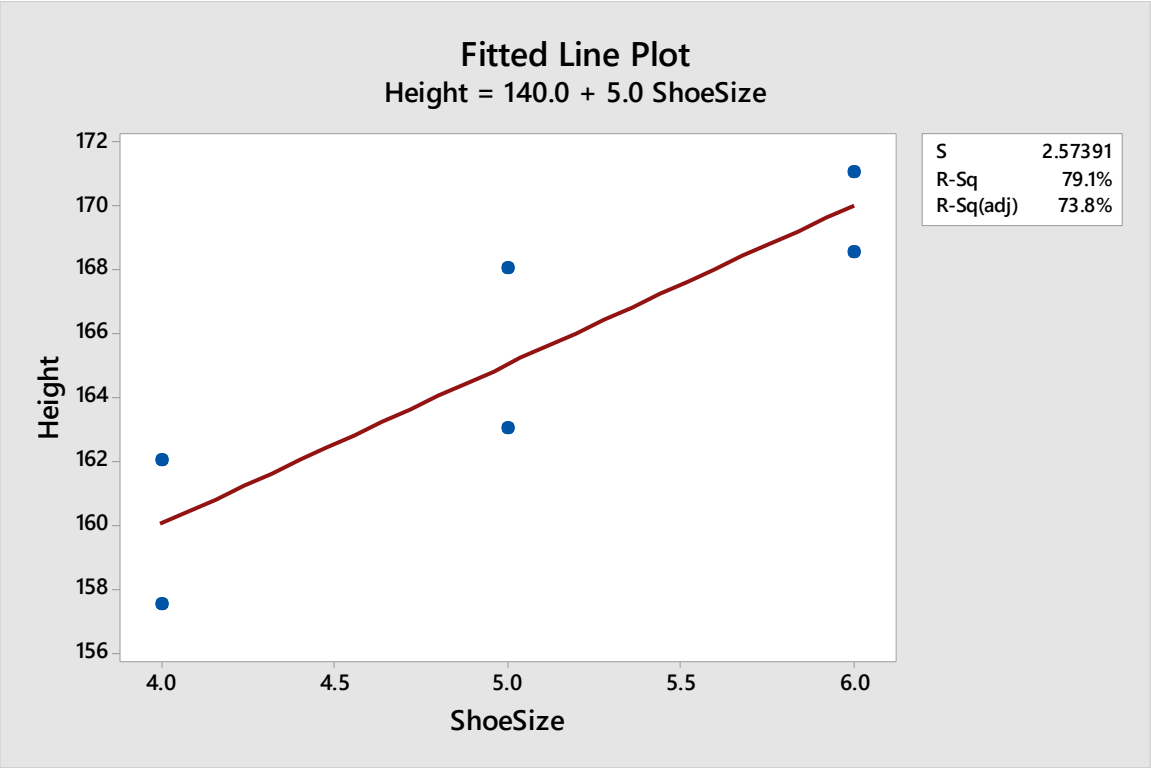
$$Y = a + b X$$

where *a* and *b* are the sample estimates of α and β respectively.

In the context of regression, the slope, *b,* is called the **regression coefficient**, and the intercept, *a,* is called the **regression constant**.

We now take a look at the regression equations for some of our earlier examples.

Fig 6.7



The regression coefficient indicates that every one unit increase in shoe-size tends to be associated with an average increase of 5 cm in height. The regression constant indicates that women who wear a size zero shoe (if such a thing existed) would be 140 cm tall, on average. Notice that the regression constant does not always make sense in isolation (because zero may not be a realistic value for *X*), but it is always meaningful as part of the regression equation.

Fig 6.8

The regression coefficient indicates that the average annual depreciation for these cars is €1340. The regression constant indicates that when new, or nearly new, cars are offered for sale, the mean price of such cars is €14,191.

Values for '*a*' and '*b*' can be found by using your calculator or statistical software.

Prediction

A regression equation can be used to predict an unknown value of *Y*, by substituting the corresponding value of *X* into the equation.

For example, the sample regression equation that relates shoe-size and height is

$$Y = 140.0 + 5.000\ X$$

or

$$Height = 140.0 + 5.000\ ShoeSize$$

Now suppose we find a size 5.5 shoe-print made by an unknown woman. We can estimate how tall this woman is by substituting 5.5 for *X* in the equation.

$$Height = 140.0 + 5.000\ (5.5)$$

$$= 167.5 \text{ cm}$$

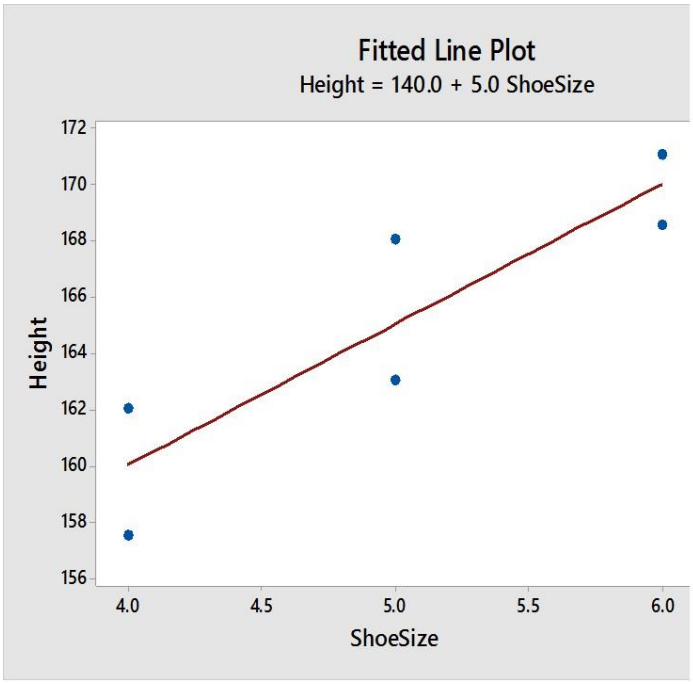
This prediction seems reasonable when we compare it with the heights of women in the sample who wear shoes of around this size. We call this type of prediction **interpolation** because the new value of *X* lies inside the range of the sample data.

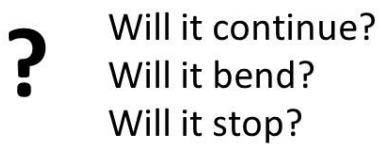
If we use the regression equation to predict the height of a woman who wears a size 14 shoe, we get a value of 210 cm for height. This type of prediction is called **extrapolation**, since this new value of *X* lies outside the range of the sample data. Extrapolation can lead to predictions that have very large errors, as a result of non-linearity in the relationship between *X* and *Y*. Our prediction assumes that the underlying relationship between height and shoe-size continues to be linear even outside the range of the sample data. The relationship may actually exhibit **curvature** rather than linearity, but this is not noticed over the limited range of the sample data.

The Irish property bubble of 2007 is an example of the dangers of extrapolation. An economic crisis occurred because developers, buyers and lenders assumed that property values would continue to increase in a linear way, in line with the levels of price inflation seen in the early years of the millennium. However property prices did not continue to rise, but fell sharply instead. The values of properties in later years turned out to be much lower than predicted and in many cases much lower than the values of the mortgages on those properties, leading to negative equity.

Extrapolation can also lead to predictions that are meaningless, as a result of **discontinuity** in the relationship between *X* and *Y*. Some shoe manufacturers do not make a size 14 shoe, so estimating the height of a woman who wears such a shoe is nonsense. It follows that extrapolation must always be used with caution.

Fig 6.9

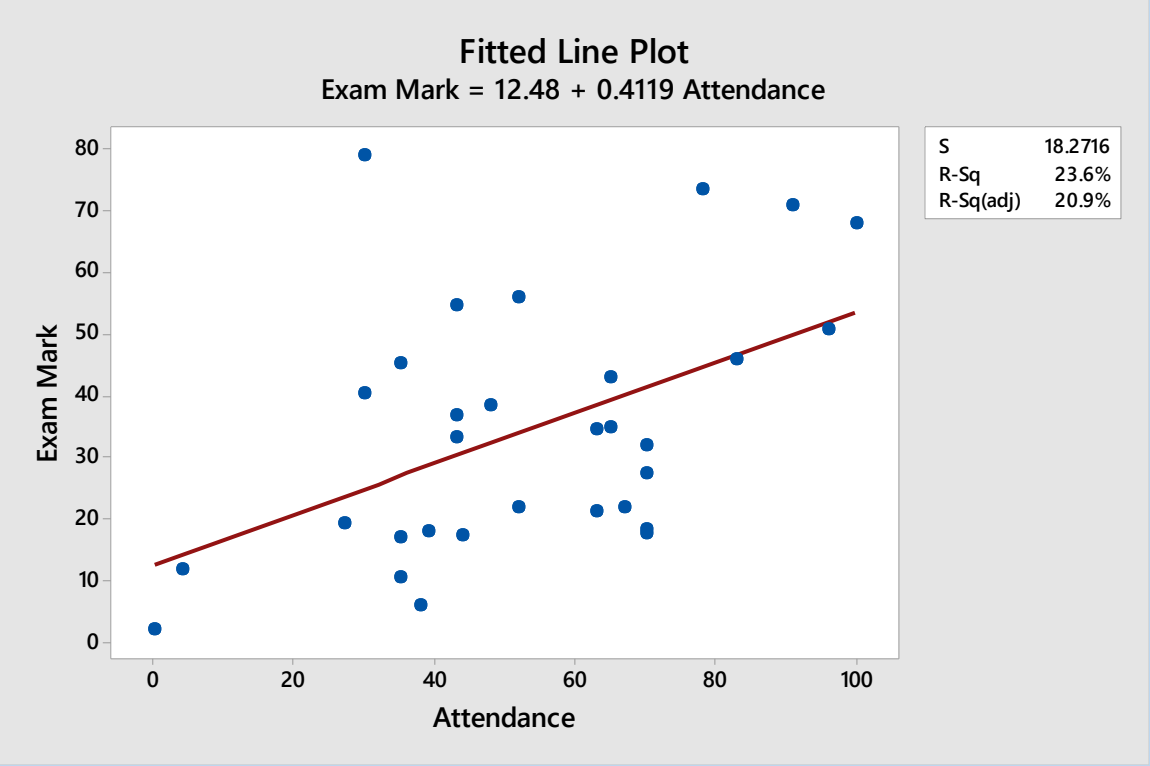




Problems 6B

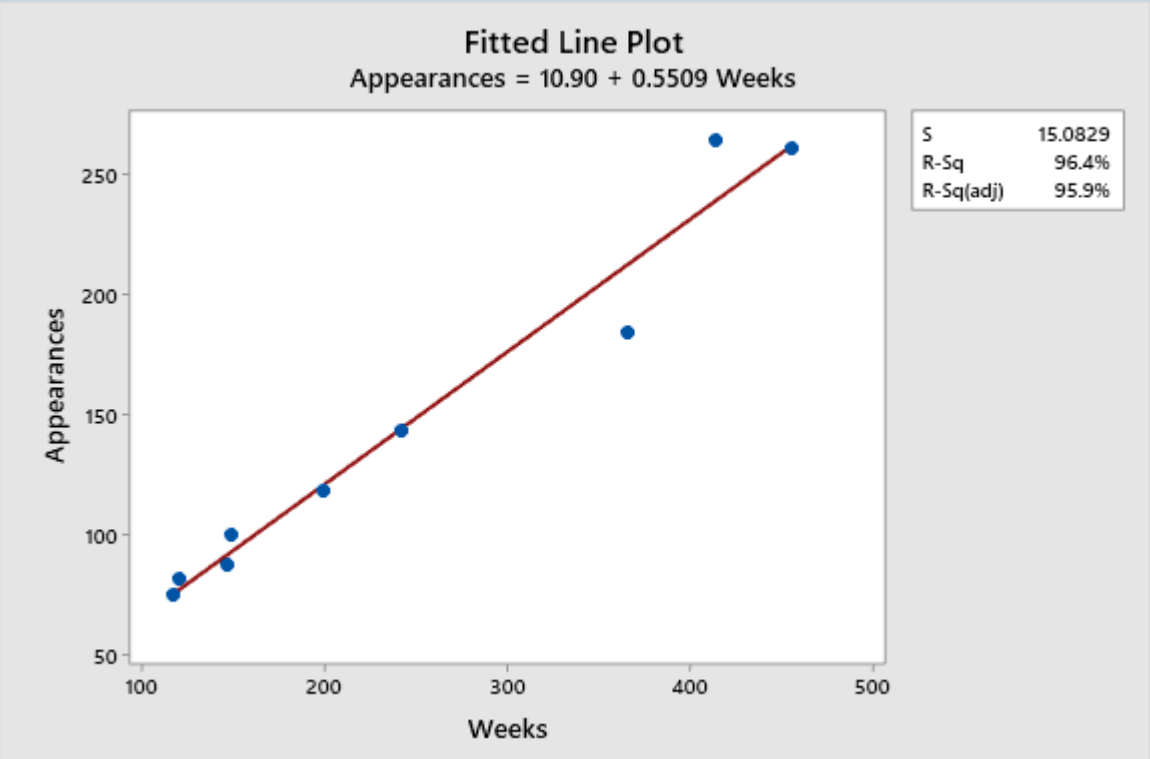
#1. A lecturer made a note of the percent attendance, and the final exam marks, of a number of students on a science course. Take a look at the fitted line plot and explain what is meant by *b, a, R-Sq(adj)* and *S*.

Fig 6.10



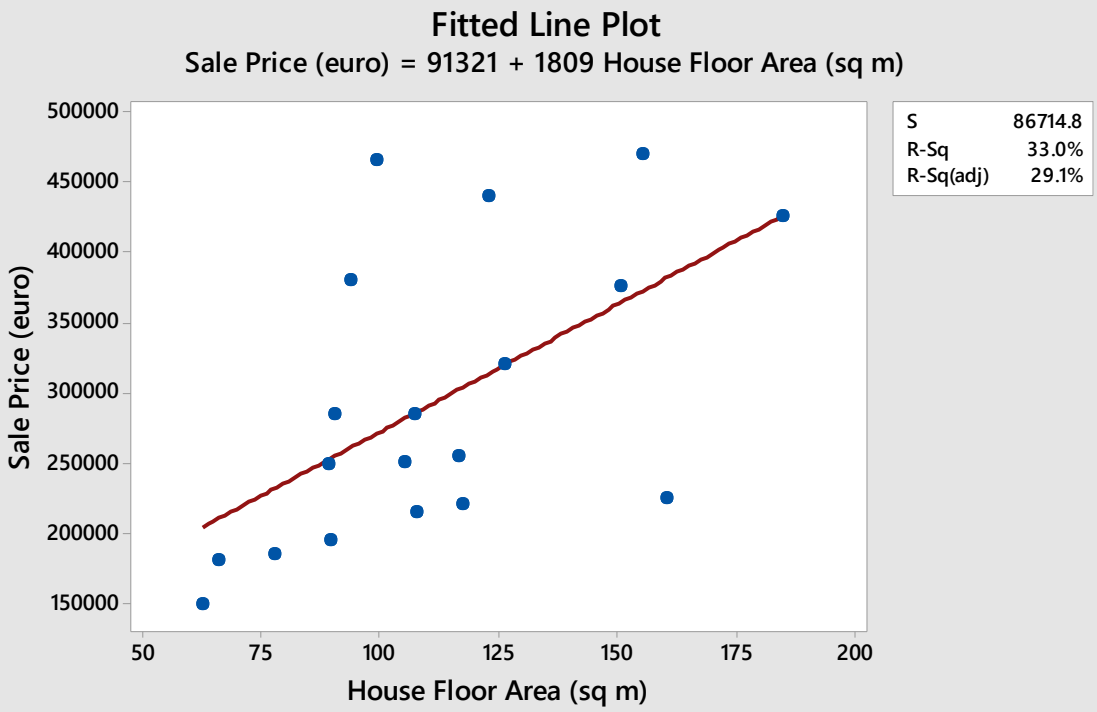
#2. A student selected a random sample of footballers from the premiership, and observed in each case the number of weeks they have been at their current club, and the number of first team appearances they have made for the club. Study the output below and explain what is meant by *b, R-Sq(adj)* and *S,* in this situation.

Fig 6.11



#3. A number of houses for sale in Dublin were sampled, and in each case the floor area in square metres and the sale price in euro were observed. Study the regression output below and predict the sale price of a house with a floor area of 140 square metres. Also, explain what is meant by *b, R-Sq(adj)* and *S,* in this situation.

Fig 6.12



#4. (Activity) Select a sample of car journeys and in each case observe the journey distance (*X*) and the journey time (*Y*).Construct a fitted line plot. Explain what is meant by *b, R-Sq(adj)* and *S,* in this situation. Identify the distance of some new journey that was not in the sample, and estimate the time for that journey.

Project 6B

Simple Regression Project

Investigate the relationship between some predictor and some response, in any application area of your choice. Collect at least 10 data pairs and write a report consisting of the following sections.

(a) Identify the population of interest, the predictor variable and the response variable.
(b) Show the data, the regression equation, and the values of *R-Sq(adj)* and *S*.
(c) Explain what the regression coefficient, *b*, tells us in this situation.
(d) Explain what the value of *R-Sq(adj)* tells us in this situation.
(e) Explain what the standard deviation, *S,* tells us in this situation.

6C. Regression Analysis

Video Lecture https://youtu.be/V91H07JGzNo

Hypothesis Testing in Regression

The most important hypothesis in regression is: H_0: $\beta = 0$

This null hypothesis states that *X* is not a useful predictor of *Y*. Graphically, this null hypothesis states that the population regression line is horizontal. If this null hypothesis is accepted, it indicates that there might be no predictive relationship at all between *X* and *Y*, and the analysis is over. But if this null hypothesis is rejected, it indicates that there is a predictive relationship between *X* and *Y*, and in that case it could be useful to construct a regression equation for predicting values of *Y*.

A second hypothesis is: H_0: $\alpha = 0$. If this null hypothesis is accepted, it means that the population regression line may pass through the origin. This would mean that *Y* is directly proportional to *X,* so that any change in *X* would be matched by an identical percentage change in *Y*. If zero is a meaningful value for *X*, it also means that the corresponding average value for *Y* is zero.

Each hypothesis can be tested using a *p*-value. The *p*-value for the null hypothesis, H_0: $\beta = 0$, can be identified by the name of the predictor variable, and it must be tested first. The *p*-value for the null hypothesis, H_0: $\alpha = 0$, will be denoted as the *p*-value for the constant.

EXAMPLE 1

Some Minitab® software output for the shoe-size and height data is as follows.

Term	Coef	SE Coef	T-Value	P-Value
Constant	140.00	6.52	21.47	0.000
ShoeSize	5.00	1.29	3.89	0.018

The *p*-value for ShoeSize is 0.018, which is less than 5%, so we reject the hypothesis that the population regression line is horizontal. We conclude that shoe-size is a useful predictor of height.

The *p*-value for the constant is 0.000, which is less than 5%, so we reject the hypothesis that the regression line passes through the origin. We conclude that height is not directly proportional to shoe-size. This makes sense: we know that women who wear a size 8 shoe are generally taller than women who wear a size 4 shoe, but they are not twice as tall.

EXAMPLE 2

The software output for the age and price of cars data is as follows.

Term	Coef	SE Coef	T-Value	P-Value
Constant	14191	239	59.28	0.000
Age	−1340.4	44.8	−29.89	0.000

The *p*-value for age is 0.000, which is less than 5%, so we reject the hypothesis that the population regression line is horizontal. We conclude that age is a useful predictor of price.

The *p*-value for the constant is 0.000, which is less than 5%, so we reject the hypothesis that the regression line passes through the origin. We conclude that price is not directly proportional to age. This makes sense, because we know that older cars cost less, not more, and we also know that new cars cannot be bought for free!

Confidence Intervals in Regression

By substituting a new value of *X* into the regression equation, we can predict the corresponding value for *Y*. This is called a **point estimate** or a **fitted value**. It is unlikely to be perfect, so a confidence interval would be useful.

EXAMPLE 1 This output predicts the height corresponding to a shoe-size of 5.5.

Fit	SE Fit	95% CI	95% PI
167.5	1.23216	(164.079, 170.921)	(159.577, 175.423)

The **fit**, i.e. the fitted value, indicates that we expect a woman who wears a size 5.5 shoe to be 167.5 cm tall. The '95% CI' is a **confidence interval** for the average height of all women in the population who wear size 5.5 shoes. The '95% PI' is a **prediction interval** that estimates the height of an individual woman who wears a size 5.5 shoe. The confidence interval estimates the height of the population regression line, when *X* = 5.5. The prediction interval estimates the height of an individual point on the graph, when *X* = 5.5.

EXAMPLE 2 This output predicts the price of a car that is 8 years old.

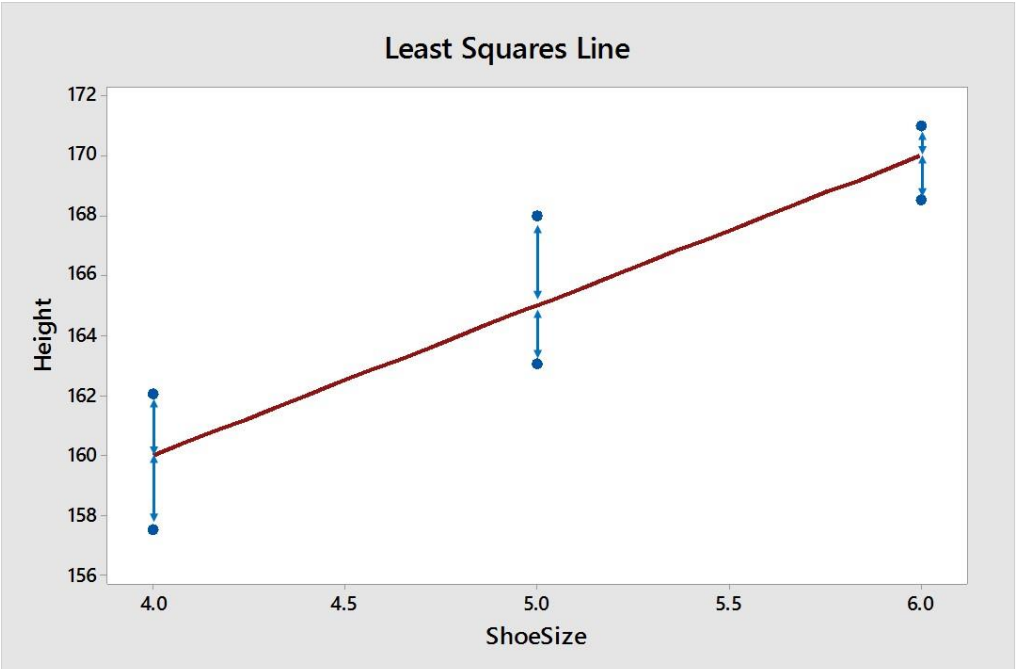
Fit	SE Fit	95% CI	95% PI
3467.73	203.008	(2999.59, 3935.86)	(2418.39, 4517.06)

The average price of 8-year old cars could be as low as €3000 or as high as €3936. But an individual 8-year old car could be offered for sale for as little as €2418 or as much as €4517.

The prediction interval, in particular, is sensitive to the normality assumption. Also, while a single confidence or prediction interval can be quoted with 95% confidence, multiple intervals that are based on the same set of data should be used cautiously, as they do not have an independent probability of 95% of being correct.

Least Squares

Fig 6.13



The purpose of a regression equation is to predict values of *Y*. Therefore the line of best fit is the line that achieves the smallest errors in predicting the *Y* values. The errors are the vertical distances from the points to the line, since the vertical distances represent the errors in *Y*. These errors (or residuals) are illustrated by the short vertical line segments on the graph. The regression line is chosen so that the sum of squared errors is a minimum and therefore it is called the least squares line.

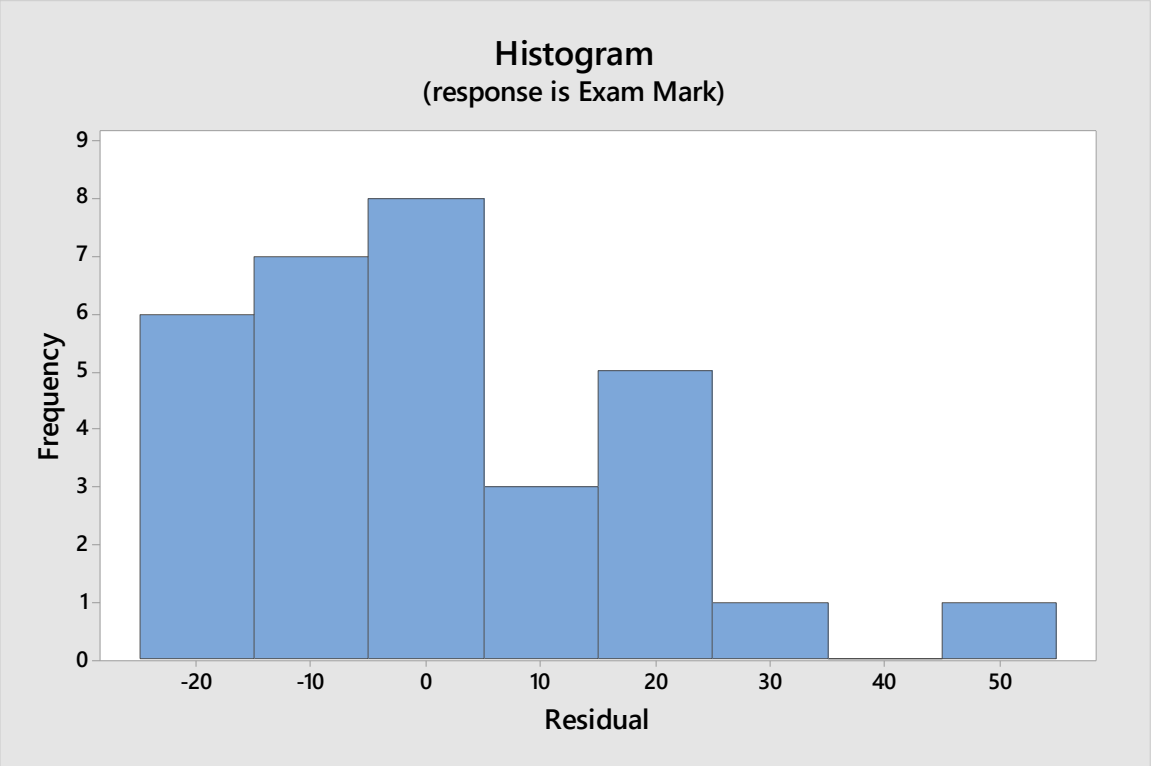
From this it follows that the least squares regression equation of *Y* on *X* is not the same as the least squares regression equation of *X* on *Y*. So if we want to predict shoe-size from height, we need to calculate a new regression equation, with height as *X* and shoe-size as *Y*.

Every *Y* value in the sample is assumed to have been drawn at random, therefore the values of *Y* must not be selected in advance, although the *X* values can be selected in advance if we wish. In a situation where the experimenter has selected the *X* values (e.g. concentration), and needs to predict *X* from *Y* (e.g. absorbance), the regression equation of *Y* on *X* should first be constructed and then transformed to provide a formula for *X*.

Assumptions

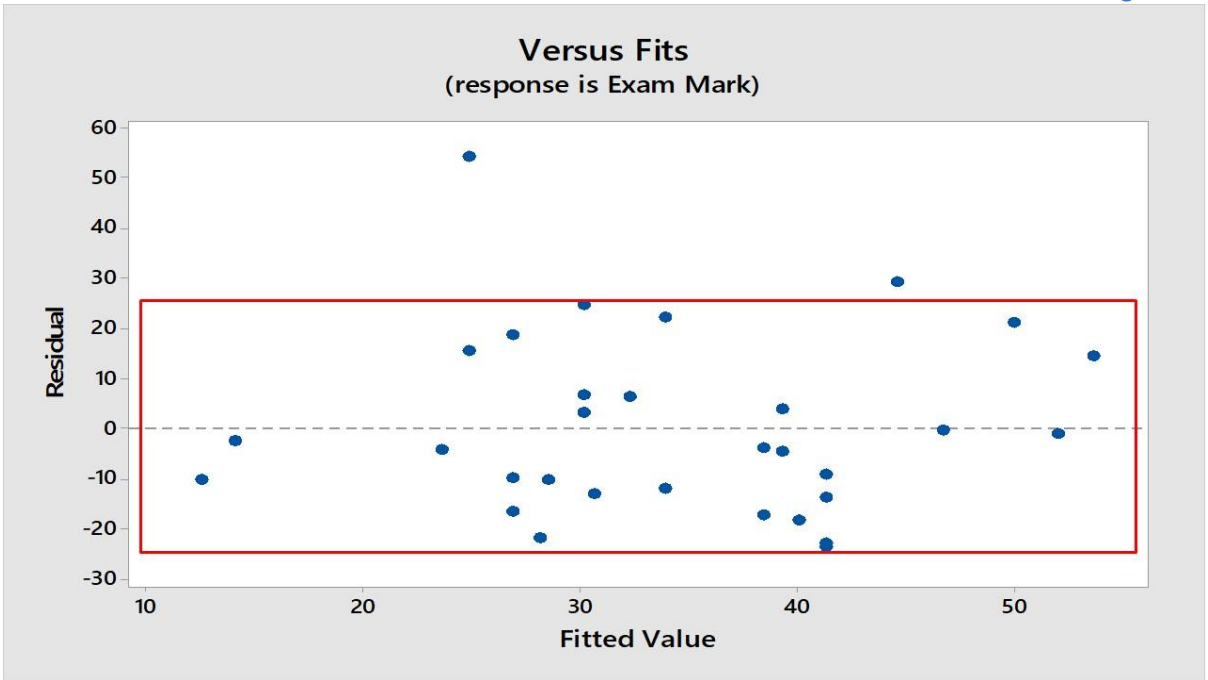
The first assumption of linear regression is that the **linear model** is the appropriate model. This means that for any value of *X*, the mean value of *Y* is given by the linear equation $Y = \alpha + \beta X$. In other words, the average values of *Y* do lie on a straight line. This assumption can be confirmed by drawing a scatterplot and observing that the points seem approximately linear without any obvious curvature (see Fig 6.10). Linear regression is often used in situations where this assumption is only approximately true. In the words of George Box, 'All models are wrong, but some are useful.'

Fig 6.14



A second assumption is usually made, that the errors are **normally distributed**. This means that for each value of *X*, the population of *Y* is normal. The normality assumption can be tested by drawing a histogram of the residuals. The histogram should look approximately normal.

Fig 6.15



A third assumption is usually made, that *Y* has a **constant variance** that does not depend on *X*. This means that for each value of *X* the standard deviation of *Y* is the same even though the mean value of *Y* is different. The constant variance assumption can be tested by drawing a plot of residuals versus fits. When viewed from left to right, the plot should show a set of points of roughly uniform depth and height, like a horizontal belt, not changing in height like, for example, a right opening megaphone.

This process of testing the assumptions is called **residual analysis** and it is illustrated here for the exam marks data presented earlier in Fig 6.10. The residual plots illustrated here are not altogether satisfactory. The histogram is not normal. Also the plot of residuals versus fits should ideally see the points falling inside the area that has been outlined on the graph in red. We should be cautious about placing too much reliance on a prediction interval in this situation.

Problems 6C

#1. Some output is shown below from a regression analysis of exam mark on attendance. Test the two null hypotheses $\beta = 0$ and $\alpha = 0$, and state your conclusions.

Term	Coef	SE Coef	T-Value	P-Value
Constant	12.48	8.03	1.56	0.131
Attendance	0.412	0.138	2.99	0.006

Also, use the output below to predict the mark of a student with 40% attendance.

Fit	SE Fit	95% CI	95% PI
28.9585	3.75129	(21.2862, 36.6307)	(-9.19061, 67.1076)

#2. Some output is shown below from a regression analysis of the number of appearances a player has made for a premiership club on the number of weeks they have been at the club. Test the two null hypotheses $\beta = 0$ and $\alpha = 0$, and state your conclusions.

```
Term                Coef    SE Coef  T-Value  P-Value
Constant            10.9       11.0     0.99    0.357
Weeks             0.5509     0.0401    13.73    0.000
```

Also, use the output below to predict the number of appearances of a player who has been at their club for 300 weeks.

```
    Fit     SE Fit          95% CI               95% PI
176.178   5.48701   (163.203, 189.153)   (138.226, 214.130)
```

#3. Some output is shown below from a regression analysis of sale price, in euro, on house floor area, in square metres. Test the two null hypotheses $\beta = 0$ and $\alpha = 0$, and state your conclusions.

```
Term              Coef  SE Coef  T-Value  P-Value
Constant         91321    72490     1.26    0.225
House Floor Area  1809      625     2.89    0.010
```

Refer to the output below and explain what is meant by the CI and the PI. The new value of X was 140 square metres.

```
   Fit    SE Fit         95% CI              95% PI
344643  26698.0   (288315, 400971)    (153216, 536070)
```

#4. (Activity) Select a sample of car journeys and in each case observe the journey distance and the journey time. Test the two null hypotheses, $\beta = 0$ and $\alpha = 0$, and state your conclusions. Identify the distance of some new journey that was not in the sample, and construct an interval estimate of the time for that journey.

Project 6C

Regression Analysis Project

Use regression analysis to investigate the relationship between some predictor and some response, in any application area of your choice. Collect at least 10 data pairs and write a report consisting of the following sections. Display the relevant software output in each case to support your answers.

(a) Identify the population of interest, the predictor variable and the response variable.

(b) Show the data, a fitted line plot, and the regression equation.

(c) Explain what r^2 (adjusted) means in this situation.

(d) Test the two hypotheses, $\beta = 0$ and $\alpha = 0$, and explain what your conclusions mean.

(e) Identify a new individual case where you could use the value of the predictor variable to predict the value of the response. Explain why that individual case is of interest to you. Use this new value of the predictor variable to construct a prediction interval for the response and explain what the prediction interval means.

6D. Multiple Regression and Non-Linear Regression

Video Lecture https://youtu.be/tLv710NO2_g

Multiple Regression

If a number of suitable predictor variables are available then a multiple regression equation can be constructed. For example, *VO2 Max*, the peak oxygen uptake, is an important measure of an athlete's endurance. It may be possible to predict VO2 Max, using age and weight as predictor variables.

The regression analysis below is based on data sampled from hurling players.

Term	Coef	SE Coef	T-Value	P-Value
Constant	56.61	4.89	11.57	0.000
Age	0.371	0.108	3.44	0.001
Weight	-0.1179	0.0540	-2.18	0.035

This analysis indicates that both Age and Weight are useful predictors of VO2 Max, since both *p*-values are significant.

The regression equation below indicates that VO2Max increases with Age but decreases with Weight.

Regression Equation

VO2Max = 56.61 + 0.371 Age - 0.1179 Weight

This is a linear equation in three dimensions, and could be represented by a plane. It is interesting to compare this multiple regression equation with the following simple regression equation that uses Age as the only predictor.

Term	Coef	SE Coef	T-Value	P-Value
Constant	47.68	2.82	16.91	0.000
Age	0.347	0.112	3.09	0.004

Regression Equation

VO2Max = 47.68 + 0.347 Age

Notice that the *p*-value of Age differs depending on whether Weight is also present in the model. The *p*-value 0.004 indicates that Age is useful on its own as a predictor of VO2Max. The *p*-value 0.001 in the first model indicates that Age is useful as a predictor of VO2Max, even if Weight is also available as a predictor.

Also, the coefficient of Age is not the same in the two models. In the simple regression equation, it signifies that every one unit increase in Age is associated with an increase of 0.347 in VO2Max. In the multiple regression equation, it signifies that every one unit increase in Age is associated with an increase of 0.371 in VO2Max, provided that the Weight remains constant.

Selection of Variables

There may be other variables that could also be used to predict the VO2Max, such as *Height, BMI*, *Bodyfat*, and so on. We are now faced with a choice among a large number of alternative regression models. We could use any one predictor variable, any two, any three, etc. How can we choose the **best subset** of predictors? A simple

criterion is to choose the model that has the highest r^2(adjusted) value. Software can be used to evaluate all possible subsets and to present the best candidates. The final headings in the output below should be read vertically, and each different model is represented by a different row of the output. An X denotes a variable that is included in the model.

Response is VO2Max

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Age	Height	Weight	Biking	Skinfat	Bodyfat	RunGrade	LungCap	Chin	Fitness	Time	Twenty
1	19.7	17.6	9.6	5.0	2.6626	X											
2	28.6	24.9	16.2	2.3	2.5425	X		X									
3	35.7	30.5	20.1	0.6	2.4453	X										X	X
4	39.3	32.5	19.9	0.7	2.4095	X		X								X	X
5	41.6	33.3	20.6	1.5	2.3963	X		X						X		X	X
6	42.0	31.7	12.5	3.3	2.4239	X		X						X	X	X	X
7	43.9	32.0	14.6	4.3	2.4195	X	X	X	X					X		X	X
8	44.7	30.9	7.8	5.9	2.4388	X	X	X	X					X	X	X	X
9	45.9	30.2	3.1	7.3	2.4508	X	X	X	X	X	X			X		X	X
10	46.2	28.3	0.0	9.1	2.4834	X	X	X	X	X	X			X	X	X	X
11	46.4	26.1	0.0	11.0	2.5225	X	X	X	X	X	X	X		X	X	X	X
12	46.4	23.4	0.0	13.0	2.5669	X	X	X	X	X	X	X	X	X	X	X	X

Adding more and more predictor variables to the model might seem like a good idea but there are a number of reasons to be cautious about adding extra predictor variables.

First, although an extra predictor variable will always increase the value of *R*-squared, unless the value of *R*-squared (adjusted) increases, the variable is probably of no use. A degree of freedom is lost every time a new variable is introduced, so it is advisable to use the statistic that is adjusted for this loss of degrees of freedom. Therefore, when comparing models with different numbers of predictor variables, use *R*-squared (adjusted) and not simple *R*-squared.

Second, if the addition of an extra predictor variable produces only a small improvement in *R*-squared (adjusted) then it may be wiser to omit the extra predictor variable, especially if it is not statistically significant. This approach, where a simpler model is chosen in preference to a more complex model, is often referred to as **Occam's razor** or the **law of parsimony**.

Third, we should be wary of using any pair of predictor variables that are highly correlated with each other. This is called **multicollinearity** and it tends to inflate the errors in the coefficient estimates. For example, if $r > 0.9$ then the variance inflation factor, ***VIF***, exceeds 5, which can be regarded as a threshold for a tolerable *VIF* value. This indicates that the variance of the coefficient estimate for this predictor is more than five times greater as a result of its correlation with other predictors.

Fourth, whenever you select a large number of predictors, there is a possibility of **overfitting** the model, i.e. including some predictor variables that are merely explaining the random variation in the sample rather than the underlying relationship in the population. A low value for ***R*-squared (predicted)** provides a warning that overfitting has occurred. The value of *R*-squared (predicted) is obtained by omitting the observations one at a time, and using the remaining observations to predict the missing observation. So *R*-squared (predicted) gives an idea of how good the model is at predicting new values of the response. It is also helpful to validate a model by applying it to a fresh set of data.

Fifth, a **Mallows' *Cp*** value that is close to $p + 1$ where p is the number of predictors, suggests a good balance between a biased model with too few predictors and an imprecise model with too many predictors.

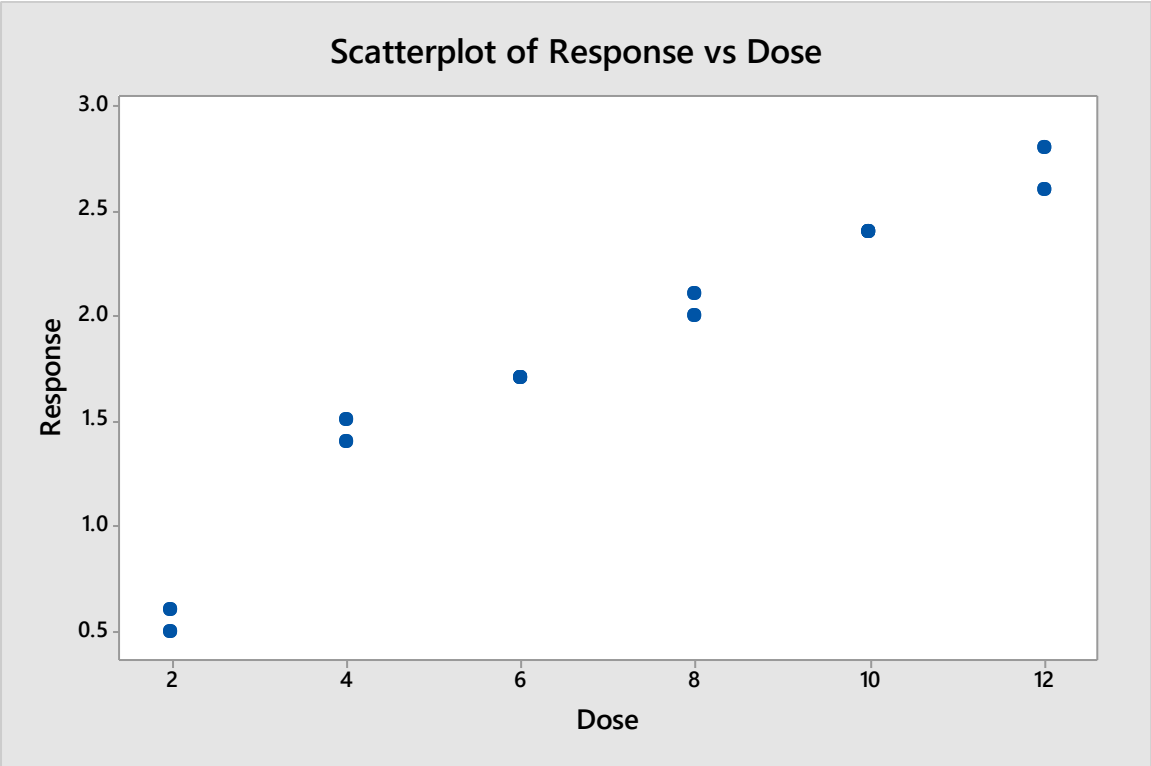
If you already have historical data from some process, best-subsets analysis is a simple way to trawl through the data for clues about which variables are useful for predicting some important response. But perhaps the most important step is to identify all the **potentially relevant variables** at the beginning and to make sure to include all these variables among the observations. For example, in this best subsets analysis, the hurling player's position was not included as a predictor, but this variable could be relevant.

Non-Linear Regression

Sometimes a scatterplot reveals a **curvilinear relationship** between two variables.

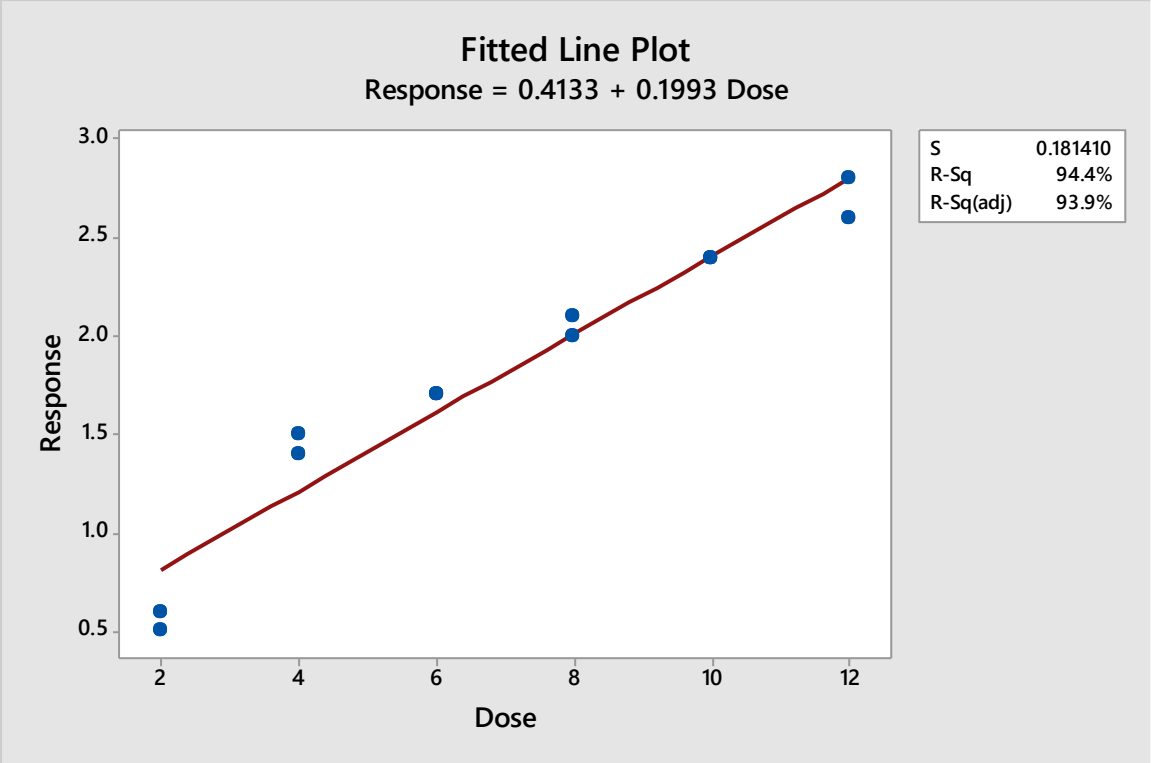
EXAMPLE Different doses of a drug were administered and the responses were observed in each case.

Fig 6.16



The curvature can be easily seen and if we went ahead and naïvely fitted a simple linear regression model, the **Lack-of-Fit** of the model would be obvious.

Fig 6.17



Look at the data points for a dose of two units. There are two replicate measurements of response and these are not identical, because the same dose will not always elicit the same response. We see this feature in every regression plot and it is called **Pure Error**. But there is another source of error present. It seems obvious that the *mean* response for a dose of two units does not lie on the line of best fit. This is called **Lack-of-Fit** and it indicates that the model used is inappropriate. The pure error measures how much the replicates differ from each other, while the lack-of-fit error measures how much the average measurements differ from the fitted values. The software tests if the lack-of-fit error is significantly greater than the pure error and provides a *p*-value to test the null hypothesis that the model is a good fit to the data.

