Queensland Government Statistician's Office

Assessing survey data

Introduction

This paper presents factors that impact on the validity of survey data. That is, the extent to which survey data accurately reflect what is happening in the real world — in local communities, regions or across the state. This information can be used when planning a survey or assessing results from others' surveys and in determining whether and how survey data can be used.

There are a number of factors which impact on validity, including the type of data — whether they're from a primary or secondary source; the data items available; whether or not the data are complete; the age of the data; and whether there are biases, errors or systemic inaccuracies due to the processes used in the creation, manipulation or presentation of data.

It is important to consider these issues and evaluate how much of a risk they might pose to the integrity of the data. Datasets are rarely, if ever, perfect; however it is helpful to develop an awareness of their limitations in order to become an informed user.

Types of data and items available

Primary data are generated by directly asking questions, e.g. surveys, or conducting trials and collating results. This can be in the form of quantitative or qualitative research. Primary data are generally collected to serve a specific need or purpose. 'Secondary data' refers to information derived from the findings from others' research. Analysis of secondary data sources may include collating results of others' research from books, reports or the internet. Selections or summaries are taken from the research to support any conclusions. When using secondary data, it is important to assess what research questions the data actually answer. Unless designed specifically for purpose, it is highly unlikely that any one dataset will answer every question of interest.

It is important to be clear about the limitations of the data and what items are available — sometimes the data required to answer a specific research question are simply not available.

Age of data

The age of a dataset should always be taken into account. Are the data twenty 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 are unlikely to be useful in making inferences about events which happened twenty years ago. Some datasets (such as those available through the Queensland Government Statistician's Office or the Australian Bureau of Statistics) are regularly updated. Data users should maintain links with the source site for these datasets, to ensure that data being used are the most recent version available.

Always use the version of a dataset that is most relevant to the period of time related to the research question. Eventually, all data will become old and potentially — depending on the nature of the dataset — out of date. If the data available are so out of date as to be likely to impact on the validity of the research, then it may be necessary to collect new data.

Biases and errors

When assessing data for bias or error it is important to consider the various survey processes and methods used. This includes whether the questionnaire has been appropriately designed and worded, the data collection method, the number and type of people included in the survey sample and whether the data represent the target population of interest.

Questionnaire

Good questionnaire design is essential. If the questionnaire is poorly worded, it is likely that the data will be problematic. Refer to the questionnaire to check that:

- jargon or technical terms have been clearly defined. Failure to do so can lead to respondents interpreting terms differently, rendering their answers non-comparable.
- time periods for question recall are reasonable. For example, asking people how many times they have visited a doctor in the last three months is much more likely to yield accurate results than asking how many times they have visited a doctor in the last year.
- if respondents are asked to order or rank a series of items, the list does not include more than five items.
- questions are worded using non-biased language. For example:

  ✔ *'Do you think the price of milk will increase in the next six months?'*

  (If YES) 'By how much?'

  ✘ *'By how much will the price of milk increase in the next six months?'*

- questions are clearly and simply worded. Questions that are confusing or difficult to understand generally yield similar response patterns, with an unusually high number of 'unable to rate' responses, and a larger than average spread of responses in the frequency distribution.
- questions use positive wording. Use of negatives or double negatives can confuse respondents, resulting in answers that are opposite to the intended answer.
- questions are suitable for the mode of administration. For example, a list of 20 options can be provided on a web survey for a respondent to read. A list of 20 options would not be as suitable in a telephone interview where they must be read out.

While it is not possible to correct question wording after the data have been collected, it helps to be aware of any possible issues that could be present.

Method of collection

Data can be collected in a variety of ways — telephone, face-to-face, mail and online are all common survey methods. It's important to consider how the data were collected, and any potential bias the collection method may introduce to the data.

