TORONTO Public Health

Survey Data Interpretation Guide

Toronto Public Health (TPH) is responsible for assessing the health of Toronto's population by measuring, monitoring and reporting health determinants and risks, health outcomes and related health inequities. 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 TPH to describe health survey data. In particular, this guide explains: samples and populations, estimates, confidence intervals, the concept of statistical significance, and the coefficient of variation.

Samples vs. Populations

Suppose we wanted to know:	**What proportion of Toronto Residents are physically active?**

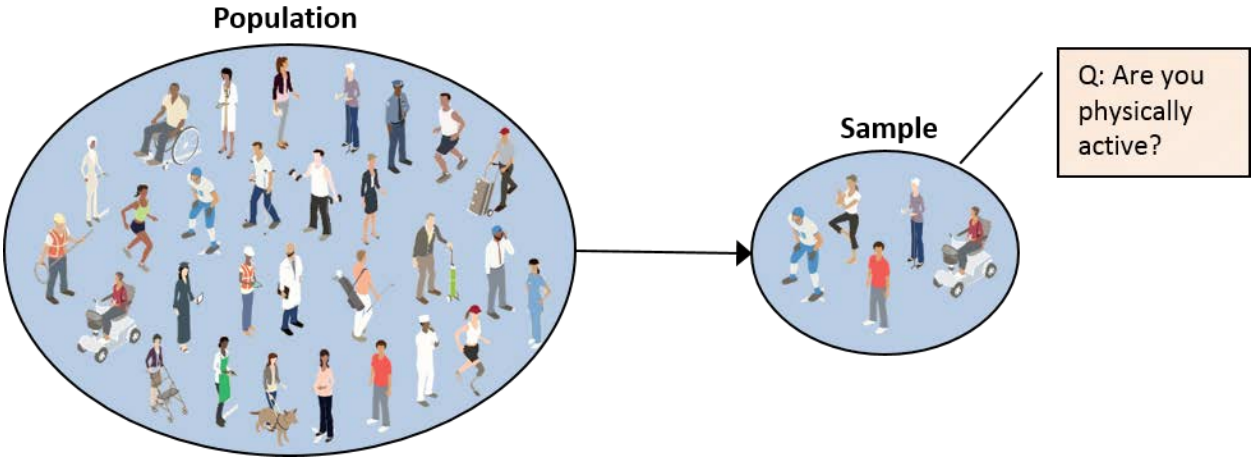
A **population** is the entire group of people we want to learn about and understand.

A **sample** is a smaller group of people from a population.

We collect information from a sample of people when we cannot gather information from every person in a population.

The information gathered from a sample is used to make conclusions about the entire population at large. Therefore, samples are more reliable the bigger they are, and the more representative they are of the entire population.

The population of Toronto is almost 2.8 million. Since it would not be practical to ask everyone about their level of physical activity, questions are asked to a **sample**: a smaller group of people selected from the broader population.

