CHAPTER

1

Investigating data

What you will learn

How often do people blink?

Blinking stops our eyes from drying out, by spreading tears across our eyeballs. With the help of eyelashes, blinking also removes dust and other small particles that might irritate the eye.

If you search for information on the rate at which people blink, you are likely to find a variety of figures, such as ‘10 times per minute’ or ‘approximately 20 to 25 times per minute’. Babies blink less because their eyes are smaller and they spend a lot of time sleeping. Children gradually increase their rate of blinking and teenagers reach the same rates as adults.

What affects our rate of blinking? If our eyes are focused on something, such as reading, our blinking might be slower. In contrast, stress and nervousness or fear may increase blinking. The numbers of blinks by politicians taking part in televised debates were counted and compared, and found to be much higher than ‘the average’. The rate of blinking is different under different conditions.

How could we collect data to investigate blinking rates? We would need to decide the conditions under which we are going to observe blinking. Do we need to keep these conditions the same for each of our ‘subjects’? If so, how? What questions or issues are we interested in investigating? <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Identify and investigate issues involving numerical data collected from primary and secondary sources **(ACMSP169)**

Pre-test

1 State whether these examples are categorical data, count data or measurement data.

- **a** The time it takes students to get to school
- **b** How long students can balance a book on their heads
- **c** Colours of cars
- **d** The number of times per minute a speaker says 'um'
- **e** Types of building materials
- **f** Brands of mobile phones
- **g** The number of pets per family
- **h** Distances between potholes along a road

2 Give some possible values for data collected in the following situations. In each case, say whether the data are categorical, count or measurement data.

- **a** Students are asked how many siblings (that is, brothers and sisters) they have.
- **b** People leaving a movie are asked to rate it.
- **c** In an experiment, how long people can hold their breath is recorded.
- **d** During storms, the numbers of lightning flashes are recorded.
- **e** In a sample of preschools, children's ages to the nearest month are recorded.
- **f** A sample of cricket ticket-holders are asked to name their favourite Australian cricketer.

3 A sample of students are asked which animal is their favourite for a pet. Which of the following graphs can be used to display these data and which are not appropriate?

a Column graph **b** Pie chart **c** Dotplot **d** Bar chart

4 A sample of students record the difference in their pulse rates before and after exercise. Which of the following graphs can be used to display these data and which are not appropriate?

a Column graph **b** Pie chart **c** Dotplot **d** Bar chart

5 A sample of students are asked how they come to school (car, bus, walk, train, cycle, other). Also recorded is whether a student is boy or girl. Name two ways these data could be presented in a report.

6 In a question on a survey, people are asked if they prefer a cat or a dog or neither as a pet. The responses are coded as 1 for a cat, 2 for a dog and 3 for neither.

- **a** What type of data are these?
- **b** What graph would you use to display these data?
- **c** What would be displayed in this graph?

Terms you will learn

categorical data
continuous variable
count data
data
experiment
experimental units
grouping
measurement data
observational units
observational study
ordinal variable
pilot study
primary data
quantitative data
random
grouping
measurement data
observational units
observational study
ordinal variable
pilot study
primary data
quantitative data
random
randomise
randomly representative data
recording sheet
secondary data
statistical data investigation process
statistical variable
subjects
survey
variation

1-1 What is the data investigation process?

How do statisticians investigate problems in the real world? They use the **statistical data investigation process**. This is how people who use statistics investigate questions in science, medicine, agriculture, business, engineering, psychology. It can be used anywhere that **data** need to be collected and where there is **variation**.

Statistical data investigation process: How real problems are tackled by statisticians and investigators conducting experiments, studies or surveys to obtain data for working statistically

Data: Information, facts, records and observations

Variation: The unpredictability of situations in which observations or measurements have different values or are not determined or not specified

Everything in nature involves variation. It happens for many reasons such as:

- people, animals, plants, materials and consumers, etc. are all different
- they don't behave or react in the same way
- conditions or circumstances change and cannot be made exactly the same all the time.

Because of variation, we need data – we need to make observations, to take measurements, to carry out **experiments**, to ask questions. Data helps us to see how much variation there is, and when it tends to be more or less, and what patterns there are in it.