- Face-to-face interviews can produce biased results as participants may respond less candidly in the presence of an interviewer, particularly to questions that could potentially reveal socially undesirable traits or acts.
- Self-completion surveys (whether mail or online) generally tend to have higher non-response rates than other types of surveys. Response rates can be improved using techniques such as well–written introductory letters, incentives for timely return of questionnaires and follow-up for those initially not responding.
- Substantial bias can be present in self-completion surveys if people who do not complete the survey have different characteristics from those who do. Methods for compensating for sampling bias are discussed later in this paper.
- Respondents may misread questions or instructions in self-completion surveys. Clear, unambiguous wording, in conjunction with bold or highlighted instructions, can help to mitigate this problem.

- In self-completion surveys there is no opportunity to clarify answers or probe for additional information.
- Online surveys require respondents to have access to a computer. This can bias the results by excluding people who do not have access to one.
- Online panels are subject to unknown but possibly considerable levels of bias because panel members may not be representative of the population they are drawn from. Panel members are those who have the time and interest to complete multiple questionnaires, on demand, and who are attracted by the monetary incentives offered. Online panel members must also have access to a computer. There are limited options open to the manager of an online panel to reduce bias. One is to use a variety of recruitment methods (e.g. emails, telephone calls, letters sent to homes etc.) in the hope of inviting as representative a sample of the population as possible to join (though a high proportion may decline). Another is to carefully manage each panel member's workload so that they are more likely to give considered answers to questions.
- Asking questions that require a respondent 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.
- A survey frame or list, which contains telephone numbers of a representative sample population, is required to conduct a telephone survey. For general population surveys, such lists are not readily available to most survey providers or they have limitations that can lead to biased results. Access to appropriate contact lists has become more of an issue in recent years, as the proportion of people who have a mobile phone only (no landline) is steadily increasing.

Sample

An exact value for an attribute of the population can only be determined by taking a census of the entire population; however this is usually inconvenient, expensive and unnecessary. Instead, it is customary to survey a sample of the population and infer the value using statistical methods.

Using a sample will reduce the cost of the survey and the type of sample chosen will impact on the accuracy of the results. A probability sample is a sample in which every unit in the population has a known chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection. Samples of modest size are often sufficient to measure an attribute with acceptable accuracy. If a probability sample is taken, a quantitative statement about the accuracy of the estimate can be made. It's important, however, to be sure there is an adequate number of responses to make the survey results statistically meaningful. This applies to both the total number of respondents, and the number of respondents in any subsets of interest (for example, 'males under the age of 30 years'). Extreme care must be exercised in drawing conclusions about subgroups of a population when the number of units captured by the sample in the subgroup is very small.

Conclusions drawn from larger sample sizes are more accurate than conclusions drawn from smaller samples. While there is no hard and fast rule on suitable sample size for a survey, it is important to consider how much uncertainty is acceptable to the end user of the report. Where major financial investment or policy decisions are involved, the accuracy of results is highly important. The accuracy of data is generally measured by examining the standard error, the relative standard error and/or the confidence intervals for the estimates.

Sampling error

Sampling error occurs as a result of using a sample from a population, rather than conducting a census of the population. It refers to the difference between an estimate for a population based on data from a sample and the true value for that population which would result if a census were taken. Sampling errors do not occur in a census, as the census values are based on the entire population. Sampling error undermines the ability of the results to be generalised to the rest of the population. It can occur when:

- the proportions of different characteristics within the sample are not similar to the proportions of the characteristics for the whole population (i.e. if 51% of the total population are women and 49% are men, then a researcher should aim to have similar percentages in their sample)

- the sample is too small, or
- the sampling method is not random.

Standard error, relative standard error and confidence intervals are all ways in which sampling error can be measured. (These are all covered further on in this paper.) Sampling error can be reduced by increasing the size of the sample, and ensuring that the sample is representative of the population.

Weighting to correct sampling error

If some subsets of the population are under-represented, as described above, then sample weights can be used to correct the bias.

