Fact Sheet

Australian Bureau of Statistics - Steps in running a survey

Further information is available on the [education section of the ABS website](http://www.abs.gov.au/websitedbs/CaSHome.nsf/Home/Senior%2BYears%2BResources).

Step 1: Planning a survey

a) Identifying your question

One of the first things you need to clarify when designing a survey is exactly what you want to find out. Start by writing your question as clearly as you can. Include as much detail as possible so that everyone else will interpret the question in the same way as you. For example, if you wanted to find out "What times do students get up in the morning?" you would need to clarify:

* Is it a normal school day, a weekend or a holiday e.g. “What time do students get up on a normal school day?”
* Will it matter if students have different school starting times? Compare the question “How long before school starts do students get up?”
* What units do you want to use to collect the data? “How many minutes before school starts do students get up?”
* How will part units be reported? Do you want data to the closest whole number? If you plan to have fractions can these be decimals?

You will also need to state some definitions:

* ‘student’ is defined as Year 7 and Year 11 Australian school students
* ‘getting up time’ is the time students get out of bed on a school day.

It is important that you maintain these definitions throughout your investigation and in any report. If your question is not clearly defined, the participants in your survey may interpret the question differently and your results won't be accurate.

b) Deciding who to include in your sample

Next you need to specify the scope of your sample. For example, are you looking at a particular age group or year level or location? In these cases, your question might be “How many minutes before school starts do Year 7 students get up compared with Year 11 students?” or "How many minutes before school starts do students in Queensland get up compared with students in South Australia?".

Sample size

Estimates are made about the total population and subgroups based on the information from the sample. Generally, larger samples will give a more accurate representation of the population. However, it can be difficult to obtain accurate information on smaller groups within the population if the sample size is small.

In addition, the level of accuracy can usually be measured. There are formulae to determine the size of the sample that should be taken depending on the level of confidence required. One of the simplest is:

Sample size = √n; (where n is the size of the population)

Randomness

To allow predictions to be confidently made about the total population, samples need to be randomly selected as well as of sufficient size. For data to be selected randomly, each data item must have the same chance of being selected as any other. Pulling data items from a hat or using the random number generator on a calculator are common ways of ensuring that data are selected randomly. Data not selected randomly may be biased towards a particular outcome.

Types of Sampling

There are a number of ways that a sample can be randomly drawn from a population. For example, you may want to ensure that each subgroup of a population is represented in the same proportion as in the general population.

Step 2: Collecting data

Once you have decided on your question, how many people will be in your sample, and have randomly selected them, you will need to consider how to collect the data. Will it be through an interview or will you collect written responses? The data you collect will also need to be in a form that is easily organised in order to analyse it. For example, you need consistency in units and fractional answers so request that the data be recorded to the nearest centimetre or half centimetre.

Interview

In an interview, a participant can ask questions if they haven’t understood something.

Written response

Written responses can be completed by many participants at the same time and are quicker than interviews.

Variables

A variable is any measurable characteristic or attribute that can have different values for different subjects – for example, eye colour, distance from school etc.

Characteristic is another way of saying variable. For example, height, age or countries of birth are all characteristics or variables of people.

Step 3: Organising data

After you have collected the data, it needs to be organised so that it is useful and ready to display.

Frequency tables

A useful way to record raw data is a tally table or frequency table.

A frequency table counts the number of times – or frequency – a value occurs in the data. For example, twenty people are asked "How many TVs do you have in your household?" If 2 households have 1 TV, the frequency of households with 1 TV is 2.

|  |  |  |
| --- | --- | --- |
| **NUMBER OF TVs** | **TALLY** | **FREQUENCY (f)** |
| 0 | l l l l | 4 |
| 1 | l l | 2 |
| 2 | l l l l l | 6 |
| 3 | l l l l l l l | 8 |

*Figure 1: Frequency table of number of TVs per household*

Frequency tables with class intervals

When a variable has a large spread, the values can be grouped together to make the data easier to manage and present. For example, if you asked students how much time it takes them to get to school each day, their responses may vary considerably. In this case, you can group the responses together in 5 minute intervals. These intervals are called class intervals. All class intervals should have an equal range. Class intervals are usually in groups of 5, 10, 20, 50 etc.

|  |  |  |
| --- | --- | --- |
| **TIME TAKEN TO GET TO SCHOOL (min)** | **TALLY** | **FREQUENCY (f)** |
| 1 - 5 | l l l l l l | 7 |
| 6 - 10 | l l l l l l l l | 9 |
| 11 - 15 | l l l l l l l l | 9 |
| 16 - 20 | 1 | 1 |

*Figure 2: Example of a frequency table showing five minute class intervals*

Step 4: Displaying information

How will you present your findings? Will you use tables, graphs or both? What sort of table or graph is most appropriate?

