Interpreting Survey Data

Health Intelligence Unit

VICTORIA State Government | Health and Human Services

Interpreting survey data

Health Intelligence Unit

To receive this publication in an accessible format phone 9096 0000, using the National Relay Service 13 36 77 if required, or email health.intelligence@dhhs.vic.gov.au

Authorised and published by the Victorian Government, 1 Treasury Place, Melbourne.

Available at https://www2.health.vic.gov.au/public-health/population-health-systems/health-status-of-victorians

Contents

Introduction

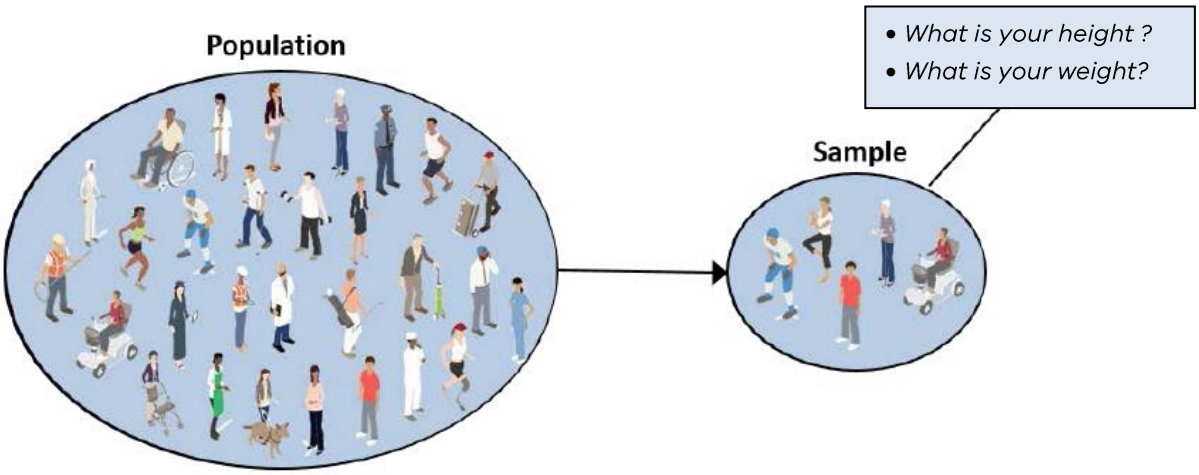
The Department of Health and Human Services (the department) is responsible for assessing the health of Victoria's population by measuring, monitoring and reporting health determinants[1], risks, health outcomes and related health inequalities. Survey data is often used to inform this work. The purpose of this guide is to provide a basic understanding of the statistical concepts and tools frequently used by the department to describe health survey data. In particular, this guide explains: samples and populations, estimates, stand error, relative standard error, confidence intervals, the concepts of statistical significance and age adjusted (or standardised) rates and testing of trends over time.

Populations vs. Samples

Imagine we wanted to know:	***What proportion of Victoria's residents are obese?***
This information would be used to help provide targeted programs and support to reduce obesity and improve the health and wellbeing of Victorians.	

A **population** is the entire group of people we want to learn about and understand.
A **sample** is a smaller group of people drawn from a population. We collect information from a sample of people when we cannot gather information from every person in a population. If we were to collect information from everyone in the population, we would call this a **census** of the population. The information gathered from a sample is used to reach conclusions about the population at large. Therefore, bigger samples are more reliable than smaller samples. Bigger samples are more representative of the entire population.

The population of Victoria is over six million. If we were interested in the proportion of the population who were obese, it would not be practical to assess everyone. Instead, we ask a sample of the population about their height and weight, from which we compute their body mass index (BMI) to determine their weight status.

