M1 Maths

S2-1 Data Collection

- data and surveys
- extracting data

Summary Learn Solve Revise Answers

Summary

If you want to find out something about a large group of people or things (the population), you can do a survey. You might ask or observe the whole population – this would be a census. But more commonly, you would ask or observe a sample of them. This survey would provide raw data, which you would then process to get the information you want.

When doing a sample survey, it is important to get a representative sample by choosing an appropriate sample size and avoiding sample bias. It is also important, in both a census and a sample survey, to word questions in a neutral way.

Sometimes the data you need has already been collected by someone else. In this case, you just have to get hold of it. This is called extracting the data.

Learn

Surveys

Imagine you are working for Toyota. They have brought out a new model of car and are going to manufacture 40 000 of them for sale in Australia. They will come in 6 colours: black, white, silver, blue, green and red. Your job is to decide how many of each colour to make.

You need to know what proportion of the people in Australia who buy cars prefer each colour, so that Toyota knows how many of each colour to make. To find out, you do a **survey**: you ask people which colour they would prefer for a car and you record their answers.

The answers you get are **data**. You then **process** this data by working out what percentage of the people you asked prefer each colour. This turns the data into **information**. Toyota then uses this information to decide how many of each colour car to make. For example, if you asked 200 people and 60 preferred black, this would be 30%. So Toyota would make 30% of the cars black. 30% of 40 000 is 12 000.

Census vs Sample Survey

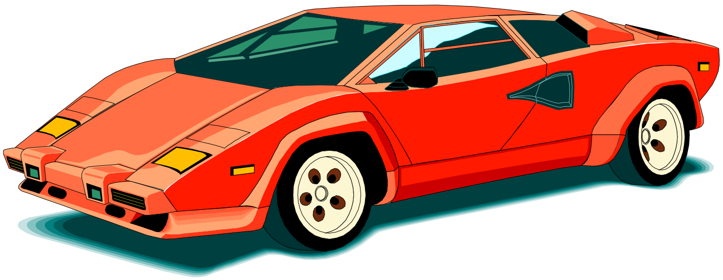
The people you want to know about are the car buyers in Australia. Car buyers include most adults between the ages of 17 and 80 with driving licenses. There are millions of them. This group of people is called the **population** for this survey. In a survey, the population is everyone you are interested in knowing about in this case all several million of them

If you asked the whole population, your survey would be a **census**. But if you don't have time to ask the whole population, you can do a sample survey. To do this, you would choose a **sample** of say 200 people and ask them, hoping that their answers will give you a reasonable idea of what percentage of the population as a whole prefer each colour. Samples which do have similar percentages to the population are called **representative samples**. The people you ask in a survey are sometimes called **respondents**, because they respond to your questions.

When doing a sample survey, there are three things that you must keep in mind: **sample size, sample bias and question wording**.

Sample Size

If you do a sample survey, asking just three people about their colour preferences wouldn't give you much of an idea about the population as a whole – it's quite possible none of your three would choose red, even though we know red is quite a popular colour. (Everyone knows red cars go faster.)



Asking 200 is a good compromise between asking millions of people and spending half your life doing it and asking just a couple and getting a fairly useless result. The number of people you ask is called the **sample size**. The choice of sample size is generally a compromise between accuracy and easiness.

If you do a census, of course, sample size isn't an issue: your sample has to be the whole population.

Sample bias

To be representative, a sample needs to be **unbiased**. Suppose you asked just members of bikie gangs and funeral directors. You would probably get a bigger percentage choosing black than you would from the population as a whole. If you asked mainly young men, you would probably get more choosing red than you would from the whole population. Both these samples would be **biased**.

You have to use a method of selecting your sample such that everyone in the population has the same chance of being chosen. You have to be careful not to make it more likely that someone is more likely to be chosen because they are a particular type of person, like a young male, an old person, a bikie or an office worker. The proportion of each type of person in your sample should be similar to the proportion in the population.

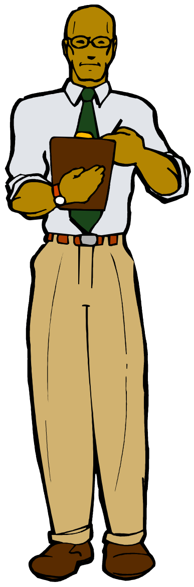
One way to do it might be to choose the first person on each of the first 200 pages of the phone book and ring them. This method should produce a reasonably representative sample of the population. Though, it could be biased towards older people because younger people are more likely not to have a land line and so not be in the phone book.