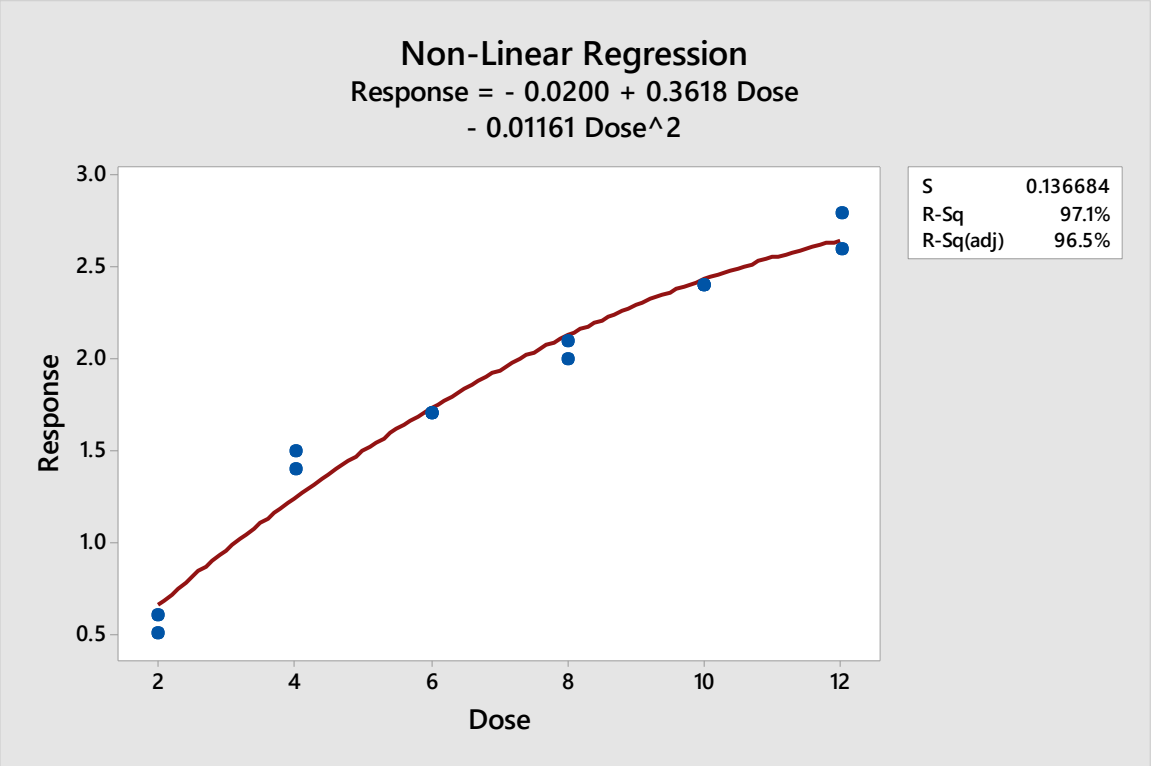
Regression Analysis: Response versus Dose

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.56007	5.56007	168.95	0.000
Dose	1	5.56007	5.56007	168.95	0.000
Error	10	0.32910	0.03291		
Lack-of-Fit	4	0.29410	0.07352	12.60	0.004
Pure Error	6	0.03500	0.00583		
Total	11	5.88917			

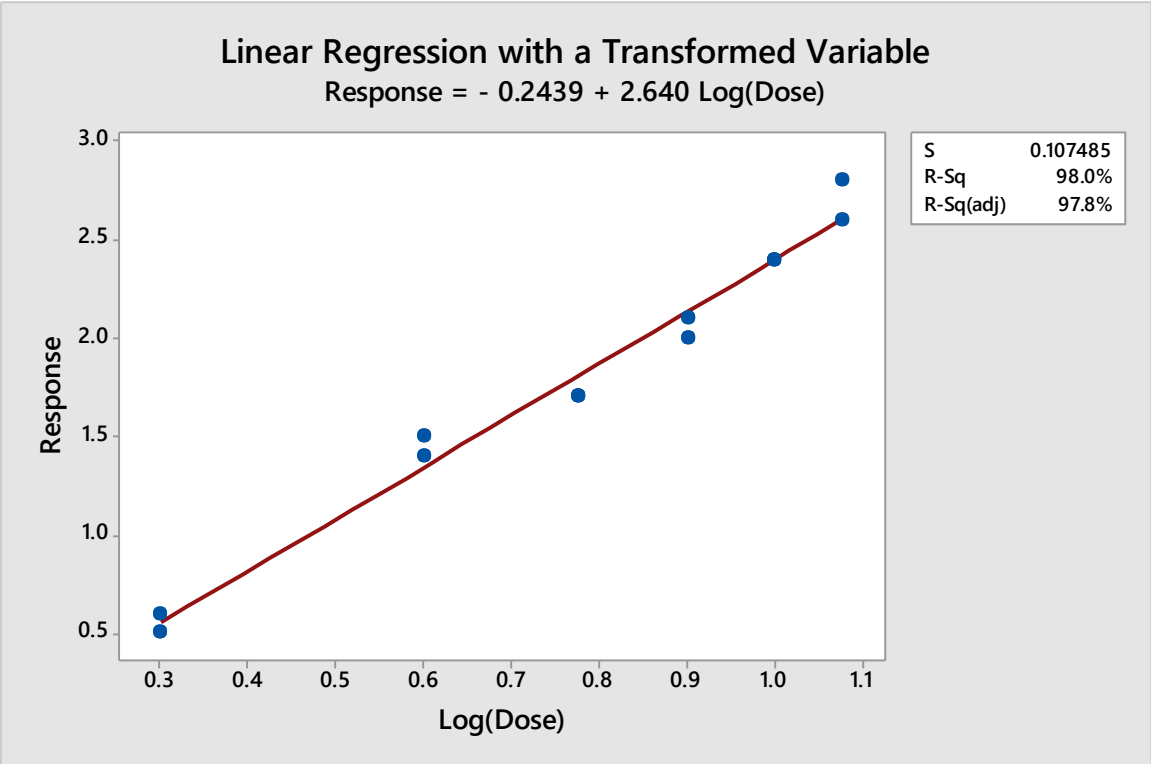
When faced with a curvilinear relationship between two variables, there are two different approaches that can be taken. The first approach is to fit a **non-linear equation**. A quadratic equation has an additional X^2 term and provides a simple curve with a single elbow, which can be either concave upwards or concave downwards.

Fig 6.18



The second approach is to apply a mathematical **transformation** which will have the effect of straightening out the curve. To begin with, the *X* variable is simply *X* itself, or X^1. This is called the identity transformation. We can stretch the *X*-axis by raising the power to X^2 or X^3. Or we can shrink the *X*-axis by using $\sqrt{X}$ or *log X* or X^{-1} or X^{-2}. The further the power is raised or lowered, the more stretching or shrinking occurs.

Fig 6.19



Having applied a linearising transformation, we are now dealing with a linear regression, and we can use all the usual techniques such as prediction intervals. Just be careful to remember that we are now dealing with a transformed variable: in this example, X now represents the log of the dose and not the dose itself, so $X = 1$ corresponds to a dose of 10 units.

If you try a transformation that bends the curve too far, or not far enough, just try again. But any transformation chosen in this way should be validated using a fresh data-set to guard against overfitting. If you have a good knowledge of the underlying process, then you may be able to identify the appropriate transformation even before looking at the data. For example, if X is the diameter of an orange and Y is its weight, then X^3 is probably the correct transformation.

The Y variable could be transformed instead of the X variable. For a Y variable (such as total bacteria count) that grows more rapidly as X increases and that also exhibits greater random variation at larger Y values, it can sometimes happen that a logarithmic transformation of Y can provide the double benefit of both linearising the relationship and stabilising the variance.

Problems 6D

#1. The output below shows a regression equation to predict the drying time of glue using the moisture content of the glue and the relative humidity of the atmosphere as predictors. Explain in words what the two coefficients mean. Time is measured in minutes, and moisture content and relative humidity are both measured in %.

```
Regression Equation
Time = -15.9 + 2.17 Moisture Content + 0.975 Relative Humidity
```

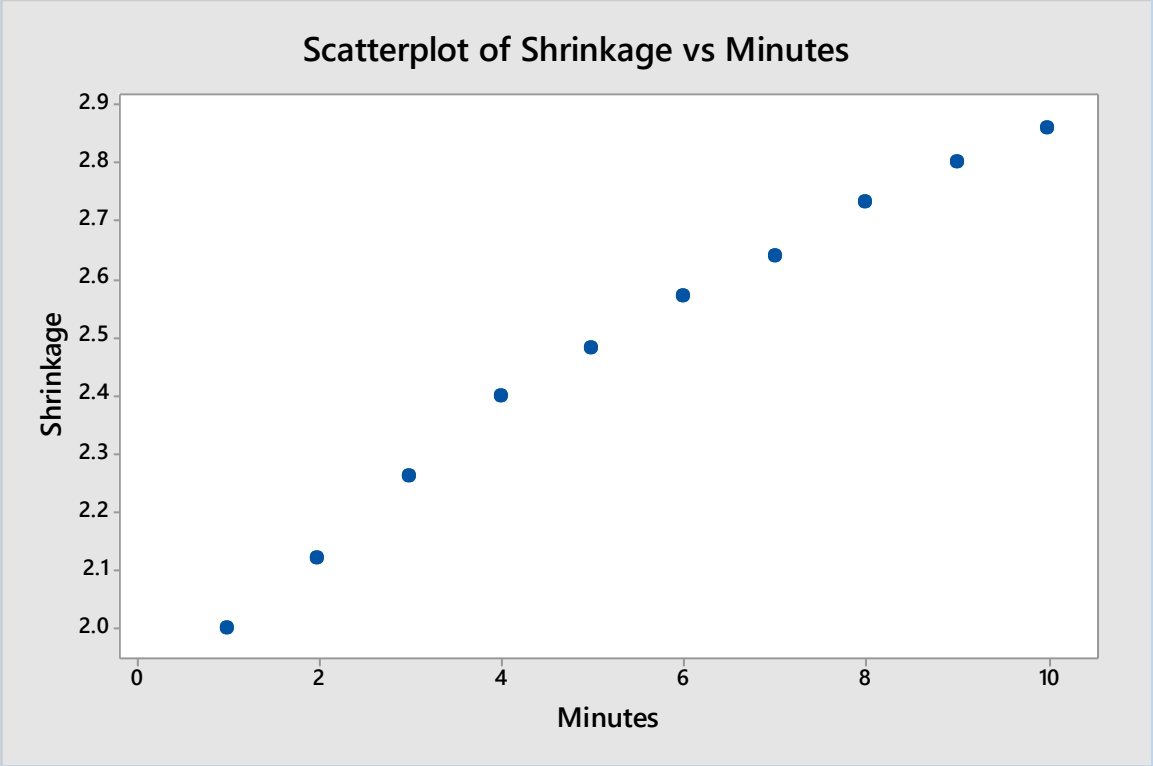
#2. Study the output below and identify the best subset of predictors for the price of a car, from among the available predictors: age, kilometres and engine size.

Best Subsets Regression: Price versus Age, Kilometers, Engine
Response is Price

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Age	Kilometers	Engine
1	99.1	99.0	98.5	1.6	407.25	X		
1	78.9	76.2	66.5	175.2	1987.1		X	
2	99.2	99.0	98.3	2.7	407.22	X	X	
2	99.1	98.9	94.8	3.5	431.06	X		X
3	99.3	99.0	95.0	4.0	417.58	X	X	X

#3. Plastic components are produced by a moulding process. The components experience shrinkage for a number of minutes afterwards. The scatterplot below shows the relationship between shrinkage, in mm, and minutes. If you wanted to construct a regression model with shrinkage as the response and minutes as the predictor, how would you proceed?

Fig 6.20



Project 6D

Multiple Regression Project

Use regression analysis to investigate the relationship between some response and at least four predictor variables, using a sample size of at least 10. Begin your investigation with a best subsets analysis. Write a report consisting of the following sections and display the relevant output in each case to support your answers.

(a) Identify the population of interest, the response variable and the candidate predictor variables.

(b) Identify the best subset of predictor variables that you will use to construct the regression model. There should be a minimum of two predictor variables in your selected model and you should justify your choice of model.

(c) Construct the regression model using your selected set of predictor variables and explain carefully in words what each coefficient in the model tells you.

(d) In relation to the regression model that you have constructed, explain carefully in words what is meant by each of the *p*-values, by *r*-squared adjusted and by *S*.

(e) Use your selected model to construct a prediction interval based on some new set of values for the predictor variables and carefully explain what this prediction interval tells you.

6E. Binary Logistic and Partial Least Squares Regression

Video Lecture https://youtu.be/YOtSNCsQk0A

Binary Logistic Regression

Binary Logistic Regression is used when the response is an attribute, i.e. a yes or no outcome. Either one or more predictor variables can be used.

EXAMPLE

A credit **scorecard** is a model that attempts to identify borrowers who would be likely to default. This example shows how a binary logistic regression model could be used to predict default, using the applicant's time at their current address, and their time in their current job. Some of the output is shown.

Analysis of Variance

		Wald Test	
Source	**DF**	**Chi-Square**	**P-Value**
Regression	2	9.05	0.011
MonthsAtAddress	1	0.62	0.432
MonthsInJob	1	8.54	0.003

The overall regression *p*-value indicates that this model is a useful predictor of default. The individual *p*-values indicate that the applicant's time in their current job, in particular, is a useful predictor of default.

Coefficients

Term	**Coef**	**SE Coef**	**VIF**
Constant	2.87	1.13	
MonthsAtAddress	-0.0169	0.0215	1.00
MonthsInJob	-0.0521	0.0178	1.00

The sign of the coefficient for *MonthsInJob* is negative. This indicates that the longer an applicant has been in their current job, the less likely they are to default. The same might be true of the time they have lived at their current address.

Odds Ratios for Continuous Predictors

	Odds Ratio	**95% CI**
MonthsAtAddress	0.9832	(0.9426, 1.0256)
MonthsInJob	0.9492	(0.9166, 0.9830)

For every extra month that an applicant has been in their current job, the estimated odds of a default reduces to 0.9492 of its previous value. Confidence intervals for the odds ratio are provided. Notice that, for the time at the current address, the confidence interval for the odds ratio includes the value 1.0 which would represent no change in the odds ratio, and this is consistent with the *p*-value for this variable which was not significant.

Prediction for Default

Variable	Setting
MonthsAtAddress	29
MonthsInJob	18

Fitted Probability	SE Fit	95% CI
0.808946	0.0910746	(0.571558, 0.930742)

The output indicates that an applicant who has been 29 months at their current address, and 18 months in their current job has a high probability of default.

When a small data set is used to construct a binary logistic regression model with a large number of predictor variables, a problem called **separation** can occur. This is a computational problem which prevents the calculation of the coefficients because the predictor variables are able to provide a perfect prediction of the response values in the sample. This problem can usually be addressed by using a larger set of sample data, or by fitting an alternative model with fewer predictor variables.

Partial Least Squares Regression

Regression methods cannot be used where there are more predictor variables than observations. Consider the extreme example of trying to draw a simple regression line on a scatter plot on which only one point is displayed. The amount of information available (one point) is simply not enough to answer all the questions about the values of the intercept and slope (two parameters) in this two-dimensional space. The same problem arises in a multidimensional situation where there are more parameters to be estimated than there are observations available.

This problem can be overcome by projecting the variables onto a smaller number of new variables so that it becomes possible to build a regression model. To take a well-known example, consider how the two variables, weight and height, can be projected onto a single variable, *BMI*, which does a better job at predicting health outcomes than either of the original variables. *BMI* could be described as a **latent variable** because it is hidden among the more obviously available variables but yet has better explanatory power.

In a similar way, a partial least squares regression model projects the original predictor variables (*X*) and response variables (*Y*) to a new space with lower dimensionality. The goal is to find the best possible latent variables, by combining the *X* variables, in order to explain as much of the variance in *Y* as possible. This technique is properly called **projection to latent structures**, but is commonly referred to as partial least squares regression (PLS regression).

PLS regression can also be used to deal with situations where the predictor variables are correlated with each other (multicollinearity), because the latent variables that arise in PLS regression are uncorrelated. This situation arises in chemometrics where spectral measurements often provide predictor variables that are correlated. A PLS regression can be constructed to predict chemical composition or other physio-

chemical properties. If there are multiple responses that are correlated, these can be modelled in a single PLS regression model. But where multiple responses are uncorrelated, it is better to use individual PLS regression models to model the different responses.

EXAMPLE

A sports scientist wishes to explore how the energy expenditure of an athlete is related to a number of variables. There are more predictor variables (17) than observations (9) so a PLS regression is used to explore the relationship. Some of the output is shown.

PLS Regression: EnergyExpend versus Age, BMI, Weight, Bodyfat, HeartrateMax, ...

Method

Cross-validation	Leave-one-out
Components to evaluate	Set
Number of components evaluated	7
Number of components selected	5

Analysis of Variance for EnergyExpenditure

Source	DF	SS	MS	F	P
Regression	5	2223044	444609	74.23	0.002
Residual Error	3	17969	5990		
Total	8	2241012			

Model Selection and Validation for EnergyExpenditure

Components	X Variance	Error	R-Sq	PRESS	R-Sq (pred)
1	0.337996	708434	0.683878	2974876	0.000000
2	0.618647	181756	0.918896	933129	0.583613
3	0.751835	70807	0.968404	932553	0.583870
4	0.832476	30664	0.986317	910520	0.593701
5	0.921339	17969	0.991982	867363	0.612959
6		7583	0.996616	939156	0.580923
7		2174	0.999030	941625	0.579822

A *p*-value of 0.002 indicates that the model selected is significant. The selected model is the one with the highest ***R-Sq*(predicted)** value (61.3%) and this model explains 92.1% of the variance in the predictors.

R-Sq(predicted) is better than *R-Sq*(adjusted) for measuring the usefulness of a model for making predictions, and it is calculated by omitting the observations one at a time from the data set. The model obtained each time is used to predict the missing observation and a cross-validation is carried out using the **PRESS** statistic (predicted residual error sum of squares).

The model can now be used to make predictions when new values of the predictor variables are provided.

Predicted Response for New Observations Using Model for EnergyExpenditure

Row	Fit	SE Fit	95% CI	95% PI
1	3613.06	65.9791	(3403.09, 3823.04)	(3289.41, 3936.72)

Project 6E

Binary Logistic Regression Project

Construct a binary logistic regression model using at least two predictor variables and with a sample size of at least 30 observations. Write a report consisting of the following sections and display the relevant output in each case to support your answers.

(a) Identify the population of interest, the response variable and the predictor variables.

(b) Show the Analysis of Variance table and explain what the p-values indicate.

(c) Refer to the variable which has the lowest p-value and explain what the sign of its coefficient indicates.

(d) Again, in relation to the variable which has the lowest p-value, explain in words what its odds ratio indicates.

(e) Use your selected model to predict the outcome based on some new set of values for the predictor variables and explain in words what your answer means.

6F. Multivariate Analysis

Video Lecture https://youtu.be/oI58PHRJHPE

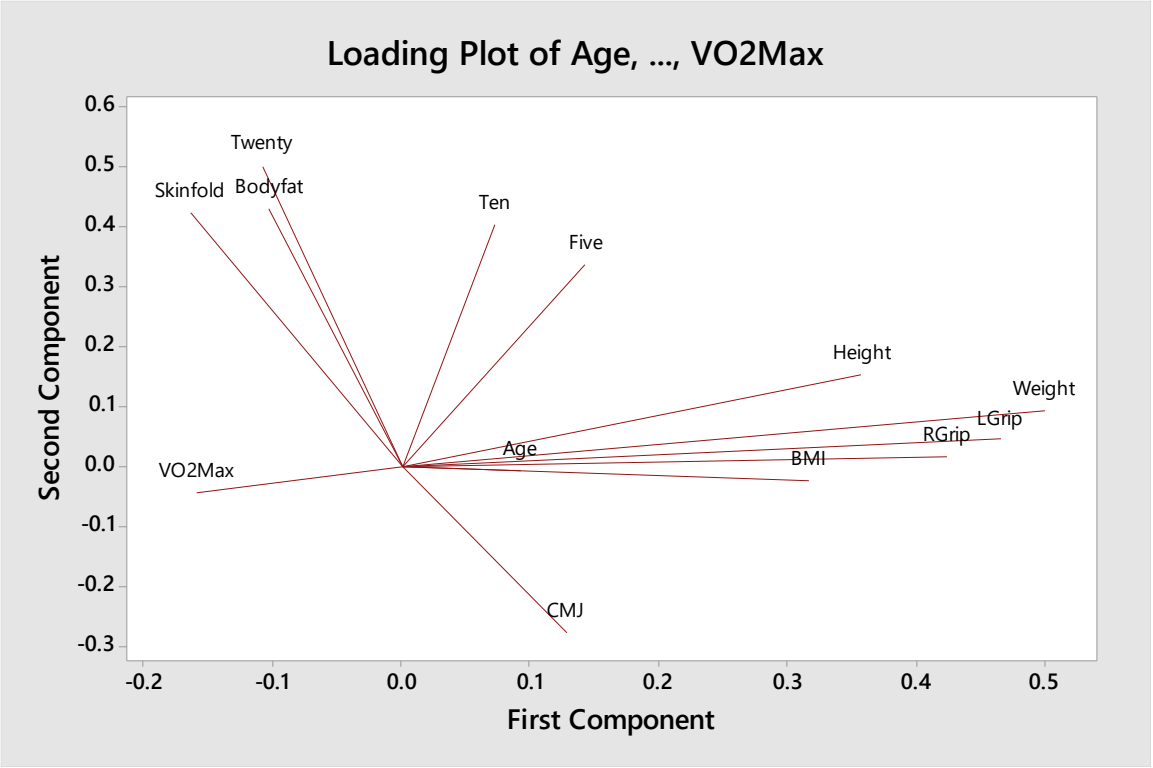
Principal Components Analysis

Principal components analysis is a technique used to replace a large number of variables with a smaller number of variables called principal components. Principal components are linear combinations of the original variables that are uncorrelated with each other. Principal components may have intuitive meanings, and may be more useful than the original variables for analysis and prediction.

When deciding how many principal components to choose, either choose all the principal components with eigenvalues greater than 1 (the **Kaiser criterion** - which indicates that they are more useful than the original variables) or else choose a sufficient number of principal components so that the cumulative proportion of explained variance exceeds some threshold such as 90%.

The interpretation of the principal components is subjective, but the coefficients of the terms may suggest an interpretation. The variables whose coefficients have larger absolute values are more important for that principal component. For example, a large number of variables measured on hurling players yielded a first component that roughly corresponds to strength and a second component that roughly corresponds to fitness, as illustrated in the loading plot in Figure 6.21.

Fig 6.21



Principal components analysis can help to deal with the problems that arise with multicollinearity. When two predictor variables are highly correlated with each other, then each of these individual variables tends to have a larger *p*-value than would arise if it were the only predictor variable in the model. The estimated coefficient for each variable may also have a surprising value and a much larger standard error because of the multicollinearity. This happens because the model 'corrects' for the presence of the second predictor variable, which is very like the variable being considered. This can give rise to a peculiar situation where each individual *p*-value may be insignificant although the model as a whole is significant. The solution is to either omit one of the variables from the model or to combine them into a smaller number of uncorrelated variables, which can be done by using principal components analysis or some other technique.

Principal components analysis focusses on the original data and seeks to reduce the number of variables. Principal components are expressed as linear combinations of the original variables. **Factor Analysis**, on the other hand, focusses on the underlying factors and how they can be used to explain the observed data. In factor analysis, the observed variables are expressed as linear combinations of the factors.

Discriminant Analysis

Discriminant analysis uses a number of variables to classify observations into different groups on the basis of a sample in which the groups are known. The analysis is useful for illustrating how the groups are related to the different variables. The model can then be used to suggest group classifications for new observations. If there are only two groups, binary logistic regression may do a better job than discriminant analysis at suggesting group classifications for new observations, e.g. a credit scorecard. If

there are more than two groups then binary logistic regression cannot be used but discriminant analysis can be used.

Discriminant analysis was used to analyse data consisting of measurements related to forty athletes who play hurling, in order to classify them into groups corresponding to four playing positions, namely: defender, forward, goalkeeper, midfielder. An extract from the output is shown.

Groups

Group	D	F	G	M
Count	16	14	4	6

Summary of Classification

	True Group			
Put into Group	D	F	G	M
D	15	0	0	0
F	0	12	0	1
G	0	0	4	0
M	1	2	0	5
Total N	16	14	4	6
N correct	15	12	4	5
Proportion	0.938	0.857	1.000	0.833

Correct Classifications

N	Correct	Proportion
40	36	0.900

The software was then used to suggest a classification for another athlete who was not included in the original data set. This could be useful for estimating the unknown playing position of an athlete, or to suggest a suitable playing position for an athlete who has just taken up this sport.

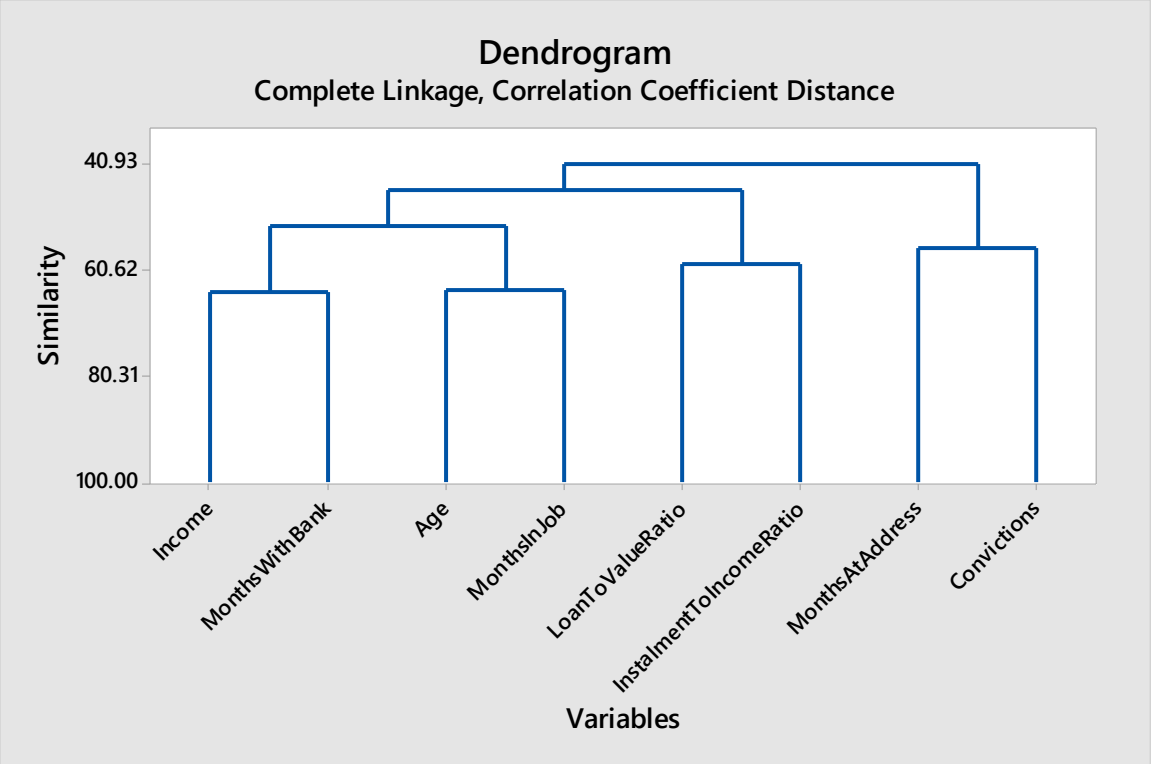
Prediction for Test Observations

Observation	Pred Group	From Group	Squared Distance	Probability
1	F			
		D	50.913	0.000
		F	14.399	0.999
		G	149.928	0.000
		M	27.660	0.001

Cluster Analysis

Cluster analysis seeks to divide a set of observations or variables into groups of similar items when no previous group classifications are provided. A cluster analysis of the variables used in a credit scorecard yielded the following dendrogram, which allows the eight variables to be separated into two, three or four groups.

Fig 6.22



You may notice that the variables are used to form clusters in a way that seems quite coherent by creating two, three or four clusters by focusing on variables (reading the variable names in the dendrogram from right to left) that seem to reflect the applicants history, whether the loan is comfortable to repay, the applicant's life profile, and the applicant's financial profile.

Problems 6F

#1. What multivariate analysis technique might be useful in each of the following situations?

(a) A tour operator wishes to use data about target customers (age, income, etc.) to classify the target customers into groups corresponding to the different types of holidays that the tour operator offers (adventure holidays, city breaks, cruises, cultural trips, beach holidays, road trips) for marketing purposes.

(b) A tutor wishes to use data about students in a large class (age, attendance, assessment marks, etc.) to divide the students into groups for discussion.

(c) A clinical researcher has a data-set consisting of twenty-five variables (age, blood-pressure, etc.) measured on each of fifteen patients, and wishes to reduce the number of variables in order to construct a regression model.

7

Designing Experiments

Having completed this chapter you will be able to:

– design experiments involving one or more factors;

– analyse experimental data using ANOVA tables;

– use general linear models and response surfaces to model processes.

The purpose of an experiment is to discover information. This information may provide insight leading to action that will improve a process. An experiment must be designed properly so that it will capture the right kind of information, capture a sufficient amount of information, and not mix this information up with something else.

Typically an experiment will consider the **effect** of a **factor** on a **response**. For example, we might want to investigate the effect of different players on the distance a football is kicked. The response (distance) is what we measure in the experiment. The factor (player) is what we change, to see how it will affect the response. In the experiment, a number of different players (**factor levels**) are studied. Notice that we are not just watching a game as spectators: we decide who kicks the ball, in what order, and how often. This is an **experiment**, and not an **observational study**. In an observational study we become familiar with the behaviour of a process, and so we can describe or predict its behaviour. For example, we might notice through observation that the goalkeeper kicks the ball further than anyone else, but this may be due to the position on the field, and not the ability, of the goalkeeper. In contrast, design of experiments (**DoE**) sets out to identify causes, and this may enable us to change the behaviour of the process.

7A. Single-Factor Experiments and ANOVA

Video Lecture https://youtu.be/Y07W4dC6MB8

A single-factor experiment considers the effect of one factor on a response, e.g. the effect of player on distance. Any other factors that could affect the distance, such as the type of football used, are kept constant during the experiment.

But what about things like the weather? Small, uncontrollable changes in the environment will cause minor fluctuations in the measured response, even when the factor level remains unchanged. It is important to design every experiment in a way that takes account of these issues, by applying the following principles.

Principles of Experimental Design

Replication

We must measure the response more than once for each factor level. This allows us to see how much **random variation** occurs in the response even when the factor level

remains the same. We call this the **error variance**. The word 'error' does not denote a mistake of any kind: it simply acknowledges that there are many small influences that cannot be identified or eliminated, such as the exact state of the weather, the grass, the ball, the player's physical and mental condition, etc. Later on we will consider exactly how many replicates are required: for now we will simply note that one observation is not enough.

When someone carries out a new task, their performance will change as they improve with practice, and eventually they will settle down at some steady level of performance. This is called the **learning effect** and it should not be confused with random variation. Random variation refers only to **unexplained variation**, and not to **explained variation**. Therefore, if an experiment aims to study people as they perform some new task, they should be allowed a sufficient number of practice runs before the experiment begins, in order to eliminate the learning effect.

Randomisation

Where an experiment consists of a series of runs that are performed one after another, the runs must be performed in a **random run order**. We do not allow one player to perform all their kicks first, with another player taking all their kicks at the end of the experiment. Instead, we make a list of all the runs to be performed and then we use a random-number generator to determine the run order. The reason for this is that there may be some progressive change in the environmental conditions, such as the wind direction, or the wind speed, or the ambient temperature. This could confer an advantage on the player who goes last, if the environmental conditions later on in the experiment are more favourable for achieving long distances. In general, anything in the experimental environment that changes over time could affect the responses, and this could affect one factor level more than others if all the runs of that factor level were scheduled one after another or at regular intervals. Randomisation addresses this concern. Note that randomisation does not guarantee rest breaks at appropriate intervals, so rest breaks should be scheduled as needed.

Sometimes, every measurement in an experiment is carried out on a different **experimental unit**, e.g. a different football for every kick. In such cases, there must be **random allocation** of the experimental units (footballs) to the different factor levels (players). We do not give the first footballs we find to one player, because the first ones may be newer or heavier or different in some other way.

Sometimes an experiment involves selecting samples from different populations in order to compare the populations. In such cases, **random selection** must be used to determine which units will be selected. For example, to investigate if the average tablet weight differs among a number of batches, a number of tablets must be randomly selected from each batch.

Blocking

Sometimes it is impossible to hold everything constant in an experiment. Suppose the number of replicates is so great that we require two days to complete the experiment. There could be some difference between the two days that would affect the responses. We would need to make sure that on any one day each player performs the same number of kicks. This keeps the experiment balanced over the two days, and within

each day the run order could still be randomised. For example, suppose each player has to perform 30 kicks, then they could each perform 20 kicks on the first day and 10 kicks on the second day. This is called blocking and each day is a **block**. In general, a block contains units that are believed to be similar to each other in responsiveness, but which may differ in responsiveness from the units in another block. Typical blocks are: in manufacturing processes, batches; in clinical trials, gender; in measurement systems, days. We will not mention blocks again until we come to deal with two-factor experiments, where the blocks can be treated as the second factor in the experiment.

Structuring your Data

A recording form can be used to display the experimental design and then to record the measurements as they arise. Notice that all the measurements are stacked in a single column. There is a column (with a sensible title) for every variable, and a row for every case. This is the best way to structure data, and it also facilitates the use of software.

In this table, a column of random numbers was generated, and these were ranked to determine the run order. When software is used to create the design, the entire table is usually sorted into run order and the random numbers are not displayed.

Table 7.1

Player	Random No.	Run order	Distance
Gareth	0.368843	5	45
Jessy	0.738684	9	48
Eden	0.474818	7	43
Gareth	0.928478	12	42
Jessy	0.534303	8	50
Eden	0.212354	2	41
Gareth	0.882270	10	46
Jessy	0.437575	6	47
Eden	0.159179	1	43
Gareth	0.890182	11	48
Jessy	0.248853	3	52
Eden	0.272851	4	48

The question of interest is: does the player have an effect on the distance? In general, in a single-factor experiment, the question of interest is: does the factor have an effect on the response?

The Null Hypothesis and the Model

The null hypothesis (H_0) states that the factor has no effect on the response. This is equivalent to saying that the population means are equal for all the factor levels. According to the null hypothesis, there is only random variation present.

The alternative hypothesis (H_1) states that the factor does have an effect on the response. If this is so, then some of the variation in the response is explained by the factor.

The model below articulates these two sources of variation, with the explained variation represented by α, and the unexplained variation represented by ε. Note that although a model looks like a formula, we do not use a model to 'get the answer', but rather to express a relationship that may exist between variables.

Model 7.1

Model for the Response in a Single-Factor Experiment

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

In this model, in generic terms:
Y represents the response at factor level *i* on occasion *j*
μ represents the population grand average response for all the different factor-levels
α represents the effect of factor level *i*, that is how much the average response at that factor level exceeds the grand average response
ε represents the random error on that occasion, that is how much the response on a single occasion exceeds the average response at that factor level.

In relation to our football example:
Y represents the distance the ball was kicked by player *i* on occasion *j*
μ represents the population grand average distance kicked for all the different players
α represents the effect of player *i*, that is how much the average distance kicked by that player exceeds the grand average distance kicked
ε represents the random error on that occasion, that is how much the distance kicked on a single occasion exceeds the average distance kicked by that player.

If the null hypothesis is correct, every $\alpha = 0$. If this is true, it does not matter who kicks the ball. According to the alternative hypothesis, there are some non-zero values of α. According to the alternative hypothesis, if you want to predict how far the ball will travel, it is useful to know who is kicking it. And if you want to control how far the ball will travel, you should choose a particular player to kick it.

ANOVA by Hand

We will now analyse the data from the experiment using **Analysis of Variance** (acronym ANOVA). As its name suggests, ANOVA involves looking at the data for evidence of every alleged source of variation. Although ANOVA is usually performed by software, it can help your understanding to work through the calculations. We begin by **unstacking** the data.

Table 7.2

Distance Gareth	Distance Jessy	Distance Eden
45	48	43
42	50	41
46	47	43
48	52	48

Next, we calculate the mean and variance of each sample.

Table 7.3

	Gareth	Jessy	Eden
Mean	45.25	49.25	43.75
Variance	2.50^2	2.217^2	2.986^2

Step 1 We look at the error variance first, i.e. how different are the distances when the player is not changed? The three sample variances 2.50^2, 2.217^2 and 2.986^2 all attempt to answer this question. Each of these variances is an estimate of the error variance, so it makes sense to combine them into a single 'pooled' estimate, by averaging them. The pooled estimate is $(2.50^2 + 2.217^2 + 2.986^2) / 3 = 6.69$, symbol S_w^2. This is the **pooled within-samples variance**, also called the error variance. Each of the individual estimates is based on 4-1 degrees of freedom, so the pooled estimate has 3(4-1) = 9 degrees of freedom.

Step 2 Next we look at variation due to the factor, i.e. how different are the distances when the players are changed? We can estimate this by calculating the variance between the sample means above. This **between-samples variance** is denoted by S_b^2. It is based on 3-1 = 2 degrees of freedom. $S_b^2 = 2.843^2 = 8.083$. This figure needs to be adjusted because it is a variance of sample means, not a variance of individual values. Individual values vary more than sample means, by a factor n, where n is the sample size. Therefore, we multiply S_b^2 by 4, in order to estimate the variance between individual distances for different players. We require $n \times S_b^2 = 4 \times 8.083 = 32.33$.

Step 3 Remember that the question of interest is: does the factor have an effect on the response? Maybe not. Perhaps the apparent variation due to the factor is just more random variation. We can test this hypothesis by comparing the variance due to the factor with the variance due to error. To see if it is significantly bigger, we compute the **variance ratio**, F, which is named after Sir Ronald Fisher who developed ANOVA. In this case $F = 32.33 / 6.69 = 4.83$.

Step 4 Finally, we consider whether the calculated value of F is so unusual that it could not have occurred by chance. From the tables of the F distribution, we see that the critical value of F, at the 5% level, having 2 and 9 degrees of freedom, is 4.256. The calculated value of F exceeds this value, and therefore its p-value is less than 5%. We reject the null hypothesis, which states that the player has no effect on the distance.

We say that the result is **significant**, i.e. the disagreement between the data and the null hypothesis is unlikely to have occurred by chance. The data assert, at the 5% level, that player does affect distance. Jessy kicks the ball further than the other players.

The ANOVA calculations are summarised in the formula below.

Formula 7.1

One-Way ANOVA

H_0: The factor has no effect on the response.
For k factor levels with n replications each, i.e. k samples of size n:

$$F = \frac{n.S_b^2}{S_w^2}$$

S_b^2 is the 'between-samples' variance
S_w^2 is the pooled 'within-samples' variance
$df = k\text{-}1, k(n\text{-}1)$

ANOVA by Software

Software will produce an **ANOVA table** which always looks like this.

One-way ANOVA: Distance versus Player

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Player	2	64.67	32.333	4.83	0.038
Error	9	60.25	6.694		
Total	11	124.92			

The p-value is less than 5%, showing that the data were unlikely to arise on the basis of the assumption that player has no effect on distance. We therefore reject this null hypothesis in favour of the alternative: player does have an effect on distance.

Every ANOVA table has the same headings: *Source* (of variation), *DF* (degrees of freedom), *SS* ('sum of squares' of the deviations), *MS* ('mean square' deviation, i.e. variance), *F* (variance ratio), and *P* (the probability of obtaining the data if the factor has no effect on the response).

ANOVA Assumptions

ANOVA is based on three assumptions.

1. The errors are **independent**. This means that if one of Gareth's kicks is a short one for Gareth, there is no reason to expect that his next kick will also be short. Gareth is not being disadvantaged in any way, and neither is any other player. That is, the experiment is free from systematic bias. Randomisation will tend to ensure that this is true.

2. The errors are **normally distributed**. The population of Gareth's distances is normally distributed, and the same can be said for the other players' distances. We rely on this assumption when we calculate *F*, because the *F* statistic assumes that we are dealing with normal populations.

3. The errors have a **constant variance**, i.e. the error variance does not depend on the factor level. Even though one player may kick the ball further on average than some other player, the variation in distances is the same for all players. We rely on this assumption when we calculate the pooled variance estimate, because by pooling the sample variances we assume that they are estimating the same quantity.

The first of these assumptions must be satisfied, since systematic bias will render the experimental results untrustworthy. The second and third assumptions are less important and ANOVA will still work well with mild non-normality or unequal variances. For this reason, we say that ANOVA is **robust**.

The ANOVA assumptions can be tested by **residual analysis**. The software calculates the estimated fitted values ($\mu + \alpha$) and residuals (the errors, ε). Then a histogram of the residuals should look approximately normal (assumption 2) and a plot of residuals versus fits should show the points forming a horizontal belt of roughly uniform height and depth (assumption 3) as illustrated previously in Fig 6.14 and Fig 6.15. The residuals can also be plotted against the run order to confirm that there was no time-related pattern.

Problems 7A

#1. Jody carried out an experiment to investigate whether the colour of a birthday candle affects the burning time, measured in seconds. The output is shown below.

One-way ANOVA: Time versus Colour

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Colour	2	4594	2297	0.58	0.586
Error	6	23571	3929		
Total	8	28165			

(a) Explain what each term in the model represents in this situation.
(b) State and test the null hypothesis, and state your conclusions in simple language.

#2. Anne carried out an experiment to investigate whether the supermarket of origin affects the weights of clementines, measured in grams. The output is shown below.

One-way ANOVA: Weight versus Supermarket

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Supermarket	1	3385.6	3385.6	27.75	0.001
Error	8	976.0	122.0		
Total	9	4361.6			

(a) Explain what each term in the model represents in this situation.
(b) State and test the null hypothesis, and state your conclusions in simple language.

#3. (Activity) Design and carry out an experiment to investigate the effect of the design of a paper airplane on its distance travelled. Five airplanes should be made using each of two different designs. One person should throw all the airplanes, in random order. If the airplanes are thrown over a tiled floor, the distance can be measured by counting the number of tiles. Analyse the data using one-way ANOVA

#4. (Activity) Use an experiment to compare the putting ability of different people. The response is the distance of a golf ball from a target when struck with a putter from a distance of 3 metres.

Project 7A

Single-Factor Experiment

Design a single-factor experiment and then carry it out and analyse the results. Choose any original area of application such as a business or industrial process, an area of academic interest, a sporting activity, or everyday life. Make sure that you have the authority to set the factor levels, and that you have a suitable instrument for measuring the response. Your report must consist of the following sections.

(a) State the purpose of your experiment.
(b) Explain why randomisation was necessary in your experiment.
(c) Write a model for the response in your experiment and carefully explain what each term in the model represents.
(d) Show the data (or a portion of the data) and the ANOVA table.
(e) State the experimental findings in simple language.

7B. Two-Factor Experiments and Interaction

Video Lecture https://youtu.be/M-ik6YCjlHQ

A two-factor experiment considers the effect of two factors on a response, e.g. the effect of driver and vehicle on fuel economy. There are two drivers (Jeremy and Ralph) and three vehicles (hatchback, saloon and SUV). Fuel economy is measured in litres per 100 km.

This experiment will allow us to determine if driver has an effect on fuel economy, and if vehicle has an effect on fuel economy. We are getting two experiments for the price of one! This is one of the advantages of including more than one factor in an experiment.

It is convenient to think about the experiment by using a table, in which the rows represent the levels of one factor, the columns represent the levels of the other factor, and each cell represents a particular **factor-level combination**. It does not matter which factor goes in the rows, and which factor goes in the columns.

Table 7.4

	Hatchback	Saloon	SUV
Jeremy	5.2	5.5	7.1
Ralph	5.7	5.4	6.9

This is referred to as a **crossed design** because every row crosses every column, showing that every level of one factor is combined with every level of the other factor, i.e. every driver drives every vehicle. The number of entries in each cell indicates the level of replication. This experiment is not fully satisfactory because there is only one observation per cell, but it will be possible to carry out a limited analysis.

The data must be presented to the software in a proper structure, with a column for every variable, and a row for every case, as shown in Table 7.5.

Table 7.5

Driver	Vehicle	Fuel Economy
Jeremy	Hatchback	5.2
Jeremy	Saloon	5.5
Jeremy	SUV	7.1
Ralph	Hatchback	5.7
Ralph	Saloon	5.4
Ralph	SUV	6.9

The ANOVA table has the usual headings but has two sources of explained variation, namely driver and vehicle.

ANOVA: Fuel Economy versus Driver, Vehicle

Source	DF	SS	MS	F	P
Driver	1	0.0067	0.0067	0.09	0.789
Vehicle	2	3.2033	1.6017	22.35	0.043
Error	2	0.1433	0.0717		
Total	5	3.3533			

Notice that the *p*-value for driver is greater than 0.05, but the *p*-value for vehicle is less than 0.05. We accept the null hypothesis that driver has no effect on fuel economy, but we reject the null hypothesis that vehicle has no effect on fuel economy. SUVs use more fuel.

In this experiment, we may not have been interested in the driver effect. We may have included two drivers in the experiment simply because there was not enough time for one driver to perform all of the experimental runs. If so, the drivers are the blocks. A block is like an extra factor that we include in an experiment because we cannot avoid doing so.

Interaction

We now consider a full two-factor experiment with replication. This experiment considers the effects of person and language on the time, in seconds, required to read a page from a novel. There are three levels of person (Samuel, Alice and Manuel) and two levels of language (English and Spanish).

Table 7.6

	Samuel	Alice	Manuel
English	159	157	326
	163	153	307
Spanish	319	306	160
	302	312	152

In this experiment, there is replication in the cells: each person reads more than one page in each language, so we can do a full analysis.

Looking at the data in the table, you will notice that the situation is not straightforward. The question, 'Who takes longer to read?' does not have a simple answer. The answer is that it depends on the language. This is referred to as an **interaction**. An interaction means that the effect of one factor depends on the level of some other factor. To put it another way, interaction is present when the effect of the combination of factor levels is not the same as what you would expect to get by adding the effects of the factors on their own.

The word 'interaction' should be used with care. It always refers to the effect of a combination of factor levels. It should never be used to refer to the effect of a single factor on its own. The phrase **'main effect'** can be used to refer to the effect of a single factor on its own.

The response in a two-factor experiment can be fully modelled as follows.

Model 7.2

Model for the Response in a Two-Factor Experiment

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \eta_{ij} + \varepsilon_{ijk}$$

In generic terms:
Y is any individual response in the table, i.e. the k^{th} observation in the cell located in row i and column j
μ is the population grand average response, averaged over all rows and columns
α is the row effect, i.e. how much the row average exceeds the grand average
β is the column effect, i.e. how much the column average exceeds the grand average
η is the interaction effect, i.e. how much the cell average exceeds what is expected: based on the additive model we would expect the cell average to be $\mu + \alpha + \beta$
ε is the error, i.e. how much an individual observation exceeds the cell average.

