Common and Rare

Carol Howald

8/12/2020

Orientation

We're continuing the story of two friends interested in how to measure poverty. You met with them at a coffee shop to show them the `NHANES` data and how to display it in the Confidence and T Little App. You had arranged the display to have `income_poverty` as the response variable and `home_type` as the explanatory variable, with the Source Package as Little App and the Data set as NHANES2. The sample size was n = 100, since that's the number of families your friends were thinking of interviewing. You had looked at the difference in mean poverty level for the "Own" group and the "Rent" group, and how it varied from one sample to another.

Your friends are a little concerned. One of the friends asks, "So every time you press New Sample, it's as if we did another set of 100 interviews on some new, randomly chosen families?"

"Right," you answer.

She asked, "And comparing the result across many such samples would let us know if our original sample really showed a difference in mean poverty?"

"Exactly," you say.

"But we can't possibly do several sets of interviews with different samples of 100 families each. We're planning to interview 100 families TOTAL. That's as much as we will have time for." She adds, "What do we do with our real data, when we just have our single sample of 100 families?"

"No problem," you respond. Let me show you what to do.

Activity

Open up the Confidence and T Little App. (See footnote[1]). Set the Source Package to `Little apps`, and the Data set to 'NHANSE2. Set the response variable to `income_poverty` and the explanatory variable to `home_type`'. Set the sample size to n = 100. Go into the App Control menu (the parallel lines icon). Check the box in the controls to show the mean values of the two groups.

1. Check the box in the controls to show the *confidence interval* on the mean. A simple test for whether there is a meaningful difference between the means of the two groups is to check whether the confidence intervals overlap or not.

   - Do the confidence intervals overlap?
   - Press New Sample several times and see whether the overlap (or lack of overlap) is consistent from one sample to another.

2. Check the "Shuffle" box next to the Apps control icon. This simulates a situation where there is no systematic relationship between the response and explanatory variables.

---

[1] https://maa-statprep.shinyapps.io/Little_App_T/

- Note whether the confidence intervals overlap. Press New Sample several times to see whether the overlap (or lack of overlap) is consistent from one sample to another.

3. Explore how sample size effects the length of the confidence intervals and whether they overlap. (Make sure the "Shuffle" box is *unchecked*, so you are looking at the actual data rather than a simulation.)

   - Set the sample size to n = 20. Measure the length of the confidence intervals using the app's measuring stick (click on the top of the confidence interval and then drag to the bottom of the interval). Do this for several new samples to get an idea of the *typical* length of the confidence intervals for a sample size n = 20. Write down your result.

   - Set the sample size to n = 200, ten times larger than the previous step. Again, measure the typical length of the confidence intervals over several new samples. Write down your result.

   - Set the sample size to n = 2000, ten times larger than the previous step and one-hundred times larger than when n = 20. Again, measure the typical length of the confidence intervals over several new samples.

   - Compare the lengths of the confidence intervals that you measures for n = 20, 200, 2000. Describe the pattern:

     – Do larger sample sizes lead to larger confidence intervals?
     – What's the ratio of the length of the n=2000 confidence interval to the n = 20 confidence interval? Can you see a simple relationship between this ratio and the ratio in sample size of 2000/20 = 100?
     – Test out your theory for the relationship between sample size and confidence interval length using the lengths of the confidence interval for n = 2000 compared to n = 200. Does your theory work?