The data investigation process starts with the first thoughts about what questions to ask. It continues through planning, collecting and exploring data, to reporting on its features. There are number of ways of describing and summarising the process. One is 'Problem, Plan, Data, Analysis, Conclusion (PPDAC)'; another is 'Plan, Collect, Process, Discuss (PCPD)'. The process provides a practical framework for tackling real problems statistically. It consists of these steps:

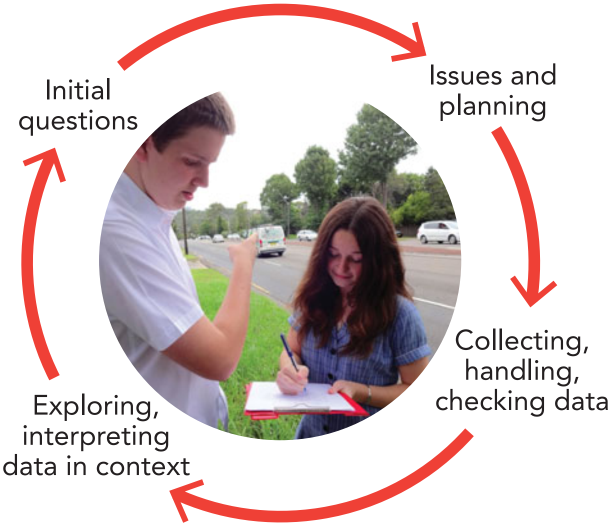
- Initial questions such as 'what do we want to find out?'
- Identifying issues (what will affect the data we collect) and planning
- Collecting, handling and checking data

Experiment: Data investigation in which investigators control conditions and measure the effect of these on some outcome(s) of interest

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press

- Exploring and interpreting data in context (in the circumstances of collecting it)
- Considering new initial questions for a further investigation.

The last step makes it a cycle, which is a process that continues to a new starting point. This can be represented by a diagram like the following.


Initial questions
Issues and planning
Collecting, handling, checking data
Exploring, interpreting data in context

In the rest of this section we look at the initial questions, and issues and planning for an investigation.

LET'S START What do we want to find out?

Sometimes a data investigation starts with a specific question, sometimes an idea or belief, sometimes a problem, and sometimes just a general situation to be investigated. Here are some examples:

- Do girls tend to blink more than boys?
- You'll see more ants in the house when rain is due.
- How well can people estimate time? Does reading or other activity make a difference?
- How do cyclists and pedestrians behave on shared paths?

Think of a question, idea, problem or situation you would like to investigate.

Key ideas

An investigation needs to be planned:

- to collect data to investigate questions or issues.
- to help us to explore any interesting or important information that might turn up.

The first step is to decide what is going to be investigated, so we can decide what data to collect.

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press

Below are some helpful questions to consider in this first step:

- What's of interest?
- What are we going to observe, record or measure?
- Is it possible or practical to collect the data we want?
- What conditions are we going to keep the same?
- Is there anything else we should observe or record in case it's useful?

Example 1: How often do people blink?

a Under the same conditions

Suppose we are interested in how often people blink under the same conditions. Our initial question is 'How often do people blink under the same conditions?'

Moving to issues and planning for this investigation, we need to ask how we collect data on this. Questions include:

- What conditions do we want?
- How can we keep conditions the same for all the people we observe?

We need to be able to observe people in the conditions we want and to count the number of times they blink. Staring at them while they are reading or watching TV might not be a good idea, as some people might notice us, and this would change the conditions. Also different people might notice us at different times, so keeping conditions constant would be very difficult.

We could perhaps count blinks for people being interviewed on TV but they'd need to be interviewed on the same topic!

Perhaps we could arrange it so that someone talks to the person being observed, while another person counts the number of blinks – without being obvious of course. We would need to have the conversation on the same topic, and keep the same people doing the talking and the observing. This could be thought of as mimicking an interview.

b Under different conditions

We may be interested in investigating the rate of blinking under different conditions. For instance we might want to compare the rate when someone is answering easy questions (e.g. what is your favourite food) to the rate when they are answering hard questions (e.g. add these three numbers in your head). This is much more difficult to plan. We would need to do the same type of thinking as in part **a** above for each set of conditions (easy or hard questions). We have the added problem that we would need to use the same subjects – that is, count blinks under each set of conditions for each person. And we would have yet another problem – do we give the same set of conditions first to each person each time? Or do we make the order **random**? We could do this by tossing a fair coin, or by using a spinner divided into two equal parts, to decide whether each person does the easy questions first or the hard questions first.

Random: Due to chance

Let's consider just one set of conditions as described in part **a**, with an 'interviewer' and an observer. What are some of the practical aspects of collecting the data? We need to keep the same person as the 'interviewer' and the same person as the observer –

ISBN 978-1-107-61421-5

Cambridge University Press

otherwise we are changing the conditions. We should keep the 'interview' on the same topics, i.e. each person gets the same set of questions. And of course, the people we pretend to interview should not know what our real purpose is. We also must choose a time interval (for example, a minute or 30 seconds) in which to count the blinks, and make sure we stick to that interval.