In relation to this particular experiment:
Y is the reading time for language i and person j on occasion k
μ is the grand average reading time for all these languages and persons
α is the language main effect, i.e. how much more time on average is required for that language, compared to the grand average
β is the person main effect, i.e. how much more time on average is required by that person, compared to the grand average
η is the interaction effect, i.e. how much more time on average is required for that particular language-person combination, compared to what would be expected
ε is the error, i.e. how much more time was taken on that occasion, compared to the average time for that language-person combination.

There are three null hypotheses of interest, namely:

every $\eta = 0$, i.e. there is no interaction, i.e. there are no unusual cells in the table, i.e. no particular factor-level combination gives an unusual response, i.e. no particular person reading any particular language takes an unusual amount of time.

every $\alpha = 0$, i.e. there is no main effect due to factor 1, i.e. there are no unusual rows in the table, i.e. language has no effect on reading time.

every $\beta = 0$, i.e. there is no main effect due to factor 2, i.e. there are no unusual columns in the table, i.e. person has no effect on reading time.

The data should be properly structured like this for analysis.

Table 7.7

Language	Person	Time
English	Samuel	159
English	Samuel	163
Spanish	Samuel	319
Spanish	Samuel	302
English	Alice	157
English	Alice	153
Spanish	Alice	306
Spanish	Alice	312
English	Manuel	326
English	Manuel	307
Spanish	Manuel	160
Spanish	Manuel	152

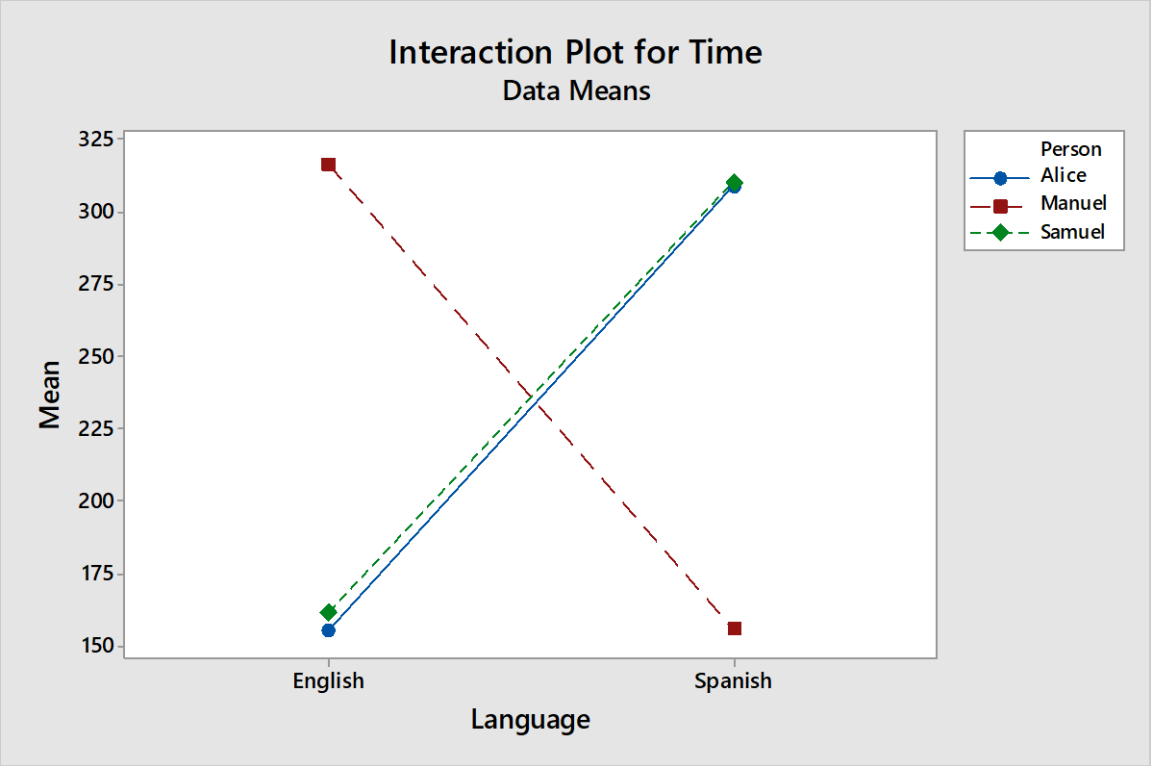
The ANOVA table has three sources of explained variation, namely language, person and interaction. The interaction term is denoted by language*person.

ANOVA: Time versus Language, Person

Source	DF	SS	MS	F	P
Language	1	6816	6816	104.60	0.000
Person	2	43	22	0.33	0.730
Language*Person	2	65010	32505	498.80	0.000
Error	6	391	65		
Total	11	72261			

The interaction *p*-value must always be investigated first. In this case, the interaction *p*-value is less than 0.05, so we conclude that there is an interaction effect. In order to find out about the details of the interaction we draw an **interaction plot**.

Fig 7.1



We can see that for Manuel, switching from English to Spanish reduces the time. But for Samuel or Alice, switching from English to Spanish increases the time.

If there is no significant evidence of interaction in an ANOVA table, then we simply go ahead and interpret the *p*-values for the main effects in the usual way.

Different kinds of interaction could arise. In the example above, some people took longer to read in English and others took longer in Spanish. With a different set of people, it could happen that all of them require extra time to read in Spanish, but the amount of extra time required could vary from person to person. This is also classed as interaction, because the effect of the combination is not obtained by adding together the main effects.

If everyone requires the same amount of extra time to read in Spanish, that is not interaction: it is a language main effect. An interaction is like an allergic reaction that occurs with a particular person-food combination. In general, that person is healthy. In general, that food does not cause problems. The unusual effect is observed only when the two combine. In the same way, we say that interaction is present if some factor-level combination gives rise to an unexpected average response.

Note that an interaction plot cannot tell us if interaction is present. Only an ANOVA table can do that. Do not be fooled into thinking that crossed lines or some other pattern on the graph proves that there is interaction. The only proof that there is interaction is a significant interaction *p*-value in the ANOVA table. After the presence of interaction has been established by the ANOVA table, an interaction plot is then used to illustrate that interaction.

It should now be apparent that a major advantage of two-factor experiments is that interactions can be discovered. In single-factor experiments, interactions cannot be discovered.

Sample Size in Experiments

As with any hypothesis test, the sample size (i.e. level of replication) should be chosen so that the test has sufficient power to detect an effect that is big enough to be considered practically important (see section 5E).

A power of 80% is recommended for research (looking for a difference), and 95% for validation (establishing equivalence). An estimate of sigma is needed to determine the sample size, and this can be found from a pilot experiment. If the sample size is not decided upon before conducting an experiment, and if an important null hypothesis is accepted, it is advisable to check the power value afterwards, to make sure that the sample size was big enough.

The determination of sample size is illustrated below for the single-factor experiment presented at the beginning of this chapter, which investigates the effect of player on distance kicked. The first panel below shows that the sample size required, to detect a difference of 5 metres with 80% power, is 7 replicates for each player. The number 5 was decided upon by asking a football manager to make a judgment about how big a difference is of practical importance. Sample size is always an integer and this is why the actual power with the recommended sample size of 7 exceeds the target power of 80%. A sample size of 6 would not be big enough.

Power and Sample Size

One-way ANOVA

α = 0.05 Assumed standard deviation = 2.587

Factors: 1 Number of levels: 3

Results

Maximum Difference	Sample Size	Target Power	Actual Power
5	7	0.8	0.852851

The sample size is for each level.

The next panel shows that the sample size required, to detect a difference of 5 metres with 95% power, is 10 replicates for each player. This is the sample size required for validation, when it is required to prove that the players are equivalent.

Results

Maximum Difference	Sample Size	Target Power	Actual Power
5	10	0.95	0.963470

The sample size is for each level.

The final panel below shows how we can retrospectively check the power after an experiment. Given that a sample size of 4 was used, and that we are interested in detecting a difference of 5 metres, we are interested to know if our experiment was powerful enough. In this case the power is only 53% which is not satisfactory. A power value of 80% or greater is desirable.

Results

Maximum Difference	Sample Size	Power
5	4	0.529900

The sample size is for each level.

Implementing Process Improvements

Laboratory experiments, performed under special conditions, do not always give a reliable indication of how a process will behave under more usual conditions. For this reason, **confirmatory runs** under actual process conditions are strongly advised, before announcing or implementing the improvements suggested by your experimental conclusions.

It may be necessary to put a system in place in order to implement and maintain the improvements that were identified in an experiment. One practical approach is to use a **mistake-proof device**, which is a physical device that prevents the less favourable factor level being used, e.g. a high barrier that excludes SUVs. Where this is not possible, **audits** can be used to ensure that the preferred factor level is used, e.g. unannounced checks to make sure that documents in Spanish are being given to Manuel to read.

Problems 7B

#1. A chef carried out an experiment to investigate the effect of food (sausage and tomato) and knife (serrated edge and straight edge) on the time in seconds taken to cut the food. The ANOVA table is shown and is followed by an interaction plot.

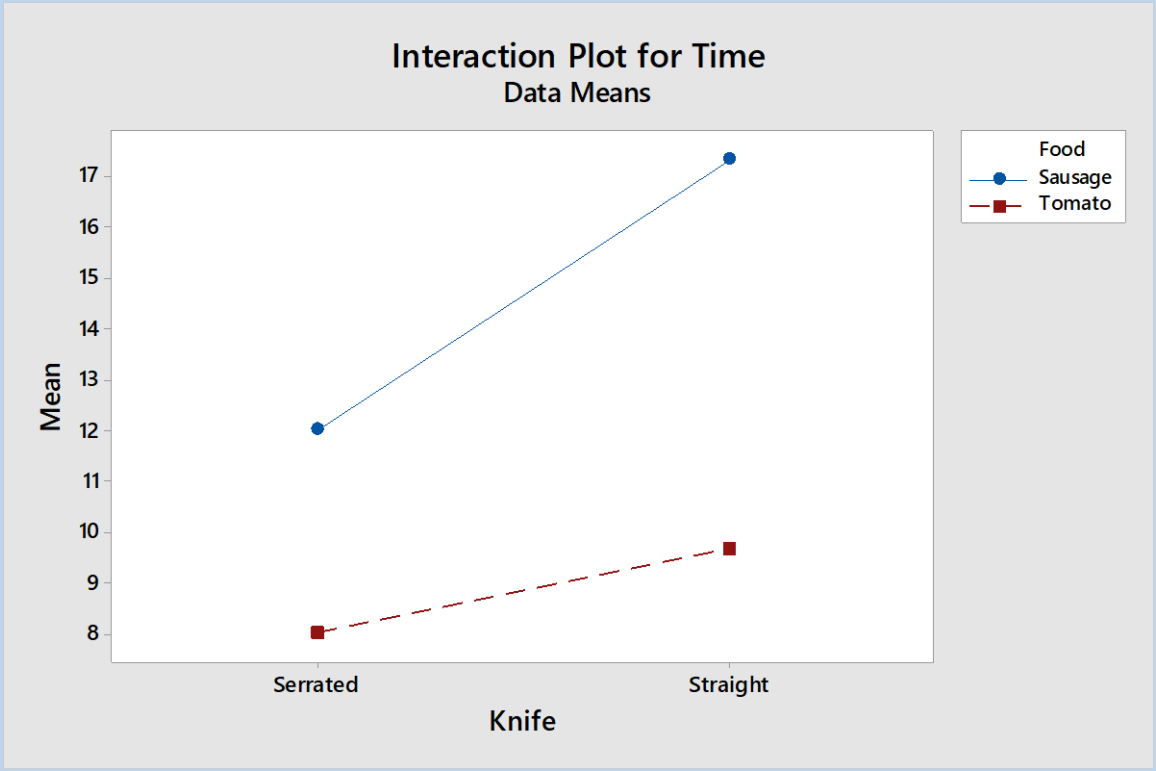
(a) Explain what each term in the model represents in this situation.

(b) State and test all the null hypotheses, and then write down your conclusions in simple language.

ANOVA: Time versus Food, Knife

Source	DF	SS	MS	F	P
Food	1	102.083	102.083	87.50	0.000
Knife	1	36.750	36.750	31.50	0.001
Food*Knife	1	10.083	10.083	8.64	0.019
Error	8	9.333	1.167		
Total	11	158.250			

Fig 7.2



#2. Darragh carried out an experiment to investigate the effect of guitar (Steel-stringed and Nylon-stringed) and method (Plucking and Plectrum) on the time required to play a scale. The ANOVA table is shown and is followed by an interaction plot.

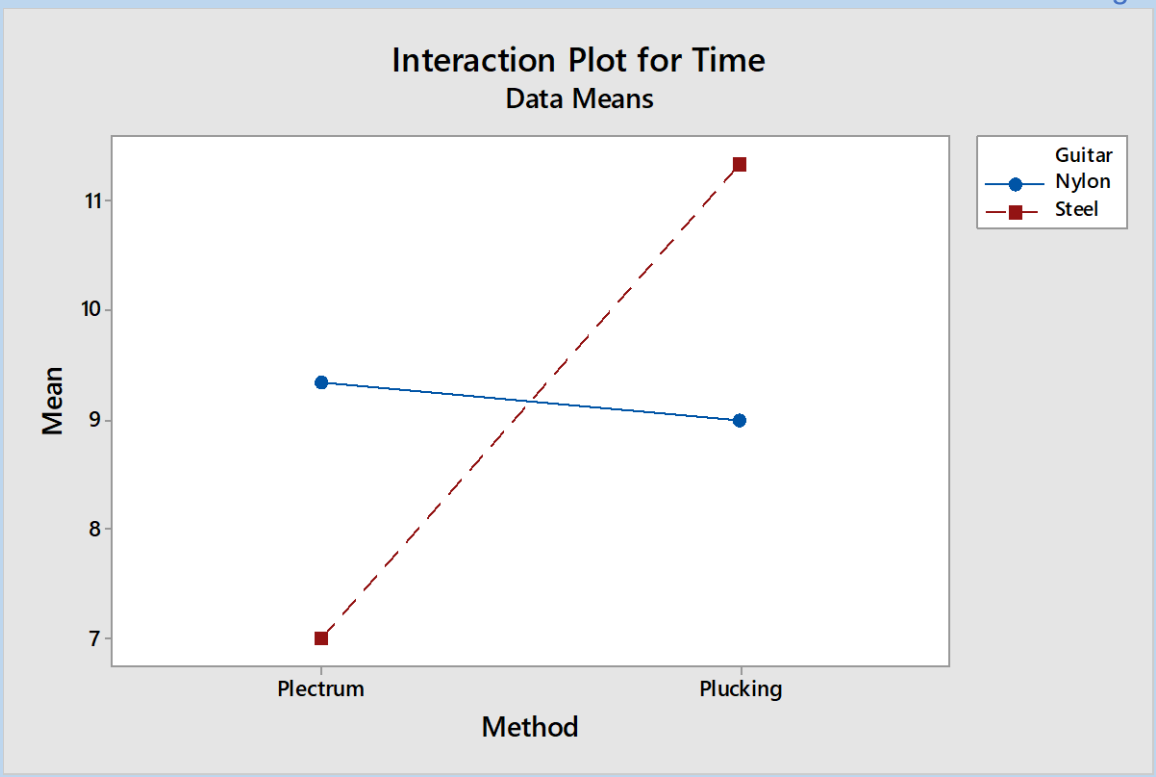
(a) Explain what each term in the model represents in this situation.

(b) State and test all the null hypotheses, and then write down your conclusions in simple language.

ANOVA: Time versus Guitar, Method

Source	DF	SS	MS	F	P
Guitar	1	0.0000	0.0000	0.00	1.000
Method	1	12.0000	12.0000	13.09	0.007
Guitar*Method	1	16.3333	16.3333	17.82	0.003
Error	8	7.3333	0.9167		
Total	11	35.6667			

Fig 7.3



#3. Ken carried out an experiment to investigate the effect of girl (Eva and Lara) and paddle type (Curved and Straight) on the time taken to travel 50 metres in a kayak. The ANOVA table is shown and is followed by an interaction plot.

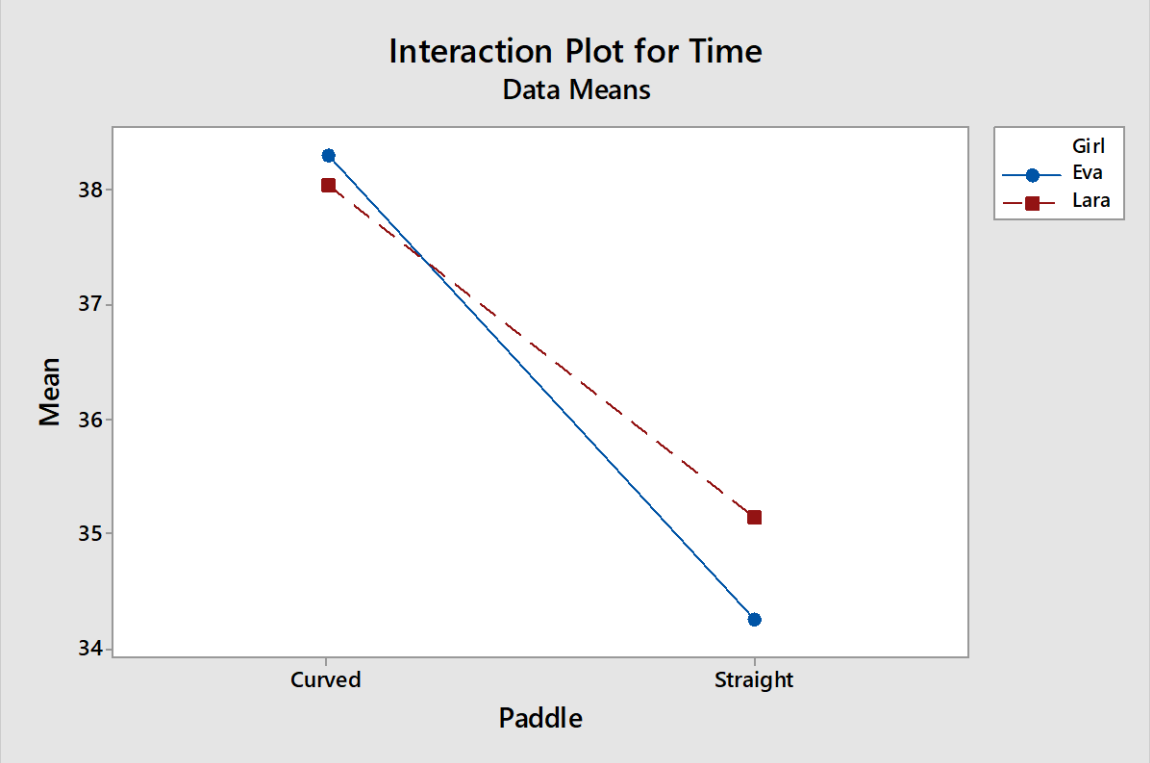
(a) Explain what each term in the model represents in this situation.

(b) State and test all the null hypotheses, and then write down your conclusions in simple language.

ANOVA: Time versus Girl, Paddle

Source	DF	SS	MS	F	P
Girl	1	0.394	0.394	0.09	0.767
Paddle	1	48.686	48.686	11.31	0.006
Girl*Paddle	1	1.328	1.328	0.31	0.589
Error	12	51.646	4.304		
Total	15	102.053			

Fig 7.4



#4. A laboratory analyst set aside a number of parts to be measured on a number of different sessions. Each session was on a different day and, during each session, each part was measured a number of times. The purpose of the experiment was to investigate if the measurement system was consistent over time. (This is called measurement system validation, and a more thorough account of this topic is given in chapter 8.) The ANOVA table is shown and is followed by an interaction plot.

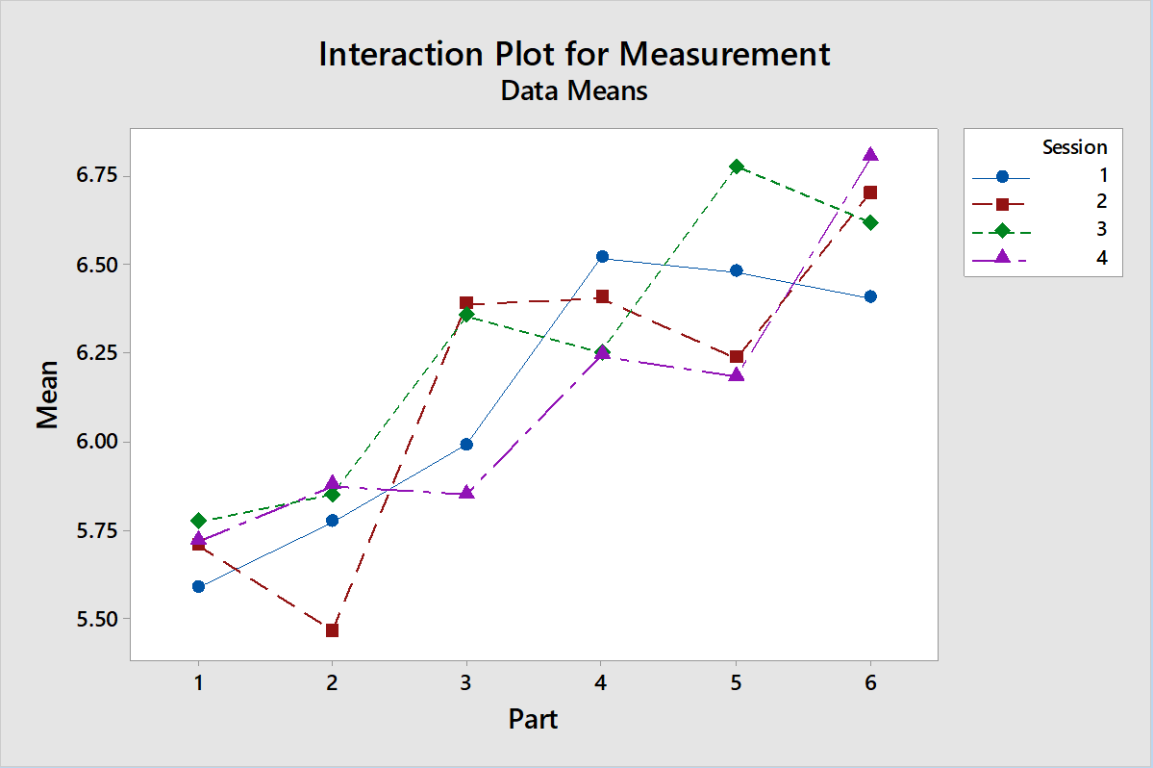
(a) Explain what each term in the model represents in this situation.

(b) State and test all the null hypotheses, and then write down your conclusions in simple language.

ANOVA: Measurement versus Part, Session

Source	DF	SS	MS	F	P
Part	5	8.6534	1.73068	18.60	0.000
Session	3	0.2846	0.09486	1.02	0.392
Part*Session	15	1.8186	0.12124	1.30	0.238
Error	48	4.4663	0.09305		
Total	71	15.2228			

Fig 7.5



#5. (Activity) Working with a classmate, design and carry out a two-factor experiment to investigate the effects of person (two levels: you and your classmate) and hand (two levels: left and right) on writing speed. Allow three replicates. Measure the writing speed by recording how long it takes to write the following sentence: 'The quick brown fox jumps over the lazy dog.' Analyse the data using an ANOVA table, illustrate the results on an interaction plot, and share your conclusions with others who have also performed this activity.

Project 7B

Two-Factor Experiment

Design a two-factor experiment with replication in any original application area of your choice. Carry out the experiment, analyse the results, and write a report consisting of the following sections.

(a) State the purpose of your experiment.

(b) Explain why randomisation was necessary in your experiment.

(c) Write a model for the response in your experiment and carefully explain what each term in the model represents.

(d) Show the data (or a portion of the data) and the ANOVA table.

(e) Draw one interaction plot.

(f) State the experimental findings in simple language.

7C. Multi-Factor Experiments

Video Lecture https://youtu.be/LADyp6Fqnag

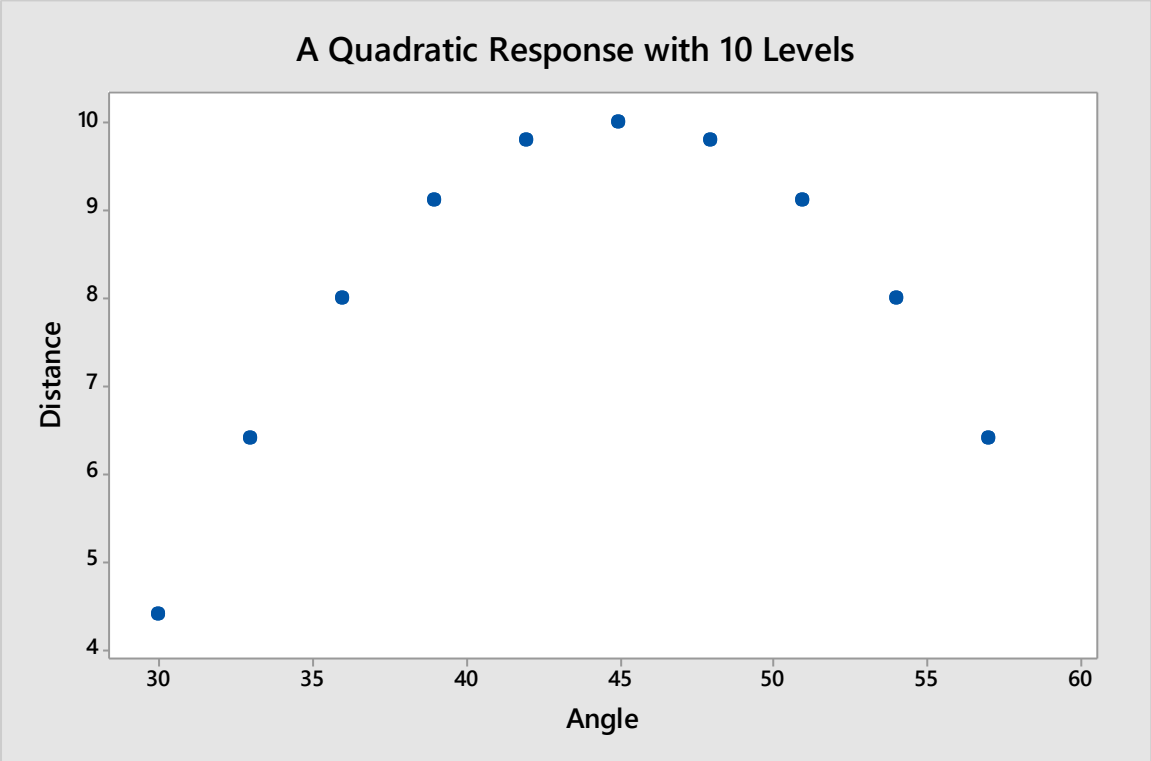
We have seen that two-factor experiments have advantages over single-factor experiments: they are more economical, and interactions can be discovered. We can also design multi-factor experiments which consider the effects of many factors on a response. However, an experiment with many factors, each at many levels, will be a very large experiment. For example, a six-factor experiment, with 10 levels of each factor, requires one million experimental runs (10^6), not counting replication! Obviously we need to consider strategies to reduce the size of multi-factor experiments.

Strategy 1: Two-Level Designs

Our first strategy is to investigate only two levels of every factor. Such experiments are called two-level factorial experiments, and for k factors they involve 2^k experimental runs. For example, a six-factor, two-level design will require 64 runs (2^6), not counting replication. The two levels of each factor are often referred to as high and low, or +1 and -1. In two-factor experiments, the difference between the mean responses at the high and low levels of a factor is called the **effect** of the factor, and the difference between the mean response at the high level of a factor and the grand mean response is called the **coefficient** of the factor.

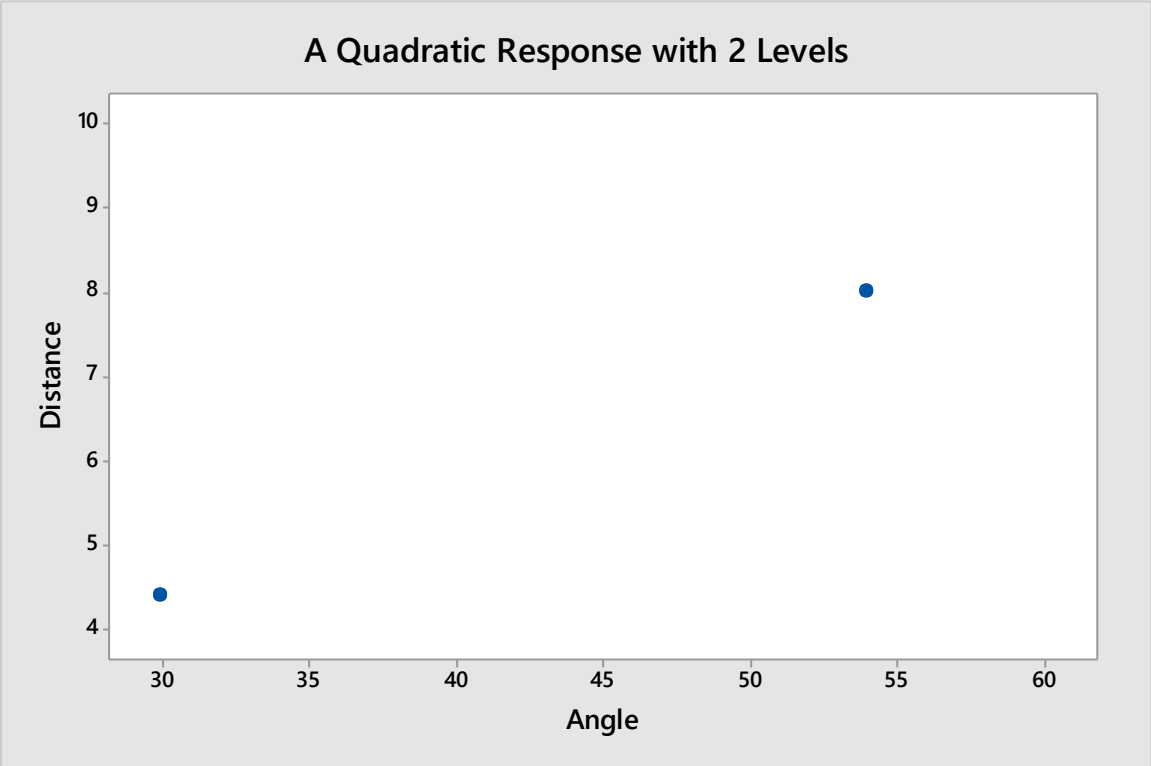
By considering only two levels of every factor, there is a risk that **curvature** will go undetected. There may be a quadratic relationship between the response and one of the factors. Observing only two levels of the factor will not allow us to see this. Consider an experiment that investigates the effect of angle on the distance travelled by a paper airplane. With a ten-level design it is easy to see curvature in the response.

Fig 7.6



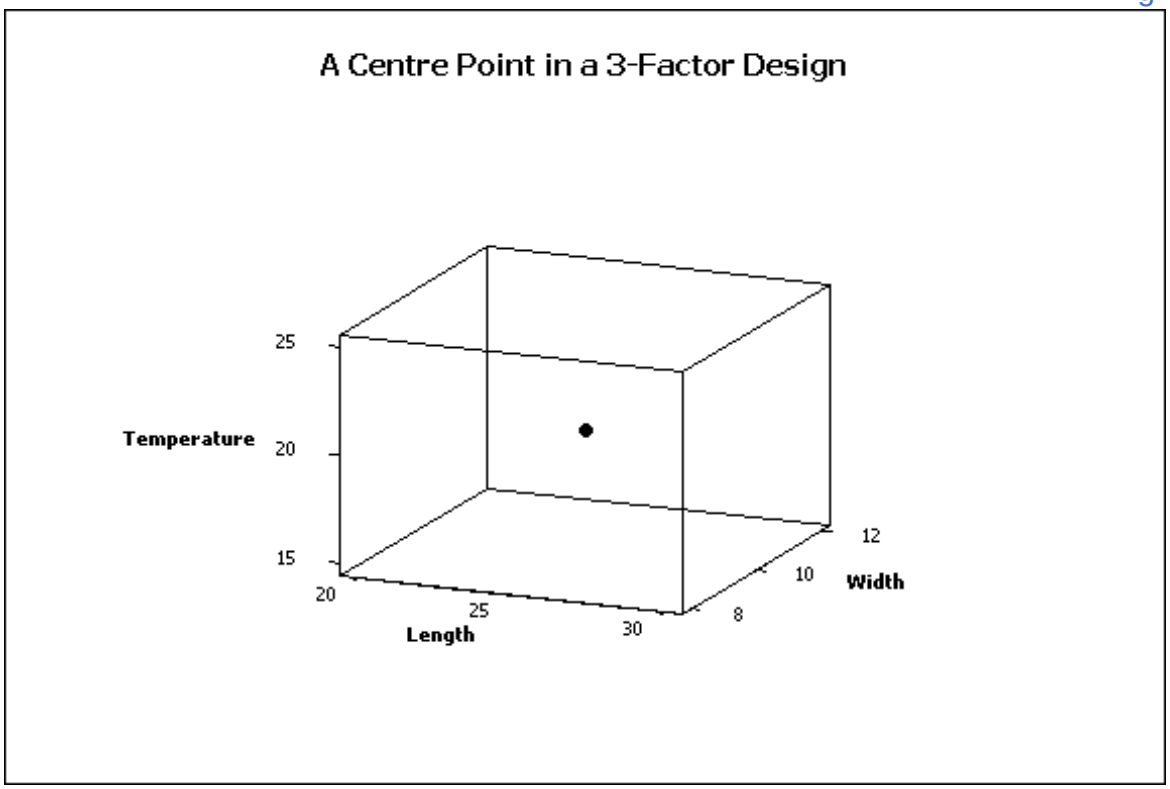
With a two-level design it is impossible to see that curvature is present.

Fig 7.7



This problem can be overcome by including a **centre point** in the design. A centre point is an experimental run at which all factors are set halfway between high and low. Irrespective of the number of factors, one centre point will provide a *p*-value for curvature and put our minds at rest about undetected curvature in the design space.

Fig 7.8



The diagram at Figure 7.8 shows a centre point in a three-factor experiment that considers the effects of length, width and temperature on the distance travelled by a paper airplane.

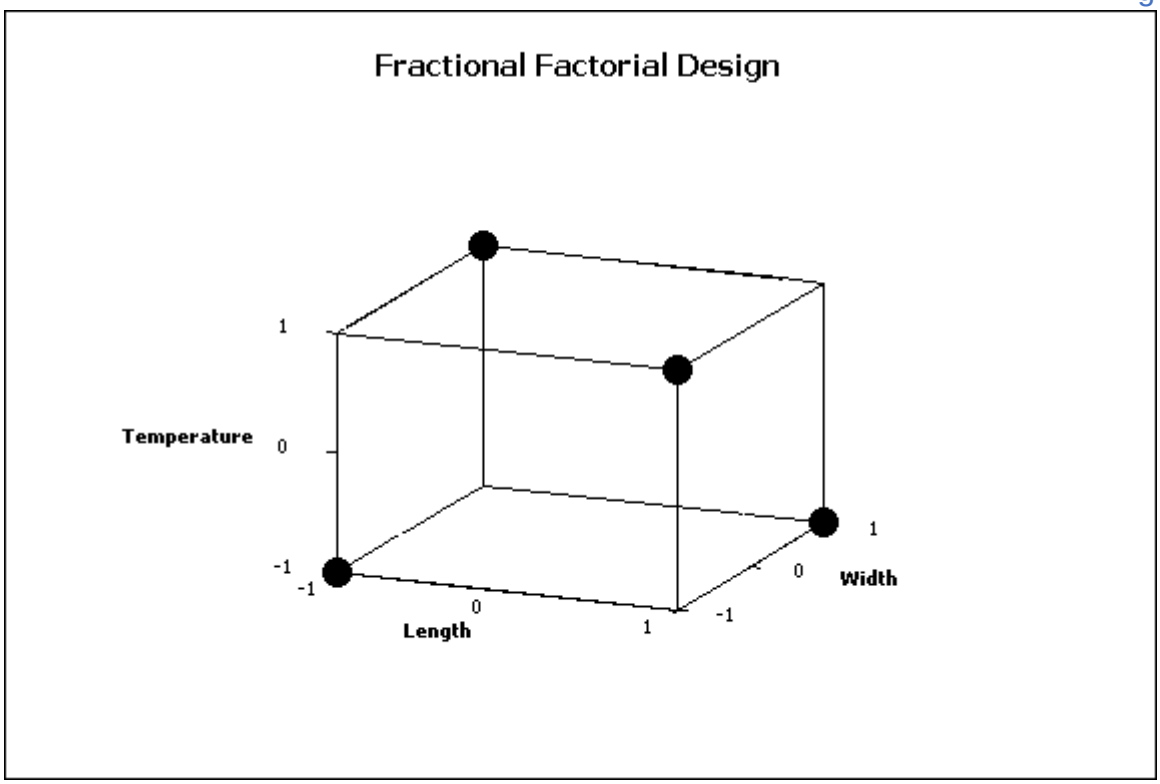
According to the curvature null hypothesis there is no curvature in any of the factors, i.e. the average response at the centre point is equal to the average response at all the corner points combined. Replication at the centre point is not essential, because there is replication at the corner points. However, if some additional time or material is available in an experiment then extra centre points will improve the error variance estimate without unbalancing the design.

A centre point makes sense only for factors with **numeric** levels. The levels of some factors are **text**, e.g. the person who throws the paper airplane is either Sean or Ben, and there is no factor setting halfway between those two levels. If we require a centre point in an experiment that includes text factors, we have no option but to duplicate the centre point at each of the two levels of every text factor.

Strategy 2: Fractional Factorial Designs

Even with every factor restricted to just two levels, an experiment may still be too large. A second strategy to reduce the size of an experiment is to perform only a fraction of the experimental runs, e.g. 1/2 or 1/4, etc. In so doing, we deliberately exclude information about certain factor-level combinations. For example, the **fractional factorial design** illustrated below uses a 1/2 fraction. Only 4 of the 8 available design points, those marked with a solid circle, are investigated.

Fig 7.9



Notice that the design is balanced: of the four factor-level combinations included, two are long and two are short, two are narrow and two are wide, two are hot and two are

cold. Also, of the two that are long, one is wide and one is narrow; and one is hot and one is cold, and so on.

However, fractional factorial designs can give rise to two problems. Firstly, information about certain interactions will be completely missing. If a three-factor interaction occurs at one of the missing combinations (e.g. long-wide-hot), we will simply not find out about it. Now, if we have a good knowledge of the process, we may be able to assert that these interactions would be zero anyway.

The second problem with fractional factorial designs is that information about different effects can become mixed up with each other. This is called **confounding** or **aliasing**. Suppose that a paper airplane needs to be well proportioned, i.e. it should be long and wide, or else short and narrow, in order to travel a long distance. This means that there is a length-width interaction. Look at each of the design points in Fig 7.8 and identify which combinations give rise to large values of distance. Do you notice something? There are two favourable combinations, and two that are unfavourable. But the two favourable combinations occur when the temperature is cold and the two unfavourable combinations occur when the temperature is hot. So it appears that temperature is the explanatory factor! The main effect of temperature has been confounded with the length-width interaction. Some confounding, especially of higher-order interactions, is inevitable in fractional factorial designs. An outline of the effects that are confounded with each other, called the **alias structure**, can be identified, so that alternative explanations can be considered for any significant effects. The confounding of temperature with the length-width interaction is indicated below by the terms C and AB appearing on the same line.

Alias Structure

```
Factor  Name
A       Length
B       Width
C       Temperature

Aliases
I - ABC
A - BC
B - AC
C - AB
```

Now when length and width match they have coded values +1 and +1 or else -1 and -1 and so the product of these values in either case is +1. This corresponds to low temperature which is the -1 level of temperature. So the negative level of temperature corresponds to the positive level of the length-width interaction and conversely the positive level of temperature corresponds to the negative level of the length-width interaction which is expressed in the final line of the alias structure as C - AB.

When terms are confounded together through aliasing, it is simply not possible to use the data from that experiment to identify which of the alternative terms is the best explanation for a significant effect that arises. So in order to formulate a conclusion, one of these approaches can be taken:

1. Select the simplest explanation. This means preferring a main effect rather than

an interaction.
2. Use your knowledge of the process to consider which explanation seems most plausible.
3. Carry out a full factorial experiment so that no terms are aliased together.

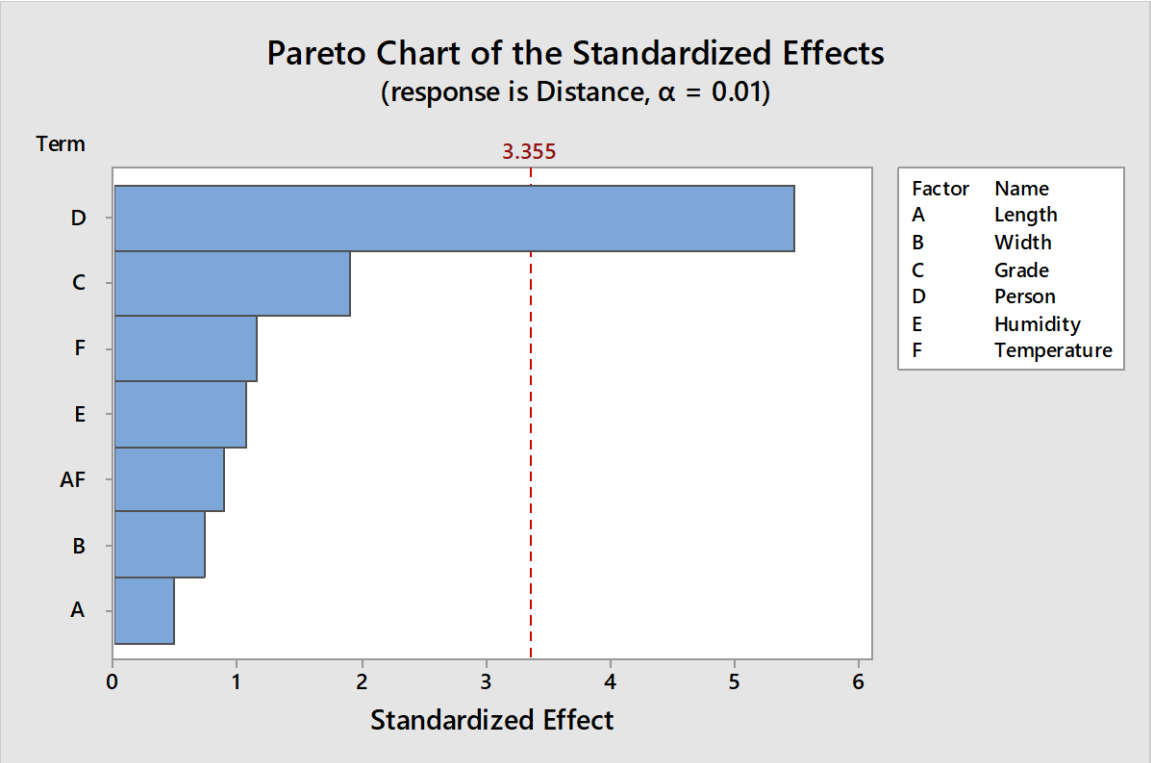
Despite their weaknesses, fractional factorial designs are amazingly powerful and very popular. They provide a lot of information in return for a small amount of data collection.

You may have noticed that Figure 7.8 shows **uncoded units** while Figure 7.9 shows **coded units**. Uncoded units are more intuitive, because the factor levels are familiar to someone who knows the process. But coded units guarantee **orthogonality** which means that every term in the model can be estimated independently of the others. Therefore, if a term that is not significant is removed, the estimates for the terms that remain in the model will not change. Coded units also allow the sizes of the coefficients (and therefore the relative importance of the factors) to be compared, if the high and low levels of the different factors are chosen so as to correspond to equally substantial changes in those factors in the opinion of the experimenter.

Strategy 3: Designs without Replication

In a multi-factor experiment, it is usually the case that most of the factors have no effect on the response, and one or two factors do have an effect. This is called the **sparsity principle**. If all the effects are plotted on a graph, the one or two important factors will stand out from the crowd because their estimated effects will be much larger than the others. Such an **effects plot** can be drawn with or without replication in the experimental design, and so the elimination of replication offers a third strategy for reducing the size of an experiment.

Fig 7.10



Multi-factor experiments are often used as **screening experiments**, with the purpose of identifying the factors that are worthy of further study in a larger experiment, where additional levels of those factors can be explored.

Notice that an alpha value of 0.01, rather than the usual 0.05 significance level, was used in the plot above. When multiple hypotheses are being tested it is advisable to reduce the value of alpha as the **family error rate** exceeds the **individual error rate**. One suggestion, attributed to Bonferroni, is to divide 0.05 by the number of hypotheses to be tested, and use the result as the alpha level. Another suggestion is to use 0.01 rather than 0.05 if five or more hypotheses are to be tested.

Problems 7C

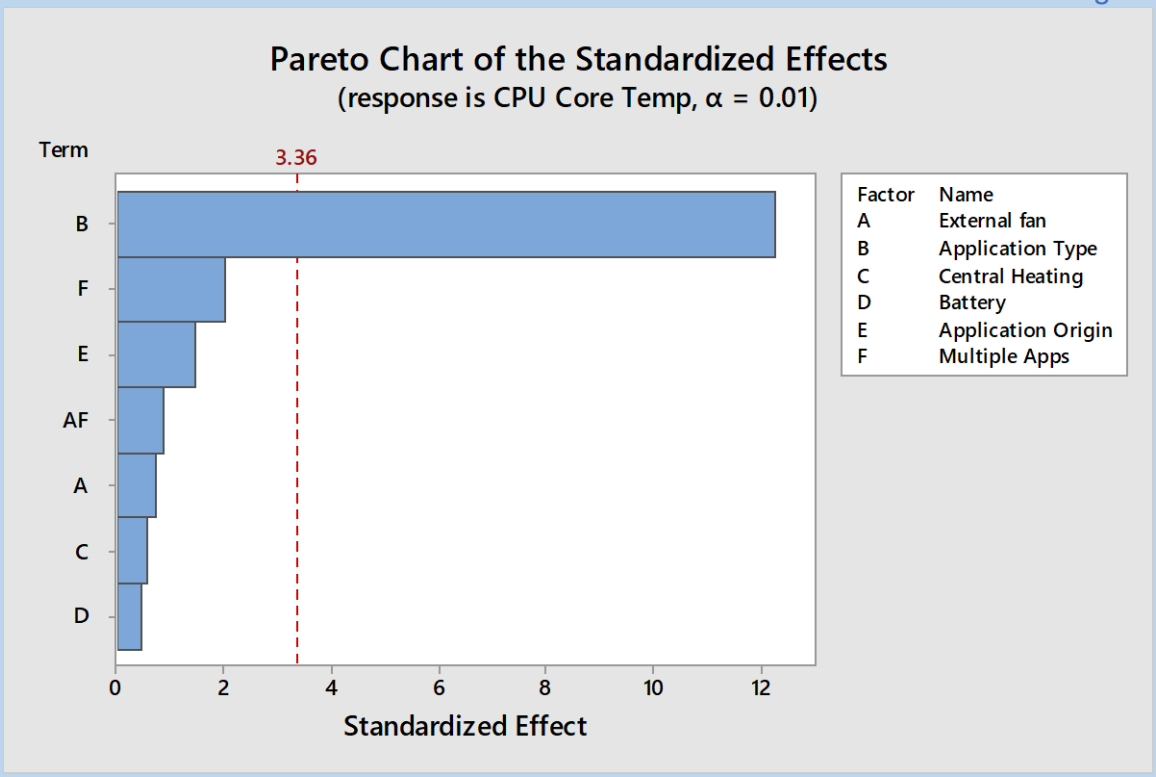
#1. A two-level experiment was carried out to investigate the effect of six factors on the Core CPU Temperature of a laptop. The factors, and their low and high levels, were: External fan (Off, On), Application Type (Browser, High-end Game), Central Heating (Off, On), Battery (In, Out), Application Origin (Internal, External), Multiple Apps (Off, On).