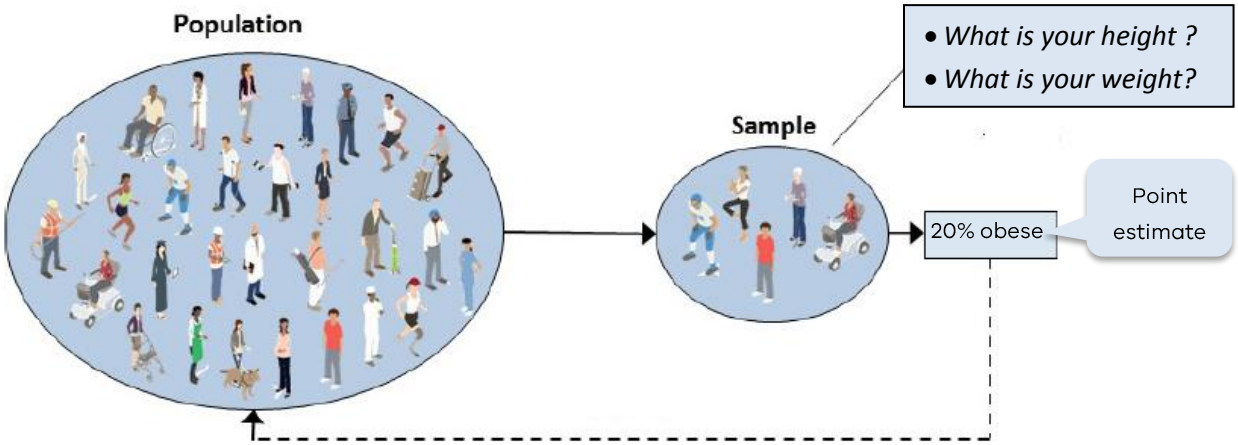


Estimates

[1] The range of behavioural, biological, socio-economic and environmental factors that influence the health status of individuals or populations.

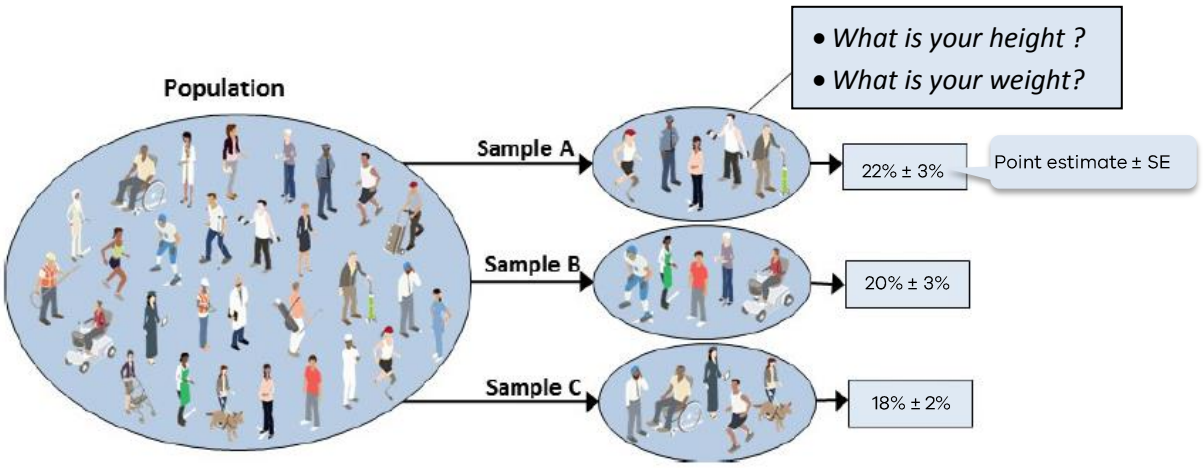
An estimate is an approximation of the unknown true value of a characteristic in the population. Estimates are created by using the data collected from a sample to make conclusions about the entire population. A point estimate is a single value from a sample that is used to describe a characteristic of the population.

What proportion of Victoria's residents are obese?
If twenty per cent of our sample were obese and our sample had been designed to represent the population, we could reasonably estimate that about twenty per cent of all people in Victoria were obese.



Standard error

If we tried to assess obesity in a number of different samples, we would likely get different estimates, due to a different mix of people in each sample. Therefore, estimates based on samples may differ from those that would have been produced if the entire population had been assessed. The most common measure of the likely difference between the true value for a characteristic in a population and the estimate based on a sample is the **standard error** (SE). The SE indicates the extent to which an estimate is likely to deviate from the true value of the characteristic being assessed in a population. The SE is expressed as a number.



Relative standard error

The **relative standard error** (RSE) is the standard error expressed as a fraction of the point estimate and is usually presented as a percentage. Estimates with RSEs of 25 per cent or more are not considered reliable for most purposes. Estimates with RSEs greater than 25 per cent, but less than or equal to 50 per cent indicate they are subject to high SEs and should be used with caution and are annotated by an asterisk (*). Estimates with RSEs greater than 50 per cent indicate very high SEs and are unreliable for general use.

The department provides an indication of the reliability of estimates from survey data in tables and charts, with a single asterisk (*) for RSEs between 25 and 50 per cent. Estimates with an RSE greater than 50 per cent are suppressed and replaced with a double asterisk (**).These estimates are considered too unreliable to be reported.