What else should we record in case it turns out to be of interest? Certainly we should record whether the person being interviewed is male or female. Should we perhaps record if they wear glasses or not in case this is important? It won't take any extra effort. Perhaps we should also record the time of day, or what each subject has just been doing. But we need to be careful not to make the recording of the data too difficult because that might interfere with the core aim: counting the number of blinks in the chosen time interval under our chosen conditions.

Example 2: How do people clasp their hands?

If you do an internet search on 'clasping hands', you will find a number of sources saying that (almost) everybody tends to clasp their hands the same way each time, with either the left or right thumb on top. Some sources say it is determined by a person's genes (information that you inherited from your parents that helps make you who you are), and that the left thumb on top is 'dominant' (to use a term from genetics). There are also sources saying that it is a myth that it is genetic, or that it is not a simple case of a dominant gene. However, there is general agreement that it is difficult to do it the other way to what you normally do. Try it and see.

One source suggests that it is more common to have the left thumb on top when clasping hands. Let's plan a data investigation to investigate this. <www.cambridge.edu.au/statsAC78weblinks>

In many ways this is an easier investigation to plan than the one about blinking in Example 1 above. This is because we need only ask people to clasp their hands without thinking about it.

Is there anything else we should consider or collect and record? Recording whether the subject is male or female is an obvious one. Another possibility to consider is whether the person is left-handed or right-handed. And this raises the interesting question of how do we define right-handedness and left-handedness. For some people, the right hand is dominant in some tasks and the left is dominant in others. For an investigation in which left-handedness or right-handedness is recorded, the definition chosen must be quite clear and clearly stated. And everyone involved in the data collection must agree and be consistent. One possibility is that the choices are either right-handedness or not, that is, at least one activity being left-handed.

ISBN 978-1-107-61421-5

Cambridge University Press

Exercise 1A

1 The topic for a data investigation is colours of cars. What are some practical issues in collecting data to investigate colours of cars? What else do you think should be recorded about each car?

2 The issue to be investigated is the amount of traffic on a main road near your school. Decide what you are going to observe, and any practical issues in your plans.

CAUTION
Safety must be considered when surveying near traffic!

3 The question is 'how often does the word 'magic' appear in *Harry Potter* books?' It is decided to investigate how often it appears per page. What are some practical issues in this investigation, and what else could be recorded in case it is useful?

4 The situation to be investigated is how students get to school – that is, what form of transport they use.

- **a** What is a practical issue in this investigation no matter what school, region or country we are considering?
- **b** Consider investigating this in your school. What are some practical issues, and what else would you record?
- **c** Check the CensusAtSchool website to see how this question is asked in their questionnaires. <www.cambridge.edu.au/statsAC78weblinks>

5 The question to be investigated is 'how long can students balance a book on their heads?'

- **a** What are the practical issues relating to the book to be used and the actual balancing activity?
- **b** Name at least two other pieces of data that could be recorded for each attempt at balancing a book.

6 How many shots does it take to win a point in tennis? To investigate this requires a number of decisions to be made in planning the investigation. Give one example of possible decisions for the following:

- **a** What is going to be observed?
- **b** Under what conditions are the data going to be collected?
- **c** What else will be recorded?

Enrichment How stretchy are jelly snakes?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

1-2 Types of data, variables and subjects

In statistics, it is very important to know what type of data you have. The main types of data are categorical, count and measurement.

Categorical data

In **categorical data** each observation falls into one of a number of distinct groups or categories. Such data are everywhere in everyday life. Some examples are:

- gender (male or female)
- direction on a road (right or left)
- type of dwelling (house, flat, room, caravan and so on)

Sometimes the categories are natural, such as with gender or direction on a road. Sometimes they are categories that we make up and describe such as type of dwelling.

Categorical data: Data that fall into categories that can be named or coded

Count data

Each observation in a set of **count data** is a count value or number. Count data occur in considering situations such as:

- the number of children in a family
- the number of vehicles passing in 2 minutes.

Count data: Record of a number of items, events, people and so on.

ISBN 978-1-107-61421-5

Cambridge University Press

Measurement data

All **measurement data** need units of measurement. Observations are recorded in chosen units of measurement. Some examples of measurement data are:

- reaction time in seconds
- age in years
- weight in kilograms of Year 7 boys.

Count data and measurement data must be described by numbers, so together they are called **quantitative data**.

All measurement data are recorded to the *nearest decimal place* – which could be whole numbers or even tens! This is usually determined by what is possible with the measuring device, or what we choose. Think a bit more about measurement data. When we say that someone is 162 cm tall, we don't mean exactly 162. We mean the height is in between 161.5 and 162.5; perhaps we mean between 161.5 and 162.4 if we are rounding 0.5 up to the next whole number. This type of situation is true for all measurement data which is really reporting data in small intervals. Sometimes these intervals are very small, and sometimes not; it depends on the measuring device or on decisions made by the data collectors/reporters. Sometimes the standard practice used for recording the data is different – for example, ages in years are not usually rounded up – and so the convention used should always be reported.