(a) Why was it not possible to include centre points in this design?

(b) What conclusion is suggested by the Pareto chart below?

(c) Given the alias structure below, what alternative conclusion could be proposed?

Fig 7.11




```
Alias Structure

Factor  Name
A       External fan
B       Application Type
```

```
C       Central Heating
D       Battery
E       Application Origin
F       Multiple Apps

Aliases
I + ABD - ACE + BCF - DEF - ABEF + ACDF - BCDE
A + BD - CE - BEF + CDF + ABCF - ADEF - ABCDE
B + AD + CF - AEF - CDE - ABCE - BDEF + ABCDF
C - AE + BF + ADF - BDE + ABCD - CDEF - ABCEF
D + AB - EF + ACF - BCE - ACDE + BCDF - ABDEF
E - AC - DF - ABF - BCD + ABDE + BCEF + ACDEF
F + BC - DE - ABE + ACD + ABDF - ACEF - BCDEF
AF - BE + CD + ABC - ADE + BDF - CEF - ABCDEF
```

#2. (Activity) Working with another person, conduct a fractional factorial experiment to test the effect on writing speed, of the six factors: Person (Max or Ben – use your own names here), Instrument (Pencil or Pen), Case (Lower or Upper), Letter (A or F), Direction (Horizontal or Vertical) and Hand (Left or Right). Allow eight runs (i.e. fraction = 1/8) with 2 replicates. Perform the experiment and analyse the results.

#3. (Activity) Carry out a fractional factorial experiment with three factors to test the effect of a number of factors on the distance that a dart lands from the centre of a dartboard. The three factors are: the person throwing the dart, the height of the dartboard, and the time allowed for the person to take aim before throwing the dart. Allow four runs (i.e. fraction = 1/2) with 3 replicates. Perform the experiment and analyse the results.

Project 7C

Multi-Factor Experiment

Design a fractional factorial experiment with at least four factors, in any original application area of your choice. Carry out the experiment, analyse the results, and write a report consisting of the following sections.

(a) State the purpose of your experiment, and list the factors and factor levels.
(b) Show the data (or a portion of the data).
(c) Show the ANOVA table and the alias structure.
(d) Draw a Pareto chart of the effects.
(e) State the experimental findings in simple language.

7D. General Linear Model

Video Lecture https://youtu.be/1iTNXxIHAG8

We used regression analysis in chapter 6 to construct models for a continuous response using one or more **continuous** predictors. And we have used ANOVA in chapter 7 to construct models for a continuous response using one or more **categorical** predictors. Both of these approaches are special cases of the **General Linear Model (GLM)**, and GLM can also combine these approaches to construct a model for a continuous response using both continuous and categorical predictors.

This is sometimes referred to as **analysis of covariance (ANCOVA)** where a **covariate** could refer to an uncontrolled continuous variable whose values are simply observed rather than assigned. The general linear model assumes that the residuals are normally distributed. (There is an even more flexible approach available, called the **Generalized Linear Model (GLiM)**, which also includes models with residuals that are not normally distributed, but these are not considered here.)

We now provide an illustration of the general linear model. Suppose that the response of interest is the endurance of hurling players, measured by VO2Max. VO2Max may be related to position, which is a categorical factor. VO2Max may also be related to age, which is a covariate. Even though players in certain positions may be younger than players in other positions, the general linear model adjusts for Age, so that the *p*-value for Position is not compromised by the effect of Age.

The output from this general linear model provides *p*-values for both age and position, a coefficient for age, and constants for each level of position.

General Linear Model: VO2Max versus Age, Position

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Age	1	14.56	14.558	13.21	0.001
Position	3	236.82	78.940	71.65	0.000
Error	36	39.66	1.102		
Lack-of-Fit	24	27.49	1.146	1.13	0.428
Pure Error	12	12.17	1.014		
Total	40	344.19			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.04962	88.48%	87.20%	84.66%

Regression Equation

```
Position
D         VO2Max = 50.88 + 0.1685 Age
F         VO2Max = 54.64 + 0.1685 Age
G         VO2Max = 46.41 + 0.1685 Age
M         VO2Max = 52.95 + 0.1685 Age
```

Fixed and Random Factors

We say that a factor is fixed when all the levels of the factor are included in an experiment. For example, an experiment considers the effect of hand and person on writing time. Hand is a **fixed factor**, because all the levels of that factor (left and right) are included in the experiment. When we analyse the data, we would like to 'pick the winner' from among the levels of that factor. A significant *p*-value tells us that the mean time is different for the two hands.

There are many different persons out there, and we may be interested in all of them. However, we cannot include every person in the experiment. Only a random sample of persons is included in the experiment, and we say that person is a **random factor**.

When we analyse the data, we are not interested in 'picking the winner' but rather in estimating the **variance component** due to that factor. A significant *p*-value tells us that the variance due to that factor is not zero, and a table of variance components will indicate how much of the variation in writing time is due to that factor.

Variance Components, using Adjusted SS

Source	Variance	% of Total	StDev	% of Total
Person	56.1235	78.53%	7.49156	88.62%
Error	15.3426	21.47%	3.91696	46.33%
Total	71.4660		8.45376	

Repeated Measures

A repeated measures design is an experiment where the same subjects (e.g. patients or athletes) are observed repeatedly at a number of different time-points, e.g. before and after an intervention. Such an experiment can be analysed using a general linear model. Subject characteristics can be included and may be continuous (e.g. blood pressure or BMI) or categorical (e.g. gender or playing position). Subject is a random factor. A significant main effect for the factor 'time' indicates that the intervention has an impact on the response. Significant interactions between time and other factors indicate that the impact of the intervention may be dependent on the subject characteristics.

Model Reduction

Model reduction involves eliminating insignificant terms from the model. This simplifies the model and improves the precision of the predictions.

We illustrate model reduction using the following model which considers the effect of maximal workload and gender on the time to exertion of athletes. The **full model** includes all potentially relevant terms and their interactions.

General Linear Model: TimeToExertion versus MaxWorkload, Gender

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
MaxWorkload	1	583.65	583.646	46.03	0.000
Gender	1	38.87	38.872	3.07	0.089
MaxWorkload*Gender	1	1.12	1.121	0.09	0.768
Error	33	418.42	12.679		
Lack-of-Fit	32	386.42	12.076	0.38	0.887
Pure Error	1	32.00	32.000		
Total	36	2132.94			

It can be seen from the ANOVA table that the interaction term is not significant. In practical terms, this indicates that the effect of maximal workload on the time to exertion does not depend on gender.

We now proceed to reduce the model by eliminating the insignificant interaction term.

General Linear Model: TimeToExertion versus MaxWorkload, Gender

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
MaxWorkload	1	862.26	862.26	69.88	0.000
Gender	1	485.07	485.07	39.31	0.000
Error	34	419.54	12.34		
Lack-of-Fit	33	387.54	11.74	0.37	0.892
Pure Error	1	32.00	32.00		
Total	36	2132.94			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.51274	80.33%	79.17%	75.92%

Regression Equation

```
Gender
0    TimeToExertion = -1.18 + 6.857 MaxWorkload
1    TimeToExertion =  6.32 + 6.857 MaxWorkload
```

There are no insignificant terms remaining in the model, so this is the **final model**.

Crossed and Nested Designs

Most of the experimental designs we have considered so far have been crossed designs, which means that every level of factor one is combined with every level of factor two, e.g. if the factors are driver and vehicle then every driver drives every vehicle. But this is not always possible. Suppose that the response is the length of laurel leaves, and the factors are tree and branch. Every branch cannot be combined with every tree because every branch belongs on a particular tree. We say that the factor 'branch' is nested within the factor 'tree', and such a design is said to be a **nested design** or **hierarchic design**.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Tree	3	1168	389.42	0.45	0.719
Branch(Tree)	12	10280	856.69	11.13	0.000
Error	16	1231	76.94		
Total	31	12680			

The *p*-value for branch is 0.000, indicating that there are significant differences between branches. The *p*-value for tree is 0.719, indicating that there are no significant differences between trees. The *F* value for tree is obtained by dividing by the mean-square for branch, not by the mean-square for error, because we are asking if there are differences between trees that are not explained by the differences between branches. In general, when considering any factor, we compare its mean-square with the mean-square of the factor that is one stage lower in the hierarchy.

Because a branch can never be combined with a different tree, there is no possibility of interaction in this experiment. Interaction cannot occur with nested factors.

Problems 7D

#1. Timber logs are dried by leaving them overnight on a storage heater or placing them in an oven for either one hour or four hours. The percentage reduction in weight of each log is observed. It is thought that the percentage reduction in weight may also depend on the weight of the log, so the original weight of each log is recorded and included in the model. Use the output provided to answer the following questions.

(a) Is the original weight a factor or a covariate? Explain.

(b) Does the weight of a log have an impact on its reduction in weight?

(c) What is the best way to dry a log? What is the next best way?

General Linear Model: %Reduction versus Original Wt, Drying Mode

```
Analysis of Variance
Source          DF  Adj SS   Adj MS  F-Value  P-Value
  Original Wt    1   24.55   24.550     6.64    0.026
  Drying Mode    2  346.96  173.480    46.89    0.000
Error           11   40.70    3.700
Total           14  419.76

Model Summary
      S    R-sq  R-sq(adj)  R-sq(pred)
1.92355  90.30%     87.66%      71.28%

Regression Equation
Drying Mode
oven 1hr        %Reduction = 9.73 - 0.00343 Original Wt
oven 4hr        %Reduction = 24.06 - 0.00343 Original Wt
storage heater  %Reduction = 12.14 - 0.00343 Original Wt
```

Project 7D

General Linear Model

Construct a general linear model with at least one factor, at least one covariate, and at least 10 data points. Write a report consisting of the following sections.

(a) State the purpose of your study, identifying the factor(s), the covariate(s), and the response.

(b) Describe how you designed the experiment and how you included randomisation.

(c) Show the data, and the ANOVA table for the full model including the interaction term(s).

(d) Test the relevant hypotheses and reduce the model, if appropriate, by removing insignificant terms from the model one at a time, beginning with the interaction term(s), until you arrive at the final model. Justify each step that you take and explain what it means. Show the ANOVA table for each model that you have considered. When you arrive at the final model, show the ANOVA table, the Model Summary, and the Regression Equation(s).

(e) Express the findings in words, using language that is accessible to a non-expert reader.

7E. Stability Studies

Video Lecture https://youtu.be/GgaOCBZv1g8

Setting Expiry Dates

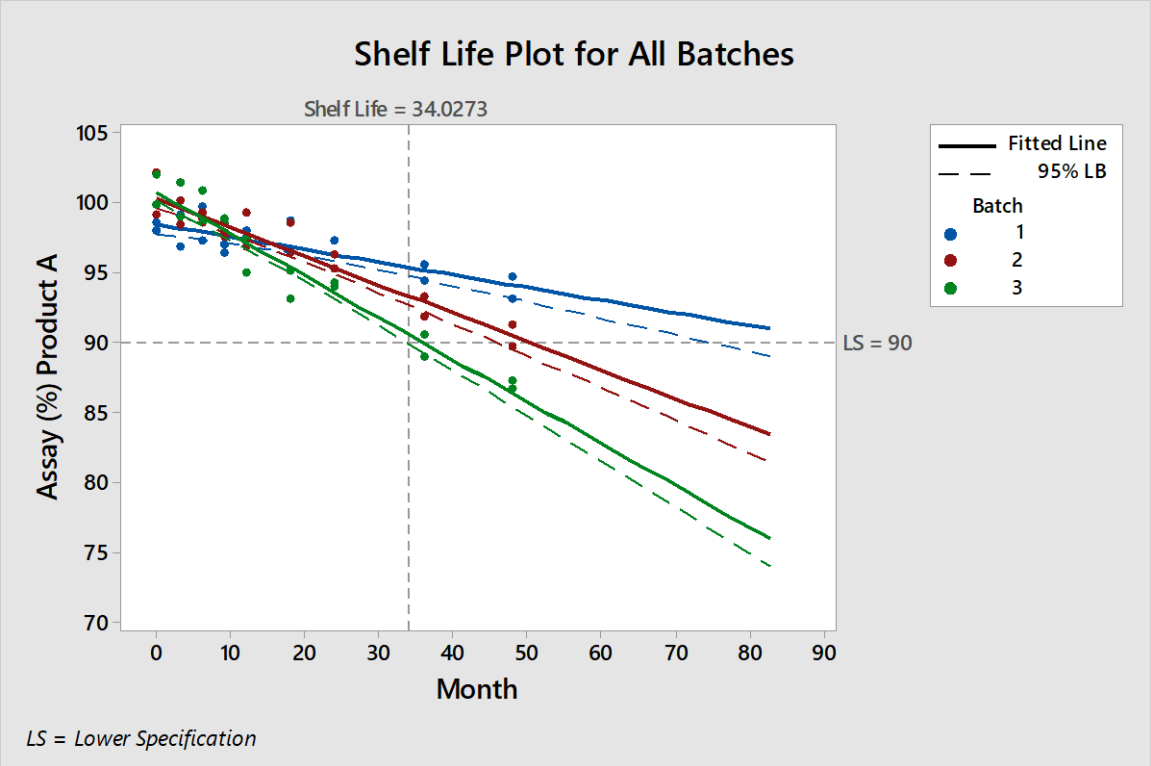
Drug products have a limited shelf life. They degrade over time, so that the remaining quantity of active pharmaceutical ingredient (API) falls continuously. Eventually, the time is reached when the product is no longer considered fit for use, because the remaining quantity of API has reached a specified limit. This lower specification limit is usually defined as a percentage of the quantity stated on the label, and for the remainder of this discussion we will assume that 90% of the label claim is the lower specification limit. So, for example, a 500 mg paracetamol tablet would have reached the end of its shelf life when the remaining quantity of paracetamol has fallen to 90% of 500 mg, which is 450 mg. An expiry date is printed on the packaging to indicate that the drug is no longer fit for use after this date. The purpose of a stability study is to identify the shelf life so that a suitable expiry date can be set.

An assay is used to determine the remaining quantity of API. The assay is repeated at a number of specified intervals, typically at 0, 3, 6, 9, 12, 18, 24, 36 and 48 months after manufacture. A regression analysis is carried out with assay percent as the response, and time as the predictor. The time in months, at which the height of the regression line falls below the lower specification limit, gives a simple indication of the shelf life of the product. Now, because a regression line represents the average value of Y for each value of X, we can expect that the average remaining quantity of API is 90% at this date, and so it can be inferred that half the doses will contain more than 90% and half the doses will contain less than 90% of the label claim at this date.

But the regression equation is based on sample data. Different samples give different results, and so it may happen that the particular sample chosen will suggest a shelf life that is too long. Therefore a confidence interval for the height of the population regression line, rather than the regression line itself, is used to determine the expiry date. And instead of using a two sided confidence interval, if it is known that the quantity of API can only decrease with the passage of time, then a one-sided 95% lower bound is constructed and this is used. The point where the 95% lower bound crosses the lower specification limit is used to determine the expiry date. Using this approach, we can be 95% confident that at least half the doses contain more than 90% of the label claim.

Data from a single batch may not represent all of the batches from a process because some batches may be different from others in terms of the rate of degradation or the initial quantity of API present. Therefore data is taken from three different batches and three regression equations are constructed. The shortest shelf life from among these three candidates is then applied to all future batches from that process.

Fig 7.12



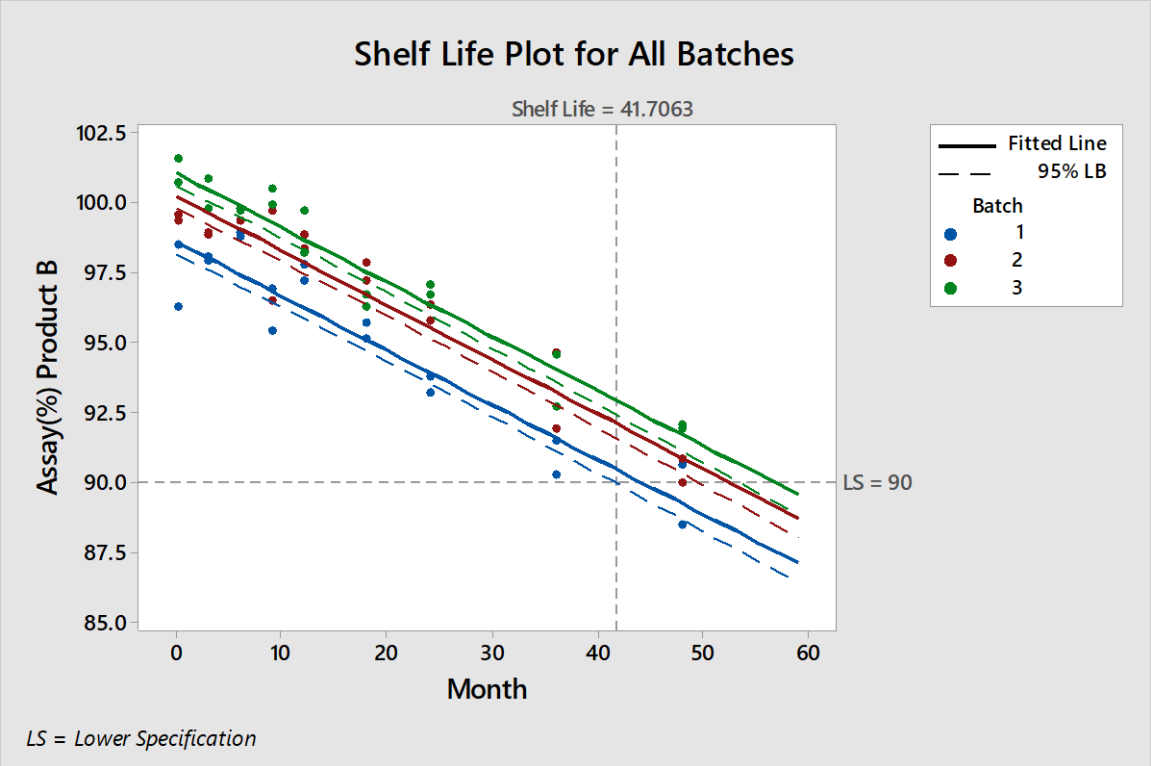
Pooling of Data

Now because the sample sizes in stability studies are quite small, the allowable shelf life calculated in this way will be unreasonably short. Therefore, data from a number of batches can be tested for **poolability** so that data from these different batches can be combined to provide a more realistic estimate of shelf life.

The first step in pooling is to use the pooled mean square error calculated from all batches to fit the regression models for every batch. This assumes that the error variance (or the standard deviation) will be the same for all three regression equations. This is always a reasonable assumption, because the standard deviation is largely due to the precision of the analytical method that is used to perform the assay, and therefore this first pooling step is always taken as a matter of course. So a pooled variance estimate, based on the three data sets, is used to fit each of the three regression equations.

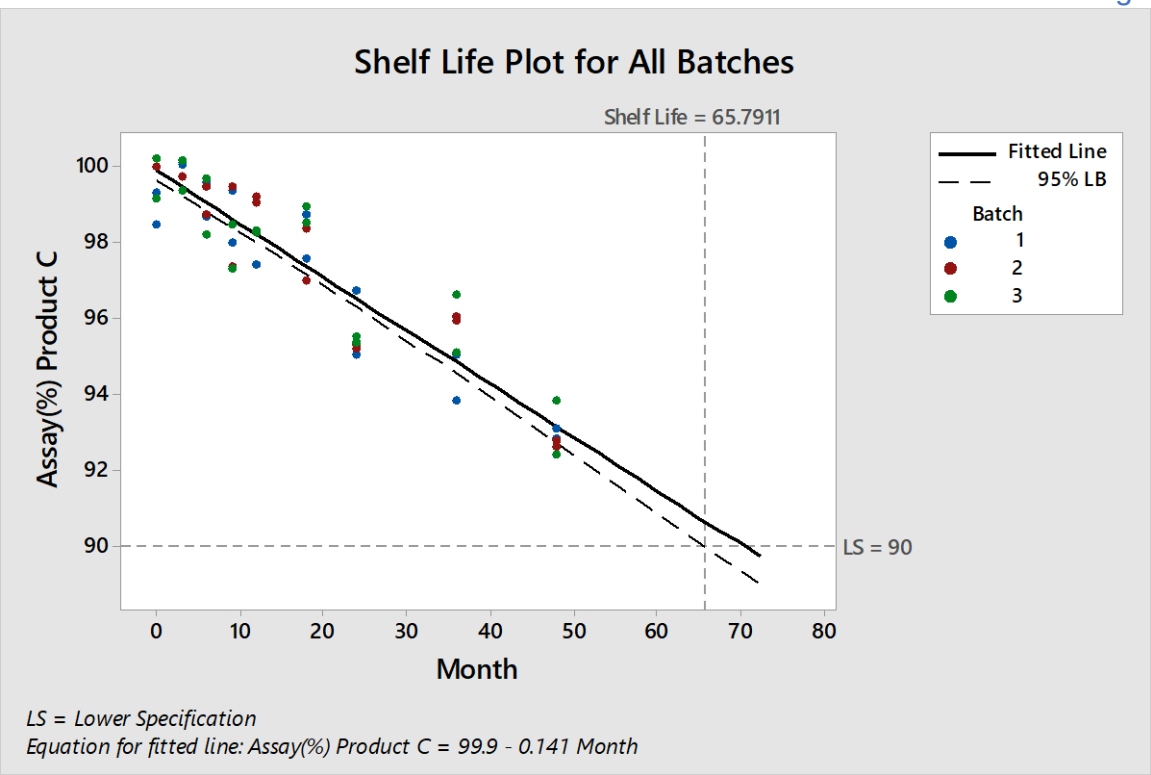
As a second pooling step, it might be reasonable to assume that, even if the batches start out with a different quantity of API, that they experience the same rate of degradation over time. This assumption is expressed by the null hypothesis that there is no interaction between batch and time. A *p*-value is used to test this interaction hypothesis and, to allow for the low power of the small sample size in a stability study, a significance level of 0.25 rather than the usual 0.05 level is used. If the null hypothesis is accepted then it is assumed that the rate of degradation is the same for all batches. Then a general linear model is used to analyse the complete set of data and a model is fitted with an individual intercept for every batch but a common slope. This is called the pooling of slopes. The shortest shelf life from among these three candidates is then applied to all future batches from that process.

Fig 7.13



The third step is the pooling of intercepts. This step assumes that all the batches started out with the same quantity of API. This assumption is expressed by the null hypothesis that batch has no effect on assay result. As in the previous step, the factor "batch" is tested at the 0.25 significance level. If this hypothesis is accepted, then it is concluded that all the batches are the same in their initial quantity of API.

Fig 7.14



In fact, all batches are now the same in every way. So now we simply fit a single regression line to all of the data, and use the lower bound from this line to identify the shelf life for all future batches.

This statistical approach of reducing the model by pooling slopes and then intercepts until a satisfactory shelf life can be confirmed is approved by the ICH (the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use).

Later pooling steps are not explored if an acceptable shelf life has already been established at an earlier stage.

Efficient Stability Study Designs

In all the models considered so far, batch is the only categorical factor. Multi-factor stability studies are often carried out using strength and container size as extra terms in the model. The amount of data increases when batches, strengths, and container sizes are combined and this typically produces narrower confidence intervals that can support longer shelf lives. The sizes of these studies are often reduced by studying only the largest and smallest levels of strength and container size and assuming that intermediate strengths or container sizes have shelf lives that are no shorter than those tested. This is called **bracketing**.

Similarly, the sizes of these studies are often reduced by testing only some of the factor level combinations at any particular time point, and this application of fractional factorial designs to stability studies is called **matrixing**.

So far we have been dealing with real-time stability tests using recommended storage conditions. Accelerated stability tests are also carried out in which a product is stored at elevated stress conditions, such as raised temperature or humidity. Such studies can be completed faster, and degradation at the recommended storage conditions can be inferred if the effect of the acceleration factor is known.

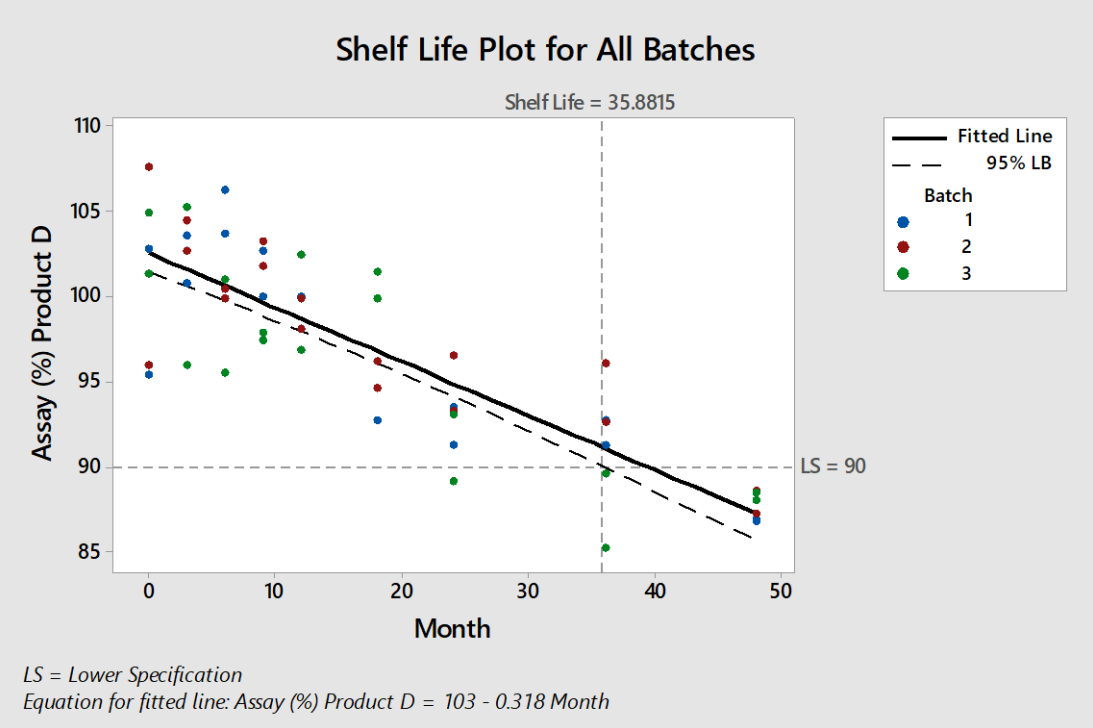
The degradation profile of a particular drug product may have a non-linear relationship with time and so an appropriate data transformation may be required.

In some stability studies, the response studied is the percentage of some degradation product, rather than the percentage of remaining active pharmaceutical ingredient. The increased quantity of some toxic degradation product may be a greater concern in relation to the shelf life than the reduced quantity of API. In such studies, it will be known that the percentage of the degradation product increases with the passage of time. Therefore, an upper specification limit is determined for the percentage of the degradation product, and a one-sided 95% upper bound is used to provide confidence in relation to the expiry date.

Problems 7E

#1. A plot from a stability study is shown in Figure 7.15. Outline the pooling steps that were followed in the analysis and explain why these steps were justified. What shelf life would you recommend for this product?

Fig 7.15



LS = Lower Specification
Equation for fitted line: Assay (%) Product D = 103 - 0.318 Month

Project 7E

Stability Study

Simulate a stability study by testing the perceived reduction in the flavour of chewing gum over time. Begin with three packs of chewing gum, which represent the three batches in the study. In order to complete this simulation in a short time, allow ten seconds to represent one month. This means that the three-month assay is performed after 30 seconds, the six-month assay after 60 seconds, and so on with further assays at 9, 12, 18, 24, 36 and 48 months (represented by 480 seconds which is 8 minutes).

Begin with batch number one. Start the timer and begin chewing the gum. At each of the time points, make a personal assessment of what you think is the level of flavour as a percentage of the initial flavour. This is the assay result. It is unlikely that you will be able to make an assessment about whether the initial flavour is a true 100%, so you can represent the initial value at zero months as a missing value. The simulation can be performed by two individuals who work together to provide two replicate measurements at each time point. So the assay results for the first batch can all be collected within an eight-minute session. If one person performs this simulation alone then it will take twice as long because the replicate measurements will have to be taken. Then the data for the second and third batches need to be collected also. Take rest breaks between batches.

When all the data are available then proceed to analyse the data and identify a suitable 'expiry' for all future batches of chewing gum from this process. Follow the protocol in regards to pooling in the usual way. And since chewing gum is not a pharmaceutical product it might be appropriate to set the lower specification considerably lower than 90%, say at 50%. The 'expiry' in this case represents the amount of time for which the gum can be chewed until its flavour falls below this minimum acceptable level.

Write a report consisting of the following sections.

(a) Specify the brand of chewing gum, the pack size, and where you got it.

(b) Show the data.

(c) Present the software output from the stability study analysis.

(d) With regard to each of the three pooling steps, say whether or not each step was taken and why.

(e) Determine a suitable 'expiry' for all future batches of gum from this process, in seconds, and explain carefully what this 'expiry' means for the user.

7F. Response Surface Methodology

Video Lecture https://youtu.be/5ELSeRFeb7k

What is a Response Surface?

When the variables that affect an important response have been identified, we may wish to explore in some detail the relationship between the variables and the response, especially if the variables are continuous variables and if curvature may be present. A **response surface** can be used to model the relationship between the input variables and the response in a process. A response surface can be used for **process characterisation** and for the identification of the **optimum settings** of the input variables that will achieve the desired value for the response. A response surface can be represented by a **contour plot** when only two input variables are considered.

Contour Plots

Contour plots use colour to represent higher and lower values of the response in the same way that maps use colour to represent higher and lower terrain, with deeper shades of blue for increasingly lower values (deep seas) and deeper shades of green for increasingly higher values (high mountains). The plots that follow illustrate the relationship between the distance travelled by a paper airplane, and the two factors: grade of paper and initial angle of flight path. The levels of grade were in the region of 80 g/m^2 and 90 g/m^2, and the levels of angle were in the region of 20 and 30 degrees to the horizontal. The first plot, in Figure 7.16, shows a number of things:

1. Greater distance tends to be achieved by using lighter paper, and larger angles.

2. The contours are roughly parallel: this indicates that the response surface is a plane, as opposed to a ridge (Figure 7.17), a trough, a peak (Figure 7.18), a valley or a saddle.

3. The optimal settings for the factors are probably outside the range of levels included in the experiment. The optimal settings may be an angle substantially greater than 30 degrees and a grade of paper substantially less than 80 g/m^2.

4. As we move outside the plotted region towards the area of optimum response, the shape of the response surface may change, e.g. there may be curvature in the surface, leading to a ridge or a peak in the response, as illustrated in the second and third plots.

5. We may be restricted from following the path to the peak by other considerations, e.g. a very low value for paper grade may optimise distance travelled but may have an adverse effect on some other target response such as product lifetime. Or a very low value for paper grade may be unattainable with the current paper manufacturing process.

Fig 7.16

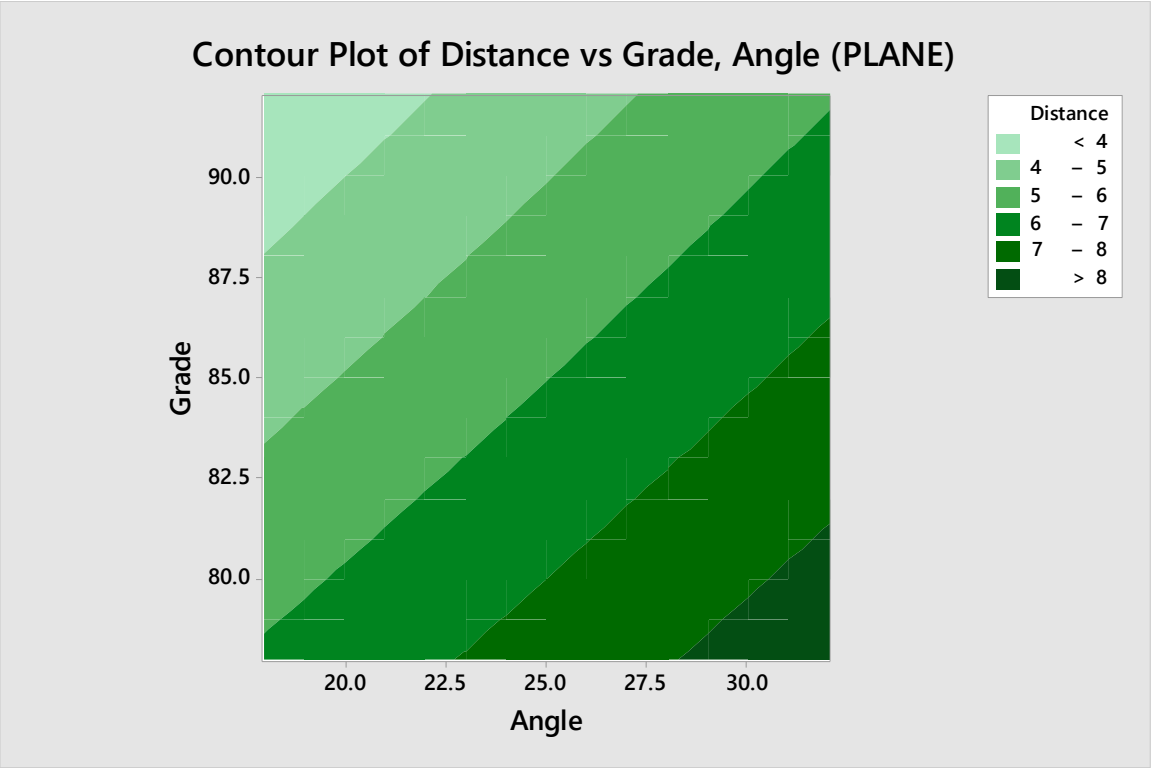


Fig 7.17

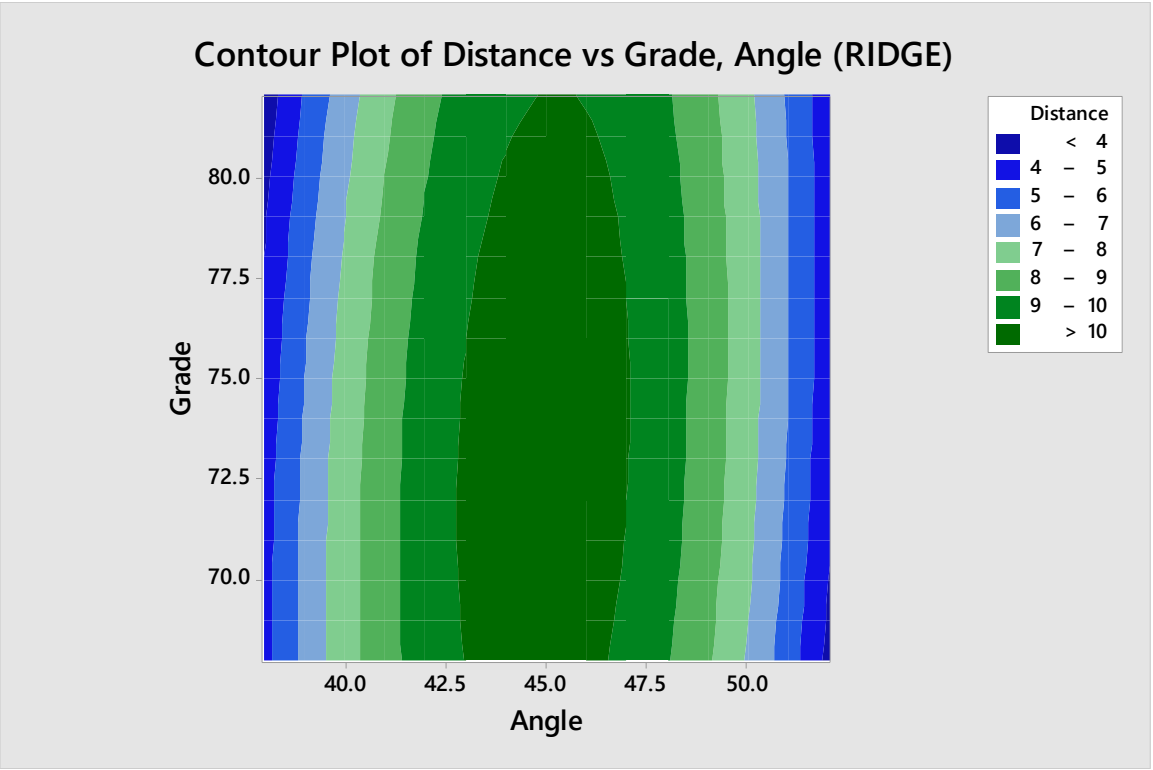
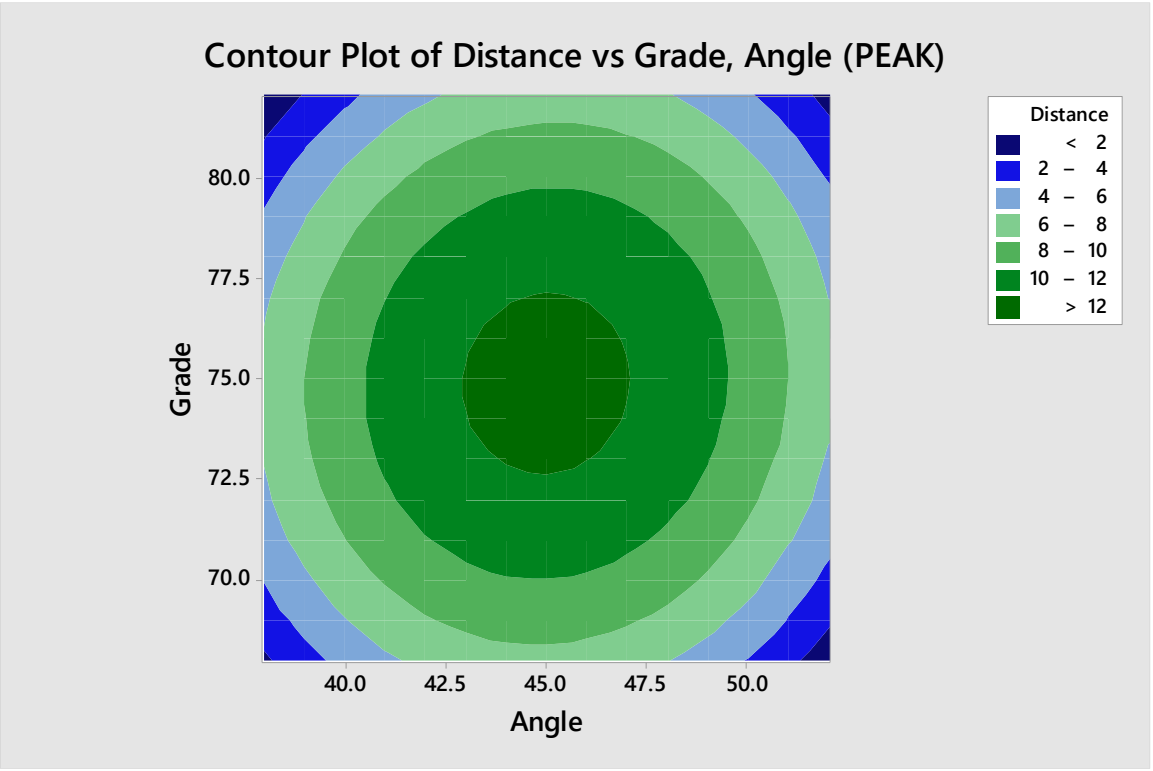


Fig 7.18



Although response surface methodology (RSM) can be used with many factors, contour plots can only show two input variables at a time, like the easting and the northing on a map. The other factors can be held at high, low, or middle settings.

Process Optimisation

The objective is usually to identify the settings of the input variables that optimise the value of the response variable, i.e. the location on the map that achieves the desired height. The optimum could mean that 'largest is best', 'smallest is best', or 'target is best'. If a number of alternative locations are equally suitable, it is best to choose a location where the contours are widely spaced. Widely spaced contours indicate a **robust region** so that variation in the process variables at such a region will cause only small variation in the response.

To fit a geometric model to the process, data can be collected from a specially designed off-line RSM experiment or from an on-line experiment using **evolutionary operation** (**EVOP**) where small sequential changes are made to nudge the process closer to its optimum settings. An off-line experiment is not restricted by the process specifications, requires fewer data, and usually delivers much faster results. The advantage of an on-line experiment is that the cost of the experiment is quite small as the product made during the experiment is saleable, but it is a slow process requiring a lot of data and repeated analysis steps.

Steepest Ascent

'The Path of Steepest Ascent' is an EVOP technique for moving the process closer to the optimum operating conditions. A simple planar model is used to approximate the response surface: this may be a useful approximation, especially if the initial experimental region is some distance away from the optimum.

The procedure is as follows:

1. Find the Path of Steepest Ascent. Use a 2-level design, with centre-points, to identify the path of steepest ascent. To travel along the path of steepest ascent, move a distance in each direction x_i proportional to its coefficient b_i. The direction to move is the sign of each b_i for ascent (bigger is better), or the opposite direction for descent (smaller is better).

2. Find the best point on this path. Measure the response at a number of incremental points along this path until the response begins to deteriorate. The experimenter can decide upon the increments, by reference to one chosen variable.

3. Find a new Path of Steepest Ascent. The point thus reached becomes the base for the next experiment, as at step 1 above. Continue in this way until the optimum is reached.

Response Surface Designs

There are certain desirable properties of response surface designs. They should be blocked when appropriate, to take account of different days, batches, operators or shifts. A design which is **orthogonally blocked** provides independent estimates of the model terms and the block effects. Such designs also minimise the variation in the

regression coefficients. A **rotatable** design provides reliable predictions, because all points that are equidistant from the centre have the same prediction variance.

Central Composite Designs

These designs are available for 2 or more factors and can be built from a 2^k factorial design. A 2^k factorial design consists of **corner points** and **centre points** and the central composite design adds **axial points** (also called **star points**) which are like centre points except for the value of one factor which places the axial point outside the cube.

Box-Behnken Designs

These designs are available for 3 or more factors and are the most efficient design for generating a response surface from scratch. They are suitable for one-off experiments and are less expensive than central composite designs for the same number of factors. They have no points outside the cube, so the experimenter has less need to worry about violating process specifications. They have no points at the corners of the cube either, so no experimental run involves all factors being set at extreme levels simultaneously.

EXAMPLE Bags are sealed by the application of a hot bar. The strength of the seal may depend on the hot bar temperature, the application time, and the material temperature. The existing process settings for these inputs are 90 ± 10, 3 ± 1 and 20 ± 10 respectively. A Box-Behnken design was used and some of the software analysis of the experimental data is shown.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.955170	99.96%	99.89%	99.47%

Coded Coefficients

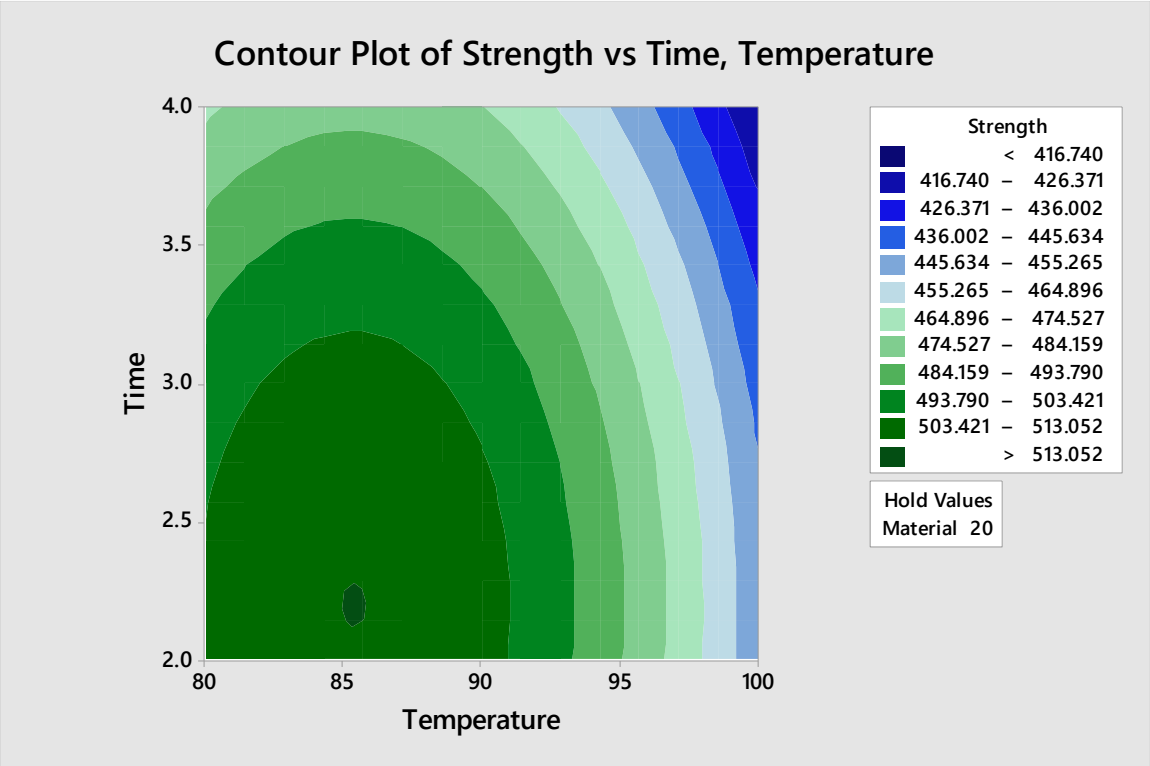
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	500.384	0.551	907.37	0.000	
Temperature	-27.828	0.338	-82.40	0.000	1.00
Time	-15.775	0.338	-46.71	0.000	1.00
Material	-0.039	0.338	-0.12	0.912	1.00
Temperature*Temperature	-30.124	0.497	-60.60	0.000	1.01
Time*Time	-9.810	0.497	-19.74	0.000	1.01
Material*Material	-0.812	0.497	-1.63	0.163	1.01
Temperature*Time	-0.107	0.478	-0.22	0.832	1.00
Temperature*Material	-0.932	0.478	-1.95	0.109	1.00
Time*Material	0.515	0.478	1.08	0.330	1.00

Regression Equation in Uncoded Units

```
Strength = -1749.9 + 51.659 Temperature + 43.02 Time
+ 1.005 Material
           - 0.30124 Temperature*Temperature - 9.810 Time*Time
- 0.00812 Material*Material
           - 0.0107 Temperature*Time - 0.00932 Temperature*Material
+ 0.0515 Time*Material
```

It can be seen that the linear and square terms in time and temperature are the significant terms. A contour plot could be constructed using these two variables, while holding material constant at its middle value.