One of the most powerful ways to communicate data is by using graphs. Data presented in a graph can be quick and easy to understand. A graph should:

* be simple and not too cluttered
* show data without changing the data’s meaning
* show any trend or differences in the data
* be accurate in a visual sense – for example, if one value is 15 and another 30, then 30 should be twice the size of 15.

Ambiguity can be reduced by

* avoiding 3D representations
* avoiding broken or uneven scales.

It is important that you give each graph a heading. Axes must be labelled and any scale must have equal intervals. A key should be included where it is needed to interpret the data.

Different graphs are useful for different types of information and it is important that the right graph for the type of data is selected.

Consider the type of data you have collected when choosing a display or graph.

|  |  |
| --- | --- |
| **Type of data** | **Appropriate display or graph** |
| Categorical | Bar graph, pie chart, dot plot |
| Numerical (discrete) | Bar graph, histogram, line graph, box and whisker plot, stem and leaf plot, age pyramid |
| Numerical (continuous) | Histogram, line graph, box and whisker plot, stem and leaf plot |


*Figure 3: Bar graph showing importance of conserving water.
Input from 0 (Not Important) to 999 (Very Important)*

A bar graph is used to represent categorical or discrete numerical data. A bar graph has a gap between each bar or set of bars and the widths of the gaps and bars are consistent. The length of the bars in a bar graph is also important: the greater the length, the greater the value. A bar graph can be either horizontal or vertical – vertical bar graphs are also known as column graphs.


*Figure 4: A horizontal bar graph*

The advantage of using a horizontal bar graph over a column graph is that the category labels in a horizontal bar graph can be fully displayed making the graph easier to read.


*Figure 5: A side by side column graph*

Side by side bar graphs are useful to compare two or more groups for the same data.


*Figure 6: A stacked bar graph*

Stacked or segmented bar graphs are sometimes used to display a breakdown of data for a particular group – for example, the breakdown of types of internet use by sex. They are most useful when the data is shown as a percentage value and the number of categories is very small. If there are too many categories, the frequency of the datasets can be difficult to read and compare.


*Figure 7: A histogram showing data with grouped intervals*

A histogram is similar to a column graph; however, there are no gaps between columns. Histograms are used for numerical data only. Data can be either discrete or continuous and grouped or ungrouped. A histogram should have equal width columns when possible.

Line Graphs

Line graphs are used for numerical data. They show how one variable changes in relation to another. The independent variable is always displayed on the horizontal axis. It is important to use a consistent scale on each axis so you can get an accurate sense of the data.


*Figure 8: A time series graph*

Time series graphs are line graphs that display a trend or a pattern in data over time. The slope can be upward or negative. Patterns can be seasonal (over a short period), or cyclic (over a longer period). Alternately, a time series can show a random variation.

To use a times series to make predictions, you can smooth fluctuations by using a suitable smoothing process such as de-seasonalising the data.


*Figure 9: A pie chart*

Pie charts, often called pie graphs, sector graphs or sector charts, show how much each sector contributes to the total. Pie charts are useful when there are only a few categories as more than five categories can make them difficult to read.

It is very important to label each sector with its value to make comparison easier.


*Figure 10: A dot plot with many to one correspondence*

A dot chart can convey a lot of information in a simple, uncluttered way. Each dot can represent one or many.


*Figure 11: Parallel box and whisker plots*

Box and whisker plots display 5 figure summary statistics: minimum, quartile 1, median, quartile 3 and maximum for a set of numerical data.

Box and whisker plots are useful for looking at the shape of a data set. In particular, parallel box and whisker plots are used to compare two data sets and are drawn on the same number line. To help identify possible outliers, 'fences' can be drawn at 1.5 x, the IQR above Q3 and below Q1.