Measurement data: Data which need units of measurement. Observations are recorded in the desired units of measurement

Quantitative data: Measurement or count data; the numerical values of the data are actual quantities

Statistical variables

When we collect or observe data, the 'what' we are going to observe is called a **statistical variable**. Statistical variables are described by words. When we consider types of data, we are also considering types of variables. So all of the above examples are examples of statistical variables, and the type of data is also the type of variable.

Measurement data and their associated measurement variables are examples of what are known in statistics as **continuous variables** because we observe their data in small intervals. This is important when we consider the graphs appropriate for data on continuous variables, and also how probabilities are estimated for, or assigned to, continuous variables.

Subjects

When we collect or observe data, the records or observations are per person, or per time interval or per family or per car or per dwelling and so on. These are the **subjects** (or **observational** or **experimental units**) of the observations or records. In planning most data investigations the type of subject is obvious, but sometimes we can choose. For example, in Example 1, the observations are for two minutes for each person, but we could choose a different interval. If we are investigating advertisements on TV, we could take observations per advertisement or per advertisement break.

Statistical variable: The 'what' we are going to observe when we collect or observe data

Continuous variables: Variables which take values in intervals – typically, continuous data have values given 'to the nearest …'

Subjects (or **observational** or **experimental units):** Individuals or objects or entities on which observations are made

LET'S START What are our subjects and variables and their types in our plan?

Consider Example 1 to investigate how often people blink. The plan is to count the number of times each person blinks in two minutes. We are also going to record the person's gender and whether they wear glasses or not. We will keep the 'interviewer' and observer the same. So the subjects are people (we will look more closely at which people in section 1-3).

We have three variables:

- number of blinks in two minutes, which is a count variable
- gender of subject, which is a categorical variable
- whether the subject is wearing glasses or not, which is a categorical variable.

If we ask them what activity they were doing just before the 'interview', we have another categorical variable. This needs some thought. It might be best for such a variable just to ask people what they were doing and decide later how to group and describe the categories.

In the plan of Example 2, what are the variables, what type are they, and what are the subjects?

Key ideas

Identification of the variables, their types and the subjects is very important in planning a data investigation, and is also very helpful in doing the planning. If we don't do this the planning is not complete, and the investigation itself might be messy and confusing.

- In planning, what are we going to observe, record or measure? These are our variables.
- Who or what are we going to collect our data on? These are our subjects.
- For categorical data, each observation falls into one of a number of categories which can be named.
- If the categories of a categorical variable are given numerical codes rather than names, this does not mean the numbers have any numerical meaning. It does not change the fact that the variable is categorical.
- Each observation in a set of count data is a count value. Count data always have a specified entity – such as a location, time or space interval, or group – within or for which the count is made.
- All measurement data need units of measurement. Observations are recorded in the units of measurement we want, and to a selected number of decimal places (or units or tens or hundreds etc.).
- A measurement variable is a statistical continuous variable. Measurement data are observed or recorded in intervals. The size of the interval depends on the measuring device or investigator's choice.
- If there are many possible categories for a categorical variable; investigators need to choose and clearly describe categories. Categories can be grouped together after data are collected, if this is appropriate or useful.
- A continuous or a count variable may be changed to a categorical variable by **grouping** possible values in broad groups or groups of unequal sizes.

Grouping: When **measurement data** or **count data** are placed in specified groups of values, or when some categories of a categorical variable are combined, observations are grouped together

Example 3: Which are the popular movies?

In a statistical investigation of this question, what are the subjects and what variables and measurements are appropriate?

The subjects are movies. The popularity of movies can be measured in a number of ways. The overall amount of money made by a movie – the gross sales or 'takings' of the movie – is one measure used. Other possible measures are the takings on the opening weekend and the run length (how long the movie runs for). Other aspects of interest in investigating this topic could be the type of movie, the rating (by critics) and the budget if this is known. Much of this information is available from the internet.

When collecting data from the internet, we need to know exactly what is meant by each of these variables, and where the data were collected.