For example, a hypothetical population might include 5 million males and 5 million females (i.e. 50% males and 50% females). If a sample survey is conducted via the internet, and responses are received from 4,000 males and 6,000 females (i.e. 40% males and 60% females), then the respondent profile does not match the population profile.

In this scenario a researcher can correct for this imbalance by attaching a weight of 1.3 for each male and 0.8 for each female. This would adjust any estimates to achieve the same expected value as a sample that included exactly 5,000 males and 5,000 females.

Usually weighting is used to correct sampling error, however sometimes it can inadvertently exaggerate an existing bias in a sample. Any existing biases in the sample should be considered in conjunction with the weighting methodology used, to help determine whether the weighting method is appropriate.

Response rates and consent rates

The response rate quoted in many research studies usually refers to the number of people who answered the survey divided by the number of people in the sample. It is usually expressed in the form of a percentage. For example:

> If a researcher sends out 100,000 emails inviting people to participate in an online survey, and 14,000 people respond, the response rate (using this method) is 14%.

Research studies may also refer to consent rates or cooperation rates. This usually refers to the number of responses divided by the number of people or households actually contacted. This is likely to be higher than the response rate, and may be considerably higher.

Organisations and agencies may vary in their methods of calculating response rates and consent (or cooperation) rates. It is important to note whether it is a response rate or consent/cooperation rate that is being quoted in a report, and to understand exactly what method has been used to calculate the quoted rate. Some methods can inflate the rates, and others may make the rates appear lower. A low response rate can be a source of error. Non-responders may have different attitudes, beliefs and behaviours to respondents. If these differing characteristics influence the answers respondents give to survey questions, estimates derived from survey responses will be biased.

When considering response rate, it is therefore important to look at the overall response rate, as well as the response rate for subgroups of interest within the population.

It is generally difficult to determine the degree and direction of any bias in estimates because usually little information about non-responders is available. However, often previous research and experience may have established that the probability of non-response is correlated with characteristics of the population being surveyed, such as age, sex, or educational attainment. If this is true, non-response bias may be reduced by weighting responses so that the age, sex and educational characteristics of the sample match those of the population the sample was drawn from. The degree of bias reduction is directly proportional to the strength of association between the so called ‘weighting variables’ and the topics covered by the survey questionnaire.

Measuring sampling error

Standard error is a statistical term for measuring the accuracy with which a statistic taken from a sample represents an entire population. The smaller the standard error, the more representative the sample will be of the overall population. The standard error is represented as a number, and decreases as the sample size increases. The standard error allows a researcher to determine how confident they can be about inferences made from survey data.

The 'archery analogy': The inference process can be likened to an archer aiming at his/her target. The bullseye on the target represents the actual value and the sample statistic is the arrow. The aim is to get the arrow as close as possible to the bullseye.

If many samples are taken, the result would be some larger estimates and some smaller. The standard error is an average distance of each estimate from the actual true value that would be obtained if the entire population was surveyed.

Remember that the reliability of an estimate is dependent to a large degree on its error. In an ideal world, all of the errors would be small, and the results would then be reliable. In reality, this is not always the case.

Relative standard error — it is important to know how far from the bullseye (actual value) an estimate is, to decide whether an estimate is accurate enough to be meaningful. A common way of deciding whether the error level present for an estimate is too high is to calculate the relative standard error (RSE). The RSE is expressed as a percentage.

Calculating relative standard error (RSE)

To calculate the relative standard error, divide the standard error by the estimate obtained, and convert it to a percentage.

For example, consider two surveys of household income that both result in a mean of $50,000. If one survey has a standard error of $10,000 then the RSE is:

%RSE = $10,000/$50,000 = 20%

If the other survey has a standard error of $5,000, then the RSE is:

%RSE = $5000/$50,000 = 10%

The survey with the lower relative standard error can be said to have a more precise measurement, since it has proportionately less variation around the mean.

Generally, if the RSE is 25% or less, results have reasonable accuracy. As the RSE increases above this threshold, more caution should be used when interpreting the results. If the RSE is greater than 50%, the result has very low reliability, and should not be used.