*Figure 12: A back to back stem and leaf plot of arm span*

Stem and leaf plots are an efficient way of recording numerical data because numbers in the stem apply to all values in the leaves. They are also very useful to show the shape of a distribution and, since they order data, can be used to identify median and quartiles.

Back to back stem and leaf plots are used to compare the distribution of two data sets.

*Figure 13: An age pyramid (Source: ABS Education Services)*

Age pyramids are used to represent a population age structure. Age pyramids are a very effective way of showing change in a country’s age structure over time or for comparing different countries. Estimates and projections of Australia's population from 1971 to 2050 are available on the ABS Animated Age Pyramid page.

Step 5: Analysing the data

You need to decide what statistics you will use and what summary information will enable you to answer your survey question.

For example:

* Will you find the mean or median or both?
* Will you exclude extreme values?

When describing a distribution, statisticians usually comment on its centre, spread and shape.

a) Measures of centre

The centre of a set of data is important. Often you want to know which value occurs most commonly or the average of a set of values.

Numerical data

**Mean**: The arithmetic mean is calculated by finding the sum of all values in the data set and then dividing it by the number of values in the set. The mean is often called the average.

**Median**: The median is the middle value of a dataset when all the values are arranged from smallest to largest number. If a dataset contains an even number of values, the median is taken as the mean of the middle two. The median is the preferred measure of centre when the data is not symmetric or contains outliers.


*Figure 14: Mean median and mode of a data set*

**Mode:** The mode is the value that occurs most often in a data set.

b) Measures of Spread

The spread of a data set, when combined with its centre, will give a more complete picture.

Range

The range is the distance from the minimum value to the maximum value in the data set. Range = maximum – minimum

*Figure 15: Calculating 5 figure summary statistics*

**Quartiles:** As the name suggests, quartiles divide data into four equal sets.

When observations are placed in ascending order according to their value, the first or lower quartile is the value of the observation at or below which one-quarter (25%) of observations lie. The second quartile is the median at or below which half (50%) of observations lie. The third or upper quartile is the value of the observation at or below which three-quarters (75%) of the observations lie.

Another way to think about this is: the median divides the data into two equal sets: the lower quartile is the value of the middle of the lower half, and the upper quartile is the value of the middle of the upper half. The difference between upper and lower quartiles (Q3 - Q1) also indicates the spread of a data set. This is called the interquartile range (IQR).

*(Further information on analysing data is available in the survey fact sheet on the ABS website).*

Step 6: Drawing conclusions

Communicating the results of your investigation is a critical part of the survey process. Ensuring the accuracy of any interpretations and avoiding misinterpretations are crucial. Keeping in mind the purpose of the investigation and your audience will help to keep your conclusions on track and avoid including unnecessary information.

Accuracy of data

All your calculations need to be accurate, verifiable from the data and clearly communicated using simple language.

Misinterpretation

To effectively communicate your results, you need to avoid any misinterpretation of the data such as using the mean when the median is more appropriate or not taking seasonal variation into account.

Stating your conclusions

With statistics, there is always a risk that the results you have do not tell the whole story. You can use the following checklist to help judge the reliability of your statistical information.

* Do your conclusions communicate the message told by the data?
* Are your conclusions based on results rather than on your opinions?
* Have you considered alternative explanations for the same results?
* Is your report set out logically including using an organisational framework such as headings and sub headings?
* Have you included the source of any information you have used or referred to?
* Have you included relevant tables and graphs?
* Are your findings clear, related to your aim and only contain necessary information?

Audience

* Have you considered your audience and used appropriate language?
* Have you anticipated questions your reader might have? For example, have you explained unusual or unexpected results? Have you justified your choice of analysis, indicated your sampling process etc?
* Can your reader check your conclusions by viewing your analysis?

Sampling error

Finally, if you intend for your results to be applicable in other contexts, it is important to understand the limits that might apply.

The difference between an estimate based on a sample survey and the true value that would result if a census of the whole population was taken is called the sampling error. Sampling error can be measured mathematically and is influenced by the size of the sample. In general, the larger the sample size the smaller the sampling error.

The way a sample is drawn is also important. In general, a random sample will result in data that is more able to be generalised to the population.