It is actually quite difficult to get a sample which is completely unbiased. We have to do the best we can and try to avoid anything which will obviously cause a major bias.

Again, if you do a census, sample bias isn't an issue.

Question Wording

Your questions need to be worded in a **neutral** way – a way that doesn't lead the respondent to be more likely to answer a particular way.



Showing a list the six colours would be neutral. Showing the actual colours would also be neutral and would probably be better because a respondent might like green, but not like the shade of green planned for the cars.

Asking 'Would you prefer this elegant silver car if it didn't cost any more than that yucky green one that looks like sheep vomit?' would probably make it more likely that the respondent would choose silver. Such a question is not worded in a neutral way.

Questions need to be neutral whether you are doing a census or a sample survey.

Processing the Data

As mentioned, the answers that you get when you do the survey are called *data*. The data might look like this:

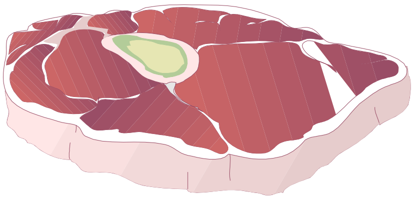
black, red, black, green, silver, red, black, green, green, blue, red, ...

The word *data* is actually the plural of *datum*. Each answer (colour) is a datum and the 200 together are data. However, most people treat data as a singular word and say things like ‘the data is reliable’ rather than ‘the data are reliable’. Either is acceptable, though.

To make it easier to use, data is generally processed in some way to present the results of the survey in a way that can be understood quickly and easily. The car colour results might be processed by counting the number of each colour in the list and presenting the numbers in a table like this:

Colour	Number	Percentage
Black	44	22%
White	52	26%
Silver	36	18%
Blue	19	9.5%
Green	11	5.5%
Red	38	19%

The processed data is sometimes called information, because it is in a form which is easily understood and tells us something useful. It informs us. Unprocessed data is sometimes called raw data to stress the fact that it hasn’t been processed.



Raw meat. Not to be confused with raw data.

The information in the table tells us at a glance that white is the most popular colour and that therefore a lot of the 40 000 cars should be white. We can go further than that and say that about 26% of people preferred white (52 out of 200 is 26%), and so 26% (i.e. 10 400) of the cars should be white. In the same way, it tells us how many of the 40 000 cars should be each of the other colours.

Data Record Template

Before collecting the data, you need to have a sheet on which to record it. This sheet is sometimes called a **data record template**. The data record template for the car colours would be quite simple, maybe just 200 cells in a column or table like the one below only longer.

We could then fill it in like this as we get answers:

black	red	black	green	silver	red
black	green	green	blue	red	

Or, to make the recording a bit quicker, we could use abbreviations and have a key to tell us what the abbreviations are, like this:

B = black S = silver U = blue G = green R = red P = pink

B	R	B	G	S	R	B	G	G	U	R	

Data from Observations

Another way to conduct the survey to find out how popular each of the colours is might be to observe cars driving down the street. You could sit by the road and write down the colour of each car that passes (ignoring the ones that are yellow or purple or any colour other than the 6 colours Toyota has chosen). We could still use the same data record template and we could still process the raw data in the same way.

So data can be answers to questions or observations.

The survey of car colours required a very simple data record template. But suppose we wanted to know the most popular TV programs for males and females of various ages. We might ask people 'What is your favourite TV program?' We might ask say 30 males and 30 females under 10 years of age, 30 males and 30 females from 10 to 19 years of age, 30 males and 30 females from 20 to 29 years of age and so on. This would be called a **stratified sample** because we are taking fixed numbers from each of a number of different subsets of the population.

The start of our data record template might then look like this:

0-9		10-19		20-29		30-39		
M	F	M	F	M	F	M	F	M

We might decide that the columns aren't wide enough to write the name of the TV program. A better design might be to have separate sheets for each age group, the sheets looking like this:

<10	
M	F

When asking someone, we would then just choose the correct sheet to write their response.

You might process the data by taking each of the categories (e.g. Females 10-19) in turn, counting the number of votes for each program and listing the three with the most votes. The results (processed data) might then look like this (with, of course, more rows).