Confidence intervals

The reliability of an estimate can also be assessed in terms of a confidence interval (CI). Instead of a single point estimate, confidence intervals provide a range of possible values (that would theoretically arise from multiple samples) for a characteristic in the population. The department prefers to use the **95 per cent CI** (95% CI). A 95% CI is a range of numbers around a point estimate that, 95 times out of 100, would contain the true value for a characteristic in the population.

Strictly speaking a 95 per cent CI means that if we were to take 100 different samples and compute a 95 per cent confidence interval for each of these 100 samples, then approximately 95 of the 100 confidence intervals will contain the true value.

The size, or width, of a confidence interval depends on three factors:

1. **Sample size**. Samples with fewer people provide less information, which results in a wider and less precise confidence interval.
2. **Level of confidence desired**. As the level of confidence increases, say from 95 per cent to 99 per cent, we are more certain that the true population value is within the confidence interval range. As the CI becomes wider, it becomes less precise at capturing the true value of a characteristic in the population.
3. **Sample variability**. Increased variation in responses creates greater uncertainty and a wider confidence interval.

Confidence intervals are formed by adding and subtracting a margin of error from the estimate. CIs can be described in text in the following ways, meaning the same thing:

As an estimate plus or minus the margin of error: **47% ± 4%**	As the upper and lower range of numbers around the estimate: **43% to 51%**

CIs can also be displayed on graphs as error bars (I-shaped) superimposed over a bar, point or line that represents the point estimate (see box on following page).

All survey data presented in reports, or presentations from the department, should include a point estimate, along with an associated standard error and/or a confidence interval.

It is important to move away from the current practice on only focussing on the point estimate and also consider the 95% CI, as the point estimate is the summary measure for a single sample drawn from the population of interest. It would likely be different, if it was computed from another sample of equal size drawn from the same population.

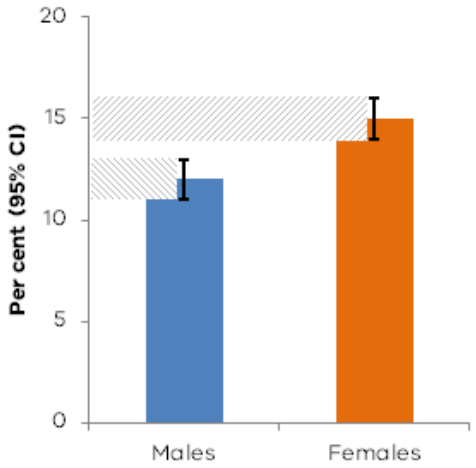
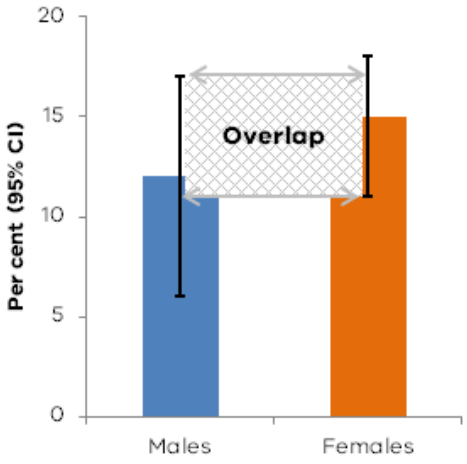
Imagine we wanted to know:	***Is the difference in the proportion of males and females who are physically active, a real difference?***
This helps decision makers know whether the variation they are seeing is meaningful and reflects a real difference, or if it is due to chance.	

Tests of statistical significance tell us whether we can be confident that a real difference exists between groups, or whether the difference observed is simply due to chance. If the probability (P value) of the observed difference occurring is less than 5 per cent ($P < 0.05$) , we usually say that the two groups are statistically different. There is a good chance that the difference, or relationship, observed in the estimates derived from the survey sample reflect a true difference, or relationship, in the wider population.

Sometimes confidence intervals are used to determine whether estimates from a sample differ from one another. With this approach, CIs that do not overlap are considered statistically significant and are most likely different from each other in the wider population. This is because each group has a completely different range of values for the characteristic being assessed and they do not overlap. In this case, the true values for each group in the population are going to be different. In contrast, when CIs do overlap, the estimates from the sample taken are not usually considered statistically different from each other.

Note that the use of CIs to determine statistical significance is a conservative method that depends on the degree of overlap in the CIs. This method is best used when there are estimates from multiple groups to be compared.

The following charts show the same proportion estimated proportion of males and females who are obese, but with different CIs.