The variables, their units if appropriate, and their types, could be:

- **Gross takings of each movie.** The units are usually US dollars. It is a measurement variable.
- **Takings of the movie on its opening weekend.** Again, the units are US dollars, and is a measurement variable.
- **Run length.** This is most likely to be in days, but might be in weeks. It is a measurement variable. This needs care in identifying – is it overall run length or maximum run length at one cinema?
- **Type of movie (comedy, action, horror, and so on).** This is a categorical variable. The categories need careful identification and description. Some of them might be combined when it comes to exploring the data.
- **Rating.** This is a categorical variable, but the order of the categories matters. Ratings are usually given numerically (for example 3 stars, 3.5 stars) but the numbers give an ordering of ratings from lowest to highest. This type of categorical variable is called an **ordinal variable** and occurs frequently in surveys collecting opinions. Note that the ratings do not measure some kind of distance; the 'distance' between a 3 and a 3.5 rated movie is not meant to be the same 'distance' as between a 3.5 and a 4 rated movie. Again this variable requires care in identifying – is it average rating across a number of critics, or the rating of a certain critic or critics' website?
- **Budget.** This is not usually as readily available as other data, but is a measurement variable, with units usually in US dollars.

Ordinal variable: Categorical variable for which the order of categories has meaning

Cambridge GO

Exercise 1B

1 A questionnaire for an investigation of households asks each one:

- What is the total number of people currently living there? (That is, it is their main place of residence.)
- What is the number of people in the household under 18 years of age?
- What is the gender and age in years of the oldest resident?

Identify the subjects, variables and their types in this investigation.

2 In each of questions 1 to 3 of Exercise 1A, identify the subjects and variables you chose in your planning. Identify the types of variables, giving (possible) categories for any categorical variables.

3 The distance between fingertips when arms are spread wide is called a span. Children's spans are approximately the same as their heights when they are small. Their spans gradually increase to more than their heights as they grow. Data are collected on spans (in cm) and heights (in cm) for boys and girls aged 4 to 18 years. Identify the subjects and variables. Identify the types of variables, giving (possible) categories for any categorical variables.

4 A statistical investigation of non-fiction books is carried out. Data are collected on their:

- price in dollars
- total number of pages
- topic
- whether their covers are hard or soft
- whether they have colour pictures or not.

Identify the subjects and variables. Identify the types of variables, giving (possible) categories for any categorical variables.

5 Coffee prices are being collected across a city. The prices of cappuccinos and flat whites are collected at cafes and restaurants in the centre of the city and in the suburbs. Identify the subjects and variables. Identify the types of variables, giving (possible) categories for any categorical variables. ★

Enrichment

How long can you hold your breath?
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press

1-3 Collecting data

Recording sheets

To collect data, we need a **recording sheet**. Whether we write it by hand or use a computer spreadsheet, the design should be the same. So for Example 1 on investigating how often people blink, each subject has a row and each variable has a column in a recording sheet or spreadsheet. The first few lines of our recording sheet might look like the following:

Recording sheet: A table or spreadsheet to record data; each variable has a column and each subject (or experimental or observational unit) has a row

Subject	Gender	Glasses	Previous activity	Number of blinks per minute
Tom	Boy	No	Reading	14
Elise	Girl	No	On computer	13
Sophie	Girl	Yes	Watching TV	24

Note that the only reason for recording names is to be able to check data for errors, and to make sure that no subjects are recorded twice.

Even if you think you have identified your variables, subjects and the categories of any categorical variables, designing your recording sheet and putting in a few test entries will help to see if anything has been forgotten.

ISBN 978-1-107-61421-5

Cambridge University Press

Pilot study

A **pilot study**, experiment or survey, is a trial run to check the plans for an investigation. These are the kinds of things that can be checked:

- The method of collecting the data will work
- The design of the recording sheet works
- The data can be collected under the planned conditions
- In an experiment, whether conditions need to be changed or better controlled
- In a **survey**, whether questions or situations need to be more carefully or clearly expressed or described
- In an **observational study**, when observations sometimes have to be made quickly, whether more help is needed
- Whether the plans need rethinking.

Pilot study: An initial trial of the investigation or preliminary experiment to check the practicalities of the planned collection

Survey: Asking questions of subjects with the data being responses; the subjects of a survey may be individuals, or groups such as companies, businesses or households

Observational study: Data investigations in which investigators observe subjects without altering or controlling conditions

Randomly representative data: Data obtained at random from a more general situation or population

Randomly representative data

Why are we collecting data? This question is at the heart of all statistical data investigations. The answer is because we want to find out information about a question or topic(s) or issue(s). That is, we want to be able to use our data to comment on a situation represented by our data.

Suppose our question is 'which sport is most popular among students in Year 7?'. We don't just ask students in one school. In another school or state or region the situation might be very different. If we want to ask 'which sport is most popular among students in our school?' and we don't have time to ask them all, we have to randomly choose students to ask. We won't get data representative of the whole school if we ask only the friends we play our favourite sport with. When we randomly choose subjects to collect the data, we can say that our data are **randomly representative** of the situation.