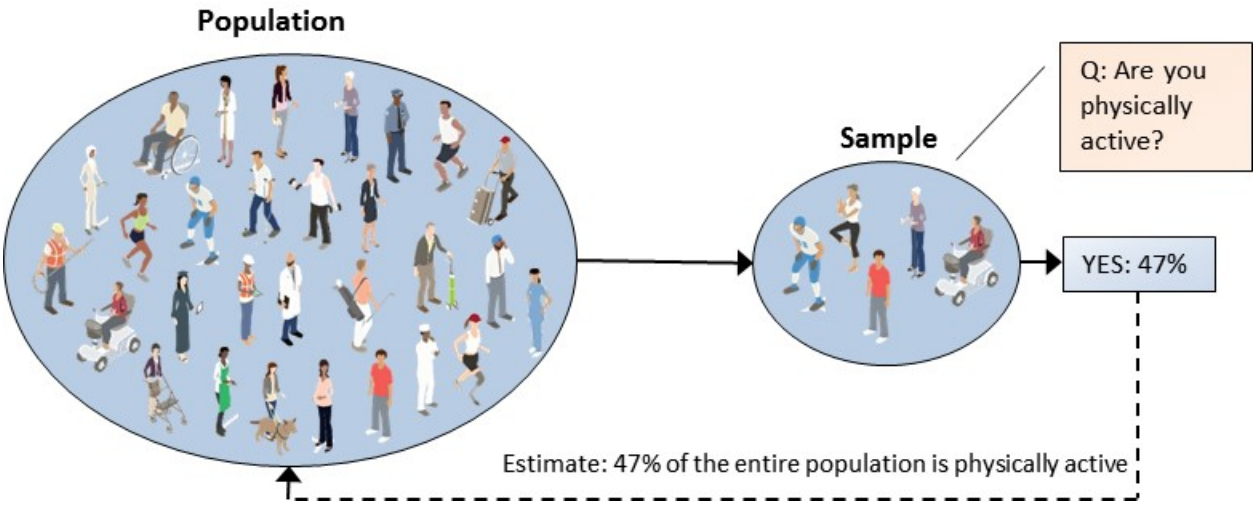


Estimates

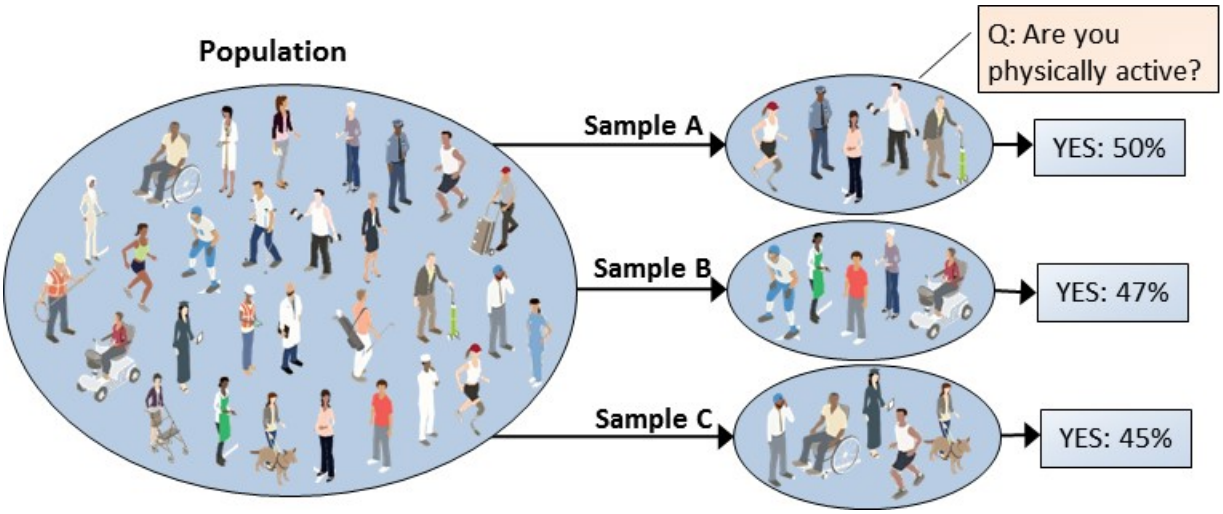
An estimate is a scientifically based approximation of the population's unknown true value. Estimates are created by using the data collected from a sample to make conclusions about the entire population. A **point estimate** is when a **single** value from a sample is used to describe the population.

What proportion of Toronto Residents are physically active?

If 47% of our single sample indicated that they are physically active and our sample has been designed to represent the population, we can reasonably estimate that about 47% of all people in Toronto are physically active.



However, if we repeated our question to different samples, we would likely get different estimates due to differences in each sample.



Confidence Intervals

Instead of a single point estimate, confidence intervals provide a **range** of possible values (that would theoretically arise from multiple samples) to create a single estimate of the population's unknown true value. Confidence intervals are also called **interval estimates**.

Most of the time, 95% confidence intervals (CIs) are used. A 95% CI is a range of numbers around the point estimate that contain the population's unknown true value 95 times out of 100. This means, if we selected 20 different samples of the same size from the population, we would expect all but one of the 95% CIs to contain the true population value.

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

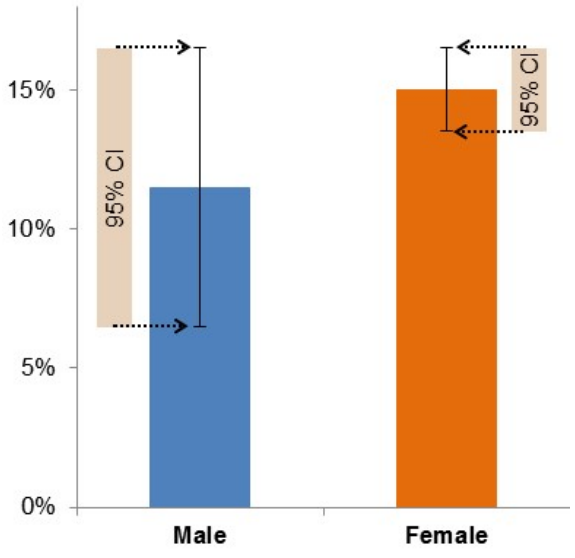
1) Sample size. Samples with fewer people provide less information, which results in a wider and less precise CI.
2) Level of confidence desired. As the level of confidence increases, say from 95% to 99%, we are more certain that the true population value is within the CI range. However, with increasing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

Confidence intervals are formed by adding and subtracting a margin of error from the estimate, which is based on the three factors above. CIs can be described in text as an estimate with a ± symbol and the margin of error value (e.g. 47% ± 4%), or as a range of numbers (e.g. 95% CI: 43% to 51%). CIs can also be displayed on graphs as error bars (I-shaped) superimposed over a bar or line that represents the point estimate (see below).

How does the proportion of Toronto residents who are physically active differ by sex?

From a Toronto sample, 46% of males (95% CI: 40% to 51%) and 48% of females (95% CI: 46% to 49%) indicated that they are physically active. For males, this means that the estimate is actually somewhere between 40% and 51%, 19 times out of 20.