	1st	2nd	3rd
Males <10	SpongeBob	Sesame Street	Bob the Builder
Females < 10	Sesame Street	Peppa Pig	Dora the Explorer
Males 10-19	Friends	Big Bang Theory	Pretty Little Liars

In summary, you do a survey if you want to find out something about a large group of people or things (the population). You might ask or observe the whole population – this would be a census. But more commonly, you would ask or observe a sample of them. You have to make sure you have a suitable sample size and an unbiased sample. If asking questions, you also need to make sure the questions are neutral. Your survey provides raw data, which you then process to get the information you want.

Practice

Q1 Match the words 1 to 12 with the meanings a to l.

1 population
2 survey
3 census
4 sample
5 data
6 raw
7 processed
8 representative
9 biased
10 data record template
11 respondent
12 neutral

a a survey of the whole population
b organised to make it easy to see the information it contains
c all of the people or things you wish to know about
d a sample which is quite like the population in terms of the proportions giving different answers or observations
e a part of the population chosen for a survey
f something on which to record answers or observations in a survey
g asking questions or making observations to find out something about a large group of people or things
h data in the form in which it is collected
i data which is not representative of the population
j answers or observations collected in a survey

k asked in a way that doesn't lead the respondent to answer a particular way

l a person who is asked questions in a survey

Q2 Imagine you want to know the most common makes of car on the roads in your area.

(a) What is the population?

(b) What would be a good sample size for a survey.

(c) Design and draw up a suitable data record template.

(d) Conduct the survey. [If that is difficult in your situation, you can make up the answers and write them into your data record template.]

(e) Process the data and present it in a form that shows the makes in order from most to least common.

Q3 Imagine you want to know the favourite colour of boys your age and of girls your age.

(a) What is the population?

(b) What would be a good sample size for a survey.

(c) Design and print a suitable data record template.

(d) Conduct the survey. [If that is difficult in your situation, you can make up the answers and write them into your data record template.]

(e) Process the data and present it in a form that shows the three favourite colours for each sex.

Extracting Data

Sometimes, you might need data to answer a question, but suitable data already exists because someone else has collected it – maybe to answer the same question.

For example, you might want to know the ages of cars on the roads. You could of course do a survey, but knowing the age of a car just by looking at it is not always easy. But the Department of Transport would have that data and (if you had a good case for wanting it), they might let you access it.

Accessing data that someone else has put together is called **extracting the data**. This is sometimes an alternative to conducting a survey.

Practice

Q4 Use the Internet to find rainfall data (how many millimetres of rain fell each day) for the past year at a location near you. A good source for Australia is the Bureau of Meteorology website (**www.bom.gov.au**). Click Climate Data Online near the bottom of the page.

Solve

Q51 When conducting a survey, it is important that people answer your questions truthfully. If you ask people which colour car they prefer, they won't generally lie to you. But suppose the question was 'Have you ever stolen anything from a shop?' In this case, people who have done so won't always want to admit it and might lie.

There is a way, however, of getting fairly reliable results for surveys which involve sensitive questions like this one. What you do is give the person you are questioning a die. Then ask them to roll it without you seeing. Then ask them to answer truthfully if they get a 1, 2, 3 or 4 and to lie if they get a 5 or a 6. That way, they can answer your question without you knowing whether they have actually stolen anything.

If no one in your sample had ever stolen anything, you would expect $^{1}/_{3}$ (33%) to answer Yes and $^{2}/_{3}$ (67%) to answer No. If everyone was a thief, you would expect $^{2}/_{3}$ (67%) to answer Yes and $^{1}/_{3}$ (33%) to answer No. If say 42% answered yes, then you can work out the approximate percentage of thieves as
$(42 - 33) \div (67 - 33) = 0.26$ or 26%.

What would be the approximate percentage of thieves if 55% answered Yes?

Revise

Revision Set 1

Q61 Re-do Practice Question Q1

Q62 Imagine you want to know the favourite singer or band of boys and of girls in your grade at school.

(a) What is the population?

(b) What would be a good sample size for a survey.

(c) Design and print a suitable data record template.

(d) Conduct the survey. [If that is difficult in your situation, you can make up the answers and write them into your data record template.]

(e) Process the data and present it in a form that shows the favourite singers/band for each sex.

Answers

Q1 1c 2g 3a 4e 5j 6h 7b 8d 9i 10f 11l 12k

Q51 65%