Suppose you want to test if fertiliser makes tomato plants grow faster. You divide your plants into two groups: group A gets fertiliser, group B gets none, and all other conditions are kept the same. You have to randomly allocate (give) test plants to each group. Your data would not be randomly representative if you put all the strong and healthy plants in group A and the smaller weaker ones in group B.

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press

You will see more of these important concepts as you learn more about statistics, but:

- when you collect data, try to collect it randomly within the conditions you have chosen <www.cambridge.edu.au/statsAC78weblinks>
- when you are exploring data or commenting on data, ask if the data can be assumed to be randomly representative with respect to the question(s) or issue(s) of interest.

Primary and secondary data

The data about movies referred to in Example 3 have been collected by other people or organisations and reported on the internet. This is an example of **secondary data**. Data we collect ourselves are called **primary data**. Notice that there are comments in Example 3 about finding out how the data are defined and/or collected. Whether we collect data ourselves, or use data collected by others, the same principles apply: the variables should be clearly identified, and exactly how, when and where the data are collected should be clear and fully reported.

Secondary data: Data collected by others

Primary data: Data collected by the investigators

LET'S START Clasping hands and folding arms

Consider Example 2 about investigating which thumb is on top in clasping hands. We've already considered that gender should also be recorded, and right-handedness or not might be of interest. People tend to fold their arms the same way each time – perhaps this could be related to handclasping. So we might also ask people to fold their arms and record which arm is on top. Now design the recording sheet.

Who are we going to ask? Because the way we clasp our hands is a characteristic that is particular to us, it does not tend to change, or to change with age. So trying to collect randomly representative data is not too difficult for this investigation, and any group of people can be assumed to be randomly representative. However, if we include folding arms, there is another consideration – what do we ask people to do first? Either we do the same for everyone or we **randomise** the order. What should we now include on our recording sheet? These practical questions are sometimes referred to as the practicalities of the investigation.

Randomise: To make random

Once the plan is done, including the design of the recording sheet and how subjects will be chosen as well as considering practical issues, carry out a pilot study. Does this tell us if there is anything that should be changed before carrying out the data collection?

The question of how many observations to collect is a big question in statistics. It takes very many observations to accurately estimate something like the proportion of people who have their left thumb on top in clasping their hands. This is why there is so much variation in the studies that have been done. <www.cambridge.edu.au/statsAC78weblinks>

Key ideas

This is a summary of things that you need to do and questions you may need to ask in planning an investigation.

- Choose the topic or issues to be investigated.
- Identify and describe the variables to be observed or recorded.

- Design the recording sheet.
- Work out how to choose subjects to be randomly representative of the situation in which we are interested.
- Consider the practicalities of collecting the data and decide exactly how to collect the data you want.
- Decide if there anything else you should observe or record in case it's useful.
- Modify the recording sheet if necessary.
- Have we forgotten anything?
- Should we do a pilot study? (The answer is yes, almost certainly!)
- What has our pilot study told us?
- Changes or problems suggested by the pilot study can then be allowed for, and the data collection plan carried out.

Example 4: How quickly does a vitamin tablet dissolve?

Some vitamin tablets are soluble. They dissolve in water, with fizzing and bubbling. A group of students decided to investigate how quickly they dissolve. They chose two brands of soluble vitamin C tablets having the same amount of vitamin C.

They decided to use three types of water (tap, filtered and mineral water). The tap water all came from the same tap, the filtered water came from the same source and the mineral water was all the same brand.

They decided to do the experiment with room-temperature water and cool water so they divided each type of water into two batches. One batch was kept in a room at 24°C and the other batch was stored in a cool place at 15°C. These were the experimental conditions.

The students decided to dissolve four tablets of each brand in each combination of experimental conditions. This meant altogether in their experiment they dissolved:

4 tablets × 2 brands × 3 types of water × 2 temperatures = 48 tablets

Because they were measuring the time the tablet took to dissolve, the students needed to decide how they would define when the tablets were completely dissolved. They chose to say it was dissolved when they could no longer see any solid bits of tablet. They used the same stopwatch, the same quantity of water each time, and the same person recorded the times in seconds.

Their subjects were tablets, and their variables and variable types were:

- brand of tablet – categorical
- temperature – measurement but because only two selected values, can be treated as categorical
- water type – categorical
- time to dissolve in seconds – measurement.

ISBN 978-1-107-61421-5

Cambridge University Press

Their recording sheet looked like this (only the first few rows are shown):

Tablet	Brand	Water temperature	Water type	Time to dissolve (in seconds)
1				
2				
3				

After conducting their pilot study, they decided that the same two people should declare the tablet dissolved.