Since the CI for females is narrower, the estimate for females is more precise than the estimate for males.



Statistical Significance

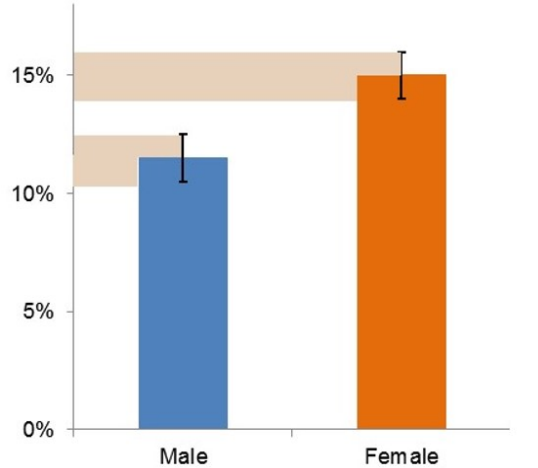
Suppose we wanted to know:	Is the difference in the proportion of males and females who are physically active, a real difference?

Tests of statistical significance tell us whether we can be confident that a real difference exists between two population groups or whether the difference is simply due to the chance nature of the people we surveyed. If the probability of the observed difference occurring is less than 5%, we usually say that the two groups are statistically different, and there is a good chance that the difference or relationship is real.

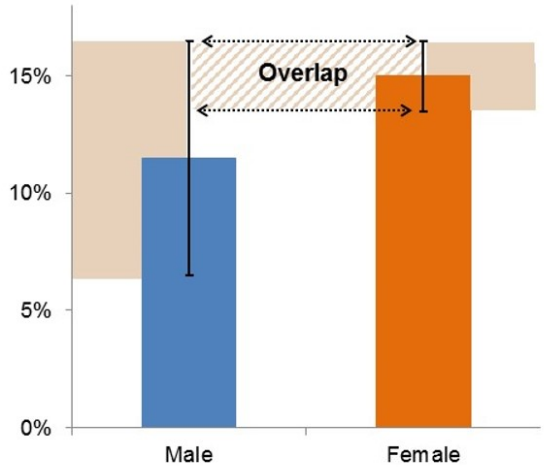
Sometimes confidence intervals (CIs) are used to determine whether estimates statistically differ from one another. With this approach, CIs that do not overlap are considered statistically significant and are most likely different from each other in the population. This is because each group has a completely different range of values for the estimate that do not overlap, so the true population value has to be different. In contrast, when CIs do overlap the estimates are not usually considered statistically different from each other. (Note: Depending on the degree of overlap in the CIs, the estimates may be statistically different. Therefore using CIs to determine statistical significance is conservative, and is best done when making multiple comparisons.)

The following figures show the same proportion of males and females who are physically active, but with different CIs. In the first figure, the CIs do not overlap. Therefore, we can conclude that females are more likely to be physically active than males.

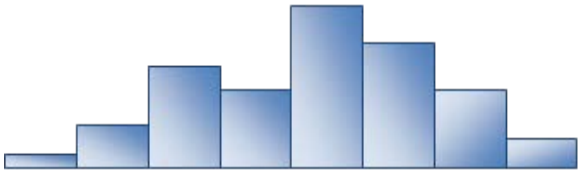
In the second figure, the CIs do overlap, so we cannot say the estimates are statistically different. Here we would say there is no difference in the proportion of males and females who are physically active.



Confidence intervals do not overlap
therefore estimates are statistically different



Confidence intervals overlap
therefore estimates are <u>not</u> statistically different



Coefficient of Variation

Suppose we wanted to know:	How much time per week do women in Toronto exercise?

If the estimate for women in Toronto who exercise is about four hours per week on average, we know that individual responses could vary from this estimate. Some women might not exercise at all, and other women might exercise daily. The coefficient of variation (CV) indicates the overall degree of variation in individual responses in relation to the average response or estimate. Therefore, the CV reflects the reliability of the estimate.

A low CV indicates that there is relatively little variation in responses which indicates a more reliable estimate.

A high CV indicates that responses differed widely, and the estimate is less reliable as an overall summary of what is happening in the population.

At TPH, the CV dictates whether estimates are releasable with the following guidelines:

A CV between 0 and 16.5 is acceptable and the estimate can be released without restriction.

A CV between 16.6 and 25 can be released with a warning that there is a high degree of variability associated with the estimate and the estimate should be interpreted with caution.

A CV greater than 25 should not be released as the estimate is unlikely to accurately describe the event within the population.

More Information

Examples of the concepts included in this document can be found in the following TPH health status reports:

The Unequal City 2015: Income and Health Inequities in Toronto
Healthy Futures 2014: Toronto Public Health Student Survey
The Global City 2011: Newcomer Health in Toronto
Toronto Health Surveillance Indicators

For further information, please contact TPH's Surveillance and Epidemiology Unit:
E: seu@toronto.ca
W: www.toronto.ca/health
T: 416-392-7450

Adapted from: Halton Health Statistics Data Interpretation (accessed August, 2015)