Fig 7.19



It can be seen that the optimum strength is greater than 513 and is achieved by process settings in the region of 85.5 and 2.2 for temperature and time.

The analysis is very useful but, as with any statistical model, it is not magic. We have fit a simple geometric model that may not be the 'true' model: perhaps linear and quadratic terms do not adequately represent the relationship. Also the values of the coefficients presented by the software are only sample estimates of the actual parameter values. In addition to all this uncertainty, we must also note that the response surface only represents the average response: individual responses will lie above and below the response surface.

Multiple Response Optimisation

A product may have many different responses which affect customer satisfaction and fitness-for-use. If we focus on optimising one of these responses then some other responses may have unacceptable values. Multiple response optimisation is a technique that allows us to consider all of these responses simultaneously. The goal of multiple response optimisation is to identify the values of the process inputs that optimise the **Composite Desirability** of the responses. You can guide the software towards an optimal solution by quantifying the following criteria.

Importance By default, all responses are considered equally important, and are assigned an importance value of 1. If you wish to attach a greater importance to any response you can assign a higher importance value, up to a maximum of 10. For responses which are less important, values as low as 0.1 may be assigned. In some

situations you may wish to repeat the analysis with different importance values, to explore the sensitivity of the solution to different points of view.

Goal You will have an opportunity to specify the goal for each response as minimise (smaller is better), target (target is best), or maximise (larger is better). If you choose minimise, for example, you will be required to specify a target value and an upper bound. The target can be either a very small value even though it may be unachievable, or a value so small that any further reduction offers no appreciable benefit.

Weight It is necessary to specify a weight for each response. The purpose of the weight is to emphasise or de-emphasise the target, compared to the bound or bounds. The default weight of 1 places equal emphasis on the target and the bounds. The weight can be increased as high as 10, placing more emphasis on the target, or reduced as low as 0.1 placing less emphasis on the target.

EXAMPLE The inputs to a welding process are Temperature, Time and Material. The process specifications for these factors are 90 ± 10, 3 ± 1 and 20 ± 10 respectively. It is required to simultaneously maximise Strength and Lifetime and minimise the Cost of these welds. The worst acceptable values for these responses are 480, 6 and 10 respectively, and their target values are 520, 12 and 0 respectively.

SOLUTION First, we create a response surface design and measure and record all the responses variables at each design point. Next, we fit a model to each response. The significant factors for each response may be different. Finally, use the response optimizer to suggest a solution such as the following.

Multiple Response Prediction

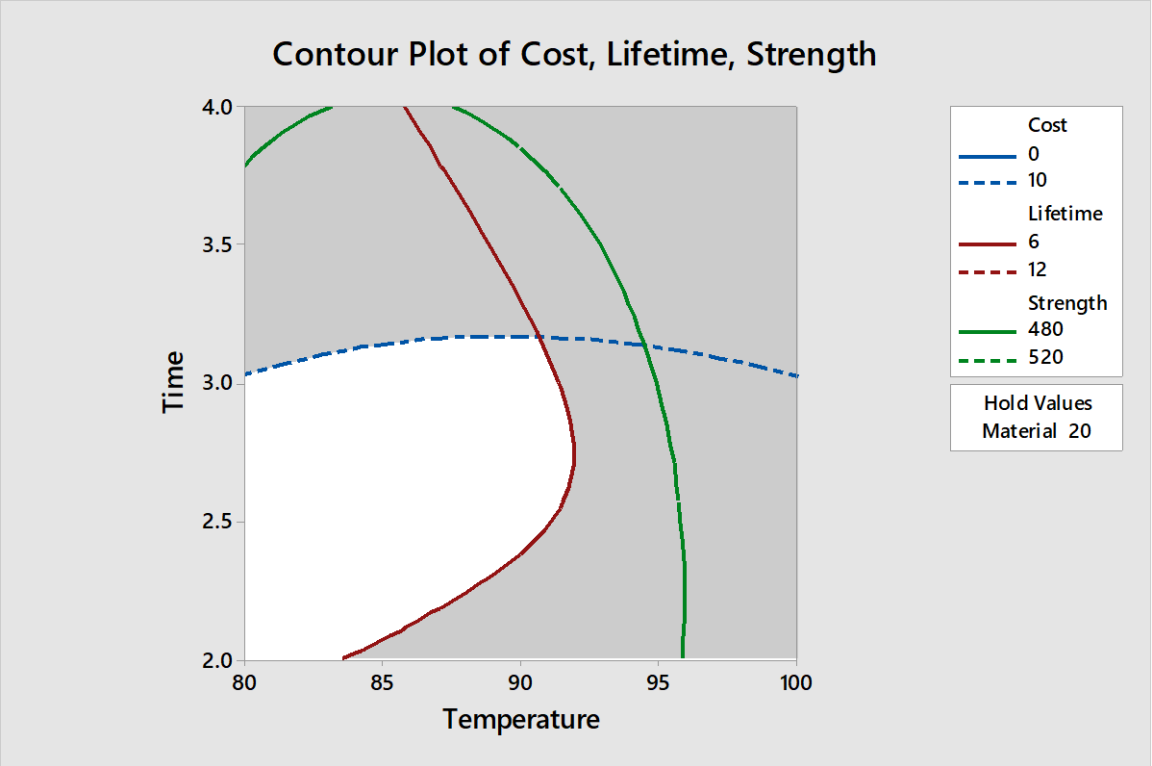
```
Variable     Setting
Temperature  81.2862
Time         2
Pressure     10

Response      Fit  SE Fit         95% CI               95% PI
Cost        9.512   0.178  (  9.150,   9.874)  (  8.833,  10.190)
Lifetime    6.581   0.248  (  6.077,   7.084)  (  5.636,   7.525)
Strength  507.809   0.469  (506.857, 508.761)  (506.023, 509.595)
```

An overlaid contour plot can be used to simultaneously display all responses in relation to any two input variables. Additional input variables not included in the plot must be held at fixed settings.

The white area in the contour plot shows the ranges of these two factors that simultaneously satisfy the specifications for all responses. Additional plots could be drawn to consider the effects of different factors.

Fig 7.20



Problems 7F

#1. (Activity) Construct a Box-Behnken Design and carry out an experiment to optimise a player's performance at throwing darts. The response is the distance from the bull's-eye. The factors are the height of the bull's-eye from the floor, the distance from the player to the dartboard, and the time that the player pauses to take aim before throwing.

Project 7F

RSM Project

Design an original RSM experiment on any process of your choice, using a Box-Behnken design. Carry out the experiment, analyse the results and write a report consisting of the following sections.

(a) State the purpose of your experiment and list the factors.

(b) Show the data.

(c) Show the relevant analysis from the software output.

(d) Draw one contour plot.

(e) State your findings in simple language and present your recommendations for further study of this process.

8

Improving Quality

Having completed this chapter you will be able to:

– investigate the capability of a process;

– use SPC to control processes;

– select and apply sampling plans;

– validate a measurement system;

– use the seven basic quality tools.

8A. Process Capability

Video Lecture https://youtu.be/pqRfd7bJUdk

Normality

Measurements from every process exhibit some variation from the target response. By measuring the process standard deviation, σ, we can quantify this variation, and then consider whether the process is able to satisfy the specifications demanded by the customer or designer. Many processes, especially well-behaved processes, are normally distributed, and so their measurements can be represented by a bell-shaped curve with a spread of $\pm 3\sigma$.

Stability

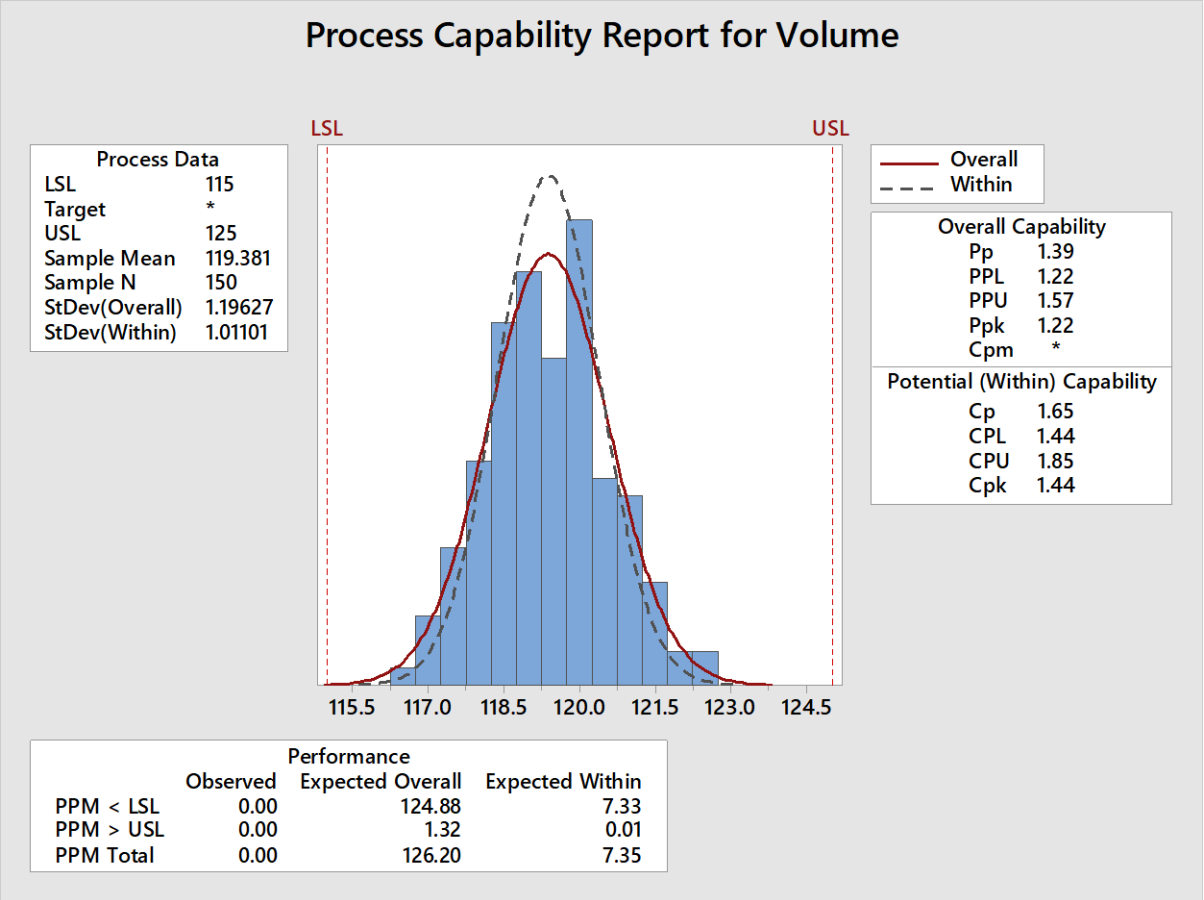
The standard deviation, σ, can have two different meanings in this context. Process measurements that are collected all at the same time give an idea of the behaviour of the process over a short time interval. The standard deviation calculated with such data represents the **short-term variation**. This is the unexplained variation in the process which can also be called the random variation or the inherent variation. But the process operates over an extended period of time, under different environmental conditions and with different lots of incoming material. These factors can cause the process to be unstable, drifting or shifting as time passes, so that the process mean is not always the same. Therefore it is more realistic to collect a set of data over an extended period, and to calculate the standard deviation that represents the **long-term variation** in the process. The long-term variation includes the unexplained variation and the explained variation, i.e. the variation about the process mean and the variation of the process mean.

The best way to carry out a process capability study is to sample a total of 150 data, collected as 30 subgroups of size 5. The subgroups are typically spread over thirty different days or batches (or else thirty different operators or thirty different machines, depending on the factor of interest). This allows both the **overall standard deviation** (i.e. long-term sigma, based on all the data) and the **within standard deviation** (i.e. short-term sigma, the pooled within-subgroups standard deviation) to be calculated.

As a rule of thumb, if the overall standard deviation exceeds the within standard deviation by more than 10%, this confirms that time-to-time variation is present, and so the process is not stable.

EXAMPLE A process fills bottles to a target volume of 120 ml. Samples of size five were taken from thirty different batches as a basis for a process capability study.

Fig 8.1



First, we check for normality. Ignore the bell-shaped curves which are drawn by the software, and look at the histogram. It peaks in the middle and tails off on both sides in a fairly symmetric pattern, so we conclude that the process is normal.

Next, we check for stability. The overall standard deviation, 1.196, is more than 10% greater than the within standard deviation, 1.011, so we conclude that this process is not stable: the mean fill-volume varies from batch to batch.

The curves on the graph represent the short-term and long-term behaviour of the process, using the dashed curve and the solid curve respectively. To use an analogy, this is like viewing a bell-shaped vehicle from behind. The dashed curve indicates the width of the stationary vehicle in a snapshot. The solid curve indicates the width of the vehicle's path over a period of time in a video, with some additional side-to-side movement as it travels. The path of the moving vehicle is wider than the stationary vehicle because of the time-to-time variation.

Centrality

Until now, we have been listening to the **voice of the process**. The process has told us what it is **able to do**, i.e. what standard deviation it is capable of delivering. We

now introduce the **voice of the customer**, i.e. what the process is **required to do**. The lower and upper specification limits are: **LSL** = 115 ml and **USL** = 125 ml.
Looking at the graph we see that the process is not centred, and this is confirmed by observing that the mean of the sample of 150 measurements, 119.381, does not coincide with the target value, 120. This means that there is a greater risk of violating the lower specification limit rather than the upper specification limit. In this case the lower specification limit can be designated as the nearest specification limit, **NSL**. For some processes, it is a simple matter to correct this asymmetry by adjusting the process mean, μ, but for certain other processes, this is not a simple adjustment at all.

Capability

We now ask whether the process is capable of satisfying the specifications, i.e. is it able to do what it is required to do? In other words, are the specification limits comfortably wider than the process variation, $\pm 3\sigma$? By dividing the difference between USL and LSL by 6σ we obtain the **process capability**, *Cp*. The minimum acceptable value for process capability is 1.33, for a process to be deemed capable. A value of 2 or greater is considered ideal. This is like comparing the width of a vehicle with the width of a road, to ensure that the vehicle fits comfortably on the road. The boundaries on the sides of the road correspond to the LSL and the USL.

This process capability is calculated using the within standard deviation, since the process has demonstrated that it is capable of achieving this standard deviation, at least in the short term. This may be rather optimistic if the process is unstable, so this is referred to as the **potential process capability**. The process capability also assumes that the process can be easily centred. If this cannot be easily done then it makes more sense to calculate a one-sided **process capability index**, *Cpk*, by dividing the difference between the process mean and NSL by 3σ. This should also have a minimum value of 1.33.

Performance

If a process is unstable, then its actual long-term performance is not as good as its potential short-term capability. Therefore, the overall standard deviation should be used to carry out the calculations. Now, dividing the difference between USL and LSL by 6σ we obtain the **process performance**, *Pp*. The process performance assumes that the process can be easily centred. If this is not realistic then it is useful to calculate a one-sided **process performance index**, *Ppk*, by dividing the difference between the process mean and NSL by 3σ, again using the overall standard deviation.

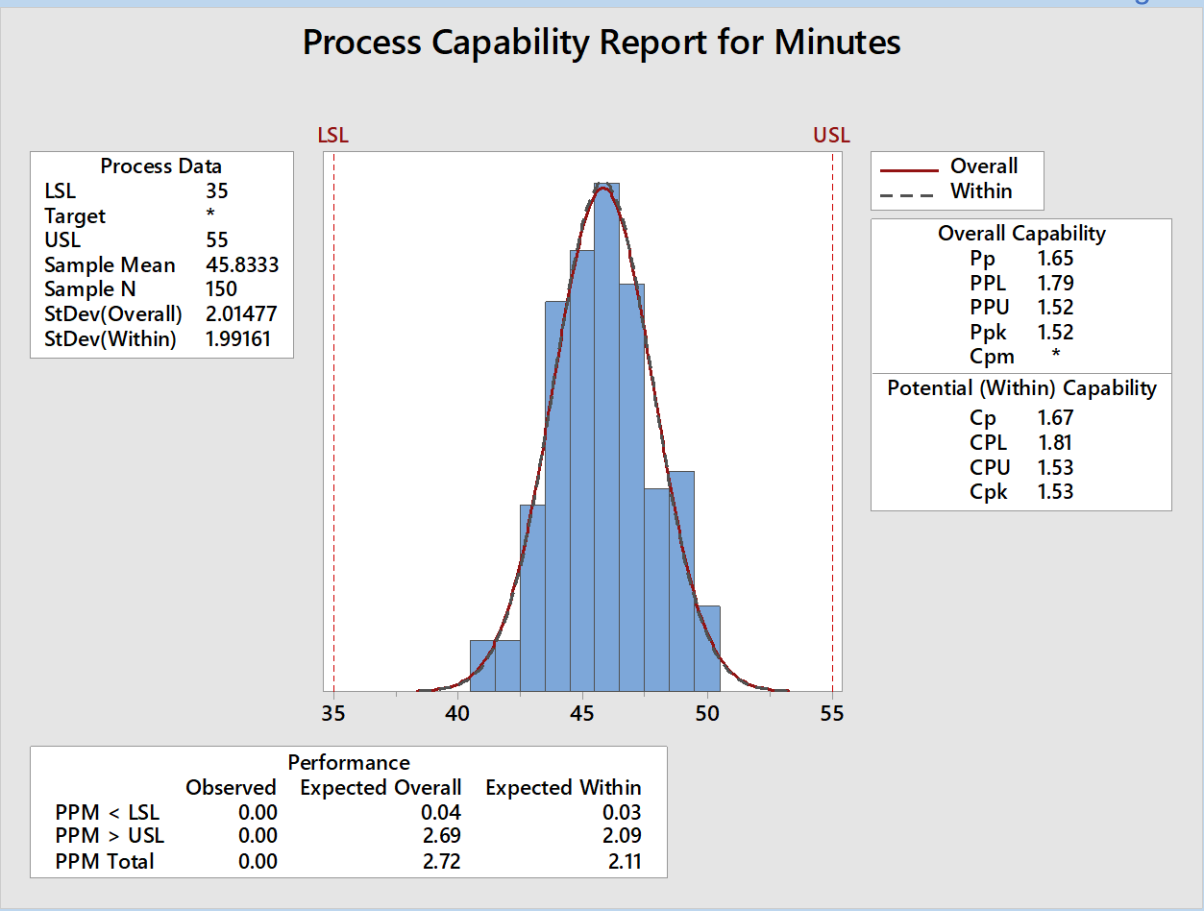
The words ‘performance’ and ‘capability’ are similar in meaning. *Cp* and *Cpk* are short-term measures of the potential process capability. The process may not be able to live up to these standards in the long-term. So the measures of performance, *Pp* and *Ppk,* are more realistic measures of what the process is able to do in the long term. So the process performance is often referred to as the **overall capability** of the process. The word ‘performance’ is also used to refer to the rate of defectives from the process in parts per million, ***PPM***.

To summarise, *Ppk* is the key measure. If the value of *Ppk* is at least 1.33 then it usually follows that all the other measures are at least 1.33 also and we can say that the process is both capable and performing.

Problems 8A

#1. The target duration of a guided castle tour is 45 minutes. The LSL and USL are 35 minutes and 55 minutes respectively. Five tours were timed on each of thirty days. The process capability report is shown below. Assess this process under each of the following headings, supporting your answers with appropriate references to the software output:
(a) normality (b) stability (c) centrality (d) capability (e) performance.

Fig 8.2



#2. Refer again to the situation described in question #1 above. If the USL was reduced to 50 minutes and the LSL remained the same at 35 minutes, would each of your answers at (a), (b), (c), (d) & (e) change or stay the same? Defend your answer in each case.

#3. (Activity) Take a pencil and a blank sheet of paper and draw a freehand line to a target length of 100 mm. Hide your first attempt and draw another freehand line. Repeat until you have a subgroup of five lines. Take a break for a few minutes and then repeat the whole exercise until you have thirty subgroups, giving a total of 150 lines. Alternatively, share the activity with 30 classmates to provide thirty subgroups, giving a total of 150 lines. Finally, measure each line with the ruler and conduct a process capability study with respect to a LSL of 70 mm and a USL of 130 mm.

#4. (Activity) How well can someone estimate a time interval of ten seconds without looking at a timer? Carry out a process capability study on this process with respect to a LSL of 7 seconds and a USL of 13 seconds.

Project 8A

Process Capability Study

Carry out a process capability study, using 30 samples of size 5 from any process. Write a report consisting of the following sections.

(a) Briefly describe the process and the measurement. Choose some values for the upper and lower specification limits, by selecting values that you think might be achievable for this process. Describe how the samples were drawn and explain what the thirty different 'batches' represent in this situation.

(b) After collecting and analysing the data, show the process capability report, including the graph and the indices. You do not need to display the raw data.

(c) What can you say about the normality, stability, centrality, capability and performance of this process?

(d) Based on this process capability study, what are the major practical insights that you have gained about this process?

(e) Based on this process capability study, what are the most promising suggestions that you can offer for improving this process?

8B. Statistical Process Control

Video Lecture https://youtu.be/lBev2yjZzaU

Statistical process control (**SPC**) observes a process to confirm that it is doing its best, or to see whether adjustments are required, in much the same way that a driver controls a vehicle by looking through the windscreen to see if steering adjustments are required. Samples of process measurements are drawn at regular intervals and some sample statistic is plotted as a time series on the SPC chart. **Control limits** are also shown on the chart. These are not specification limits: they represent values that are unlikely to be violated as long as the process remains **in-control**, i.e. continues to do its best. The limits are placed three sigma above and three sigma below the centre line, so the probability of a point falling outside a control limit is approximately one in a thousand. Control limits can be compared to rumble strips along the edges of a road to alert drivers who are drifting out of their lane.

Ideally, measurements are made on the process rather than the product, since the product will be right if the process is right. For example, in baking a cake, SPC would be applied to the oven temperature, rather than to the moisture content of the finished cake. This facilitates timely intervention: if the process goes **out-of-control**, and prompt **corrective action** is taken, the product will be OK. The idea is to correct problems before any defective units are produced. Even if a defective unit of product does occur, the focus of SPC is not on fixing that unit, but on identifying the cause and taking action to prevent any recurrence of the same problem.

Small variations in process measurements are inevitable. SPC charts distinguish between such unexplained variation, due to **common causes**, and more serious

explained variation due to **special causes**. Common causes include the many small, unavoidable changes in the environment and the process that give rise to normally distributed process measurements. A special cause is a single factor that has a large impact on the process, and that must be identified and dealt with. Special causes will typically be one of the 6 Ms:

Men, e.g. operators that are untrained or unmotivated or fatigued
Material, e.g. defective or different material being introduced
Machines, e.g. damage, maladjustment, or overdue maintenance
Methods, e.g. procedures are inappropriate, unspecified or ignored
Milieu (environment), e.g. weather, utilities
Measurements, e.g. instruments giving false signals, such as indicating that an in-control process is out-of-control or vice versa.

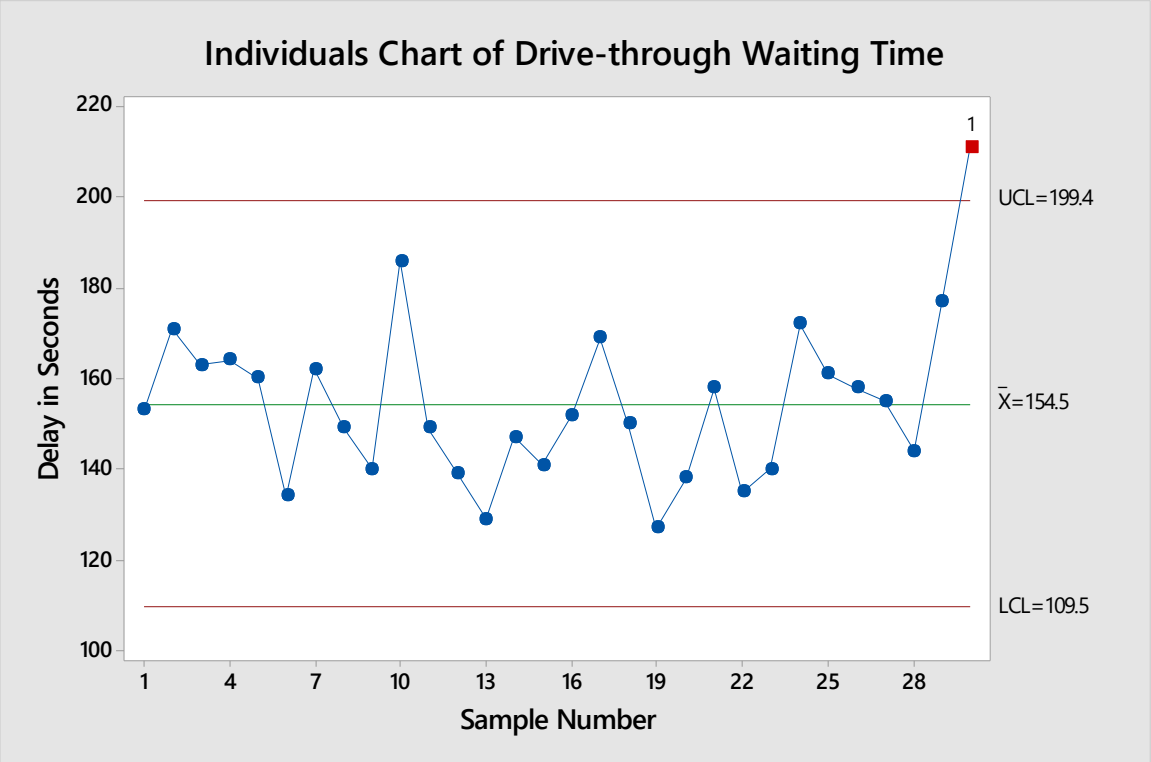
When an SPC chart gives an out-of-control signal, it cannot tell us what has caused the problem. We must use our knowledge of the process to interpret the SPC chart, and to pursue the cause and the solution. The time at which a signal appeared on the chart is a useful clue, although a problem will often not show up in a control chart until after a number of samples have been taken.

Good violations are also possible, where an out-of-control signal points to an improvement, e.g. the waiting time or the number of defectives may have fallen. In such instances the cause should be identified and action taken to ensure recurrence, and new control limits should be calculated to reflect the improved process capability.

Individuals Chart

An individuals chart is the simplest control chart. Individual process measurements, assumed normally distributed, are sampled and plotted.

Fig 8.3



In this example, customers at a drive-through restaurant were sampled and the waiting times experienced by these customers were recorded in seconds. For the first 30 customers sampled, the data were plotted on the individuals chart.

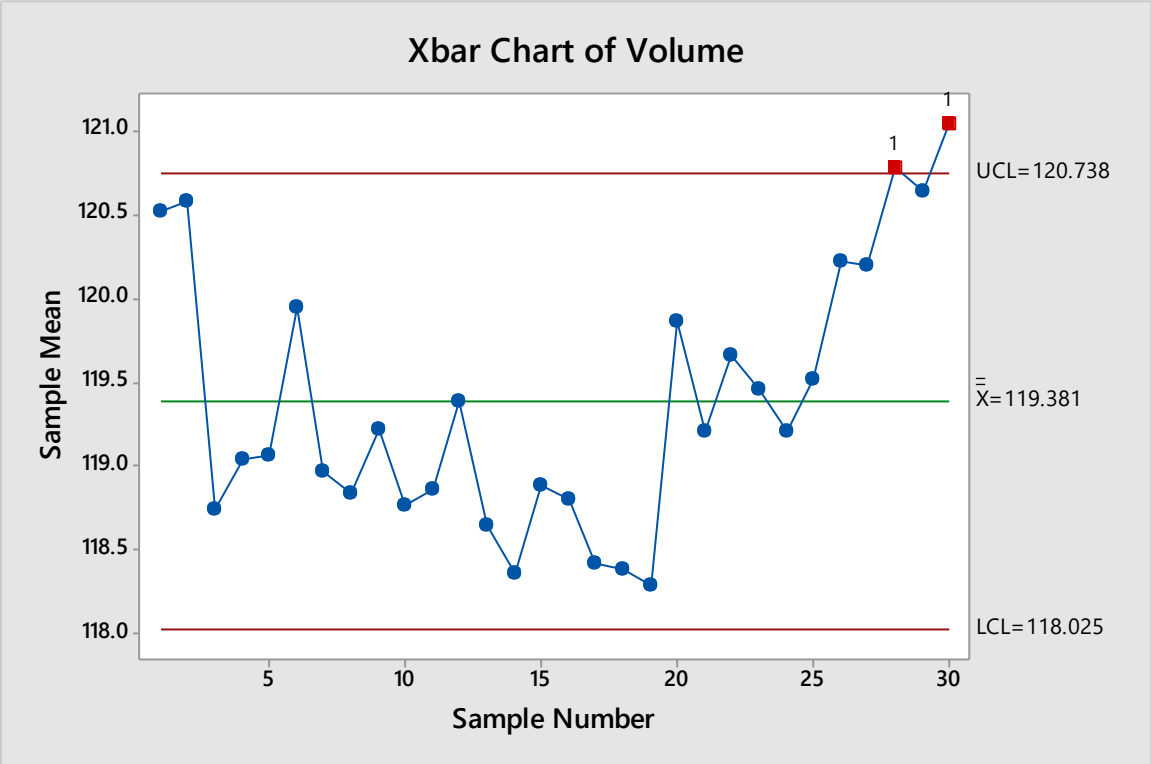
The final point indicates that the process is out-of-control. The reason could be an untrained worker (men), a shipment of lettuce heads instead of the usual lettuce leaves (material), a broken drink dispenser (machines), failure to start cooking the food before pouring drinks (methods), a reduction in the electrical supply voltage that affects the cooking time (milieu), or measuring the time until the customer drives away instead of the time until the order is delivered (measurements). The cause must be identified and corrected. Suppose the USL is 240 seconds: then corrective action can be taken without any violations having occurred.

Xbar Chart

An **Xbar chart** uses a plot of sample means to track the process mean. The process measurements do not need to be normally distributed because sample means are normal anyway. The specification limits for individual values must not be shown on the chart, or compared with values on the chart, because these are sample means.

In this example, samples of five bottles each were taken from thirty different batches of a filling process, and the fill volume of each bottle was observed.

Fig 8.4



The default approach is to use the StDev(Within) to calculate the control limits but if the process is unstable then it may be more realistic to use the StDev(Overall). The final few points indicate that this process is out-of-control.

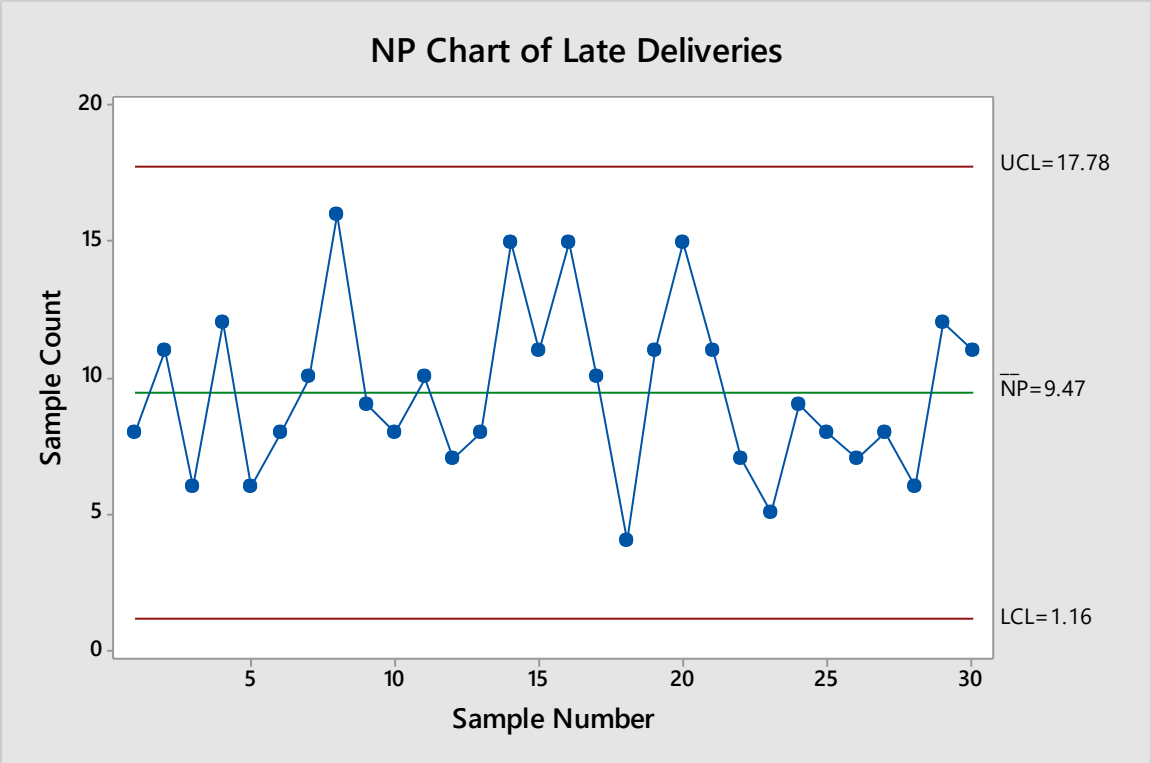
Sample standard deviation charts, ***S* charts**, or sample range charts, ***R* charts**, are sometimes used to accompany Xbar charts. These charts give a signal when the spread of values rises or falls significantly.

NP Chart

We now consider control charts for attributes. The number of **defectives** in a sample can be monitored using an **NP chart**. This involves drawing a sample of fixed size, *n*, at regular intervals, counting the number of defectives in the sample, and plotting this number on the chart. Every unit in the sample is wholly classified as defective or non-defective. The parameter *P* represents the proportion defective in the process, and this can be estimated from the data. Sample sizes must be large, because an average of at least 5 defectives per sample is required. Violation of an upper control limit indicates an increase in the proportion of defective units, but violation of a lower control limit indicates improvement in the process. NP charts are based on the binomial distribution.

In this example, the proportion of late deliveries made by a courier was monitored by sampling 50 deliveries each day and counting how many of these were late. The sample counts for thirty consecutive days were plotted on an NP chart.

Fig 8.5



The chart indicates that the process is in control. This means that the proportion of late deliveries is consistent, with no evidence of special causes.

In this example, the proportion of late deliveries is high, and this might not be acceptable to some customers. But control charts do not exist to ensure that products conform to customer specifications, but to ensure that processes perform to the highest standard of which they are capable.

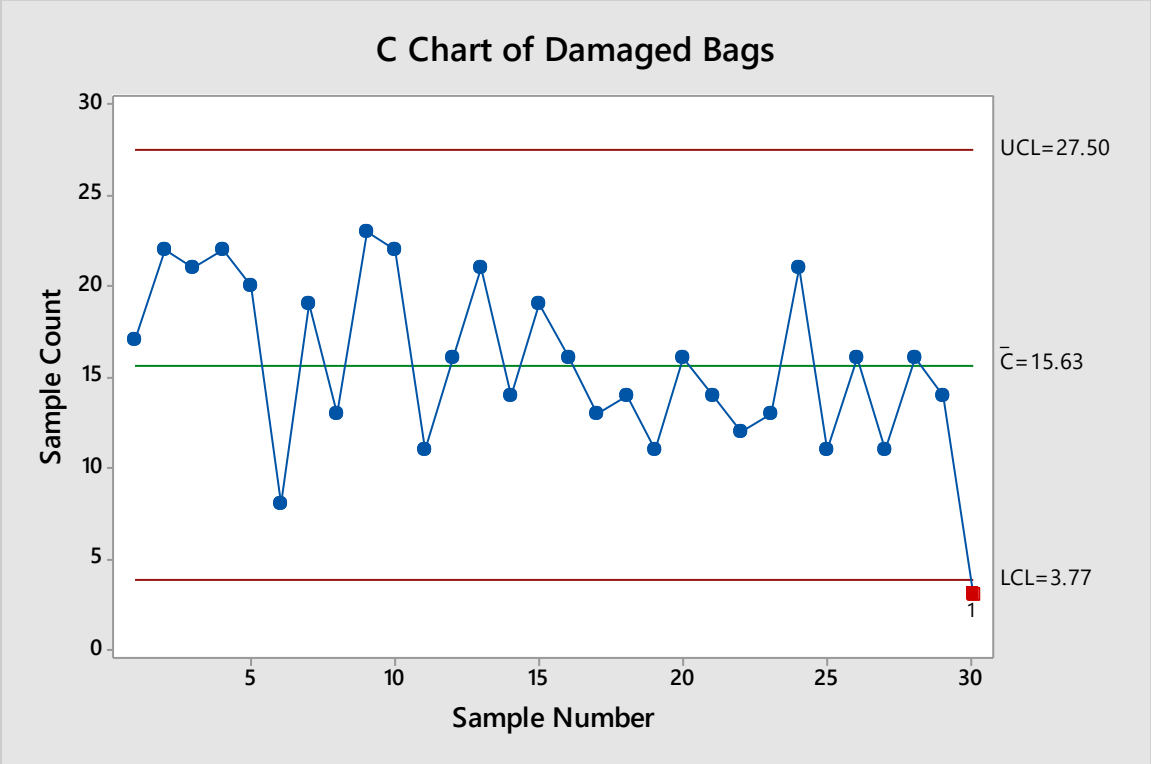
C Chart

The next attribute control chart, the **C chart**, tracks the number of **defects** in a sample. A defect is not the same thing as a defective unit, because we can count many defects on a single unit. The sample must be defined: it could consist of one unit of product or a number of units, or it could be an interval of time, a length, an area, or a volume. The number of defects in the sample (e.g. missed calls, potholes, scratches, bubbles, etc.) is plotted on the chart.

Sample sizes must be large, because an average of at least 5 defects per sample is required. Violation of an upper control limit indicates an increase in the level of defects, but violation of a lower control limit indicates improvement in the process. C charts are based on the Poisson distribution.

In this example, the number of bags damaged at an airport each day was recorded for thirty consecutive days. The results were plotted using a C chart. Notice that a sample is defined as a day.

Fig 8.6



The chart indicates that the process has improved. The cause should be identified and action taken to ensure recurrence. Possible causes include: a new worker taking greater care with the bags (men), a reduction in the maximum allowed bag weight that makes damage to bags less likely (materials), installation of a new carousel (machines), a revised procedure for the handling of bags (methods), better lighting in the baggage reclaim area (milieu), or a failure to record all the damaged bags on a particular day (measurements).

Setting up an SPC System

The questions to be addressed when setting up an SPS system are as follows.

1. Which process parameter needs to be controlled?

In a filling process, the critical parameter could be the fill volume, the fill height, the fill weight, the gross weight, or the proportion of bottles that are overfull, or underfull.

2. Which statistics should be plotted?

A chart for controlling fill volume could plot individual volumes, sample means, or sample means and standard deviations.

3. How big should the sample be?

'Little and often' is best. A 'handful' of about five measurements, drawn frequently, is preferable to a larger sample drawn less frequently, because this facilitates a faster response when problems occur.

4. How will the samples be drawn?

Samples should form **rational subgroups**. This means that the units within a sample should be as alike as possible, so they should be taken close together in time and not mixed from different sources. Unexplained variation should be the only source of variation within a sample.

5. How often will samples be drawn?

This depends on whether the process tends to drift or shift when it goes out-of-control. For a process that is prone to shift suddenly out of specification, the economic principle is that the cost of defective product should be the same as the cost of sampling inspection. Therefore if defective product is costing more than inspection, you should draw samples more frequently. In the case of a process that drifts at a predictable rate, the interval between samples should be short enough so that there is not enough time between samples for the process to drift out of specification.

6. How will the control limits be calculated?

It is usual to estimate the parameters from the process data, and use these values to calculate the control limits. If the values of the process parameters are known with greater certainty from historic data, these values can be used instead. It is not wise to use target parameter values for calculating the limits, as this can lead to two problems. Firstly, if the process is not achieving the targeted standard, there will be multiple violations on the SPC chart and no corrective actions may be available. Secondly, if the process is performing to a higher standard than required, the level of control will be relaxed, and problems will be detected later rather than earlier.

7. What response will be made to an out-of-control signal?

This is the most important question. The purpose of control charts is to provoke action to deal with causes when required so that the process remains in control. Therefore it is necessary to identify who is responsible for taking action, how they will be made aware of control chart violations, and what resources are available to them for exploring and implementing corrective actions. The action taken should be visible to the personnel who maintain the control charts, so that they remain motivated.

Problems 8B

#1. The Xbar chart of volume, at Fig 8.4, indicates that the process is out-of-control. Can you suggest some possible causes, based on the 6 Ms?

#2. An SPC chart is required to monitor the number of scratches on polarising filter screens. The process mean is 1.7 scratches per screen. What type of chart should be used, and what sample size would you recommend?

#3. (Activity) Construct an Xbar chart using the data arising from the activity described at Problems 8A #3. Identify whether the process is out-of-control and, if so, recommend some corrective action.

#4. (Activity) Construct an Xbar chart using the data arising from the activity described at Problems 8A #4. Identify whether the process is out-of-control and, if so, recommend some corrective action.

Project 8B #1

Xbar Chart

Construct an Xbar chart, using 30 samples of size 5 from any process. Write a report consisting of the following sections.
(a) Briefly describe the process and the measurement. Describe how the samples were drawn and explain what the thirty different 'batches' represent in this situation.
(b) Draw the Xbar chart. You do not need to display the raw data.
(c) Based on the evidence in the chart, is this process in control? Explain carefully what your answer means.
(d) Select one of the 6 Ms which you consider to be the most relevant factor for the control of this process. Explain why this factor is important and how attention to this factor could improve this process.
(e) Select another one of the 6 Ms which you consider to be the next most relevant factor for the control of this process. Explain why this factor is important and how attention to this factor could improve this process.

Project 8B #2

Attribute Control Chart

Construct an attribute control chart, either an NP chart for defectives or a C chart for defects, using 30 samples from any process. Write a report consisting of the following sections.
(a) Briefly describe the process and explain what is meant by a defective, or a defect, in this particular context. Describe what the samples consist of, and what the thirty different 'batches' represent in this situation.
(b) Draw the chart. You do not need to display the raw data.
(c) Based on the evidence in the chart, is this process in control? Explain carefully what your answer means.
(d) Select one of the 6 Ms which you consider to be the most relevant factor for the control of this process. Explain why this factor is important and how attention to this factor could improve this process.
(e) Select another one of the 6 Ms which you consider to be the next most relevant factor for the control of this process. Explain why this factor is important and how attention to this factor could improve this process.

8C. Acceptance Sampling

Video Lecture https://youtu.be/CdekSfjGPzM

Sampling inspection by attributes provides a method for determining the acceptability of a series of batches from a process. It confirms that the product delivered by the process satisfies stated requirements.

Of course, it is better to manufacture products **right first time** rather than using resources to inspect out the bad units afterwards. If it is inevitable that some defective units will be made, then effective **100% inspection** can be used to identify and remove all of these defectives. If 100% inspection is too costly, then **sampling inspection** can be used to ensure that the level of defectives is below some specified threshold.

Sampling Plans

Sampling inspection is carried out by following a **sampling plan**. A sampling plan consists of a rule for selecting the sample and a rule for making a decision based on the sample.

EXAMPLE Sampling Plan: n = 8, Ac 1, Re 2.
This plan calls for a random sample of 8 units to be selected from the batch and inspected. If one or fewer defectives are found in the sample, then all the rest of the batch is accepted. If two or more defectives are found in the sample, then the entire batch is rejected. Rejected batches undergo **rectification**, which involves 100% inspection and replacement of all defectives by non-defectives.

How Sampling Plans Work

Whether or not a particular batch will be accepted depends partly on the quality of the batch, and partly on the luck of the draw. The quality of the batch is measured by the proportion of defectives, ***p***, in the batch, called the **incoming quality level**. For any given value of p, there is a certain chance that the batch will be accepted, called the **probability of acceptance**, P_A. The probability of acceptance is the proportion of all such batches that would be accepted, in the long run.

For a 'good' batch, with only 5% defective, P_A = 94% with this sampling plan. The producer and consumer might have agreed in advance that any value of p up to 5% is acceptable as a process average. This figure, the highest acceptable process average per cent defective, is called the **acceptable quality level**, **AQL**. Batches that have a proportion defective equal to the AQL should be accepted, and we see that 94% of such batches are accepted. The remaining 6% of such batches are simply unfortunate: their samples of 8 units just happen to contain two or more defective units, although the proportion defective in the batch is only 5%. Hence, the probability of rejection in such a case (6%, in this example) is called the **producer's risk**, ***α***. Note that the AQL does not specify the maximum acceptable per cent defective in a batch, but the average acceptable per cent defective in the process. Therefore, some individual batches with per cent defective higher than the AQL may be accepted.

For a 'bad' batch, with 40% defective, P_A = 11% with this sampling plan. The producer and consumer might have agreed in advance that 40% defective is unacceptable even in a single batch. Such a figure is called the **lot tolerance per cent defective** (**LTPD**),

or **limiting quality** (**LQ**), or **rejectable quality level** (**RQL**). Batches with a proportion defective equal to the LTPD should not be accepted, and we see that only 11% of such batches are accepted. These 11% are fortunate: their samples of 8 units just happen to contain fewer than two defective units, although the proportion defective in the batch is 40%. Hence, the probability of acceptance in this case (11%, in this example) is called the **consumer's risk**, ***β***.

