Sampling bias and the confidence interval

Daniel Kaplan

6/12/2020

Orientation

Note: The app used here is in the original style of the Little Apps. It has not yet been translated into the current system used for the Apps

The lesson on *What is a confidence interval?* showed how confidence intervals are supposed to behave. Specifically, a 95% confidence interval should cover the corresponding population parameter for 95% of possible samples. Achieving this performance of course depends on the confidence interval being calculated with an appropriate procedure. But it also depends on the sample being *unbiased*.

An *unbiased sample* is one where each member of the population has exactly the same probability of being included in a sample. Conversely, in a *biased sample*, some members of the population will be more likely to be included in a sample than others.

There are many real-world situations that lead to sampling bias. For example, surveys and polls almost always involve sampling bias because some people choose not to participate. This is called *self-selection* bias. A sample used in a study of an experimental cancer treatment might be biased because people who are sicker might be more likely to volunteer for the experiment than people whose cancer is in remission. Studies about exercise and health can be biased because people who like exercising might be more engaged and volunteer for the study.

When there is considerable sampling bias, it may happen that confidence intervals fail to cover the population parameter with the right frequency. I write "may happen" because there is an additional consideration, namely that the mechanism producing the bias happens to be correlated with the value of the response variable.

In this activity, you're going to create sampling bias by ...

Activity

1. Open up the Confidence intervals and sampling bias. (See footnote[1]). As described in the lesson on *What is a confidence interval?*, the graph that will be displayed is very simple: A jittered point plot of the response variable from a random sample from the population, a line indicating the population mean of the response variable, and a confidence interval corresponding to the sample being displayed in the graph.

   By default, the sample displayed in the Little App is unbiased. More precisely, the population is defined by a large data set. Each row from the data set is equally likely to be included in a new sample.

   - Press "New Sample" many times in a row, noting after each press whether the confidence interval generated by the new sample includes the population parameter. Since the default confidence level

[1] https://dtkaplan.shinyapps.io/LA_sampling_bias

is 95%, it's to be expected that the vast majority of the time the confidence interval will include the population parameter.

2. Now to introduce sampling bias. Since this is a simulation, you get to control how such bias comes about. You do that by selecting a *biasing variable*. When the simulation selects a sample, it looks at the individual levels of the biasing variable. Using the sliders, you can set the relative probability that rows with one level of the biasing variable will be included in the sample. When the sliders are all at the same level, there is no sampling bias. But when you move one or more sliders to a different level from the others, you cause the simulation to create sampling bias.

   - Choose a biasing variable, say, `sex`. Then set the slider for "males" to be zero. This will eliminate males from the sample. You can confirm this by looking at the colors of dots in the graph, which show the level of the biasing variable for each point. You can also see this in the "statistics" tab underneath the main plot, which shows the frequency in the population of the different levels compared to the frequency in the sample itself.

   Move the slider for males to 50%. Record what fraction of the sample consists of males. . . .

   Is it near 50% or less? Explain why.

3. Sampling bias doesn't necessarily cause faulty confidence intervals. To do this, there has to be some relationship between the response variable and the factors that account for the sampling bias. To make it easier to see the sampling bias, for this step and the following ones, set the sample size large, say, to n = 2000.

   - To show an extreme example ... use the Source package as `Little Apps`, `NHANES2` as the data set and `testosterone` as the response variable. With no sample biasing variable, the confidence interval on mean testosterone will behave as expected, covering the population parameter (the mean level of testosterone among people) in the appropriate proportion of sampling trials. Confirm that this is the case.

   - Turn on `sex` as the biasing variable. Sex is, of course, strongly correlated with testosterone. Turn down the relative frequency of men in the sample until you reach a situation where the confidence interval hardly ever includes the population parameter. Then arrange the sliders so that men are sampled at a higher rate than women. Depending on the direction of the sampling bias, the confidence intervals will be systematically too high or too low.

   Which leads to confidence intervals that are too high: bias toward men in the sample or bias toward women? Explain why, in everyday terms. .

   - Let's look for more natural examples where sampling bias plays out in faulty confidence intervals. Turn to the `Births_2014` data set and set mother's age (`age_mother`) as the response variable. Set `pay` to be the biasing variable. (`pay` refers to whether the mother is covered by government insurance or not.) If you used government insurance records to study the age at which women give birth, you would of course be biasing your sample to women included in the government records. Simulate this to find out whether such biasing leads to faulty confidence intervals on the mean of `age_mother`.

   Are the confidence intervals too low or too high? .

4. An important potential source of sampling bias is missing information. For example, in the `NHANES` survey data, questions regarding general health were only asked of adults. You can see this very clearly with response variables such as `height`, `weight`, and so on where the response variable is strongly correlated with whether the person is a child.

   - Keeping the sample size large, set the response variable to `height` and the biasing variable to "none". Observe that the confidence interval covers the population parameter (the mean height) appropriately.

- Now set the biasing variable to `health_general`. Even with the sliders all set to the same level, the confidence interval no longer covers the population parameter. This is because no children have been included in the sample, since there is no data available for them for the biasing variable.

Does the omission of children from the data bias the confidence interval to be low or high?

Version 0.3, 2020-08-14