They chose the test tablets at random from the packets they'd bought, and they tested the combinations in random order. Even though they were careful in keeping other conditions constant, randomising the order ensures there was nothing else unintentionally introduced into the experiment. They did this by tossing a six-sided die. To choose which brand of tablet for the first test, they tossed the die, with an even number meaning brand A, and an odd number meaning brand B. The same rule is used to decide the other variable with two categories – temperature. For water type with three categories, throwing a 1 or a 2 means tap water, a 3 or a 4 means filtered water and a 5 or a 6 means mineral water. That means that of the 48 tablets in the investigation there were:

- 24 of each brand (48 tablets ÷ 2 brands)
- 24 tested at each temperature (48 tablets ÷ 2 temperatures)
- 16 tested in each water type (48 tablets ÷ 3 types of water).

For full details of ways of randomising the order of experiments, see Cambridge GO.

Exercise 1C

1 Data were collected at a bus station to investigate people's use of the lifts. The bus platforms can be accessed only from an overbridge, which can be reached by stairs or a lift for either direction. During the morning and evening peak times on two weekdays (Tuesday and Thursday), the choice of lift or stairs, the direction (up or down), and gender (male, female) of all passengers were recorded. At the end all the observations from all the recording sheets were combined in an overall spreadsheet.

a Give the column headings of this spreadsheet.

b Suggest at least one problem that may have shown up in a pilot study.

c What other information would you like to know about the data collection?

d Can we reasonably assume that the data is randomly representative of the situation being investigated? What further information is needed to be able to judge this?

2 In a study of usage of a new pedestrian and cycle bridge, the numbers of pedestrians and cyclists travelling in each direction were recorded for morning and afternoon periods of 15 minutes each, on each day of a week.

a Describe the recording sheet, including the subjects and the variables.

b Suggest any further information that should be provided about the data collection.

c Suggest any practical problems.

3 During the study described in question 2, in the same time periods, a second data collection was made. Pedestrians and cyclists were selected at random, and their direction of travel, their speed and gender were recorded. Perform the steps **a**, **b**, and **c** described in question 2 on this second data collection.

4 In an experiment to investigate whether colour makes a difference to how long candles burn, candles of four different colours from one brand were chosen at random from a packet. Their lengths and diameters were measured. They were then placed upright in the same environment with no wind, lit and allowed to burn for 5 minutes before being extinguished. Their lengths were measured again.

a What would the recording sheet look like?

b What else could the experimenters do in carrying out their experiment to make sure no other effects (such as human factors or variables) were unintentionally included in the experimental conditions?

5 Your school wants to find out what families want for a school newsletter, including:

- whether they want one weekly or fortnightly
- whether they want to have the option of receiving it online as well as in print
- if they want each child to receive a copy or just the eldest child.

a Write the questions for this survey, giving the choices of responses.

b How do you suggest the survey be carried out?

c Is there another question you'd like to include?

Enrichment

The ABS survey on Children's Participation in Cultural and Leisure Activities
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press

Chapter summary

What is the data investigation process?

- Initial questions and issues
- Planning, designing
- Collecting, handling and checking data
- Exploring and interpreting data in context
- Considering new questions and issues to be investigated.

What's of interest?

- What to observe, record or measure?
- Can we collect the data we want?
- What conditions to keep consistent?
- Anything else we should observe or record?

Types of data and variables

- Categorical data
  - Observations fall into distinct named categories
  - Categories may be combined if appropriate
  - Numerical codes are just labels.
- Count data
  - Observations take count values
- Measurement data
  - Observations have units of measurement and a number of decimal places
  - Measurement variables are examples of continuous variables.

Subjects and observational or experimental units

- The people, objects or items on which data is collected
- Chosen or allocated as randomly as possible.

Recording sheet

- Each variable has a column
- Each subject or observational or experimental unit has a row.

Pilot study

- Trial collection plans
- Adjust plans if necessary.

Primary and secondary data

- Primary data are collected by investigators
- Secondary data are collected by others
- For both, must know or report *what*, *how*, *when*, *where*.

Multiple-choice questions

1 To investigate people's views on daylight saving, which of the following is **not** likely to be of interest in a survey?

A Their gender
B Where they live
C What they had for breakfast
D Their age

2 To investigate whether people jaywalk, which of the following is **not** likely to be of interest?

A Their gender
B What pets they have
C Their occupation
D Their age

ISBN 978-1-107-61421-5

Cambridge University Press

3 To investigate a new plant species for growth, which of the following is **not** part of the data to be collected?

A The soil **B** Where it comes from

C Who is taking care of it **D** Its name

4 What type of variable is the number of goals scored in a football match?

A Categorical

B Count

C Measurement/continuous

5 What type of variable is the variety of a tree?

A Categorical

B Count

C Measurement/continuous

6 What type of variable is the age in years of a tree?

A Categorical **B** Count

C Measurement/continuous

7 In a poll, people are asked to choose from the following for their approval rating of a new mayor: 1 = strongly disapprove, 2 = disapprove, 3 = neutral, 4 = approve, 5 = strongly approve. What type of variable is the response to this question?