What Sampling Plans Do

Firstly, sampling plans can identify a very bad batch (LTPD or worse) immediately. Secondly, sampling plans can identify an unacceptable process (worse than AQL) sooner or later: as batches from such a process are frequently submitted for sampling inspection, some of them will eventually get caught.

There are some things that sampling plans cannot do for us. Firstly, sampling plans cannot take the place of corrective action: when batches are rejected because of poor quality, corrective action is essential – just continuing with sampling inspection will not improve the process. Secondly, sampling plans cannot guarantee a correct decision about any one particular batch. Thirdly, sampling plans cannot guarantee **zero defects**: either 'right first time' or effective 100% inspection is required for this.

Sampling Plan Calculations

Let n, p and c represent the sample size, the incoming quality level, and the acceptance number, respectively. The probability of acceptance is the probability of obtaining c or fewer defectives. This is the cumulative probability of c for a binomial distribution with parameters n and p.

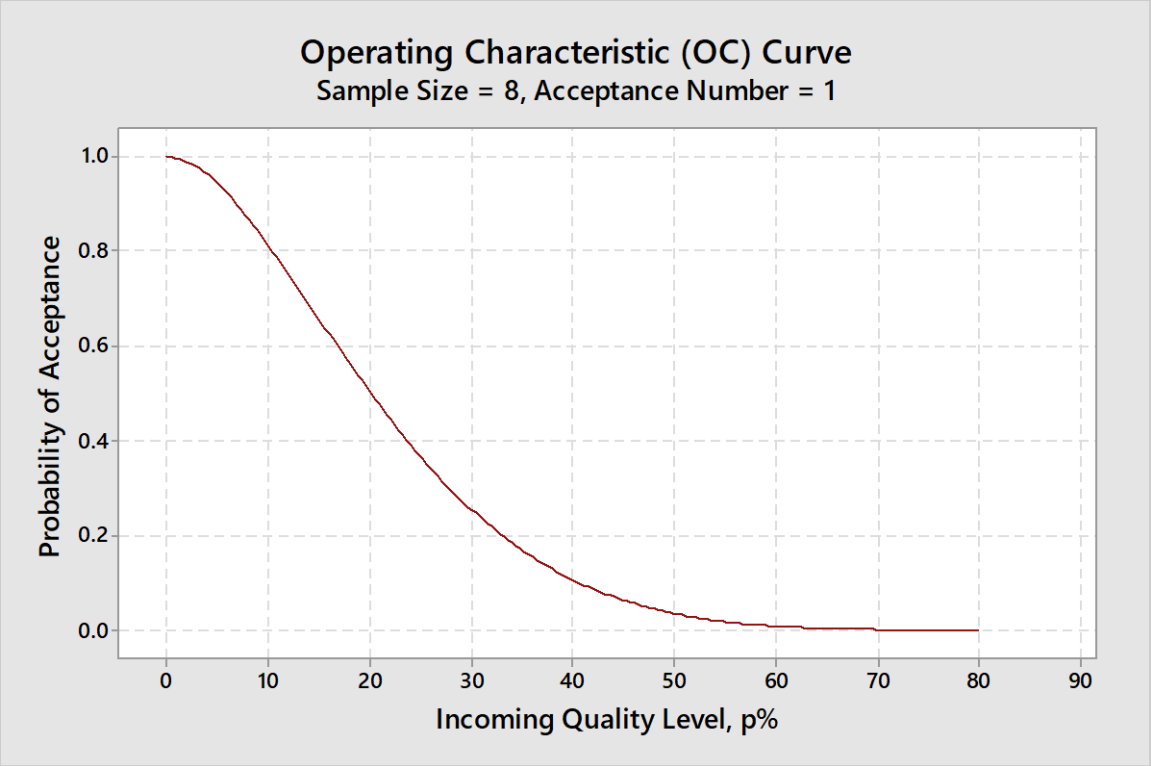
The table below shows the probability of acceptance, corresponding to a number of different values of the incoming quality level, for the sampling plan n = 8, Ac 1, Re 2.

Table 8.1

p (%)	P_A
0	1
5	0.9428
10	0.8131
15	0.6572
20	0.5033
25	0.3671
30	0.2553
35	0.1691
40	0.1064
45	0.0632
100	0

A graph of P_A versus p is called an **operating characteristic curve**.

Figure 8.7



The OC curve is very useful. It shows the level of protection that the plan offers against any value of the incoming quality level. Before selecting a sampling plan for use, its OC curve should be studied, to see if it is satisfactory. For example, the above plan could be useful in a situation where the LTPD is 40%, because P_A = 11%, but it would not be useful if the LTPD was 20% because P_A = 50%. The value of p that has a probability of acceptance of 50% is called the **indifferent quality level**. The indifferent quality level in this example is p = 20%.

Outgoing Quality

Average outgoing quality (AOQ) is the proportion of defective parts downstream from sampling inspection. If no rectification is carried out on the rejected batches, then AOQ is only marginally better than the incoming quality level. (The rejected batches are removed, and in the case of accepted batches, any defective part found while sampling is replaced.) If rectification is carried out, the AOQ is better than the incoming quality level of the submitted batches. Theoretically, AOQ is low when incoming quality is very good and also when incoming quality is very poor, because poor quality leads to a great deal of rectification of batches. The peak on the AOQ curve is known as the **average outgoing quality limit** (AOQL). The AOQL provides a 'worst case' guarantee for the average outgoing quality of a continuous series of batches from a stable process, assuming that a rectification scheme is in place. Typically, the AOQL is marginally greater than the AQL if large samples are used. If the sampling plan uses smaller samples, the AOQL may be around two or three times the AQL.

Double Sampling Plans

Double sampling involves taking two smaller samples rather than one sample. If the first sample is very good or very bad, the batch can be accepted or rejected straight away, without any need for a second sample. Otherwise the second sample is drawn and a verdict is reached based on the cumulative sample. The advantage of double sampling plans is that they typically involve smaller sample sizes. The disadvantages are that they are more complex to operate and in some situations it can be very troublesome to have to return for a second sample. **Multiple sampling** plans are similar to double sampling plans, but involve taking up to seven small samples: after each sample it may be possible to make a decision about the batch without drawing any further samples. **Sequential sampling** takes this concept to the limit: after each sampling unit is drawn, a decision is made to accept the batch, to reject the batch, or to continue sampling.

Practical Guidelines for Sampling Plans

Before choosing a sampling plan, the AQL and the LTPD must be identified. A sampling plan that meets your requirements can then be found using software or published tables or an online sampling plan calculator. All samples should be drawn randomly from the entire batch, and this includes the first and second samples in a double sampling plan. Avoid sampling only from the start of a run, or selecting clusters of units that are packaged together. If shortcuts are taken with random sampling then you cannot be sure what level of protection your sampling plan provides.

There are a couple of things that are useful to know if you are in a position to choose the LTPD or the batch size. Firstly, for any fixed value of the AQL, you can reduce the sample size by choosing the largest possible value for the LTPD. Secondly, if you choose a large batch size then fewer samples will be needed but when a batch fails there will be a large rectification task to be completed.

Problems 8C

#1. (Activity) For a cellular telephone service, a dropped call happens when your phone gets disconnected from the network during a conversation. Let us regard a call as a unit of product, a dropped call as a defective unit, and a day as a batch.
(a) Identify the values of the AQL and LTPD that you consider appropriate.
(b) Use software to identify a suitable sampling plan.
(c) Explain what the producer's risk and the consumer's risk mean in this context.
(d) Use the OC Curve of your selected plan to identify the indifferent quality level.

8D. Measurement System Validation

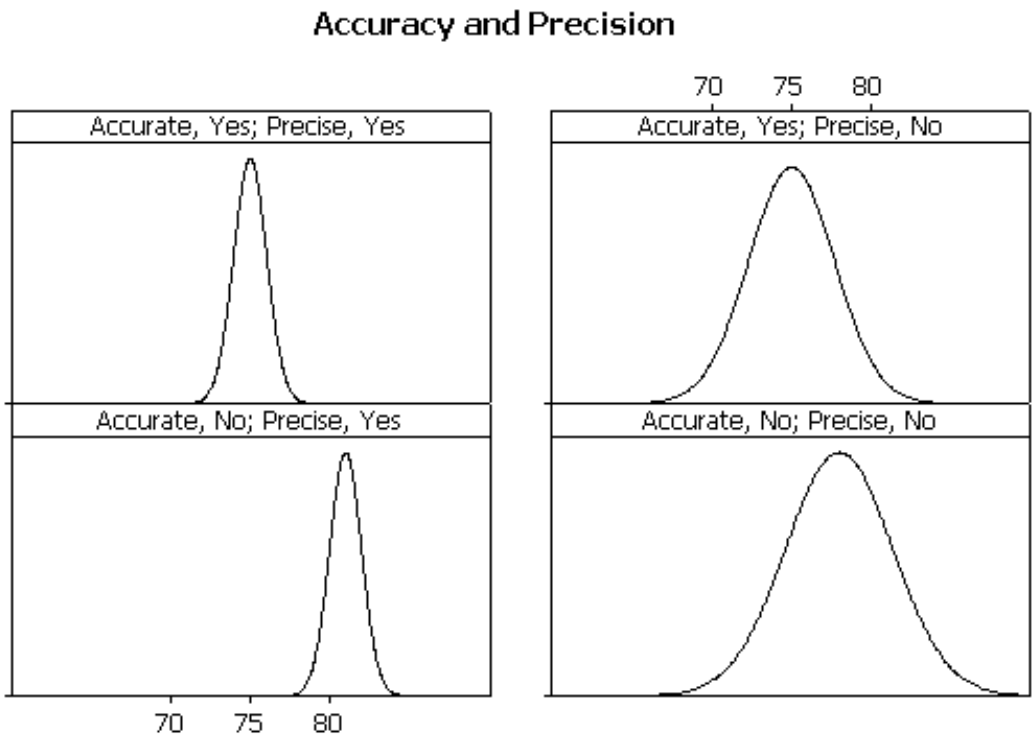
Video Lecture https://youtu.be/qVoAwvHZh6s

Accuracy and Precision

A measuring instrument can be tested by repeatedly measuring a fixed quantity, and comparing the results with the correct answer. The instrument is said to be **accurate** (**unbiased**) if the mean is on target. It is said to be **precise** if the standard deviation is small, i.e. if the results are close together.

EXAMPLE A person who weighs 75 kg is weighed repeatedly on a scale. Any of the following patterns could arise.

Fig 8.8



Gage Linearity and Bias Studies

A linearity and bias study estimates the average **bias** of an instrument, and investigates whether the bias changes as bigger objects are measured (**linearity**). In this context, the instrument is often referred to as a **gage**, and the objects are referred to as **parts**.

A number of parts are selected, covering the full range of the instrument. The **reference value** for each part is obtained from a trusted authority such as a reference laboratory or an expert opinion. Then each part is measured a number of times with the instrument. The individual deviation of each measurement from its reference is calculated, and also the average deviation for each part.

EXAMPLE An analogue kitchen scale was tested by weighing each of five consumer products twice. The data are shown in Table 8.2.

The mean of all the individual deviations is the average bias.
Average bias = 83.5
This scale tends to overstate the weight by 83.5 grams on average.
The regression coefficient of average deviation on reference estimates the linearity.
Average Deviation = 53.5 + 0.0673 Reference
Multiply the regression coefficient by 100 to get the %Linearity.
%Linearity = +6.73%
For every 100 grams increase in the reference measurement, the bias tends to increase by 6.73 grams.

Table 8.2

Part	Reference	Measurement	Deviation	Average Deviation
1	200	260 270	60 70	65.0
2	300	375 340	75 40	57.5
3	400	500 505	100 105	102.5
4	625	725 740	100 115	107.5
5	700	780 790	80 90	85.0

Components of Variance

When a number of parts are measured, the total variation in the measurements arises from two different sources. There is **material variation** because the parts are not all the same, and **measurement system variation** because even when the same part is measured repeatedly, the results are not the same, because the instruments and operators that take the measurements are not perfect. These two sources of variation are assumed to be independent, so the following additive model applies.

Model 8.1

$$\sigma^2_{Total} = \sigma^2_{Mat} + \sigma^2_{MS}$$

Care should be taken that any subtle material variation is not attributed to the measurement system. Examples include time-to-time variation, such as shrinkage, or within-part variation, such as different diameters on the same part.

It may be possible to explain some of the measurement system variation by pointing out that the measurements arose on different **sessions**, i.e. different operators, different laboratories, different instruments, or different days. But even on a single session, there may be some unexplained variation. The explained variation between sessions is called **reproducibility** (rpd) and the unexplained variation within sessions is called **repeatability** (rpt). These two sources of variation are also assumed to be independent, so an additive model can be used again.

Model 8.2

$$\sigma^2_{MS} = \sigma^2_{rpd} + \sigma^2_{rpt}$$

A single model can be used to represent all three sources of variation in a set of measurements: differences between parts (material variation), differences between sessions (reproducibility), and unexplained variation (repeatability).

Model 8.3

$$\sigma^2_{Total} = \sigma^2_{Mat} + \sigma^2_{rpd} + \sigma^2_{rpt}$$

Gage R&R Studies

An **R&R study** is a designed experiment, in which a number of typical parts are measured a number of times in succession (to investigate repeatability), on a number of different sessions (to investigate reproducibility). If measurement is destructive, replication can be achieved by using indestructible surrogate parts, or sub-samples of very similar parts.

EXAMPLE A measurement system is used to measure the deflection of timber beams, for which LSL = 20 mm and USL = 100 mm. An R&R study involved measuring 15 randomly chosen beams twice each, on each of four days. The output is shown below.

Gage R&R Study - ANOVA Method

Two-Way ANOVA Table With Interaction

Source	DF	SS	MS	F	P
Part	14	1982.97	141.641	27.0788	0.000
Session	3	6.52	2.174	0.4156	0.743
Part * Session	42	219.69	5.231	0.5894	0.964
Repeatability	60	532.44	8.874		
Total	119	2741.63			

Two-Way ANOVA Table Without Interaction

Source	DF	SS	MS	F	P
Part	14	1982.97	141.641	19.2085	0.000
Session	3	6.52	2.174	0.2948	0.829
Repeatability	102	752.13	7.374		
Total	119	2741.63			

Gage R&R

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	7.3739	30.52
Repeatability	7.3739	30.52
Reproducibility	0.0000	0.00
Session	0.0000	0.00
Part-To-Part	16.7834	69.48
Total Variation	24.1572	100.00

Process tolerance = 80

Source	StdDev (SD)	Study Var (6 × SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	2.71549	16.2929	55.25	20.37
Repeatability	2.71549	16.2929	55.25	20.37
Reproducibility	0.00000	0.0000	0.00	0.00
Session	0.00000	0.0000	0.00	0.00
Part-To-Part	4.09675	24.5805	83.35	30.73
Total Variation	4.91500	29.4900	100.00	36.86

Number of Distinct Categories = 2

The analysis consists of an ANOVA in which the response is 'measurement' and the factors are 'part' and 'session'. A preliminary ANOVA table with interaction is used to confirm that there is no significant interaction. This would mean that some parts gave higher results in a particular session.

The main analysis consists of an ANOVA table without interaction, which is used to estimate variance components, which are also presented as standard deviations. Remember that variances are additive while standard deviations are not.

We see that 'part' is significant, and contributes 69% of the total variance. This is not at all surprising since it is to be expected that parts are different.

The remaining 31% of the variation is contributed by the measurement system, and all of this seems to be unexplained variation due to repeatability. This is unfortunate because explained variation due to reproducibility can sometimes be corrected by addressing the causes. If reproducibility was the main problem, corrective actions could include training (if sessions denote different operators), or standard operating procedures (if sessions denote different laboratories), or calibration (if sessions denote different instruments), or environmental controls (if sessions denote different days). When repeatability is the main problem there may be no corrective actions available: perhaps the measuring instrument is simply not good enough.

The following metrics are often used to summarise the results of a gage R&R study.

1. Number of Distinct Categories

This is the number of different groups into which the measurement system can divide the parts. It is calculated by taking account of the variation among the parts, and comparing this to a confidence interval for a single measurement. Anything less than two groups (big and small) is obviously useless for distinguishing between parts. In the automotive industry, at least five distinct categories are required for a measuring system to be considered adequate. In this example, the number of distinct categories is 2.

2. Signal-to-noise ratio

The signal-to-noise ratio (**SNR**) also considers whether the measurement system is able to tell the difference between the parts that are presented to it. The SNR is the ratio of the material standard deviation to the measurement system standard deviation. In the example above:

SNR = 4.09675 ÷ 2.71549 = 1.5

Threshold values for the SNR are as follows:

> 10	good
3 – 10	acceptable
< 3	unacceptable

The SNR in our example is unacceptable. The measurement system is unable to distinguish between the parts that are presented to it. This may be because the parts are very alike. Strangely, as production processes improve, the SNR goes down: it appears that the measurement system is deteriorating when in fact the parts are simply more alike, and therefore it is more difficult to tell them apart.

3. Precision to tolerance ratio

Although the measurement system in our example cannot distinguish between the parts that are presented to it, it may be able to tell the difference between a good part and a bad part. The **precision to tolerance ratio**, also called the **%R&R**, or the **%Tolerance**, considers how well a measurement system can distinguish between good and bad parts, by dividing the study variation by the tolerance.

In our example the %R&R = 20.37%
Threshold values for the %R&R are as follows:

< 10%	good
10 – 30%	acceptable
> 30%	unacceptable

So this measurement system is able to distinguish between good and bad parts.

Problems 8D

#1. A map was used to estimate a number of distances by road, in kilometres, and these estimates were compared with the actual distances measured by carrying out the journey. A gage linearity and bias study yielded the following regression equation:

Average Deviation = -0.950 - 0.0550 Reference

Calculate the %Linearity and explain what it means.

#2. Suzanne carried out a Gage R&R study in which three operators each measured twelve parts three times using a vision system. For these parts, LSL = 42 mm and USL = 48 mm. Use the output below to identify the values of the Number of Distinct Categories, the SNR and the %R&R and explain what these values indicate.

Gage R&R

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	0.06335	2.94
Repeatability	0.02525	1.17
Reproducibility	0.03809	1.77
Session	0.03809	1.77
Part-To-Part	2.09107	97.06
Total Variation	2.15441	100.00

Process tolerance = 6

Source	StdDev (SD)	Study Var (6 × SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	0.25169	1.51013	17.15	25.17
Repeatability	0.15891	0.95348	10.83	15.89
Reproducibility	0.19518	1.17106	13.30	19.52
Session	0.19518	1.17106	13.30	19.52
Part-To-Part	1.44605	8.67631	98.52	144.61
Total Variation	1.46779	8.80675	100.00	146.78

Number of Distinct Categories = 8

#3. (Activity) Create two rulers by photocopying a ruler twice with different magnification settings each time. Conduct a Gage R&R study by using these rulers to repeatedly measure a number of rubber bands.

Project 8D

Gage R&R Study

Design and carry out a Gage R&R Study on any measurement system of your choice.

For example, you could use a ruler to measure the lengths of a number of potatoes or you could use a stopwatch to measure the durations of a number of lines from a recorded song. You simply need any set of objects ('parts') of different sizes that are difficult to measure so that differences will arise even when the same object is measured repeatedly. The different sessions could be different people or different days.

You will not be assessed on how well you take the measurements but rather on how well you assess the measurement system. Use at least five parts and at least two sessions. You will also need at least two replicates. You must also choose values for the lower spec and the upper spec, and you may choose any two values you like.

After you have collected the measurements and analysed the results, write a report consisting of the following sections.

(a) Identify the measurements, the parts and the sessions.

(b) Display the data.

(c) Show the ANOVA table and the rest of the Gage R&R Study output.

(d) Interpret the values of the number of distinct categories, the SNR and the %R&R.

(e) Based on the software output, present your recommendations for improving this measurement system.

8E. The Seven Basic Quality Tools

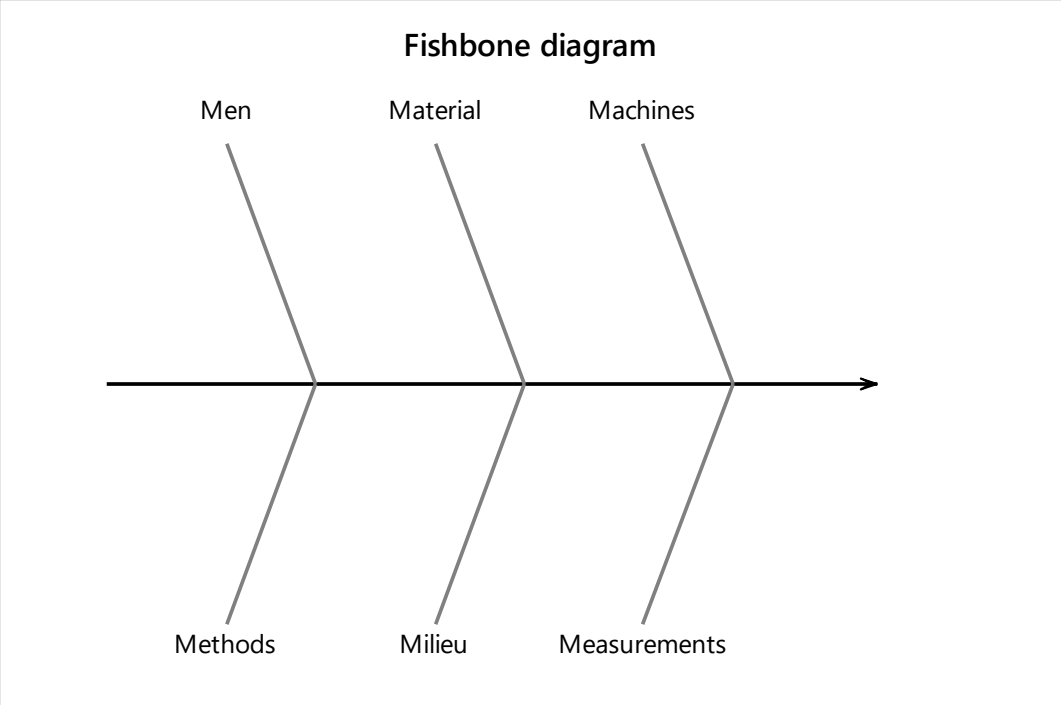
Video Lecture https://youtu.be/5_1jENiJIwc

Many problems that arise with processes can be solved using simple statistical tools. These seven basic quality tools can be understood and used by people with only a little training. More advanced techniques and more skilled practitioners can be called upon when required.

#1. Fishbone diagram

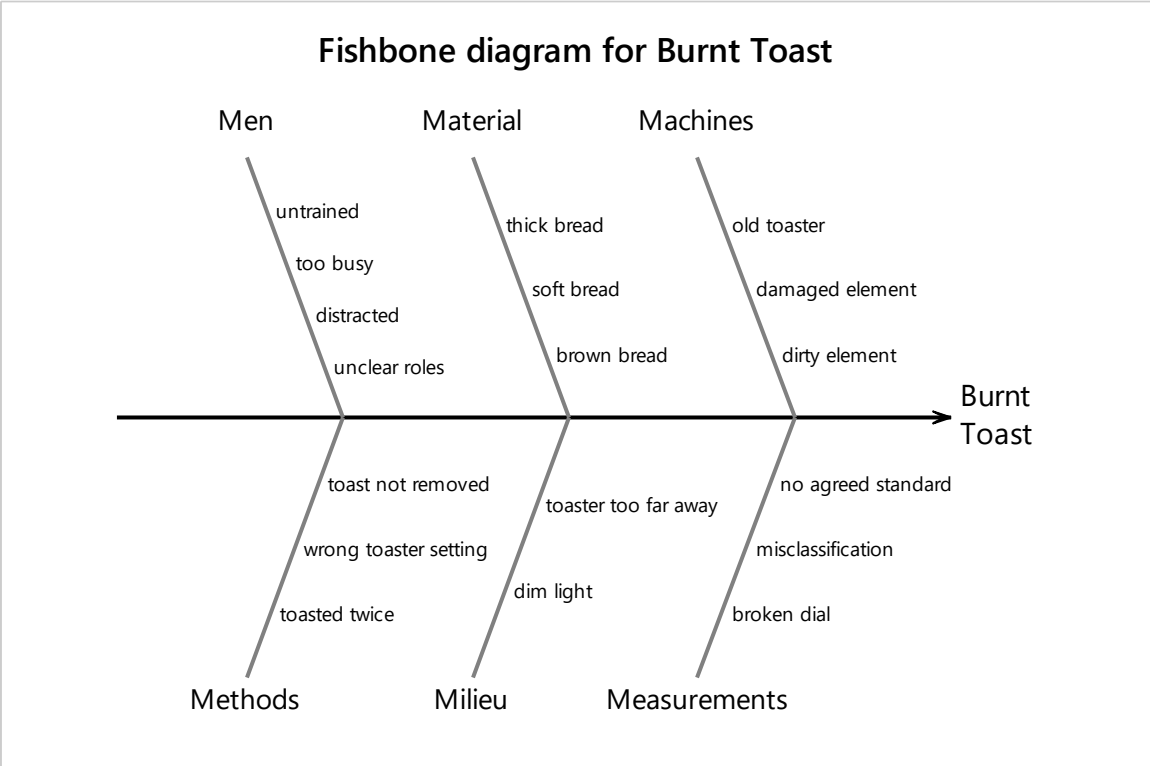
A fishbone diagram is used to identify **what** is causing a problem. It can also be called an **Ishikawa diagram** after its inventor, or a **cause-and-effect diagram**. The main six categories of causes (Men, Material, Machines, Methods, Milieu and Measurements) branch out from the spine, like a fishbone.

Fig 8.9



The problem to be addressed is written at the right hand side, and then a brainstorming session is held. Different causes are suggested and these are written on the diagram as sub-branches in the hope of identifying the root cause of the problem. The fishbone diagram in Fig 8.10 could be used to identify the cause of 'Burnt Toast'.

Fig 8.10



Any one of these suggestions might prove to be the key to solving the problem. For example, perhaps the 'dim light' in the toaster area makes it difficult to judge when the toast is ready and so installing a brighter light might solve the problem.

#2. Check sheet

A check sheet is used to identify **where** defects are occurring. In its simplest form it consists of a table that lists different categories of defects. The defects are then observed and recorded as they arise, by making a check mark on the sheet to represent each occurrence. Every fifth check mark is usually entered as a strikethrough in order to make the chart easier to read. A check sheet of the types of defects encountered by a motor breakdown service is shown below. The most frequently occurring type of defect becomes obvious from the check sheet.

Table 8.3

Fuel	Puncture	Keys	Battery	Mechanical	Accident	Other
/	//// ///	//	//// ////	//// ///	///	/

Instead of categorising the defects by type, they could be categorised by location.

Fig 8.11

Rather than writing the different locations in a table, a map or diagram can be used to build up a picture of where the defects are occurring, and this is called a **defect concentration diagram**. Examples include the location of traffic accidents on a city road network map, the location of injuries on a diagram of the human body, or the location of perforations on a diagram of a supposedly waterproof glove. A famous historical example is a diagram, shown at Fig 8.11, used by Dr John Snow to record the addresses of victims of the 1854 London cholera outbreak. This diagram led him to conclude that the source of the outbreak was the water from the Broad Street pump.

A check sheet can also be used to discover the distribution of numerical data and in this case it is called a **tally sheet**. The tally sheet below was used to record the shoe-sizes of students in a class.

Table 8.4

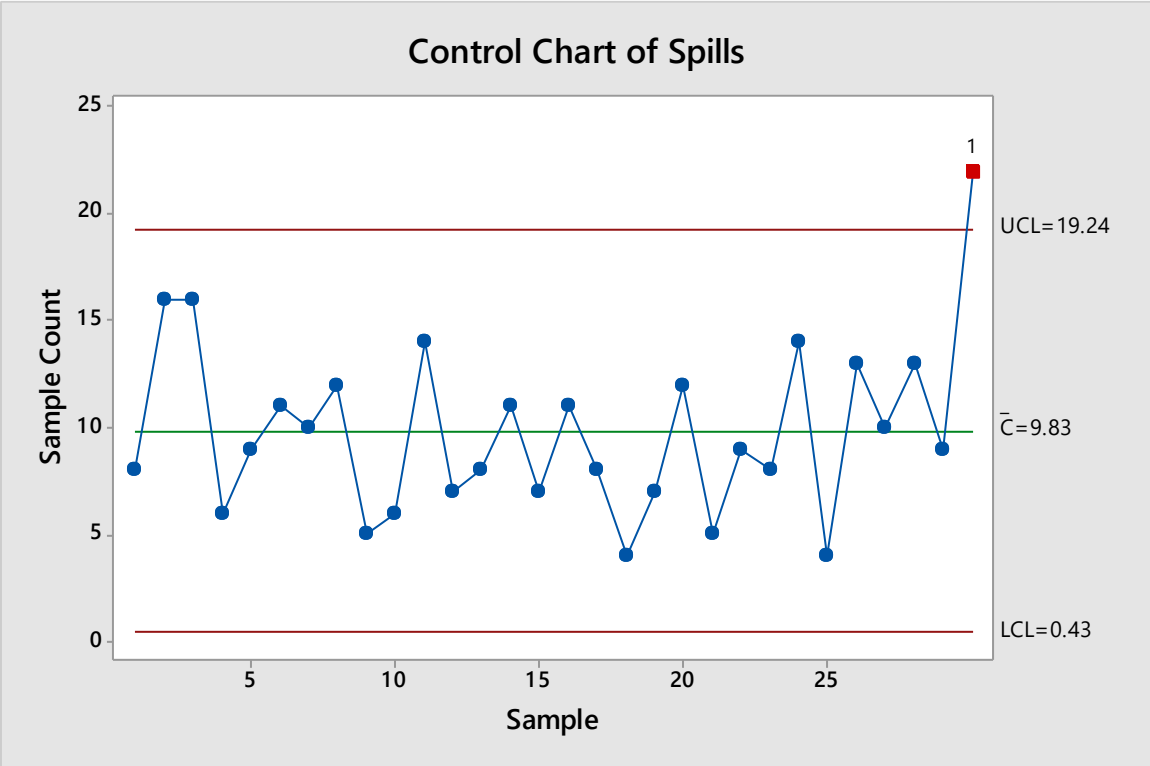
Shoe-size	Tally
2	𝍸
3	𝍸𝍸
4	𝍸𝍸𝍸
5	𝍸𝍸𝍸𝍸
6	𝍸𝍸𝍸
7	𝍸𝍸
8	𝍸𝍸𝍸
9	𝍸𝍸𝍸𝍸
10	𝍸𝍸𝍸𝍸
11	𝍸𝍸𝍸
12	𝍸𝍸
13	𝍸

The tally sheet reveals that the distribution is bimodal.

#3. Control chart

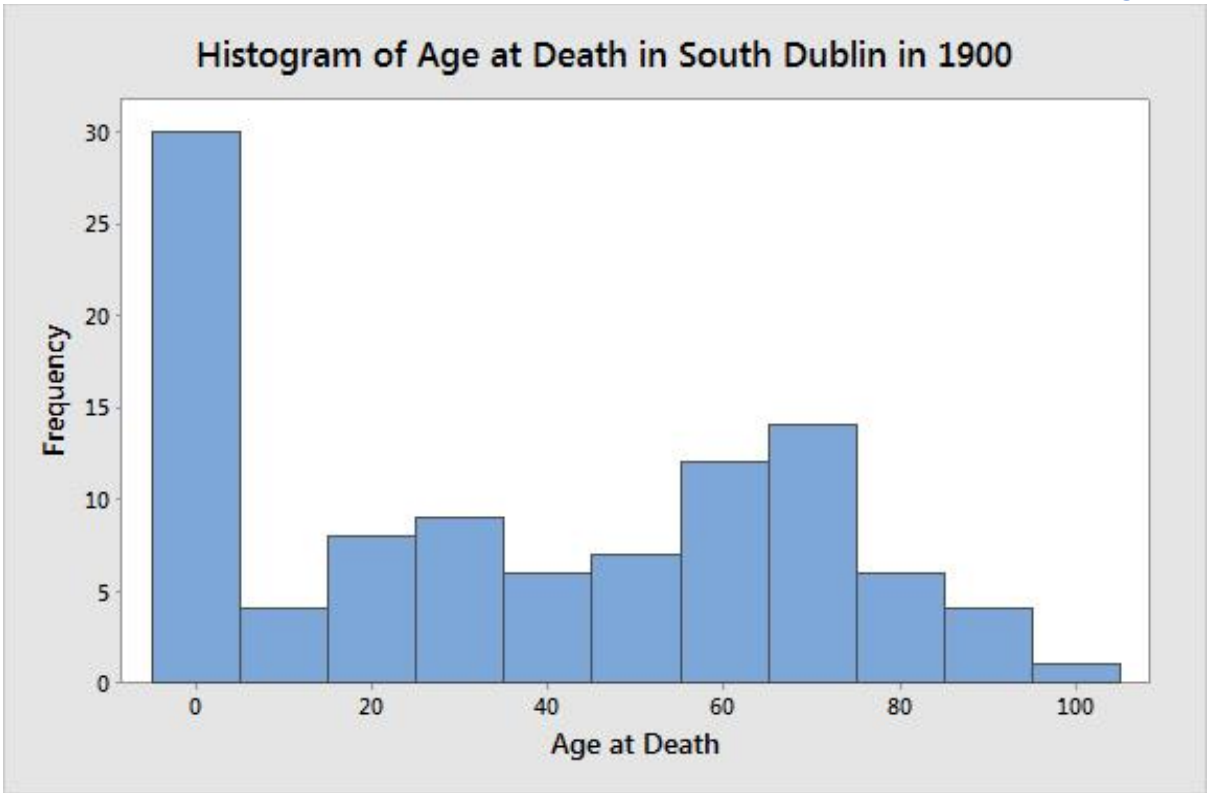
A control chart is used to identify **when** adjustment is needed in order to keep a process in control. The control chart gives a simple picture of how the process is behaving over time, and provides a clear signal when the process requires intervention to identify and correct the cause of any problem. Control charts look like a hospital patient's temperature chart. There are many different types of control chart but they typically consist of a time series plot of points going from left to right to represent some feature of a process, with control limits above and below that are unlikely to be crossed while the process continues to perform at its best. When a point is plotted outside a control limit, the cause (Men, Material, Machines, Methods, Milieu or Measurements) is sought out and corrective action is put in place to prevent recurrence of the problem. As long as the points fall inside the control limits, the process is left alone because too much adjustment harms process performance. Control charts are described in greater detail earlier in this chapter in section 8B. The control chart below monitors the number of spills per hour on a filling line. It is obvious from the last point plotted that the average number of spills per hour has changed. Mopping up these spills does not solve the problem. The cause of the change must be identified and action must be taken to address the cause so that the problem will not recur.

Fig 8.12



#4. Histogram

Fig 8.13



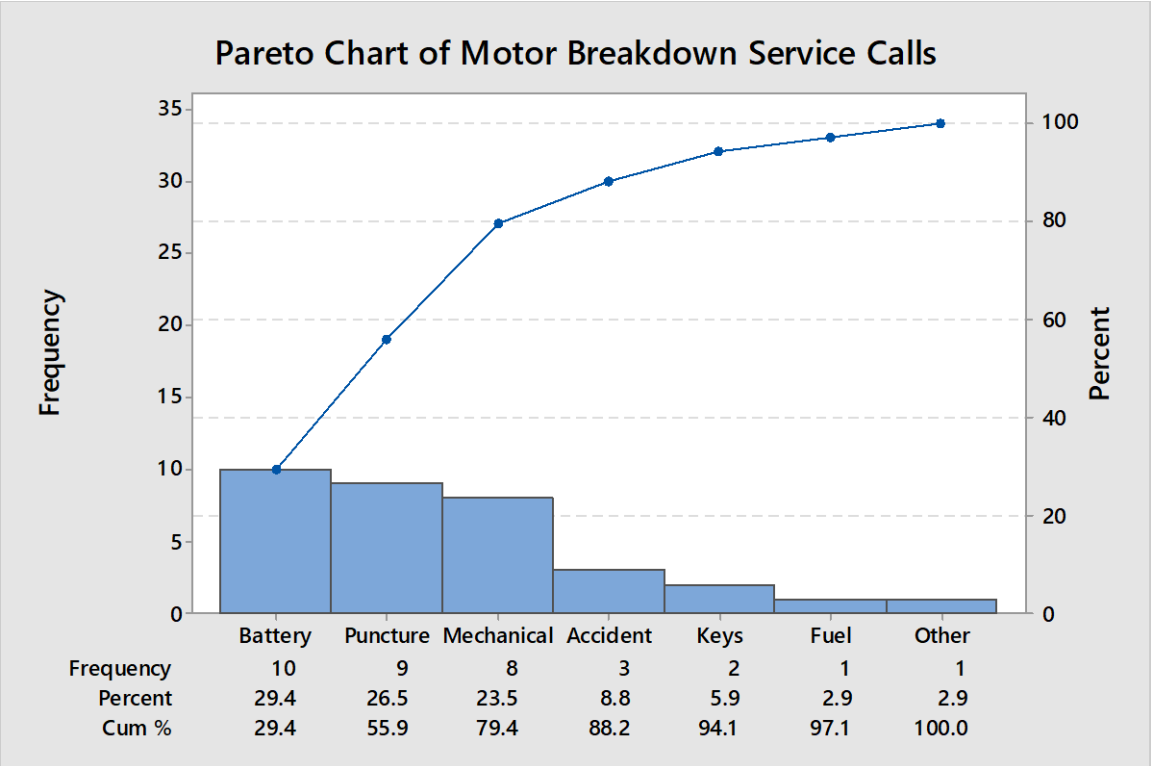
A histogram is used to show **how** a variable is distributed, and this can reveal a lot about the underlying process. A detailed discussion of histograms is presented in chapter 1 where some of the different shapes are explained. A histogram is a simple

but powerful tool that should not be set aside in favour of more complex forms of analysis that may be used later on as required. A large sample of at least 100 data should be used to construct a histogram, because histograms constructed using small samples can be useless or misleading.

The histogram at Fig 8.13 shows the age at death in South Dublin in 1900. The histogram has three peaks. The first peak, close to zero, reveals the very high level of infant mortality at that time. The next tallest peak, at around age seventy, corresponds to people who died in old age. The third peak, at around age 30, may relate to women who died in childbirth.

#5. Pareto chart

Fig 8.14

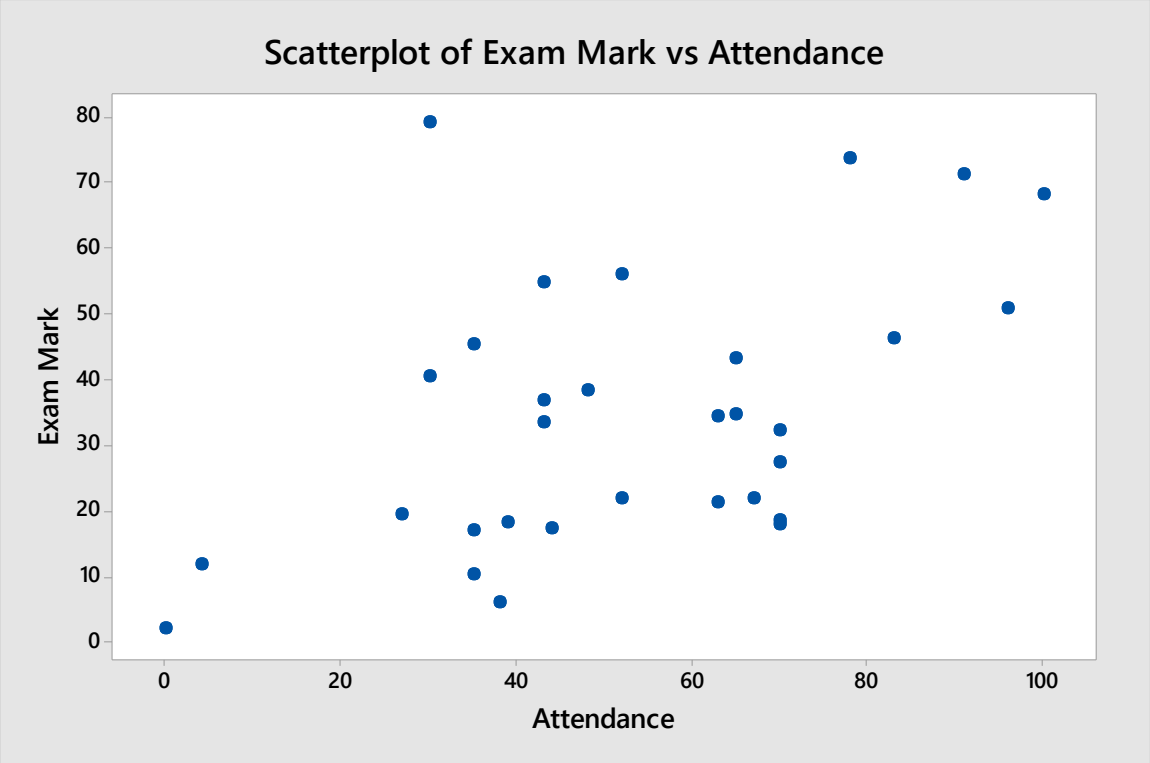


A Pareto chart is used to identify **which** problems should be addressed first. A Pareto chart consists of a bar chart, with each bar representing a certain defect category, and with the bars presented in order of decreasing height. The bars usually represent the number of times that each defect category occurred, but they can alternatively be used to represent the cost incurred by each defect category. A successful quality improvement project which targets the first category on the Pareto chart guarantees a bigger win than a project that targets any other category. A check sheet can be used to gather data before inputting it into a Pareto chart. The chart at Fig 8.14 suggests that 29.4% of motor breakdown service calls could be eliminated by addressing the battery issue, and 55.9% of service calls could be eliminated by addressing the battery and puncture issues.

#6. Scatter plot

A scatter plot can be used to explain **why** a process output variable takes certain values, by using a process input variable to explain its behaviour. A number of values of the input and output variables are observed, and these data pairs are plotted, with the input variable on the horizontal axis and the output variable on the vertical axis.

Fig 8.15



The plot indicates whether a relationship exists between the two variables. The plot will also suggest what value of the input variable will lead to the most desirable value of the output variable, although this assumes that there is a causative relationship between the variables. The scatter plot shown illustrates the relationship between attendance and examination mark for a number of students on a college course. The plot shows that a relationship does exist, and that higher examination marks tend to arise for students with higher attendance. Scatter plots are studied in detail in chapter 6 as part of a comprehensive treatment of regression analysis.

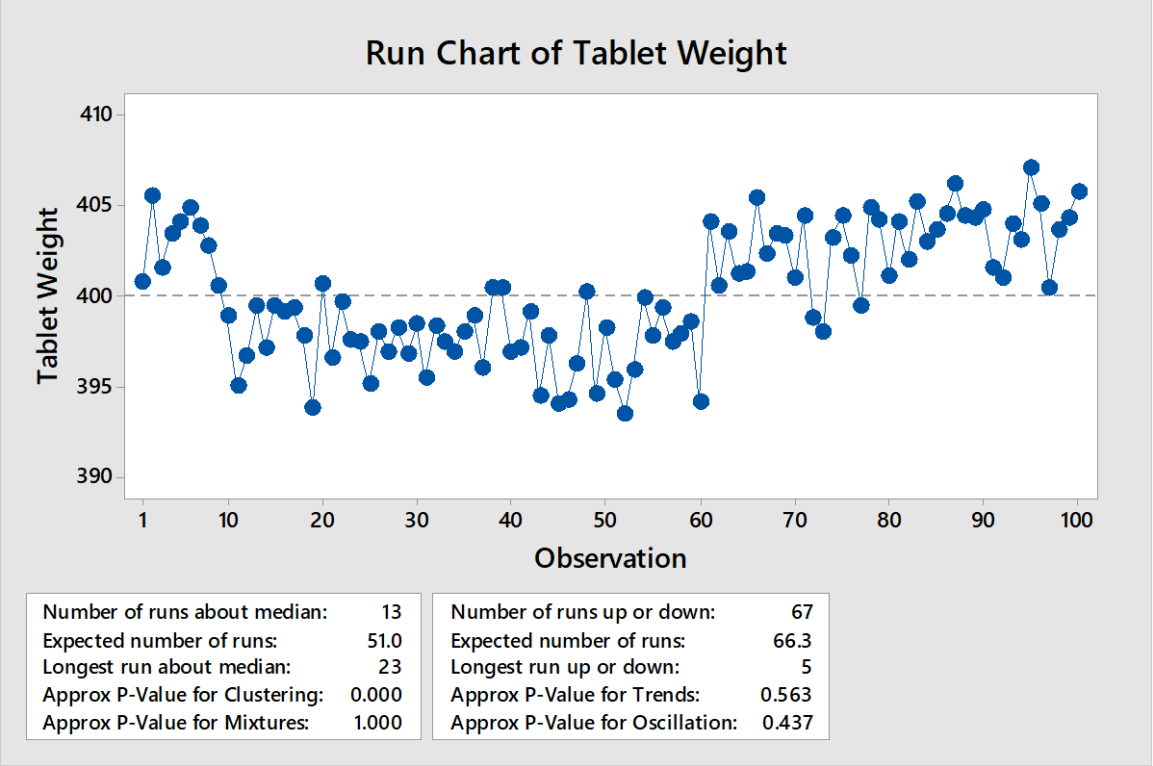
#7. Run chart

A run chart is used to identify **what kind** of variation occurs in a process. It consists of a time series plot, with the median value also shown. A random plot suggests that no special causes of variation are present. Typical patterns consist of runs of points on one side of the median, or runs of points moving up or down.

Clusters: Long runs on one side of the median suggest that there is variability due to the way the process is set-up at different times.
Mixtures: A scarcity of points near the median suggests that data have been combined from two sources.

Trends: Long runs of points slowly going up, or slowly going down, suggest that there is a progressive underlying change such as wear, and give a warning that the problem may get worse if not addressed.
Oscillation: Points that flip-flop up and down quickly suggest that successive observations are dependent, with every unit somehow compensating for the previous unit. Perhaps there is an over-zealous operator who is adjusting the process after each unit: this is called **chasing the process** and it makes the process worse, not better, because it is based on responding to random variation.

Fig 8.16



This run chart shows evidence of clustering. Tablets are coming from one source for a while, and then another. These sources could be different tablet presses or different batches.

Problems 8E

#1. Use a fishbone diagram to identify possible causes of each of these problems:

(a) parcels delivered late

(b) dents occurring on cars parked at a shopping mall.

#2. Design a check sheet that could be used to explore each of these problems:

(a) the location of scratches on the lenses of spectacles

(b) which questions are answered incorrectly by people filling out an online form?