Caution should be taken when looking at RSEs for proportions or percentages, particularly when percentages for the estimates are close to zero or 100%. Consider the following example:

To estimate the percentage of Queenslanders who have visited a state or national park in the last year, a sample is taken, and an estimate obtained. The sample estimate shows that 80% of respondents have visited a state or national park in the last year, with a standard error of 8%. The RSE is therefore:

%RSE = 8/80 x 100 = 10%, which is considered to be quite acceptable.

If the sample estimate showed that only 9% of respondents had visited a state or national park in the last year, with the same standard error of 8%, then the RSE would be 89%

%RSE = 8/9 x 100 = 89%, which would not be considered to be acceptable.

In this example, while both estimates have a standard error of 8%, the smaller estimate has a much bigger RSE.

RSEs can become inflated simply because the percentage estimate is small, and due to the nature of the calculation. In cases where estimates are close to zero or 100%, seek the advice of a skilled statistician in regards to the reliability of the estimate.

Confidence intervals for proportions and confidence levels — a confidence interval is another measure of the accuracy of an estimate. Confidence intervals provide a range of values around the estimate, within which the true value can be expected to fall. The smaller the confidence interval is for a particular estimate, the more precise the estimate is.

Normally, confidence intervals are based on what is called the 95% confidence level. A 95% confidence level means that the reader can be 95% sure that the true value lies somewhere between the stated confidence interval. Another way of saying this is that if 100 samples were taken from the population, it would be expected that 95 out of 100 times the true value would fall within the stated confidence interval.

A confidence interval consists of two numbers — a lower and upper limit. For example:

> A total of 10,000 respondents are asked whether they have solar panels installed on the roof of their home. Thirty percent (30%) of respondents report that they do have solar panels. For the sample size of 10,000, the estimate of 30% has a confidence interval of ± 0.9%. So the lower limit (L) is 29.1%, and the upper limit (U) is 30.9%. This means there is a 95% chance that the true value lies between 29.1% and 30.9%.
>
> If the same question was asked of 200 respondents, and 30% reported that they had solar panels, the confidence interval for the estimate is ± 6.4%. So the lower limit (L) is 23.6%, and the upper limit is 36.4%. This means there is a 95% chance that the true value lies between 23.6% and 36.4%. With the smaller sample size, the accuracy of the estimate is lower.

The size of the confidence interval is affected both by the sample size, and the proportion of respondents giving a particular response. The confidence interval will be bigger for small sample sizes than for larger sample sizes. The confidence interval will be larger for estimates closer to 50%, and smaller for estimates further away from 50%. Examining the confidence limits provides valuable information on the accuracy of the data.

Statistical significance is a statement about the likelihood of findings being due to chance. Confidence intervals are one way for researchers to determine whether or not the difference between two estimates is statistically significant, rather than being purely due to chance. If confidence intervals do not overlap, then the difference between the estimates is likely to be statistically significant 99 times out of 100. For example:

> Imagine that respondents to a survey were asked whether they have water-saving shower heads installed in their home. Sixty-nine percent (69%) of respondents who are living alone state that they do, and 56% of respondents who are living in houses with others report that they do not. At first glance, it may appear that respondents who live alone are more likely to have water saving shower heads than respondents living with others.