The CIs *do not overlap*. Therefore, the proportion of obese females was significantly higher than the proportion for males.	**The CIs *overlap*, so estimates are <u>not</u> statistically different. There is no statistically significant difference in the proportion of males and females who were obese.**
 **Confidence intervals do not overlap, therefore, estimates are statistically different**	 **Confidence intervals overlap, therefore, estimates are <u>not</u> statistically different**

Directly age-adjusted rates

When comparing across geographic areas, some method of age adjusting is typically used to control for area-to-area differences in health events that can be explained by differing ages distributions in the populations. For example, an area that has an older population will likely have higher **crude rates** for cancer, even though its exposure levels and cancer rates for specific age groups are the same as those of other areas. This is because cancer is more common among older people than younger people. **Age-adjusted** rates control for age effects, allowing better comparability of rates across populations from different areas.

Testing for trends across time

The department uses a statistical procedure to test for trends across time, called ordinary least-squares regression analysis. Ordinary least squares linear regression is a statistical technique used for the analysis and modelling of linear relationships between a response variable, e.g. smoking prevalence , and one, or more, predictor variables, e.g. time (in years). If the relationship between two variables appears to be linear, then a straight line can be fitted to the data in order to model the relationship.

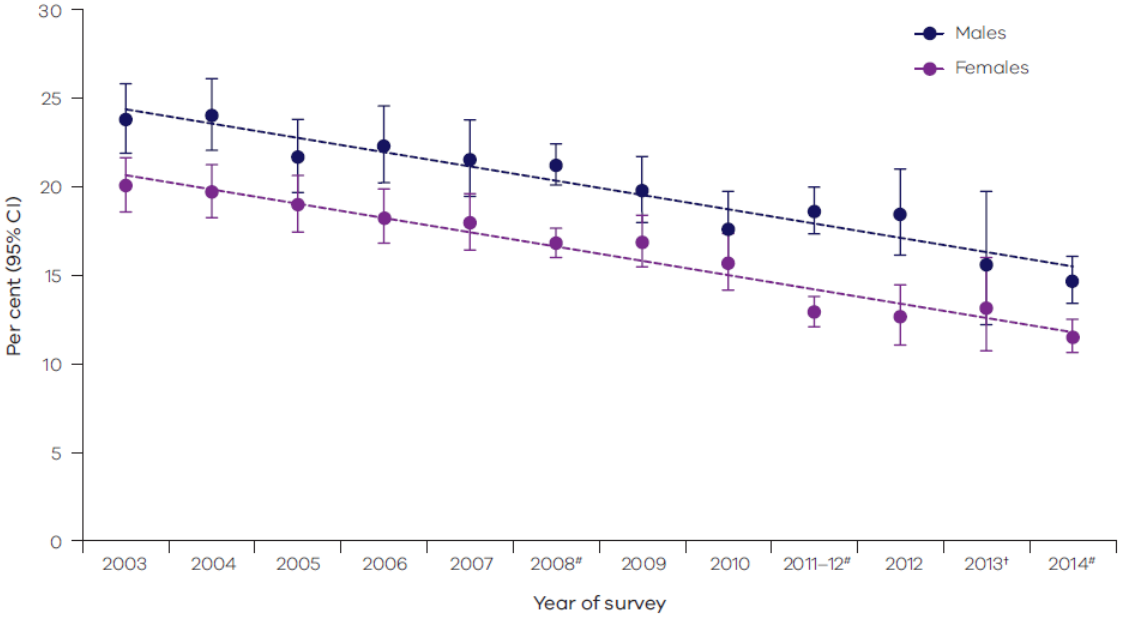
The linear equation (or equation for a straight line) for a bivariate (or two variables, such as 'smoking prevalence' and 'time') regression takes the following form:

$$y = mx + c$$

where **y** is the response (dependent variable: smoking prevalence) variable, **m** is the gradient (slope), **x** is the predictor (independent variable: year of survey) variable, and **c** is the intercept (or constant).

In the figure below, the modelling shows that smoking prevalence (expressed per cent) over time has decreased. This statistical approach considers prevalence estimates for all time points and calculates the line that best fits the data and the slope of the line, to estimate the average annual change in prevalence. The trend over time is considered statistically significant if the 95 per cent confidence interval of the slope (i.e. the regression coefficient) does not include the value zero (0).

Proportion (%) of current smokers, by survey year and sex, Victoria, 2003–2014



For further information, please contact the **Health Intelligence Unit**:
e-mail: health.intelligence@dhhs.vic.gov.au
Website: Health status of Victorians
Adapted from: Toronto Public Health Survey Data Interpretation Guide (accessed January, 2018).