A Categorical **B** Count

C Measurement/continuous

8 What type of variable is the price per kilogram of apples?

A Categorical **B** Count

C Measurement/continuous

9 What type of variable is the number of accidents per week in a state?

A Categorical **B** Count

C Measurement/continuous

10 In collecting real-estate data for properties sold, its selling price, suburb, land size, number of bedrooms, bathrooms and car spaces are recorded, and whether it has a pool or not. What are the subjects of these data?

A Properties **B** Sales

C Suburbs **D** Buyers

Short-answer questions

1 In collecting daily data on the weather at a chosen location, three of the variables recorded are maximum temperature, wind direction and amount of rainfall.

a Give two other variables that may be of interest.

b State the types of all five variables.

2 The colours of some small sweets are often of interest because people like different colours.

a Below are two possible questions of interest. How would you suggest collecting data for each of these?

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press

  - **i** The topic of interest is the allocation of colours by the manufacturer – that is, in what proportions are the different colours made?
  - **ii** The topic of interest is the number of blue sweets in the smallest packets of sweets produced by the manufacturer – that is, how many blue sweets are in each of the smallest packets of sweets?

- **b** What are the subjects in each of the above?

3 Data on the usage of computer games by Year 7 students are to be obtained. It is decided to ask approximately how long (in hours) a student spends playing computer games in a week and the gender of the student.

- **a** Give at least two other variables that may be of interest.
- **b** State the types of all variables.
- **c** Suggest how to choose a sample of students to be asked to obtain a reasonably randomly representative group.

4 Some music fans think that songs of some styles tend to be longer than those of other styles. To investigate this, they look at the top 100 songs on a particular chart, and collect the length of each song (in minutes and seconds), the genre of the song, and the nationality of the performer(s) and whether the song is by a solo artist or a band.

- **a** State the types of all variables.
- **b** Suggest no more than eight names for classifying song genre. Suggest groupings to change that to five.
- **c** Suggest no more than four names for classifying nationality.
- **d** Suggest a problem in using the top 100 songs on a particular chart.

5 Students decide to do an experiment to investigate the performance of three different paper plane designs. They decide to use two different types of paper for each design. Two students each made four planes for each design and each paper. These groups of four planes were divided into two, and thrown by two throwers in a large indoor space. The students recorded the flight time (in seconds), the distance travelled by the plane (in cm) and whether it landed upright or not.

- **a** What was the total number of throws?
- **b** Name the variables and state their types.

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press

c As you can see, the students have designed this experiment very carefully and thoughtfully. What else should they do in carrying out the experiment to avoid introducing any unwanted effects?

d Show what the data recording sheet would look like by giving the headings and a few possible rows of data.

6 Real-estate data are collected on houses sold. For each house sold, its selling price, region, land size, numbers of bedrooms, bathrooms and car spaces are recorded, and whether it has a pool or not.

a State the types of all variables.

b It is decided to record numbers of bedrooms as <3, 3, 4, 5, >5, and bathrooms as 1, 2, 3, >3. How does this change your answers to part **a**?

c Show what the data recording sheet would look like (for part **a**) by giving the headings and a few possible rows of data.

Extended-response questions

1 The topic for a data investigation is family pets. Below are two possible main aspects to investigate. For each of these, identify practical issues and how you would handle them, and anything else you would choose to record.

a The number of pets a family has

b Whether they have cats and dogs as pets

2 Reactions can be measured in many different ways. A classic way is catching a ruler that is dropped. See also the CensusAtSchool questionnaire(s) to see how reactions are measured. This question in CensusAtSchool is an international common question that may be included in the CensusAtSchool questionnaires of the United Kingdom, Canada and New Zealand. <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

a It is decided to measure reactions by catching a ruler that is dropped, and measuring the distance on the ruler where it is caught. List at least two practical considerations and at least two other pieces of data that should be recorded. Choose and name your variables, state their types, and show what the data recording sheet would look like by giving the headings and a few rows of possible data.

b Cambridge GO gives two internet games to test reaction time. Write at least two instructions for players to help provide consistent conditions for investigating people's reaction times. <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

3 Find a news story that reports information from a survey.

a Does the report say how many observations were collected?

b Does the report say exactly how the data were collected?

c Does the report give all the questions that were asked?

d Do you think the report is justified in its statements?

e Do you think the report gives sufficient justification for its statements?

ISBN 978-1-107-61421-5 © MacGillivray & Petocz Cambridge University Press