#3. A provider of home TV and broadband services offers a telephone helpline to customers for resolving technical issues. The times taken by a particular helpline operator to deal with successive calls are plotted on a run chart. Suggest one reason in each case why this run chart might exhibit each of the following patterns:

(a) clusters

(b) mixtures

(c) trends

(d) oscillation.

#4. Which tool would you use in each of the following investigations?

(a) where to place protective covers on hire cars to prevent stone damage

(b) how recent results affect the performance of a football team

(c) where to start improving a restaurant that has received many complaints.

Project 8E #1

The Seven Basic Quality Tools

Use your initiative to identify seven opportunities in your workplace, or on your campus, to apply the seven basic quality tools. In each case, identify the process concerned, the problem to be addressed, the measurements that should be observed, and the quality improvement that would arise from using this tool.

Do not include graphs in your answer. Be creative and remember that these tools can be applied to manufacturing or service processes, to administrative or accounting processes, to recruitment or auditing, or in the canteen or car park.

Your report should consist of the following sections.

(a) Fishbone diagram

(b) Check sheet

(c) Control chart

(d) Histogram

(e) Pareto chart

(f) Scatter plot

(g) Run chart

Project 8E #2

Advanced Statistics Project

Design and carry out a suitable advanced statistics project that will allow you to demonstrate competency in the areas that you have studied. Collect the measurements, analyse the results and write a report consisting of the following sections.

(a) Describe the study design, including the choice of factors and covariates, the choice of factor levels, and the use of randomisation and blocking.

(b) Explain how the responses were chosen for observation. Explain how the measurement system was validated and how missing data would be handled.

(c) Show the data and the software analysis. Explain how the model was constructed. Analyse the residuals.

(d) Outline the assertions and assumptions that are supported by the analysis. Discuss any alternative explanations for the data, and any additional insights that have arisen from the residual analysis or otherwise. Discuss the implications of sample size.

(e) Present your findings in non-technical language for an audience that includes colleagues, clients and the general public. Include a description of the knowledge gained, the action recommended, and the further research ideas generated. Be prepared to respond to any questions with answers that are clear, coherent, correct, complete, creative and convincing.

Appendix 1: Computer Labs with Minitab®

Read this introduction first

Each of these Labs corresponds to a particular section in the book, e.g. Lab 1A corresponds to Section 1A in chapter 1. But before you begin you need to put some things in place by completing the following three steps.

Step 1: Download all the Data-sets
Go to https://zenodo.org/record/6778595#.Yrw-eHbMKUl and download the ***Applied Statistics data-sets***. Then right-click on the ***Applied Statistics data-sets*** file in your downloads folder and **Extract All** the files to a folder of your choice on your computer for future use. Take note of this folder name and location.

Step 2: Check that you have Minitab® software
If you are using a computer at college or at work it may already have Minitab installed. If you are using your own computer the best option is to get the Minitab software and licence from your college or employer. If that is not possible you can get a free 30-day trial of Minitab at http://www.minitab.com

Step 3: Open a Saved Worksheet
Launch Minitab on your computer. Now open the saved worksheet **1a Heights of Irishmen** from the Minitab main menu by selecting **File > Open…** and browsing to the folder on your computer where you saved the ***Applied Statistics data-sets***. Choose the sub-folder **Data Sets > Chapter 1** and then select the File name **1a Heights of Irishmen**, click **Open** and click **OK**. (Earlier versions of Minitab require you to select **File > Open Worksheet…** and then specify the file type, i.e. ***Excel***). Remember how to open a worksheet because you will need to do it again many times.

Lab 1A

#1. Draw a Histogram
With the worksheet **1a Heights of Irishmen** open in Minitab, from the main menu choose **Graph > Histogram…** and select **Simple**. Click **OK**. Double-click on the relevant variable name, ***Height*** in this case, in the left pane of the dialog box (or single-click and click **Select)**. Click **OK**. The graph can be viewed in the output window.

#2. Draw a Time Series Plot
Open the data set **1f Bakery Revenue**. From the main menu choose **Graph > Time Series Plot…** Select **Simple**. Click **OK**. Double-click on the variable name, ***Bakery Revenue*** in this case. Click on **Time/Scale…** to open the sub-dialog box and select the **Index** radio button. Click **OK**. Click **OK** again.

#3. Practice
Draw more histograms using the data sets in the worksheets ***1b, 1c, 1d, 1e, 1k, 1l***. There is a shortcut button called **Edit Last Dialog** which is useful when you want to repeat the same steps.
Draw more time series plots using the data sets in the worksheets ***1g, 1h, 1i, 1j, 1m***.

Lab 1B

#1. Select Numbers at Random
To select numbers at random from a list of integers, click on **Calc > Random Data > Integer...** Suppose you want to select 10 ID numbers at random from a list of 1000 ID numbers from 1 to 1000. Then type ***10, Sample, 1*** and ***1000*** in the four fields and click **OK.**

#2. Draw a Multi-Stage Sample
Suppose you want to select 20 items at random from a shopping website which has 153 pages with 30 items on each page. Generate 20 random page numbers by clicking **Calc > Random Data > Integer...** and then typing ***20, Page, 1*** and ***153*** in the four fields and clicking **OK.** Now generate 20 random item numbers by clicking **Calc > Random Data > Integer...** and then typing ***20, Item, 1*** and ***30*** in the four fields and clicking **OK.** Each row of the worksheet now specifies a particular item on a particular page. The same approach can be used to select players at random from a league by randomly selecting team numbers and squad numbers.

#3. Sort IDs into Random Order
First create a list of as many ID numbers as you need, say 1000, by clicking on **Calc > Make Patterned Data > Simple Set of Numbers.** Type ***ID, 1,1000*** into the first three fields and click **OK**. Now assign a random number to each ID by clicking **Calc > Random Data > Uniform...** and typing ***1000*** and ***Random*** in the first two fields and clicking **OK**. Finally, sort the whole list into random order by clicking **Data > Sort…** and selecting ***Random*** as the **Column** at **Level 1** and clicking **OK**.

Lab 1C

#1. Calculate Summary Statistics
Open the data set **1n Heights of Trees**. From the menu choose **Stat > Basic Statistics > Display Descriptive Statistics…** Click on the variable name (***Height*** in this case) and click **Select**. Click **OK**. The results are displayed in the output window. You can choose which statistics will be displayed by opening the **Statistics…** sub-dialog box before clicking **OK**. Simply select or de-select from the list and then click **OK** and **OK** again.

#2. Practice
Use software to solve ***Problems 1C, #1 #2*** and ***#3***. The data are saved in worksheets ***1o, 1p, 1q*** and ***1r***.

Lab 1D

#1. Calculate Internal Consistency
Open the data set **1s Hotel Satisfaction.** From the menu choose **Stat > Multivariate > Item Analysis...** and in the variables field select all of the variables ***Enjoy-Speakers*** and click **OK**. The output window shows the value of Cronbach's Alpha and also shows that its value would greatly improve if the *Speakers* item was omitted.

Lab 2B

#1. Calculate Factorials
To calculate 3! click on **Calc > Calculator…** In the field **Store result in variable** type any name you like such as ***Letters***. Scroll down through the **Functions** until **Factorial** is highlighted and click **Select**. Type the number ***3*** to replace the blue highlighted **number of items** and click **OK**. The result appears in the output window in a column called **Letters**.

#2. Calculate Permutations
To calculate ${}^{7}P_{3}$ click on **Calc > Calculator…** In the field **Store result in variable** type any name you like such as ***Horses***. Scroll down through the **Functions** until **Permutations** is highlighted and click **Select**. Replace the numbers in brackets with (***7,3***) and click **OK**. The result appears in the output window in a column called **Horses**.

#3. Calculate Combinations
To calculate ${}^{47}C_{6}$ click on **Calc > Calculator…** In the field **Store result in variable** type any name you like such as ***Lottery***. Scroll down through the **Functions** until **Combinations** is highlighted and click **Select**. Replace the numbers in brackets with (***47,6***) and click **OK**. The result appears in the output window in a column called **Lottery**.

Lab 3B

#1. Calculate Normal Probabilities
Choose **Calc > Probability Distributions > Normal.** Select the radio button for **Cumulative Probability.** For **Mean** insert ***175*** and for **Standard deviation** insert ***10***. Select the radio button for **Input constant** and insert ***182.5***, and click **OK**. The answer given in the output window is 0.7734 but the answer represents the probability that a man is shorter than 182.5 so the probability that a man is taller than 182.5 is found by subtracting this answer from 1, giving 0.2266.

Lab 3D

#1. Calculate Binomial Probabilities
Choose **Calc > Probability Distributions > Binomial.** Select the radio button for **Probability.** For **Number of Trials** insert ***6***. For **Event Probability** insert ***0.10***. Select the radio button for **Input constant** and insert ***2***, and click **OK**. The answer, 0.098415, is given in the output window.

#2. Calculate Poisson Probabilities
Choose **Calc > Probability Distributions > Poisson.** Select the radio button for **Cumulative Probability.** For **Mean** insert ***3*** and select the radio button for **Input constant** and insert ***2***, and click **OK**. The answer given in the output window is 0.423190 and represents the probability of receiving two or fewer complaints. The probability of receiving more than two complaints is found by subtracting this answer from 1.

Lab 3E

#1. Draw a Normal Probability Plot
Open the worksheet **3h Heights of Irishmen**. Click on **Graph > Probability Plot…** Select **Single** and click **OK**. Click on the required variable in the pane on the left side of the dialog-box (***Heights of Irishmen*** in this case) and click **OK**. Notice that the plot is quite linear. A *p*-value below 0.05 would indicate that the data do not fit the selected distribution.

#2. Draw a Non-normal Probability Plot
Open the worksheet **3i Taxi Waiting Times.** Click on **Graph > Probability Plot…** Select **Single** and click **OK**. Click on the required variable in the pane on the left side of the dialog-box (***Taxi Waiting Times*** in this case) and click **OK**. Notice that the plot is not linear.
Now repeat the first step above by clicking on **Graph > Probability Plot…** and selecting **Single** and clicking **OK**, but this time click on **Distribution…** and select ***Lognormal*** instead of ***Normal*** and click **OK** and **OK** again. This distribution seems to be a good fit.
Rather than trying out different distributions one at a time, you can use **Stat > Quality Tools > Individual Distribution Identification** to check out up to 16 distributions at once. Lower values of the Anderson Darling statistic (AD) indicate a better fit. Low *p*-values indicate a poor fit.

#3. Practice
Repeat this analysis using the data in worksheet ***3j***.

Lab 4B

#1. Estimate a Population Mean (sigma known)
Open the worksheet **4b Cycle Times**. Choose **Stat > Basic Statistics > 1-Sample Z.** Place the cursor in the second field (i.e. the first field that is blank) and then select the variable ***Cycle Times***. Insert ***2*** in the **Known standard deviation** field and click **OK**.

#2. Estimate a Population Mean (sigma unknown)
Open the worksheet **4c Flight Times**. Choose **Stat > Basic Statistics > 1-Sample t.** Place the cursor in the second field (i.e. the first field that is blank) and then select the variable ***Flight Times***. Click **OK**.

#3. Estimate a Population Proportion
Choose **Stat > Basic Statistics > 1 Proportion…** From the drop-down menu select ***Summarised data***. For **number of events** insert ***16***, and for **number of trials** insert ***50***. Select **Options…** and from the drop-down **Method** menu choose ***Normal approximation***. Click **OK** twice. The answer corresponds to the answer that would be obtained by hand-calculation. A better answer is obtained by leaving **Method** at the default ***Exact*** setting.

#4. Practice
Use software to solve ***Problems 4B***.

Lab 4C

#1. Find a Sample Size for Estimation
Click **Stat > Power and Sample Size > Sample size for estimation.** From the drop down menu, choose the relevant parameter that you want to estimate. Enter the relevant information and click **OK.**

#2. Practice
Use software to solve ***Problems 4C***.

Lab 4D

#1. Estimate a Standard Deviation
Open the worksheet **4k Loaves**. Choose **Stat > Basic Statistics > Graphical summary. Select *Loaves*** and click **OK**.

#2. Practice
Use software to solve ***Problems 4D #2***.

Lab 4E

#1. Compare Means Using Independent Samples
Open the worksheet **4o Apartment Rentals**. Choose **Stat > Basic Statistics > 2-Sample t.** From the drop-down menu select **Each sample is in its own column,** and for **Sample 1** and **Sample 2** select ***Lucan*** and ***Naas*.** Click on **Options…** and in the sub-dialog box check the box for **Assume Equal variances** and click **OK** and **OK** again.

#2. Compare Proportions
Choose **Stat > Basic Statistics > 2 Proportions**
From the drop-down menu select **Summarised data**. For **Sample 1** enter ***98*** events and ***200*** trials. For **Sample 2** enter ***34*** events and ***300*** trials. Click **OK**.

#3. Practice
Use software to solve ***Problems 4E***.

Lab 5B

#1. Test a Mean (sigma known)
Open the worksheet **5a Peanuts.** Choose **Stat > Basic Statistics > 1-Sample Z.** Place the cursor in the second field (i.e. the first field that is blank) and then select the variable ***Bag Weights***. Enter ***3*** for the **Known standard deviation.** Check the box for **Perform hypothesis test** and enter ***25*** as the **Hypothesised mean**. Open the **Options…** sub-dialog box and change the **Alternative hypothesis** to **Mean < Hypothesised mean**. Click **OK** and **OK**.

#2. Test a Mean (sigma unknown)
Open the worksheet **5a Peanuts.** Choose **Stat > Basic Statistics > 1-Sample t.** Place the cursor in the second field (i.e. the first field that is blank) and then select the variable ***Bag Weights***. Check the box for **Perform hypothesis test** and enter ***25*** as

the **Hypothesised mean**. Open the **Options…** sub-dialog box and change the **Alternative hypothesis** to **Mean < Hypothesised mean**. Click **OK** and **OK**.

#3. Test a Proportion
Choose **Stat > Basic Statistics > 1 Proportion…** From the drop-down menu select **Summarised data**. For **number of events** insert ***30***, and for **number of trials** insert ***315***. Check the box for **Perform hypothesis test** and enter ***0.08*** as the **Hypothesised proportion**. Open the **Options…** sub-dialog box and change the **Alternative hypothesis** to **Proportion > Hypothesised proportion**. For **Method** select **Exact** for the best result, but if you want to mimic calculations performed by hand then select **Normal approximation.** Click **OK** and **OK**.

#4. Practice
Use software to solve ***Problems 5B***.

Lab 5C

#1. Test Means using Paired Samples
Open the worksheet **5e Vitamin C in Peppers**. Choose **Stat > Basic Statistics > Paired t...** For **Sample 1** and **Sample 2** select ***Fridge*** and ***Room Temperature.*** Click **OK**.

#2. Test Means using Independent Samples
Open the worksheet **5f Groundwater**. Choose **Stat > Basic Statistics > 2-Sample t.** From the drop-down menu select **Each sample is in its own column,** and for **Sample 1** and **Sample 2** select ***Influent*** and ***Effluent***. Open the **Options…** sub-dialog box and change the **Alternative hypothesis** to **Difference > hypothesised difference** and check the box for **Assume Equal variances.** Click **OK** and **OK** again.

#3. Practice
Use software to solve ***Problems 5C.***

Lab 5D

#1. Analyse a 2×2 Contingency Table
Open the worksheet **5j Smartphones.** Choose **Stat > Tables > Chi-square test for Association…** From the drop-down menu select **Summarised data in a two way table.** In the next field select the columns ***iPhone*** and ***Android***. As the **Labels for the table**, in the **Rows** field select ***Gender***, and leave the **Columns** field empty. Click on the **Statistics…** button and in the sub-dialog box check the final tick-box **Each cell's contribution to chi-square** and do not uncheck any boxes. Click **OK** and **OK** again.

#2. Analyse a 2×2 Contingency Table using raw data
Open the worksheet **5jj Smartphones - raw data.** Choose **Stat > Tables > Chi-square test for Association…** From the drop-down menu select **Raw data (categorical variables).** For **Rows** select ***Genders*** and for ***Columns*** select **Smartphones**. Click on the **Statistics…** button and in the sub-dialog box check the final tick-box **Each cell's contribution to chi-square** and do not uncheck any boxes. Click **OK** and **OK** again

#3. Test for Association
Open the worksheet **5k Titanic Survival.** Choose **Stat > Tables > Chi-square test for Association…** From the drop-down menu select **Summarised data in a two way table.** In the next field select the columns ***Saved*** and ***Lost***. As the **Labels for the table**, in the **Rows** field select ***Class***, and leave the **Columns** field empty. Click on the **Statistics…** button and in the sub-dialog box check the final tick-box **Each cell's contribution to chi-square** and do not uncheck any boxes. Click **OK** and **OK** again.

#4. Practice
Use software to solve ***Problems 5D***. The data are saved in worksheets ***5l, 5m, 5n, 5o*** and ***5p***.

Lab 5E

#1. Determine the Sample Size for Testing
Choose **Stat > Power and Sample Size** and then select the analysis that is envisaged, e.g. choose **1-Sample t...** if you plan to test a claim about the mean weight of bags of peanuts. Leave the **Sample sizes** field blank. For **Differences** enter the minimum difference that is of practical importance, e.g. -***2***. For **Power values** insert ***0.80*** if looking for a difference. Provide a value for the **Standard deviation** based on a pilot sample, e.g. ***2.28***. Open the **Options…** sub-dialog box and make any necessary changes, e.g. select ***Less than***. Click **OK.** To avoid graphical output, open the **Graphs…** sub-dialog box and un-check the box. Click **OK** and **OK** again.

Lab 5F

#1. Test a Variance
Open the worksheet **5q Bagel Diameters**. Choose **Stat > Basic Statistics > 1 variance**. From the drop-down menu select **One or more samples each in a column** and in the next field select ***Bagel Diameters***. Check the box for **Perform hypothesis test** and enter ***2*** as the **Hypothesised standard deviation**. Open the **Options…** sub-dialog box and change the **Alternative hypothesis** to **Standard deviation > hypothesised standard deviation**. Click **OK** and **OK** again.

#2. Test the Difference Between Two Variances
Open the worksheet **5r Paint Thickness**. Choose **Stat > ANOVA > Test for Equal Variances**. From the drop-down menu select **Response data are in one column for all factor levels**. In the **Response** field select ***Thickness*** and in the **Factors** field select ***Mode***. Click on **Options...** and check the box for **Use test based on normal distribution.** Click **OK** and **OK** again.

#3. Test a Goodness-of-Fit
Open the worksheet **5s Fair Die.** Choose **Stat > Tables > Chi-Square Goodness-of-Fit Test (One Variable)…** For **Observed Counts** enter ***Frequency*** and for **Category Names** enter ***Outcome***. To avoid graphical output, open the **Graphs…** sub-dialog box and un-check the boxes. Click **OK** and **OK** again.

#4. Practice
Use software to solve ***Problems 5F***. The data are saved in worksheets ***5t, 5u, 5v, 5w, 5x*** and ***5y***.

Lab 5G

#1. Draw a Survival Plot
Open the data set **5zc Survival Data**. Choose **Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Nonparametric Distribution Analysis…** In the variables field select ***Treatment Placebo***. Click on **Censor…** and in the sub-dialog box select the radio button for **Time Censor at:** and enter ***10*.** Click **OK**. Click on **Graphs…** and in the sub-dialog box check the option for **Survival Plot**. Click **OK**. Click **OK** again.

#2. Analyse Clinical Trial Data
Use software to repeat the analyses that are shown in the text in section ***5G Clinical Trials***. The data are saved in worksheets ***5za*** and ***5zb***.

#3. Practice
Use software to solve ***Problems 5G***. The data are saved in worksheets ***5zd*** and ***5ze***.

Lab 6A

#1. Draw a Scatter Plot
Open the worksheet **6a ShoeSize and Height**. Click on **Graph > Scatterplot** and select **Simple** and click **OK.** For the **Y Variable** in row 1 select ***Height*** and for the **X variable** select ***ShoeSize***. Click **OK**.

#2. Find a Correlation Coefficient
Open the worksheet **6a ShoeSize and Height**. Click on **Stat > Basic Statistics > Correlation…** and in the **Variables** field select ***Height ShoeSize***. Click **OK**.

#3. Practice
Draw a scatter plot and find the correlation coefficient for each of the data-sets in worksheets ***6b*** and ***6c***.

Lab 6B

#1 Draw a Fitted Line Plot
Open the worksheet **6a ShoeSize and Height**. Select **Stat > Regression > Fitted Line Plot**. Select the appropriate variable for **Response** (***Height*** in this case) and **Predictor** (***ShoeSize*** in this case) and click **OK**.

#2. Practice
Draw a fitted line plot for each of the data-sets in worksheets ***6b, 6d, 6e*** and ***6f***. Then solve ***Problems 6B***.

Lab 6C

#1. Carry out a Regression Analysis
Open the worksheet **6a ShoeSize and Height**. Click on **Stat > Regression > Regression > Fit Regression Model**. Select the appropriate variables for **Responses** (***Height*** in this case) and **Continuous Predictors** (***ShoeSize*** in this case) and click **OK.**

#2. Make a Prediction
Having already carried out the regression analysis, now click on **Stat > Regression > Regression > Predict**. Type the new value of the predictor (***5.5*** in this case) underneath the name of the predictor (**ShoeSize** in this case) and click **OK.**

#3. Carry out a Residual Analysis
Open the worksheet **6d Attendance and Exam Mark**. Click on **Stat > Regression > Regression > Fit Regression Model**. Select the appropriate variables for **Responses** (***Exam Mark*** in this case) and **Continuous Predictors** (***Attendance*** in this case). Click on the **Graphs...** sub-dialog box and check the boxes for **Histogram of residuals** and **Residuals versus fits** and click **OK** and **OK** again.

#4. Practice
Carry out a regression analysis for each of the data-sets in worksheets ***6b, 6d, 6e*** and ***6f***. Make a prediction in each case using a new value of the predictor. Then solve ***Problems 6C***.

Lab 6D

#1. Carry out a Multiple Regression Analysis
Open the worksheet **6g Hurling.** Click on **Stat > Regression > Regression > Fit Regression Model**. Select the appropriate **Response** (***VO2Max*** in this case) and the **Continuous Predictors** (***Age Weight*** in this case) and click **OK.**

#2. Select a 'Best Subset' of Variables
Open the worksheet **6g Hurling.** Click on **Stat > Regression > Regression > Best Subsets**. Select the appropriate **Response** (***VO2Max*** in this case) and list all the variables that may be useful as **Free Predictors** (***Age-Twenty*** in this case**).** Click **OK**.

#3. Practice
Select a 'Best Subset' of variables for ***Price*** using the data in worksheet ***6j*****.** Select a 'Best Subset' of variables for ***Final Mark*** from among the other two sets of marks provided in worksheet ***6n*****.**

#4. Carry out a Quadratic Regression
Open the worksheet **6h Dose-Response Curve**. Select **Stat > Regression > Fitted Line Plot**. Select the appropriate variable for **Response** (***Response*** in this case) and **Predictor** (***Dose*** in this case). For **Type of Regression Model**, click the **Quadratic** radio button and click **OK**.

#5. Carry out a Regression using a Transformation
Open the worksheet **6h Dose-Response Curve**. Use **Calc > Calculator** to transform either the predictor or response variable. Then construct a simple linear regression model using the transformed variable. For example choose **Calc > Calculator** and in the **Store result in variable** field type ***RootDose***. Enter ***SQRT('Dose')*** in the Expression field and click **OK**. Now select **Stat > Regression > Fitted Line Plot**. Select the appropriate variable for **Response** (***Response*** in this case) and **Predictor** (***RootDose*** in this case). For **Type of Regression Model**, click the **Linear** radio button and click **OK**. Now try another transformation by choosing **Calc > Calculator** and in the **Store result in variable** field type ***LogDose***. Enter ***LOGTEN('Dose')*** in the Expression field and click **OK**. Now select **Stat > Regression > Fitted Line Plot**.

Select the appropriate variable for **Response** (***Response*** in this case) and **Predictor** (***LogDose*** in this second case). For **Type of Regression Model**, click the **Linear** radio button and click **OK**. Which model is better? Use the value of *R-Sq(adj)* to help you decide.

#6. Practice
Using the data in worksheet ***6k Shrinkage*****,** find a model for predicting ***Shrinkage*** using either quadratic regression or a suitable transformation.

Lab 6E

#1. Perform a Binary Logistic Regression
Open the worksheet **6l Credit Scorecard**. Choose **Stat > Regression > Binary Logistic Regression > Fit Binary Logistic Model**. In **Response** enter ***Default***. In **Continuous Predictors** enter ***MonthsAtAddress MonthsInJob***. Click **OK**. Now, to make a prediction, choose **Stat > Regression > Binary Logistic Regression > Predict.** The **Response** is **Default**. Enter ***29*** and ***18*** for **MonthsAtAddress** and **MonthsInJob** respectively. Click **OK**.

#2. Perform a Partial Least Squares Regression
Open the worksheet **6m Energy Expenditure**. Choose **Stat > Regression > Partial Least Squares**. In Responses, enter ***EnergyExpenditure***. In Model, enter ***Age-Steps***. Click **Options…** and choose **Leave-one-out**. Click **OK**. Click **Graphs**… and uncheck all boxes. Click **OK**. Click **OK** again.

Lab 6F

#1. Perform a Principal Components Analysis
Open the worksheet **6g Hurling**. Choose **Stat > Multivariate > Principal Components…**. In **Variables**, enter ***Age- VO2Max***. Click on **Graphs…** and check the box for **Loading Plot for first 2 components**. Click **OK**. Click **OK** again.

#2. Perform a Discriminant Analysis
Open the worksheet **6g Hurling**. Remove the final row of data for validation later on. Choose **Stat > Multivariate > Discriminant Analysis**. In **Groups**, enter ***Position***. In Predictors, enter ***Age-VO2Max***. Now click on **Options**... and in the field **Predict group membership for:** enter the final row values that were removed. Click **OK** and **OK**.

#2. Perform a Cluster Analysis
Open the worksheet **6l Credit Scorecard**. Choose **Stat > Multivariate > Cluster Variables...** In the **Variables or distance matrix** field enter ***Income-Convictions***. Check the box for **Show dendrogram** and click **OK**.

Lab 7A

#1. Create a Single-factor Experimental Design
The first step is to list the Factor Levels: select **File > New > Worksheet > OK.** Select **Calc > Make Patterned Data > Text Values…** In the first field, type the name of the factor, e.g. ***Design***. In the second field, type the names of the factor levels, separated

by spaces, e.g. ***Red Blue Green***. In the third and fourth fields, type the number of repetitions you require, e.g. ***1*** and ***4***. Click **OK**.
The second step is to randomise the run order: select **Calc > Random Data > Uniform…** and specify the **Number of rows of data to generate**, e.g. ***12***. In the second field, type a name for this column, e.g. ***Random*** and click **OK**. Now select **Data > Rank…** and in the first field select ***Random,*** and type ***RunOrder*** in the second field. Click **OK**.
The third step is to prepare a column for the response data: go to the first blank column of the worksheet and, in the title cell, type the name of the response, e.g. ***Distance***.
You are now ready to perform the experiment. Carry out the runs in the order specified in the **RunOrder** column, and record the response each time in the response column. When you are finished, analyse the data using One-Way ANOVA.

#2. Carry Out a One-Way ANOVA
Open the worksheet **7a Distance Kicked.** Select **Stat > ANOVA > One-Way…** In the dialog box, for **Response** enter ***Distance*** and for **Factor** enter ***Player*** and then click **OK**. The ANOVA table is displayed in the output window.

#3. Practice
Repeat this analysis using the data in worksheets ***7b*** and ***7c***.

Lab 7B

#1. Create a Two-factor Experimental Design
Select **Stat > DOE > Factorial > Create Factorial Design…** Select the last radio button, i.e. **General Full Factorial Design**. Open the **Designs…** sub-dialog box, and specify the **Name** of each Factor, the **Number of Levels** and the **Number of replicates**, e.g. ***Person*** and ***Hand*** for the Names, ***2*** and ***2*** for the levels and ***3*** for the replicates. Click **OK** once. Open the **Factors…** sub-dialog box, and change the **Type** of each factor from ***numeric*** to ***text*** in this case. Change the **Level Values** of **Person** to read ***Wallace*** and ***Gromit*** (use your actual names here), and the **Level Values** of **Hand** to read ***Left*** and ***Right***. Click **OK** and **OK** again. The experimental design is automatically created, and randomised, and displayed in a new worksheet. Enter a name for the response, e.g. ***Time***, at the head of the first blank column. The experiment can now be carried out, and the data analysed by a balanced ANOVA.

#2. Carry Out a Balanced ANOVA
Open the worksheet **7e Reading Time.** Select **Stat > ANOVA > Balanced ANOVA…** In the dialog box, for **Responses** select ***Time*** and enter all the terms you want in the **Model** field, e.g. ***Person Language Person*Language***. Click **OK**. The ANOVA table is displayed in the output window. (Note: '*Person ! Language'* is a shorthand for '*Person Language Person*Language'.)*

#3. Draw an Interaction Plot
Open the worksheet **7e Reading Time.** Select **Stat > ANOVA > Interaction Plot…** For **Responses** enter ***Time*** and for **Factors** enter ***Person Language***. Click **OK**. (It is best to select the factors in logical order, e.g. *person* and then *language* rather than the other way around. The lines on the graph will show what happens with each person when the language changes. For a simpler graph, do not check the box for a full interaction plot matrix.)

#4. Create a Main Effects Plot
If there is no significant interaction but there are one or more significant main effects then a main effects plot is preferable to an interaction plot. Open the worksheet **7h Kayaking.** Use ANOVA to confirm that there is no significant interaction but there is a significant main effect. Now select **Stat > ANOVA > Main Effects Plot…** For **Responses** select ***Time*** and for **Factors** select ***Paddle***. Click **OK**.

#5. Practice
Repeat these analyses using the data in the worksheets ***7f, 7g, 7h*** and ***7i***.

#6. Find the Sample Size Required for an Experiment
Select **Stat > Power and Sample Size** and then select the analysis that is envisaged, e.g. **One-Way ANOVA.** Suppose you are planning a second experiment to investigate the effect of player on distance kicked, as at the beginning of chapter 7. For Number of levels insert ***3***. Leave the **Sample sizes** field blank. For **Values of the maximum difference between means** insert a value that is of practical importance, e.g. ***5***. For **Power values** insert ***0.80*** if looking for a difference. For **Standard deviation** insert ***2.587***, i.e. the pooled standard deviation value that arose in the earlier, pilot experiment. To avoid graphical output, open the **Graph…** sub-dialog box and un-check the box. Click **OK** and **OK** again. The output calls for 7 replicates for each player. Now, by repeating these steps and changing the power value to ***0.95***, you can identify the sample size required to prove equivalence, which turns out to be 10 replicates for each player. Alternatively, if the sample size has already been decided it can be entered in the **Sample sizes** field, and the power value can be calculated by leaving the **Power values** field blank, e.g. a sample size of 4 is associated with a power value of 0.5299. To find the sample size for experiments with more than one factor, use **Stat > Power and Sample Size** and then choose either **General Full Factorial Design** or **2-Level Factorial Design** as appropriate.

#7. Practice
Repeat the sample size analysis using the data in worksheets ***7b*** and ***7c***.

Lab 7C

#1. Create a Multi-factor Experimental Design
Select **Stat > DOE > Factorial > Create Factorial Design…** Select the 1st radio button, i.e. ***2-level factorial***. Specify the **Number of factors** in the drop-down list, e.g. ***6***. Open the **Designs…** sub-dialog box, and specify the **Design** *(e.g.* ***1/8 fraction****)*, the **Number of center points** (e.g. ***0***) the **Number of replicates** *(*e.g. ***2***) and the number of blocks (e.g. ***1***). Click **OK** once. Open the **Factors…** sub-dialog box, and specify the **Name** and **Type** of each Factor. Provide a name or number for the low and high level of each factor. Click **OK** and **OK** again. The design is automatically created, and randomised, and displayed in a new worksheet. Enter a name for the response at the head of the first blank column. You are now ready to carry out the experiment.

#2. Analyse Data from a Multi-factor Experiment
If the design was created in Minitab then the analysis can be performed very simply. After entering the response values, select **Stat > DOE > Factorial > Analyze Factorial Design…** Select the response. Click on **Options…** and change the confidence level from 95 to 99 and click **OK** and **OK** again.

Lab 7D

#1. Fit a General Linear Model
Open the worksheet **7n Hurling.** Choose **Stat > ANOVA > General Linear Model > Fit General Linear Model…** For **Responses** enter ***VO2Max***, for **Factors** insert ***Position***, and for **Covariates** insert ***Age***. Click on the **Model...** sub-dialog box and, in the pane at the top left, select all the factors and covariates to be included in an interaction term (holding the **Ctrl** button on your keyboard while clicking) and then click the button on the right of the dialog box to **Add** those interactions to the model. Click **OK.** (Note: You can remove **Terms in the model** by clicking on the red X symbol.)

#2. Fit a Nested Model
Open the worksheet **7m Laurel Leaves.** Choose **Stat > ANOVA > General Linear Model > Fit General Linear Model…** For **Responses** enter ***Length***, for **Factors** insert ***Tree Branch***. Now open the **Random/Nest…** sub-dialog box and indicate that ***Branch*** is nested in the specified factor ***Tree***, and that the **Type** of each factor is ***Random*** rather than ***Fixed***. Click **OK** once. In the **Results** sub-dialog box, see that the **Variance Components** box is checked. Click **OK** and **OK** again.

Lab 7E

#1. Carry out a Stability Study
Open the worksheet **7p Assay Product A.** Click on **Stat > Regression > Stability Study > Stability Study…** In the **Response** field select ***'Assay (%) Product A'***, in the **Time** field select ***Month***, in the **Batch** field select ***Batch***. In the **Lower Spec:** field enter ***90*** and click **OK**.

#2. Design a Stability Study
Click on **Stat > Regression > Stability Study > Create Stability Study Worksheet.**

#3. Practice
Carry out Stability Studies using datasets ***7q, 7r*** and ***7s*** (Products B, C and D).

Lab 7F

#1. Draw a Contour Plot
Open the Minitab worksheet **7t Plane**. Click on **Stat > DOE > Response Surface > Analyse Response Surface Design**. Click **Yes**. In **Continuous factors** select ***Angle Grade***. Click on the **Low/High…** sub-dialog box and click **OK**. Click **OK** again. For **Responses** insert ***Distance***. Click **OK**. Now choose **Stat > DOE > Response Surface > Contour Plot**. Click **OK**. Repeat the exercise to draw contour plots using the data in the worksheets ***7u Ridge*** and ***7v Peak***.

#2. Create a Box-Behnken Design
Choose **Stat > DOE > Response Surface > Create Response Surface Design.** Choose **Box-Behnken.** Choose a **Number of continuous factors,** e.g. ***3.*** Click on **Designs…** make any changes or none, and click **OK.** Choose **Factors…** and enter the names of the factors, e.g. ***Temperature, Time*** and ***Material***, for **A**, **B** and **C**, and their low and high levels, e.g. ***80*** & ***100***, ***2*** & ***4***, ***10*** & ***30***. Click **OK.** Click **OK** again. Now, to simulate the data collection, launch Excel and open the file **7w Seal Strength**

RSM TOOL 1. Paste the matrix of Temperature-Time-Material measurements from your experimental design into this worksheet and the experimental results will appear in the **Strength** column. Copy and paste the results back into your Minitab worksheet.

#3. Analyse a Box-Behnken Design
After entering the column of response values in the worksheet, choose **Stat > DOE > Response Surface > Analyze Response Surface Design**. Enter the **Responses** (***Strength*** in this case) and click **OK.** In this example you will notice that Temperature and Time are significant but Material is not. You can now draw a contour plot by clicking **Stat > DOE > Response Surface > Contour Plot.** Click on the contours sub-dialog box and click **Number** and enter ***11*** and then click **OK** and **OK.** Only two factors can be included in a contour plot: additional factors are held at chosen values.

#4. Carry out Multiple Response Optimisation
Create and store the design by following the steps at #2 above but this time use the Excel file **7x Welding Cost RSM TOOL 2** to simulate the data collection for the three responses ***Strength, Lifetime*** and ***Cost***. Fit a model to each response by clicking **Stat > DOE > Response Surface > Analyse Response Surface Design** and for **Responses** selecting ***Strength Lifetime Cost*** and clicking **OK**. Choose **Stat > DOE > Response Surface > Response Optimiser.** Select ***minimise*** for **Cost** and ***maximise*** for **Lifetime** and for **Strength**. Click **Setup…** and for each response, where relevant, specify the **Lower, Target** and **Upper, *0*** & ***10, 6*** & ***12, 480*** & ***520,*** for **Cost, Lifetime** and **Strength** respectively. Leave all the values of **Weight** and **Importance** at ***1*** for now. Click **OK** and **OK** again.

#5. Draw an Overlaid Contour Plot
Choose **Stat > DOE > Response Surface > Overlaid Contour Plot.** Select all available responses. Click on **Contours** and insert **Low** and **High** values, i.e. ***0*** & ***10, 6*** & ***12, 480*** & ***520,*** for **Cost, Lifetime** and **Strength** respectively. Click **OK.** Click **OK** again.

Lab 8A

#1. Carry out a Process Capability Analysis
Open the worksheet **8a Bottle Filling**. Select **Stat > Quality Tools > Capability Analysis > Normal.** In the **Single Column** field, select ***Volume***. In the **Subgroup Size** field, enter ***5***. For **Lower spec** and **Upper spec** enter ***115*** and ***125***. Click **OK**.

#2. Practice
Repeat this analysis using the data in worksheet ***8b Castle Tours***.

Lab 8B

#1. Draw an Individuals Chart
Open the worksheet **8c Drive-through.** Select **Stat > Control charts > Variables chart for Individuals > Individuals**. In the **Variables** field, select ***Time.*** Click **OK**.

#2. Draw an XBar Chart
Open the worksheet **8a Bottle Filling.** Select **Stat > Control charts > Variables chart for Subgroups > Xbar.** Place the cursor in the second field (i.e. the first blank field) and then select the variable ***Volume***, and for **subgroup sizes** enter ***5***. Click **OK**.

#3. Draw an NP chart
Open the worksheet 8d Late Deliveries**. Select** Stat > Control charts > Attributes charts > NP. **In the** Variables **field, select** *Late Deliveries***, and for** subgroup sizes **enter** *50***. Click** OK.

#4. Draw a C chart
Open the worksheet **8e Damaged Bags.** Select **Stat > Control charts > Attributes charts > C.** In the **Variables** field, select ***Damaged bags***. Click **OK.**

Lab 8C

#1. Create a Sampling Plan
Select **Stat > Quality Tools > Acceptance Sampling by Attributes.** Enter the **AQL, LTPD**, **Alpha** and **Beta** and click **OK.**

Lab 8D

#1. Carry out a Gage Linearity and Bias Study
Open the worksheet **8i Kitchen Scale.** To find the linearity, select **Stat > Regression > Fitted Line Plot**. For **Response**, select ***'Average Deviation'*** and for **Predictor** select ***Reference*** and click **OK**. The slope, 0.0673, multiplied by 100, gives the % linearity, +6.73%. To find the bias, select **Stat > Basic Statistics > Display Descriptive Statistics** and for **Variables** select ***Deviation***. The mean in the output, 83.5, is the average bias.

#2. Carry out a Gage R&R Study
Open the worksheet **8j Timber Beams.** Select **Stat > Quality Tools > Gage Study > Gage R&R Study (Crossed)**. Select ***Part, Session*** and ***Measurement*** in the three fields of the dialog box. Open the **Options...** sub-dialog box and enter *20* for ***lower spec*** and *100* for ***upper spec***. Click **OK** and **OK** again. The %Tolerance is displayed in the output window, and the SNR can be calculated by dividing the SD for Part-To-Part by the SD for Total Gage R&R.

#3. Practice
Repeat #1 using the worksheet ***8k Map Distances*** and #2 using ***8l Vision System***.

Lab 8E

#1. Draw a Fishbone diagram
Open the worksheet **8m Burnt Toast.** Select **Stat > Quality Tools > Cause and Effect...** Under the heading **Causes** select the six variable names in the order they appear in the left pane. In the column **Label** change the entries to match the words in the **Causes** column. Click **OK**.

#2. Draw a Pareto Chart
Open the worksheet **8n Motor Breakdown.** Select **Stat > Quality Tools > Pareto chart…** Enter ***Motor Breakdown*** and ***Frequency*** in the first two fields and click **OK.**

#3. Draw a Run Chart
Open the worksheet **8s Tablet Weight.** Select **Stat > Quality Tools > Run Chart…** Enter ***Tablet Weight*** and ***1*** in the first two fields and click **OK.**

Appendix 2: Workshops with SPSS®

Read this introduction first

Each of these Workshops corresponds to a particular section in the book, e.g. Workshop 1A corresponds to Section 1A in chapter 1. But before you begin you need to put some things in place by completing the following three steps.

Step 1: Download all the Data-sets
Go to https://zenodo.org/record/6778595#.Yrw-eHbMKUl and download the ***Applied Statistics data-sets***. Then right-click on the ***Applied Statistics data-sets*** file in your downloads folder and Extract All the files to a folder of your choice on your computer for future use. Take note of this folder name and location.

Step 2: Check that you have IBM® SPSS® Statistics software
If you are using a computer at college or at work, it may already have IBM® SPSS® installed. If you are using your own computer the best option is to get the software and licence from your college or employer. If that is not possible you can get a free 14-day trial of SPSS at http://www.presidion.com/software/ibm-spss-trial-downloads/

Step 3: Open a Saved Worksheet
Launch IBM® SPSS® on your computer. Now open the saved worksheet **1a Heights of Irishmen** from the main menu by selecting **File > Open > Data…** and browsing to the folder on your computer where you saved the ***Applied Statistics data-sets***. Choose the sub-folder **Data Sets > Chapter 1**, and in the **Files of Type** field select ***Excel***. Select the **Worksheet** called **1a Heights of Irishmen**, click **Open** and click **OK**. Remember how to do this because you will need to open saved worksheets again.

Workshop 1A

#1. Draw a Histogram
With the worksheet **1a Heights of Irishmen** open, from the main menu choose **Graphs > Chart Builder > OK.** In the **Gallery** tab, click on **Histogram**, and some icons will appear representing different chart styles. Click on the first icon which represents a simple histogram, and then drag it upwards into the canvas where it says **Drag a Gallery chart here...** Drag the variable ***Height*** into the **X-Axis** and click **OK**.

#2. Draw a Time Series Plot
Open the data set **1f Bakery Revenue**. From the main menu choose **Graphs > Chart Builder > OK.** In the **Gallery** tab, click on **Scatter/Dot**, and some icons will appear representing different chart styles. Click on the first icon which represents a simple scatter, and then drag it upwards into the canvas where it says **Drag a Gallery chart here...** Drag the variable ***Time*** into the **X-Axis** and drag the variable ***Bakery Revenue*** into the **Y-Axis**. Click **OK**. Double-click on the graph to bring up the **Chart Editor** and then click on **Elements > Interpolation Line > Close.**

#3. Practice
Draw histograms using the data sets in the saved worksheets ***1b, 1c, 1d, 1e, 1k, 1l***, and time series plots using the data sets in the saved worksheets ***1g, 1h, 1i, 1j, 1m***.

Workshop 1C

#1. Calculate Summary Statistics
Open the data set **1n Heights of Trees**. From the menu choose **Analyze > Descriptive Statistics > Descriptives…** Use the arrow to select the variable ***Height*** and click **OK**.

#2. Practice
Use software to solve ***Problems 1C, #1 #2*** and ***#3***. The data are saved in worksheets ***1o, 1p, 1q*** and ***1r***.

Workshop 1D

#1. Calculate Internal Consistency
Open the data set **1s Hotel Satisfaction**. From the main menu choose **Analyze > Scale > Reliability Analysis...** and select all of the items ***Enjoy-Speakers***. Select ***Alpha*** as the **Model** and click **Statistics...** Select **Scale if item deleted** and then click **Continue** and click **OK**.
The output shows the value of Cronbach's Alpha and also shows that its value would greatly improve if the "Speakers" item was omitted.

Workshop 4B

#1. Estimate a Population Mean (sigma unknown)
Open the worksheet **4c Flight Times**. From the menu choose **Analyze > Compare Means > One-Sample T-Test...** Use the arrow to select the variable ***Flight Times*** and click **OK**.

Workshop 5B

#1. Test a Mean (sigma unknown)
Open the worksheet **5a Peanuts.** From the menu choose **Analyze > Compare Means > One-Sample T-Test...** Use the arrow to select the variable ***Bag Weights.*** Enter ***25*** in the **Test Value** field and **c**lick **OK**.

Workshop 5D

#1. Analyse a Contingency Table using Raw Data
Open the worksheet **5jj Smartphones - raw data**. To create the table and analyse for contingency, click on **Analyse > Descriptive Statistics > Crosstabs...** Select one variable for the **rows** (***Genders*** on this occasion) and the other variable for the **columns** (***Smartphones*** on this occasion). Click **Statistics...** and check **Chi-square**, then click **Continue**. Click on **Cells...** and ensure that **Observed** and **Expected** are both checked. Click **Continue.** Click **OK**.

#2. Analyse a Contingency Table using Frequency Data
Open the worksheet **5kk Titanic Survival - freq**. To analyse for contingency using frequency data, click on **Data > Weight Cases**. Now select **Weight Cases by** and choose ***Frequency*** as the **Frequency Variable** and click **OK**. Now click on **Analyse > Descriptive Statistics > Crosstabs...** Select one variable for the **rows (*Classes*** on this occasion) and the other variable for the **columns** (***Outcomes*** on this occasion). Click **Statistics...** and check **Chi-square**, then click **Continue**. Click on **Cells...** and ensure that **Observed** and **Expected** are both checked. Click **Continue.** Click **OK**.

#3. Practice
Use software to solve ***Problems 5D***. The data are saved in worksheets ***5ll, 5mm, 5nn, 5oo*** and ***5pp***.

Workshop 6A

#1. Draw a Scatter Plot
Open the worksheet **6a ShoeSize and Height**. Select **Graphs > Legacy Dialogs > Scatter/Dot...** Select **Simple Scatter > Define** and select the appropriate variable for **Y Axis** (***Height*** in this case) and **X Axis** (***ShoeSize*** in this case) and click **OK**.

#2. Find a Correlation Coefficient
Open the worksheet **6a ShoeSize and Height**. Click on **Analyse > Correlate > Bivariate…** Use the arrow to select the two variables ***Height*** and ***ShoeSize*** and click **OK**.

#3. Practice
Draw a scatter plot and find the correlation coefficient for each of the data-sets in worksheets ***6b*** and ***6c***.