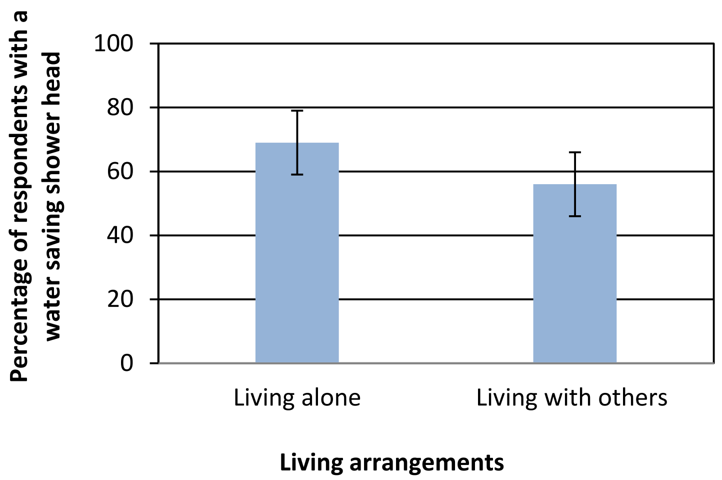
The next step would be to examine the confidence interval (see table below). When the confidence interval is taken into account, then it can be seen that the two estimates are in fact NOT statistically significantly different. The 'true' value for the estimate of 69% could fall anywhere between 59% and 79%, while the true value of the estimate of 56% could fall anywhere between 42% and 70%.

Table 1: Percentage of respondents who have water-saving shower heads installed in their home

		95% Confidence Interval	
Living arrangements	**Estimate**	**Lower**	**Upper**
Living alone	69%	59%	79%
Living with others	56%	42%	70%

The overlap can be easily seen when represented graphically. The graph below shows the upper and lower bounds of the confidence intervals through the use of error bars. It can be seen that there is overlap in the error bars, indicating that there is no statistically significant difference in the estimates.

Figure 1: Percentage of respondents who have water-saving shower heads installed in their home



Other methods for determining statistical significance include a variety of statistical tests (t-tests, F-tests etc) chosen according to the particular question being asked. When examining data, make sure that it is possible to determine whether any apparent differences are statistically significant, or are merely due to chance.

Non-sampling error

Non-sampling error is caused by factors other than those related to sample selection. It refers to the presence of any factor, whether systemic or random, that results in the data values not accurately reflecting the 'true' value for the population. Although non-sampling errors cannot be measured in the same way as sampling errors, they are just as important.

The following table lists common sources of non-sampling error and some questions to consider when assessing the accuracy of available data. It is possible that the answers to some of these questions will remain unknown, particularly when secondary data are being used. Nonetheless, it is important to be aware of the various sources of error, so that potential limitations of the data can be understood.

Source of error	Examples	Questions to consider
Survey methods	Inappropriate method (e.g. internet survey for people who are unlikely to have access to the internet).	Was an appropriate method used? Refer to *Method of collection* section of this paper.
Questionnaire	Loaded, misleading or ambiguous questions, poor layout or sequencing.	Does the questionnaire use plain English, clear questions and logical layout? Was it tested thoroughly? Refer to *Questionnaire* section of this paper
Coverage	Poor access to respondents in the target population	Do all members of the target population have an equal chance of being interviewed? Are all members (or a representative sample) of the target population included on the list of people to contact?
Interviewers	Leading respondents, making assumptions, misunderstanding or inaccurately recording answers.	Does the questionnaire include clear interviewer instructions? Was appropriate training provided, including field supervision?
Respondents	Refusals, accidental or intentional provision of inaccurate responses, rounding answers, protecting personal interests or integrity.	Have interviewers received adequate training? If a web or postal survey, is the introduction well written and likely to encourage participation? Does the introduction provide reassurances around confidentiality?
Processing	Errors in data entry, coding or editing.	Were processing staff adequately trained and supervised? Has a sample of each person's work been checked?
Estimation	Incorrect weighting, errors in calculation of estimates.	Did skilled statisticians undertake the estimation? Are appropriate quality procedures in place to ensure adequate checking?

Further information and advice

The Queensland Government Statistician's Office has extensive experience in the design and conduct of statistical surveys and is able to provide specialist consultation and advice on request. For more information please see the QGSO website, www.qgso.qld.gov.au, or email govstat@treasury.qld.gov.au.

Queensland Government Statistician's Office
Queensland Treasury
http://www.qgso.qld.gov.au/