

Questions to ask yourself

1. Do the data I have retrieved answer my original question? (If not, what am I missing?)

Remember when you first started looking for your data. What was the reason you were searching for data? What was the question you were trying to address? Did you have any preconceived ideas about the type of data you wanted to retrieve? Try making a list of the following questions, and answering them as honestly as you can:

- ✓ Why did I start looking for data/information?
- ✓ What was I hoping to find?
- ✓ Where did I search?
- ✓ What have I found?
- ✓ What are the gaps between the data I've found, and what I originally wanted to find? (if any)
- ✓ Are there any ways I can address these gaps?

Once you have actually retrieved some data, it's a good idea to pause and take stock. Do these data actually answer your original question? It's easy to get distracted when searching for information and it's easy to retrieve data which we think answers our question, but in fact it may not.

Sometimes the data you want to find are not available. So you may have retrieved some information which is as close as possible to what you are trying to find out, but there are still some pieces missing. Try to be very clear with yourself about the limitations of your data. No dataset will answer every question. You still need to think about what you are doing and where you are searching, and try to identify what's missing.

Often, what is missing can spark research questions of your own. Or, at the very least, what is missing may indicate recommendations you can make about how to improve the data available in a certain area.

2. How good are the data? How much error is in the data?

One of the problems with data is that data usually incorporates error. Error can stem from a variety of sources, and usually some indication of error is provided with data. These errors may be in the form of a 'standard error' or 'relative standard error'.

It is extremely important to evaluate the quality of your data with one eye firmly fixed on the standard error or relative standard error. Here is some brief information about the relative standard error:

Relative standard error

In an ideal world, all of our errors would be small, meaning that our results are reliable. However, due to problems of low sample sizes and other issues, small standard errors are not always the case. This raises the question of how big an error is too big?

To answer this question, it is normal practice to compare the standard error with the actual estimate. To make this comparison, we divide the standard error by the estimate obtained, and convert it to a percentage.

For example, in order to estimate the percentage of Queenslanders who have visited a state or national park in the last year, I take a sample and obtain an estimate of 71% who have visited a state or national park in the last year, with a standard error of 8%. The relative standard error is therefore:

$$\%RSE = \frac{8}{71} \times 100 = 11\% , \text{ which is quite acceptable.}$$

Generally, if the relative standard error is 25% or less, results have reasonable accuracy. However, as the relative standard error increases above this threshold, more caution needs to be taken when interpreting the results. The Office of Economic and Statistical Research usually highlights unreliable results by the use of asterisks (*). For example, if a result has a relative standard error greater than 25% but less than or equal to 50%, we place one asterisk next to the value. If the relative standard error is great than 50%, we place two asterisks next to the value, and we advise not using the estimate due to its high unreliability.

When you obtain a set of data, try not to regard its contents as universal truth – datasets do not “tell” us anything. The most that we can do with data is make inferences, assuming that the data are reliable. Remember that the reliability of your dataset is dependent to a large degree on its error, which you can assess by calculating the relative standard error.

3. How up-to-date are the data?

Eventually, all data will become old and, potentially (depending on the nature of the dataset) out-of-date. If you are working with datasets, it’s a good idea to keep in mind when the data were collected. Are these data 20 years old? If so, what might have changed since they were collected? Conversely, if the data are very recent (e.g., collected in the last three months), they may not be useful in making inferences about events which happened 20 years ago.

Remember to take into account the ‘age’ of your dataset when you are using it. Also, some datasets are regularly updated, so maintain links with the site where you found your data, to ensure that you can keep up-to-date with later versions of your dataset as they are released.

4. How were these data collected? How might this impact on the nature of the data?

Data can be collected in a variety of ways. Common methods of collection involve sampling a given population using techniques such as surveys, interviews, and measurements. Other methods of data collection utilise the whole population, known as conducting a census. It's important to think about how your data were collected. Were they originally a sample of a population, and you are now trying to use the sample to estimate characteristics of a population? Or, do you have complete Census data?

In addition, different issues affect different types of data collection methods. Here are a few examples:

- Surveys and interviews can be affected by participant non-response, which can bias results.
- Measurements of other forms (e.g., measurement of plant leaf growth) assume high technical measurement reliability, which may or may not be present.
- Face-to-face interviews can produce biased results as participants may behave differently in the presence of an interviewer.
- Asking people questions requiring them to recall information from the past (known as “retrospective reporting”) often produces inaccurate information. People develop memory biases and gaps over time, which can lead to the distortion of remembered information.

It's important to consider these issues and evaluate how much of a risk they pose to your data. Datasets are rarely (if ever) perfect; however it is helpful to develop an awareness of their limitations so that you can become an informed user.

5. Have you referenced this data?

Data obtained from the Data Hub are provided free of charge to all Queensland public sector employees. The costs associated with purchasing this data are absorbed by the Queensland Government. However, as a part of the licensing agreement with data suppliers such as the Australian Bureau of Statistics, users are required to acknowledge (reference) the source of their data.