Workshop 6B

#1 Draw a Fitted Line Plot
Open the worksheet **6a ShoeSize and Height**. Select **Graphs > Legacy Dialogs > Scatter/Dot...** Select **Simple Scatter > Define** and select the appropriate variable for **Y Axis** (***Height*** in this case) and **X Axis** (***ShoeSize*** in this case) and click **OK**.
Double-click on the scatterplot to bring up the **Chart Editor**. Then choose **Elements > Fit line at Total > Close.**

#2. Practice
Draw a fitted line plot for each of the data-sets in worksheets ***6b, 6d, 6e*** and ***6f***. Then solve ***Problems 6B***.

Workshop 6C

#1. Carry out a Regression Analysis
Open the worksheet **6a ShoeSize and Height**. Click on **Analyze > Regression > Linear**. Select the appropriate variables for **Dependent** (***Height*** in this case) and **Independent** (***ShoeSize*** in this case) and click **OK.**

#2. Make a Prediction
Open the worksheet **6b Age and Price**. Type a new value, ***8***, at the end of the **Age** column. Now click on **Analyse > Regression > Linear**. Select the appropriate variables for **Dependent** (***Price*** in this case) and **Independent** (***Age*** in this case). Now click on **Save…** and under the heading **Predicted Values** check the box for **Unstandardized**, and under the heading **Prediction Intervals** check the boxes for **Mean** and **Individual**. Click **Continue** and then click **OK**. The fitted value is displayed in the data window, along with a confidence interval and a prediction interval.

#3. Carry out a Residual Analysis
Open the worksheet **6d Attendance and Exam Mark**. Click on **Analyze > Regression > Linear**. Select the appropriate variables for **Dependent** (***Exam Mark*** in this case) and **Independent** (***Attendance*** in this case) and click on **Plots...** and select ***ZRESID*** for **Y:** and ***ZPRED*** for **X:** and also check the **Histogram** box. Click on **Continue** and then click on **OK**.

#4. Practice
Carry out a regression analysis for each of the data-sets in worksheets ***6b, 6d, 6e*** and ***6f***. Make a prediction in each case using a new value of the predictor. Then solve ***Problems 6C***.

Workshop 7A

#1. Carry Out a One-Way ANOVA
Open the worksheet **7a Distance Kicked.** Click on **Analyze > General Linear Model > Univariate…** In the dialog box, select the appropriate **Dependent Variable** (***Distance*** in this case) and **Fixed Factors** (***Player*** in this case) and click **OK**.

#2. Practice
Repeat this analysis using the data in worksheets ***7b*** and ***7c***.

Workshop 7B

#1. Carry Out a Balanced ANOVA
Open the worksheet **7e Reading Time.** Click on **Analyze > General Linear Model > Univariate…** In the dialog box, select the appropriate **Dependent Variable** (***Time*** in this case) and **Fixed Factors** (***Language*** and ***Person*** in this case) and click **OK**.

#2. Draw an Interaction Plot
Open the worksheet **7e Reading Time.** Click on **Analyze > General Linear Model > Univariate…** In the dialog box, select the appropriate **Dependent Variable** (***Time*** in this case) and **Fixed Factors** (***Language*** and ***Person*** in this case). Click on **Plots...** Now select the appropriate factor for **Horizontal Axis** (***Language*** in this case) and **Separate Lines** (***Person*** in this case) and click on **Add** so that ***Language* Person*** appears in the **Plots** field. Click on **Continue** and click **OK**.

#3. Practice
Repeat these analyses using the data in the worksheets ***7f, 7g, 7h*** and ***7i***.

Appendix 3: Exercises with Excel

Read this introduction first

Each of these Exercises corresponds to a particular section in the book, e.g. Exercise 1A corresponds to Section 1A in chapter 1. But before you begin any of these exercises you need to put some things in place by completing the following four steps.

Step 1: Download all the Data-sets
Go to https://zenodo.org/record/6778595#.Yrw-eHbMKUl and download the ***Applied Statistics data-sets***. Then right-click on the ***Applied Statistics data-sets*** file in your downloads folder and **Extract All** the files to a folder of your choice on your computer for future use. Take note of this folder name and location.

Step 2: Check that you have Excel installed
If you are using a computer at college or at work it may already have Excel installed. If you are using your own computer the best option is to get the Excel software and licence from your college or employer. If that is not possible you can get a free 1 month trial of Microsoft 365, which includes Excel, at https://products.office.com/en-ie/try

Step 3: Check your Excel setup
Click on the **Data** tab. If you can see a **Data Analysis** button then you have everything you need already. If you can't see a **Data Analysis** button then you need to add in the **Analysis ToolPak** by following these steps. Click on **File > Options > Add-Ins** and in the **Manage** box select ***Excel Add-ins*** and click **Go...**. Check the checkbox for **Analysis ToolPak** and click **OK**.

Step 4: Open a Saved Worksheet
Launch Excel on your computer. Now open the saved worksheet **1a Heights of Irishmen** from the Excel main menu by selecting **File > Open…** and browsing to the folder on your computer where you saved the ***Applied Statistics data-sets***. Choose the sub-folder **Data Sets > Chapter 1** and then select the File name **1a Heights of Irishmen**, and click **Open**. Remember how to open a worksheet because you will need to do it again many times.

Exercise 1A

#1. Draw a Histogram
With the worksheet **1a Heights of Irishmen** open in Excel, click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **Histogram** and click **OK**. In the Histogram dialog window, click in the Input Range field and select the cells containing the data, ***$A$1:$A$101***. Check the ***Labels*** box because the first cell contains a label. Check the **Chart Output** box and click **OK**. The graph will appear but it needs to be improved. Double-click on the title, ***Histogram***, and type a more complete title. Also double-click on the label, ***Bin***, and type a more suitable name for this label.

#2. Draw a Histogram with Integer Values
Open the worksheet **1q Letters per Word.** Before drawing a histogram, we need to arrange for the bins to correspond to the integers 2, 3 ,4, 5, 6, 7, 8, 9, 10 because these are the only values that make sense. Type the title ***Number of Letters*** into cell B1 and in the cells underneath this, cells B2:B10, type the integers 2, 3 ,4, 5, 6, 7, 8, 9, 10 . Now click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **Histogram** and click **OK**. In the Histogram dialog window, click in the Input Range field and select the cells containing the data, ***$A$1:$A$28*** and in the Bin Range field select the cells ***$B$1:$B$10.*** Check the ***Labels*** box because the first cell contains a label. Check the **Chart Output** box and click **OK**. Double-click on the title, ***Histogram***, and type a more complete title.

#3. Draw a Time Series Plot
Open the worksheet **1f Bakery Revenue**. Highlight the cells that contain the data, ***A1:B31***. Click on the **Insert** tab and from **Charts** select the **Scatter** option that shows points connected by straight line segments. Double-click on the Y axis to reveal the **axis options** and change the minimum from zero to something more suitable in this case such as 200. Click on the plus sign beside the graph to add axis titles and type a suitable name on each axis.

Exercise 1B

#1. Select Numbers at Random
Suppose you want to select 100 ID numbers at random from a list of 1000 ID numbers ranging from 1 to 1000. Click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **Random Number Generation** and click **OK**. In the Random Number Generation dialog window, for **Number of Variables** enter ***1***, for **Number of Random Numbers** enter ***100***, for **Distribution** select ***Uniform***, and for **Parameters** enter **Between *1* and *1001***. Click on the **Output Range** radio button and then click into this field and choose where you would like the random numbers to appear, e.g. starting in cell ***$M$2***, and then click **OK.** If you would like to have the numbers in increasing order then click on the **Data** tab and then the **Sort** button and click **OK**. Some numbers may occur more than once.

#2. Sort IDs into Random Order
First create a list of as many ID numbers as you need, say 1000, by typing ***1*** and ***2*** into cells A1 and A2, then dragging down from cell A2 to cell A1000, and from the **Auto Fill Options** menu choosing ***Fill Series***. Next, generate 1000 random numbers with the output range starting in cell B1. Now highlight the two columns, A and B, and click on the **Data** tab and then the **Sort** button. Choose to **Sort by** column B and click **OK.**

#3. Draw a Multi-Stage Sample
Suppose you want to select 100 items at random from a shopping website which has 153 pages with 30 items on each page.
First, generate 100 random page numbers as follows. Click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **Random Number Generation** and click **OK**. In the Random Number Generation dialog window, for **Number of Variables** enter ***1***, for **Number of Random Numbers** enter ***100***, for **Distribution** select ***Uniform***, and for **Parameters** enter **Between *1* and *153***. Click on the **Output Range** radio button and then click into this field and choose where you would like the random numbers to appear, e.g. starting in cell ***$M$2***, and then click **OK.** To display the numbers as integers, click on **Home** and **Number** and then keep clicking on the **Decrease Decimal** button until all the decimal places disappear. To display the page numbers in increasing order, click on the **Data** tab and then the **Sort** button and click **OK.**
The next step is to generate 100 random item numbers as follows. . Click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **Random Number Generation** and click **OK**. In the Random Number Generation dialog window, for **Number of Variables** enter ***1***, for **Number of Random Numbers** enter ***100***, for **Distribution** select **Uniform**, and for **Parameters** enter **Between *1* and *30***. Now it is important this time to choose a DIFFERENT place where the random numbers will appear so click on the **Output Range** radio button and then click into this field and specify that the numbers are starting in cell ***$N$2***, and then click **OK.** To display the numbers as integers, click on **Home** and **Number** and then keep clicking on the **Decrease Decimal** button until all the decimal places disappear. Do NOT sort the item numbers into increasing order. You can type ***Page*** in cell M1, and ***Item*** in cell N1.
The same approach can be used to select players at random from a league by randomly selecting team numbers and squad numbers.

Exercise 1C

#1. Calculate Summary Statistics
Open the data set **1n Heights of Trees**. Click on the **Data** tab and then the **Data Analysis** button. Highlight the **Descriptive Statistics** entry in the list of Analysis Tools and click **OK**. Select the cells containing the data, ***$A$1:$A$6*** in this case. Check the box for ***Labels in first row.*** Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$C$1***. Check the box for **Summary statistics** and then click **OK.**

#2. Practice
Use software to solve ***Problems 1C, #1 #2*** and ***#3***. The data are saved in worksheets ***1o, 1p, 1q*** and ***1r***.

Exercise 3B

#1. Calculate Normal Probabilities
This exercise reproduces the calculations in the first example in section 3B of the textbook. Open Excel and in a blank cell type ***=normdist(182.5,175,10,1)*** and press return. The answer given is 0.7734 but the answer represents the probability that a man is shorter than 182.5 so the probability that a man is taller than 182.5 is found by subtracting this answer from 1, giving 0.2266.

Exercise 3D

#1. Calculate Binomial Probabilities
This exercise reproduces the calculations in the first example in section 3D of the textbook. Open Excel and in a blank cell type ***=binomdist(2,6,0.1,0)*** and press return. The answer, 0.098415, represents the probability of obtaining exactly two brown eggs in a carton of six eggs filled randomly from a population in which 10% of eggs are brown.

#2. Calculate Poisson Probabilities
This exercise reproduces the calculations in the last example in section 3D of the textbook. Open Excel and in a blank cell type ***=poisson.dist(2,3,1)*** and press return. The answer, 0.4232, represents the probability of receiving two or fewer complaints on a particular day if the mean number of complaints per day is 3. The probability of receiving more than two complaints is found by subtracting this answer from 1.

Exercise 4B

#1. Estimate a Population Mean (sigma unknown)
Open the data set **4c Flight Times**. Click on the **Data** tab and then the **Data Analysis** button. Highlight the **Descriptive Statistics** entry in the list of Analysis Tools and click **OK**. Select the cells containing the data, ***$A$1:$A$6*** in this case. Check the box for ***Labels in first row.*** Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$C$1***. Check the box for **Confidence level for mean** and then click **OK.** The margin of error is displayed.

Exercise 5C

#1. Test Means using Independent Samples
With the worksheet **5f Groundwater** open in Excel, click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **t-Test: Two-Sample Assuming Equal Variances** and click **OK**. In the **Variable 1 Range** enter ***$A$1:$A$4*** and in the **Variable 2 Range** enter ***$B$1:$B$9*** and enter zero for **Hypothesized Mean Difference**. Check the **Labels** box because the first row contains labels. Select the **Output Range** radio button and then click into this field and select a location for the output such as ***$P$1*** and click **OK**. The test output will appear.

#2. Test Means using Paired Samples
With the worksheet **5e Vitamin C in Peppers** open in Excel, click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **t-Test: Paired Two Sample for Means** and click **OK**. In the **Variable 1 Range** enter ***$B$1:$B$5*** and in the **Variable 2 Range** enter ***$C$1:$C$5*** and enter zero for **Hypothesized Mean Difference**. Check the **Labels** box because the first row contains labels. Select the **Output Range** radio button and then click into this field and select a location for the output such as ***$P$1*** and click **OK**. The test output will appear.

Exercise 5F

#1. Test the Difference Between Two Variances
With the worksheet **5r Paint Thickness** open in Excel, click on the **Data** tab and then the **Data Analysis** button. Scroll through the Data Analysis menu and select **F-Test Two-Sample for Variances** and click **OK**. In the **Variable 1 Range** enter ***$B$2:$B$7*** and in the **Variable 2 Range** enter ***$B$8:$B$15*** and do NOT check the **Labels** box. Select the **Output Range** radio button and then click into this field and select a location for the output such as ***$P$1*** and click **OK**. The test output will appear.

Exercise 6A

#1. Find a Correlation Coefficient, Adjusted *R* Square and *S*
Open the worksheet **6a ShoeSize and Height**. Click on the **Data** tab and then the **Data Analysis** button. Highlight the **Regression** entry in the list of Analysis Tools and click **OK**. Select the cells containing the *Y* data, ***$B$1:$B$7*** in this case, and the *X* data, ***$A$1:$A$7*** in this case. Check the box for ***Labels.*** Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$L$1***. Click **OK.** The **correlation coefficient** (called *Multiple R* here) and values for the **Adjusted *R* Square** and ***S*** are shown under **Regression Statistics**.

Exercise 6B

#1. Draw a Fitted Line Plot
Open the worksheet **6a ShoeSize and Height**. Highlight the cells that contain the data, ***A1:B7***. Click on the **Insert** tab and from **Charts** select the **Scatter** option that includes no lines. Click on the plus sign beside the graph to add axis titles and then type a suitable name on each axis. Now right-click on one of the points and select **Add Trendline**. In the **Format Trendline** panel check the box for **Display Equation on chart**.

#2. Regression Equation
Open the worksheet **6a ShoeSize and Height**. Click on the **Data** tab and then the **Data Analysis** button. Highlight the **Regression** entry in the list of Analysis Tools and click **OK**. Select the cells containing the *Y* data, ***$B$1:$B$7*** in this case, and the *X* data, ***$A$1:$A$7*** in this case. Check the box for ***Labels.*** Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$L$1***. Click **OK.** Values for the intercept and slope are shown under **Coefficients**.

Exercise 6C

#1. Regression Analysis
Open the worksheet **6a ShoeSize and Height**. Click on the **Data** tab and then the **Data Analysis** button. Highlight the **Regression** entry in the list of Analysis Tools and click **OK**. Select the cells containing the *Y* data, ***$B$1:$B$7*** in this case, and the *X* data, ***$A$1:$A$7*** in this case. Check the box for ***Labels.*** Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$L$1***. Click **OK.** *P*-values for the intercept and slope are shown under ***P*-value**.

Exercise 6D

#1. Carry out a Multiple Regression Analysis
This exercise reproduces the regression equation shown in Problems 6D #1. Open the worksheet **6i Glue**. Click on the **Data** tab and then the **Data Analysis** button. Highlight the **Regression** entry in the list of Analysis Tools and click **OK**. Select the cells containing the *Y* data, ***$A$1:$A$11*** in this case, and the *X* data, ***$B$1:$C$11*** in this case. Check the box for ***Labels.*** Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$L$1***. Click **OK.**

Exercise 7A

#1. Carry Out a One-Way ANOVA
Open the worksheet **7aa distance unstacked.** Click on the **Data** tab and then the **Data Analysis** button. Highlight the **ANOVA Single Factor** entry in the list of Analysis Tools and click **OK**. Select the cells containing the data, ***$A$1:$C$5*** in this case. Check the box for ***Labels in first row.*** Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$L$1***. Click **OK.** The ANOVA table and summary statistics are displayed.

Exercise 7B

#1. Carry Out a Balanced ANOVA
Open the worksheet **7ee Reading unstacked.** Click on the **Data** tab and then the **Data Analysis** button. Highlight the **Anova: Two-Factor With Replication** entry in the list of Analysis Tools and click **OK**. For **Input Range**, select the cells containing the data, ***$A$1:$D$5*** in this case. For ***Rows per sample*** enter 2. Click on the **Output Range** radio button and then click into this field and choose where you would like the output to appear, e.g. starting in cell ***$J$1***. Click **OK.** The ANOVA table and summary statistics are displayed.

Appendix 4: Answers to Problems

Answers 1A

1. The weights of the wine gums seem approximately normal. This indicates that the wine gums are fairly similar in weight with a typical weight of 5.3 or 5.4 grams. Some wine gums are heavier or lighter than this, but the values are increasingly rare as they depart further from the average.

2. The histogram has multiple peaks at regular intervals, with troughs in between. This pattern is called a 'comb' distribution, because the peaks resemble the teeth on a comb. This tells us that there are a number of popular engine sizes and these differ from each other in increments of 0.2 litres. Similar patterns arise with prices of motor fuel at different outlets, which differ in increments of 1 cent, and prices of clothing and shoes, which differ in increments of 5 or 10 euro.

3. In January and July every year, when new registration plates are issued, there is a sharp increase, and the figures tend to decrease over the following months. There is a slump in the figures that begins early in 2020, at the beginning of the Covid-19 pandemic. A recovery in the figures can be seen midway through 2021 where, contrary to the usual pattern, the July figure exceeds the previous January figure.

Answers 1B

1. (a) This is a self-selecting sample. Outpatients who have been waiting a long time will be more interested in this topic and therefore may be more likely to contact the show. Also, outpatients with longer waiting times are more likely to be waiting at the time the radio show is aired than outpatients with shorter waiting times.

(b) A single school is a cluster and may not be representative of all the families in the town. Even worse, this researcher sampled children but should have sampled families: larger families will tend to be over represented in this sample, because there is a better chance of selecting a student from a larger family than a smaller one.

(c) The first 50 passengers is not a random sample and may include fewer of those people who are unhappy with the timetable because they have difficulty arriving in time for the train. The first number of items in any list do not constitute a random sample and may often lead to biased estimates.

(d) People who are at home in the afternoon tend to be older and may have different voting preferences than people who are not at home. Also, homeowners may have different voting preferences than people who are not homeowners. Also, mentioning the name of one candidate out of many candidates suggests that candidate's name to respondents as a potential answer.

Answers 1C

1. (a) $\bar{X} = 8$, $S = 1$ (b) $\bar{X} = 8$, $S = 3.633$

2. $\bar{X} = 4.296$, $S = 2.2503$.

3. (a) Mean, 28. Mode, 19. Median, 20.

(b) The mean is not useful. It is inflated by the outlier and returns a value that does not properly represent the ages of any friends.

(c) The mean will change to an unknown number greater than 28. Mode: no change. Median: no change.

4. (a) The median. (b) The mode. (c) The mean.

Answers 1D

(a) Binary scale. Provide 'Yes' or 'No' check-boxes.

(b) Ratio scale. Invite the respondent to write a number.

(c) Likert Scale. For each item, invite a selection from different levels of satisfaction.

(d) Open-ended. Provide some blank space for the respondent to write as they wish.

Answers 2A

1. (a) If the die is rolled many times, then a 'six' occurs on one-sixth of all rolls. On a single future roll, 'one-sixth' is a measure of how likely it is that a six will occur, i.e. it is more likely not to occur.

(b) 90% of all invoices are paid within 30 days. 90% is a measure of how likely it is that a single future invoice will be paid within 30 days, i.e. very likely.

(c) No pigs fly. A pig is certain not to fly.

2. (a) $p = 2/6 = 1/3$

(b) Draw a random sample. Then $p = r / n$ when r occurrences are observed on n trials.

(c) Make a guess, but do not guess 0 or 1.

3. (a) 1/26 (b) 5/26 (c) 6/26

4. (a) 1/2 (b) 1/6 (c) 1/12 (d) 7/12

5. (a) 1/36 (b) 25/36 (c) 5/36 (d) 11/36 (e) 10/36

6. (a) 1/8 (b) 7/8

7. (a) 28% (b) 18% (c) 42% (d) 12% (e) 82%

Answers 2B

1. 40,320

2. (a) 1 / 6840 (b) 1 / 1140

3. 10,737,573 ways. Every time we choose six numbers to 'take away', we are choosing 41 numbers to 'leave behind', and so the answers are the same.

4. (a) 1 / 10,737,573 (b) 246 / 10,737,573 (c) 12,300 / 10,737,573

5. ${}^nP_n$ means the number of ways of arranging n things, taken from among n things, i.e. all n things are arranged. The definition of $n!$ has the same meaning.

6. To take n things from among n things, you have to take them all; you have no choice. There is only one way to do it, hence ${}^nC_n = 1$

Answers 2C

1. (a) vi (b) x (c) viii (d) i (e) x

2. 94.85%

3. The posterior odds are 754 to one that the owner of the shoes is male.

Answers 2D

1. (a) 0.6840

(b) 0.9001

(c) 0.948024

2. 0.971

3. (a) 0.726750

(b) 0.8357625

(c) 0.91933875

(d) Add a third paddle. System reliability 0.937325812 rather than 0.927696375.

(e) 4 paddles

4. (a) 0.58212

(b) 0.814968

(c) 0.9081072

(d) 0.9702

(e) 5

Answers 3A

1. (a) {0,1} (b) Uniform (c) 2

Answers 3B

1. (a) 0.1587
(b) 0.6247
(c) 0.7734
(d) 0.1314 or 0.1292
(e) 0.8643
(f) 0.0994

(g) 0.1015
(h) 0.5
(i) The answer is zero. The tables do not provide probabilities for *z*-scores exceeding 3.09 because the tail area is tiny.
(j) The answer is zero. No corn stalk is exactly 17.000... cm tall. The sketch shows a line, not an area.
(k) 0.95 or 95%.

2. 'stated with 95% confidence'... 'between 12.08 cm and 19.92 cm.'

3. 32.32% small, 47.06% medium, 14.88% large, 5.74% other.

4. (a) 0.0082

(b) 0.8185

(c) 0.1574

(d) Probability = 0, because $z = 10$. Values greater than this will virtually never occur.

(e) Between 60.2 and 79.8 seconds.

5. (a) 0.3085

(b) 0.3830

(c) 0.1359

(d) Probability = 0, because an exactly pre-specified value in a continuous distribution will virtually never occur.

(e) 172.66

Answers 3C

1. Q, R, U and Z only.

2. N, O, R, S and T only.

Answers 3D

1. 0.0625 + 0.25 + 0.375 + 0.25 + 0.0625 = 1, because it is certain ($p = 1$) that the number of heads will be 0 or 1 or 2 or 3 or 4.

2. (a) 0.0305 (b) 0.0023 (c) 0.9977 (d) 0.9672 (e) 0.0328

3. 0.9664

4. 0.1314

5. (a) 67.03% (b) 32.17% (c) 0.80%

6. (a) 0.30% (b) 98.62%

Answers 3E

1. The 3-Parameter Weibull distribution is the best fit out of the candidate distributions considered in the graph. Its probability plot shows some random scatter without obvious curvature.

Answers 4A

1. (a) 10 grams (b) 2 grams
2. (a) 8
(b) 1.8257
(c) Normal.

Answers 4B

1. The mean weight of all the bags of rice filled for that customer lies between 497.478 and 498.278 grams, with 95% confidence.

2. The mean diameter of all the plastic tubes made by that process today lies between 12.010 and 12.098 mm, with 95% confidence.

3. The mean journey distance for all service calls by that engineer lies between 6.7 and 17.3 km, with 95% confidence.

4. The mean expenses incurred on food and accommodation, by all the delegates attending that conference, lies between 744 euro and 922 euro, with 95% confidence.

5. Between 24% and 36% of all the booklets in the consignment have defective binding, with approximately 95% confidence.

6. Between 13.7% and 19.5%, of all the students at that university campus, walk from home to their classes, with approximately 95% confidence.

Answers 4C

1. 79

2. 139

3. 97

4. 733

Answers 4D

1. 0.410

2. Between 0.017 and 0.082

Answers 4E

1. The tape measurements are between 15.48 and 1.52 shorter than the laser measurements, on average.

2. Kevin is between 7.44 and 5.06 kg lighter than Rowena.

3. The 10 a.m. mean minus the 11 a.m. mean is between -6.05 and 5.25. There might be no difference.

4. Between 4.02% and 13.73% more summer visitors than winter visitors had prepaid tickets.

Answers 5A

1. A type 1 error means that the product conforms to specifications but the test wrongly asserts that the product does not conform. A type 2 error means that the product does not conform to specifications but the test fails to identify this nonconformity. A type 2 error is virtually always more costly. The shoelaces will be relied upon 'in the field' to contribute to some larger purpose. This purpose may be thwarted by defective shoelaces, resulting in considerable loss for the user, additional trouble and expense to arrange replacement, and loss of reputation for the supplier. A type 1 error will lead to the disposal of a batch of perfectly good shoelaces, but the losses are limited to the cost of production and the cost of disposal.

2. There are two major problems with the manager's approach. Firstly, having formed a hunch and formulated a hypothesis to be tested, the procedure states that a fresh random sample should then be drawn to test the hypothesis, rather than going back and using the pessimistic sample that gave rise to the hunch in the first place. This mistake is called HARKing, i.e. Hypothesising After the Results are Known.

Secondly, a cluster sample of waiting times taken together like this are not independent: one hold-up in the queue may cause a number of later waiting times to be prolonged also. A random sample, rather than a cluster sample, should always be used to test a hypothesis.

Answers 5B

1. H_0, $\mu = 7$, is accepted, $z = -1.25$, critical $z = \pm 1.96$

2. H_0, $\mu = 50$, is accepted, $t = -2.0042$, critical $t = -2.015$

3. H_0, $\mu = 80$, is rejected, $t = -4.24$, critical $t = \pm 2.365$

4. H_0, $\pi = 0.5$, is rejected, $z = -2.427$, critical $z = -1.645$

5. H_0, $\pi = 0.25$, is rejected, $z = -2.12$, critical $z = \pm 1.96$

Answers 5C

1. Yes. Reject H_0, $t = 3.207$, critical $t = 2.132$

2. Yes. Reject H_0, $t = -2.705$, critical $t = -1.943$

3. Yes. Reject H_0, $t = -16.69$, critical $t = -2.353$

Answers 5D

1. Yes. Reject H_0. Chi square = 8.879, $df = 1$, critical Chi-Sq = 3.841. Men on ADT are less likely to be infected with COVID-19.

2. Yes. Reject H_0, Chi-Sq = 45.214, critical Chi-Sq = 5.991. The proportion of female students tends to be lower on engineering courses, and higher on science courses.

3. No. Accept H_0, Chi-Sq = 4.185, critical Chi-Sq = 7.815. The Lusitania sank quickly and the deployment of the lifeboats was chaotic, so being able to swim or being able to climb into a lifeboat in the water was relevant to survival. Class was not relevant.

4. Yes. Reject H_0, Chi-Sq = 40.40, critical Chi-Sq = 3.841

5. Yes. Reject H_0, Chi-Sq = 10.89, critical Chi-Sq = 3.841. Boxes packed by machine are more likely to include some broken biscuits.

Answers 5E

1. Validation is like requiring a defendant in a trial to prove their innocence. Rather than assuming their innocence, it is a case of 'guilty until proven innocent'.

2. Ask a medical doctor, or other competent person, to identify the minimum increase in average blood pressure that is of practical importance.

Answers 5F

1. Yes. Reject H_0, Chi-Sq = 0.8, critical Chi-Sq = 2.733

2. No. Accept H_0, F = 5.33, critical F = 6.388

3. No. Accept H_0, F = 1.20, critical F = 9.01

4. No. Reject H_0, Chi-Sq = 17.55, critical Chi-Sq = 16.919

5. Yes. Accept H_0, Chi-Sq = 4.4, critical Chi-Sq = 11.070

6. No. Reject H_0, Chi-Sq = 2124, critical Chi-Sq = 11.070. After.

Answers 5G

1. No significant difference. Chi-Sq = 0.312, critical Chi-Sq = 3.841

2. No significant difference, t = -0.59, critical t = 2.306

Answers 6A

1. The coefficient of determination estimates the proportion of the variation in the journey times (of all such journeys) that is explained by the variation in the journey distances.

Answers 6B

1. Every 1% increase in attendance tends to be associated with an average increase in final exam mark of 0.4119%.

Students who don't attend at all would be expected to obtain an average final exam mark of 12.48%.

20.9% of the variation in the final exam marks, of all the students on this science course, can be explained by the variation in their attendances.

Students with the same level of attendance will have final exam marks that differ by 18.27%, typically, from the average final exam mark of students with this level of attendance.

2. Every one extra week tends to be associated with an increase of 0.55 in the number of appearances.

95.9% of the variation in the number of appearances can be explained by the variation in the number of weeks.

The actual number of appearances by a individual player differs by 15, typically, from the average number of appearances of all players who have been at their club for that same number of weeks.

3. The predicted sale price is €344,581. Every one square metre increase in the floor area tends to be associated with an average increase of €1809 in the sale price.

29.1% of the variation in the sale prices, of all the houses for sale in Dublin, can be explained by the variation in their floor areas.

Houses with the same floor area will have sale prices that differ by about €87000 typically, from the average sale price of houses with this floor area.

Answers 6C

1. Beta: reject, because $p = 0.006$, so attendance is a useful predictor of final exam mark. Alpha: accept, because $p = 0.131$, so attendance could be directly proportional to final exam mark. With 95% confidence, a student with 40% attendance would obtain a final exam mark between zero and 67%.

2. Beta: reject ($p = 0.000$). Weeks is a useful predictor of appearances.
Alpha: accept ($p = 0.357$). Appearances may be directly proportional to weeks.
After 300 weeks, we can say with 95% confidence that a player would have made between 138 and 214 appearances.

3. Beta: reject, because $p = 0.010$, so floor area is a useful predictor of sale price. Alpha: accept, because $p = 0.225$, so price could be directly proportional to floor area. With 95% confidence, the average sale price of houses with a floor area of 140 square metres, lies between €288,315 and €400,971. With 95% confidence, the sale price of an individual house with a floor area of 140 square metres is between €153,216 and €536,070.

Answers 6D

1. Every 1% increase in moisture content is associated with an average increase of 2.17 minutes in the drying time, provided that the relative humidity remains constant. Every 1% increase in relative humidity is associated with an average increase of 0.975 minutes in the drying time, provided that the moisture content remains constant.

2. Age, by itself, is the best predictor. The law of parsimony applies here.

3. Either fit a quadratic regression equation, or transform the minutes variable, with something like log(minutes).

Answers 6F

(a) Discriminant analysis could be used because the groups are known in advance.

(b) Cluster analysis could be used because there are no predetermined groups.

(c) Principal components analysis can be used to reduce the number of variables, or else partial least squares regression if values of the response variable are available.

Answers 7A

1. (a)

μ represents the population grand average burning time for all the different colours.

α represents the effect of colour i, that is how much the average burning time of that colour exceeds the grand average burning time.

ε represents the random error on that occasion, that is how much the burning time of a single candle exceeds the average burning time for that colour.

(b) H_0: Colour does not affect burning time. Accept H_0, because $p = 0.586$. The data do not prove, at the 5% level, that colour affects burning time.

2. (a)

μ represents the population grand average weight of clementines for both supermarkets.

α represents the effect of supermarket i, that is how much the average weight of clementines in that supermarket exceeds the grand average weight.

ε represents the random error on that occasion, that is how much the weight of a single clementine exceeds the average weight of clementines in that supermarket.

(b) H_0: Supermarket does not affect weight. Reject H_0, because $p = 0.001$. The data prove, at the 5% level, that supermarket affects weight. One supermarket stocks clementines that are heavier, on average.

Answers 7B

1. (a)

μ is the grand average time taken, averaged over all foods and knives.

α is the food main effect, i.e. how much more time is taken to cut that food on average, compared to the grand average.

β is the knife main effect, i.e. how much more time is taken using that knife on average, compared to the grand average.

η is the interaction effect, i.e. how much more time is taken on average for that particular food and knife combination, compared to what would be expected.

ε is the error, i.e. how much more time was taken on that occasion, compared to the average time taken for that particular food and knife combination.

(b) A tomato is quicker to cut than a sausage. It's quicker to cut these foods with the serrated knife than the straight knife. Switching from the serrated knife to the straight knife increases the cutting time for a sausage more than for a tomato.

2. (a)

μ is the grand average time, averaged over all guitars and methods.

α is the guitar main effect, i.e. how much more time on average is required to play a scale on that guitar, compared to the grand average.

β is the method main effect, i.e. how much more time on average is required to play a scale using that method, compared to the grand average.

η is the interaction effect, i.e. how much more time on average is required for that particular guitar-method combination, compared to what would be expected.

ε is the error, i.e. how much more time was taken on that occasion, compared to the average time for that guitar-method combination.

(b) With a steel-string guitar, switching from using a plectrum to plucking the strings increases the time required to play a scale, but with a nylon-string guitar, it does not make much difference to the time.

3. (a)

μ is the grand average time, averaged over all paddles and girls.

α is the paddle main effect, i.e. how much longer the time is, on average, using that paddle, compared to the grand average.

β is the girl main effect, i.e. how much longer the time is, on average, when travelled by that girl, compared to the grand average.

η is the interaction effect, i.e. how much longer the time is, on average, for that particular paddle-girl combination, compared to what would be expected.

ε is the error, i.e. how much longer the time was on that occasion, compared to the average time for that particular paddle-girl combination.

(b) The paddle affects the time. The straight paddle is faster.

4. (a)

μ is the grand average measurement, averaged over all parts and sessions.

α is the part main effect, i.e. how much higher the measurement is, on average, for that part, compared to the grand average.

β is the session main effect, i.e. how much higher the measurement is, on average, during that session, compared to the grand average.

η is the interaction effect, i.e. how much higher the average measurement is for a particular part during a particular session, compared to what would be expected.

ε is the error, i.e. how much higher the measurement was on a single occasion, compared to the average measurement for that part during that session.

(b) The parts are different but there is no evidence that the session has a significant effect on the measurement.

Answers 7C

1. (a) All the factors have text levels. No centre points are available.

(b) Application type affects temperature. (High-end game raises the temperature.)

(c) The combination of battery in with fan off, or the combination of multiple apps with central heating on, are also possible explanations for high temperature. AD represents external fan and battery. The plus sign before AD indicates that either both are at high (+ 1 and + 1) or both are at low (- 1 and - 1) levels, which in either case would give rise to a positive product (+AD) when factor B is at the high level. Having the fan off (which is the low level of factor A) is the more plausible explanation for high temperature because the purpose of a fan is to reduce temperature. Therefore, the combination of fan off and battery in could explain the significant experimental result. This can also be confirmed by examining the experimental data. Similarly for CF, where the high level of both factors is a plausible explanation for high temperature.

Answers 7D

1. (a) It is a covariate, because it is a continuous variable.

(b) Yes. $p = 0.026$

(c) Put it in the oven for 4 hours. Next best, leave it on a storage heater overnight.

Answers 7E

1. First came the pooling of the mean square error from all batches: no justification is required for this. Next came the pooling of slopes: this step required $p > 0.25$ for the test of equality of slopes. Next came the pooling of intercepts: this step required $p > 0.25$ for the test of equality of intercepts. A recommended shelf life is 24 months. A shelf life of 36 months is not supported by the analysis.

Answers 8A

1. (a) Normality. The process appears normal, based on a visual appraisal of the histogram. (b) Stability. The process appears to be stable based on the values of StDev(Overall) and StDev(Within) (c) Centrality. The process appears central since the estimated process mean is very close to the target. (d) Capability. The process appears capable since the *Cpk* value exceeds 1.33. (e) Performance. The process is performing since the *Ppk* value exceeds 1.33.

2. (a) Normality. The process is normal, because the histogram will not change. (b) Stability. The process is stable because the values of StDev(Overall) and

StDev(Within) will not change (c) Centrality. The process is no longer central since the estimated process mean is not close to the new target. (d) Capability. The process is not capable since the *Cpk* value will fall below 1.33. (e) Performance. The process is not performing since the *Ppk* value will fall below 1.33.

Answers 8B

1. Possible causes include: an untrained worker misjudging the fill volume (men), a less viscous batch of liquid (materials), damage to the filling head (machines), incorrect sequencing of the steps involved in the fill cycle (methods), a change in the voltage of the power supply used to pump the liquid into the bottles (milieu), or incorrect measurement of the fill-volume due to instrument calibration issues (measurements).

2. A C chart should be used, because it is required to monitor the number of scratches (defects) and not the number of scratched screens (defectives). A sample size of three screens is required so that the mean number of scratches per sample exceeds five.

Answers 8D

1. %Linearity = -5.5%. For every 100 km increase in the reference measurement, the bias tends to decrease by 5.5 km. (The map underestimates more on long journeys.)

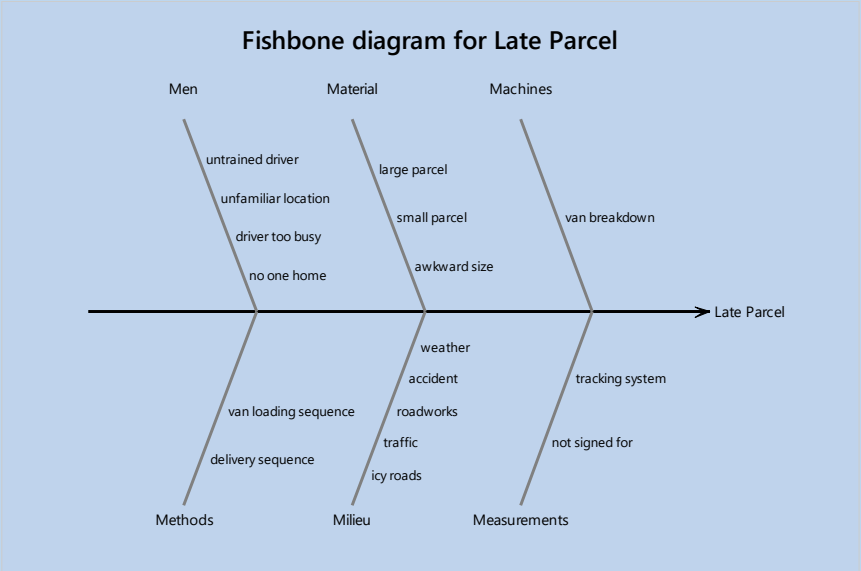
2. The Number of Distinct Categories = 8. This indicates that the measurement system is able to separate the parts into eight different groups. This is adequate for distinguishing different parts because the number of groups is five or more.

SNR = 1.44605 ÷ 0.25169 = 5.7. Since the SNR > 3, the measurement system is acceptable. It is able to distinguish between the parts that are presented to it.

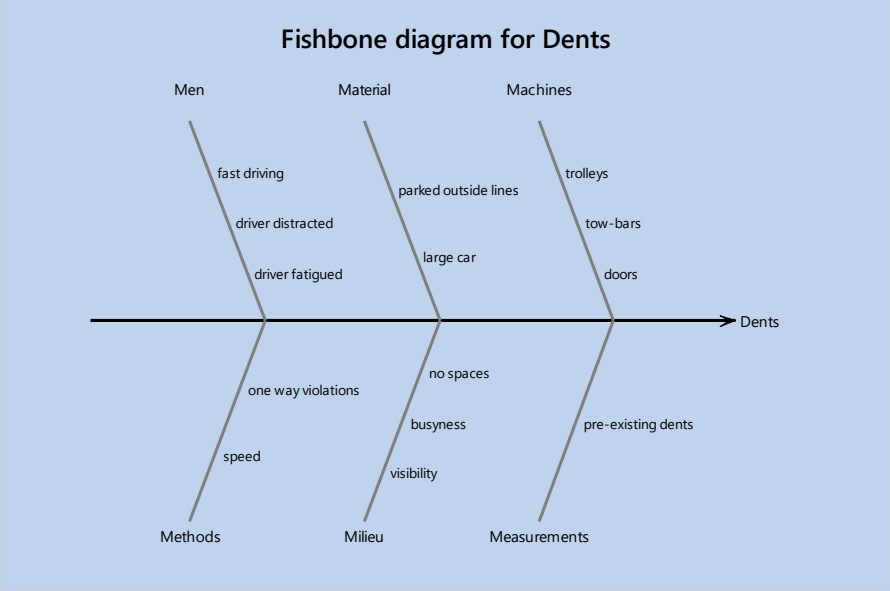
%R&R = 25.17% which is acceptable because it is less than 30%. This measurement system is able to distinguish between good and bad parts.

Answers 8E

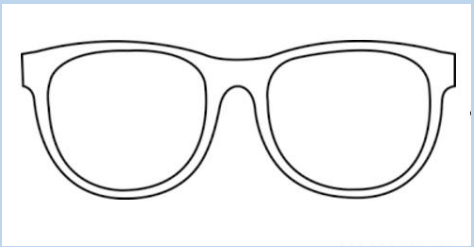
1. (a)



(b)



2. (a)



(b)

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10

3.
(a) Clusters: At a certain times there may be a particular problem with the service which affects many customers. If this problem can be resolved quickly then there will be a cluster of short duration calls. At another time there may be a more complex problem requiring longer calls to resolve.

(b) Mixtures: Customer calls will be concerned with either TV or broadband problems. Broadband issues may take longer to resolve.

(c) Trends: Operator fatigue could give rise to progressively longer calls.

(d) Oscillation: After spending a long time on one call, the operator may make a special effort to deal with the next call quickly in order to meet productivity targets.

4. (a) Check sheet (defect concentration diagram) based on a diagram of a car body with a check mark for every incidence of stone damage.
(b) Run chart.
(c) Pareto chart.

Appendix 5: Statistical Tables

Statistical Table 1

Normal Distribution: Cumulative Probability

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8079	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

The table provides the majority probability corresponding to any *z*-score, either positive or negative. To obtain the minority probability, subtract the tabulated value from one.

Statistical Table 2
Percentage Points of the *t*-distribution

df	5%	2.5%
1	6.314	12.706
2	2.920	4.303
3	2.353	3.182
4	2.132	2.776
5	2.015	2.571
6	1.943	2.447
7	1.895	2.365
8	1.860	2.306
9	1.833	2.262
10	1.812	2.228
11	1.796	2.201
12	1.782	2.179
13	1.771	2.160
14	1.761	2.145
15	1.753	2.132
16	1.746	2.120
17	1.740	2.110
18	1.734	2.101
19	1.729	2.093
20	1.725	2.086
21	1.721	2.080
22	1.717	2.074
23	1.714	2.069
24	1.711	2.064
25	1.708	2.060
26	1.706	2.056
27	1.703	2.052
28	1.701	2.048
29	1.699	2.045
30	1.697	2.042
40	1.684	2.021
60	1.671	2.000
120	1.658	1.980
∞	1.645	1.960

$z = t_{\infty}$

Statistical Table 3

Percentage Points of the Chi-Square Distribution

df	97.5%	95.0%	5.0%	2.5%
1	0.0010	0.0039	3.841	5.024
2	0.0506	0.1026	5.991	7.378
3	0.2158	0.3518	7.815	9.348
4	0.4844	0.7107	9.488	11.143
5	0.8312	1.146	11.070	12.833
6	1.237	1.635	12.592	14.449
7	1.690	2.167	14.067	16.013
8	2.180	2.733	15.507	17.535
9	2.700	3.325	16.919	19.023
10	3.247	3.940	18.307	20.483
11	3.816	4.575	19.675	21.920
12	4.404	5.226	21.026	23.337
13	5.009	5.892	22.362	24.736
14	5.629	6.571	23.685	26.119
15	6.262	7.261	24.996	27.488
16	6.908	7.962	26.296	28.845
17	7.564	8.672	27.587	30.191
18	8.231	9.391	28.869	31.526
19	8.907	10.117	30.143	32.852
20	9.591	10.851	31.410	34.170
21	10.283	11.591	32.671	35.479
22	10.982	12.338	33.924	36.781
23	11.689	13.091	35.172	38.076
24	12.401	13.848	36.415	39.364
25	13.120	14.611	37.653	40.647
26	13.844	15.379	38.885	41.923
27	14.573	16.151	40.113	43.195
28	15.308	16.928	41.337	44.461
29	16.047	17.708	42.557	45.722
30	16.791	18.493	43.773	46.979
49	31.555	33.930	66.339	70.222
50	32.357	34.764	67.505	71.420
99	73.361	77.046	123.225	128.422
100	74.222	77.929	124.342	129.561
119	90.700	94.811	145.461	151.084
120	91.573	95.705	146.567	152.211
149	117.098	121.787	178.485	184.687
150	117.985	122.692	179.581	185.800

Statistical Table 4
Upper 5% Points of the *F*-Distribution

df	1	2	3	4	5
1	161.4	199.5	215.7	224.6	230.2
2	18.51	19.00	19.16	19.25	19.30
3	10.13	9.55	9.28	9.12	9.01
4	7.709	6.944	6.591	6.388	6.256
5	6.608	5.786	5.409	5.192	5.050
6	5.987	5.143	4.757	4.534	4.387
7	5.591	4.737	4.347	4.120	3.972
8	5.318	4.459	4.066	3.838	3.687
9	5.117	4.256	3.863	3.633	3.482
10	4.965	4.103	3.708	3.478	3.326
11	4.844	3.982	3.587	3.357	3.204
12	4.747	3.885	3.490	3.259	3.106
13	4.667	3.806	3.411	3.179	3.025
14	4.600	3.739	3.344	3.112	2.958
15	4.543	3.682	3.287	3.056	2.901
16	4.494	3.634	3.239	3.007	2.852
17	4.451	3.592	3.197	2.965	2.810
18	4.414	3.555	3.160	2.928	2.773
19	4.381	3.522	3.127	2.895	2.740
20	4.351	3.493	3.098	2.866	2.711
21	4.325	3.467	3.072	2.840	2.685
22	4.301	3.443	3.049	2.817	2.661
23	4.279	3.422	3.028	2.796	2.640
24	4.260	3.403	3.009	2.776	2.621
25	4.242	3.385	2.991	2.759	2.603
26	4.225	3.369	2.975	2.743	2.587
27	4.210	3.354	2.960	2.728	2.572
28	4.196	3.340	2.947	2.714	2.558
29	4.183	3.328	2.934	2.701	2.545
30	4.171	3.316	2.922	2.690	2.534

The top row represents the numerator degrees of freedom, and the left column represents the denominator degrees of freedom.