Variation and the 95% summary interval

StatPREP Class Lesson

Orientation

Why is the word *variable* used to describe a column of a data frame? In mathematics, a variable is an *unknown* quantity. In math class, when they say "solve for x," they mean figure out what the value of unknown x is.

In statistics, a *variable* is different. It's a known quantity – known because you measured it, known because there it is, right in the data frame. This known quantity is called a variable because it takes on various different values in one row or another of the data frame.

An important technical task in statistics is to *measure* how much variation there is in a variable. The reason this is important is that we are often interested in *explaining* a response variable using one or more explanatory variables. In deciding whether a potential explanation is a good one or not, it's helpful to know how much of the variation in the response can be attributed to the explanatory variables.

There are several ways to measure variation of a quantitative variable. The most commonly used in statistics is the *standard deviation*. Another measure is the *95% summary interval*, which is perhaps easier to understand than the standard deviation.

The purpose of this lesson is to help you understand the 95% summary interval and why it is used by statisticians.

Activity

Open up the Center and spread Little App, and select the `Births_2014` data frame. Set the sample size to n = 200.

1. Set the response variable to be `age_mother`, the age of the mother at the time of birth. You'll see that there is variability, that is, not every mother is exactly the same age.

2. Perhaps the most natural way to measure the amount of variation is the length of the interval between the smallest value and the largest value.

   - Use the measuring stick built in to the app to find the length of the interval from smallest to largest age. How much is it? What are the units?

3. Leaving the measuring stick from (2) on the display, press the New Sample button. For the new sample, are the maximum and minimum values the same as in the original sample?

   - Press New Sample several times to get an idea of how much the maximum and minimum value vary from sample to sample.

- Do that again 10 times, but each time you do so use the measuring stick to find the length of the max-to-min interval for that sample. Write it down. In the end, you'll have 10 values for the length of the interval. You can get an idea for how much they vary.

4. Sometimes variables are such that the values of the maximum and minimum depend strongly on the sample. Sometimes not. One at a time, set each of these variables to be the response variable: mother's weight, baby's weight, APGAR score.

   - For each of the variables, press New Sample several times to get a quick idea of how much the length of the max-to-min interval changes from sample to sample.

5. You might have noticed that only two of the points in each sample determine the max-to-min interval. For this reason, there is a lot of variation in the interval length from sample to sample. To avoid this, statisticians have invented another kind of interval to describe variation called the "95% summary interval." The idea is to avoid making the interval depend on just two points and instead let several points contribute. The 95% interval is designed to cover *almost all* of the data points, where "almost all" means 95%.

   - Turn on the "summary interval" display and make sure the "interval level" is set to 95%. Keep the sample size at n = 200. How many of the data points are *outside* the summary interval? Explain why this number makes sense. (Hint: Think how the 95% interval is defined.)
   - Press New Sample many times to get a sense for whether the 95% interval or the max-to-min interval varies more from sample to sample.

6. In statistics, you're almost always working with a sample but your real interest is not that particular sample, but the broader population from which the sample was taken. Statisticians like to work with summaries that are informative but which also behave "nicely" with respect to sampling. In particular,

   - The summary shouldn't vary excessively from sample to sample.
   - The summary shouldn't change systematically as the sample size gets larger. You've already looked at how the max-to-min interval and the 95% summary interval vary from one sample to another, holding the sample size constant. (We used n = 200.) Now you're going to see what happens as the sample size get's bigger and bigger.

   a. Set the sample size to n = 200 and turn on the display of the 95% summary interval. Quickly press New Sample several times to get an idea of a typical value for the maximum and minimum value of the variable in

the sample. Then, use the measuring stick to mark out roughly the range from the typical maximum to the typical minimum.

b. Similarly, find the typical length of the 95% summary interval over many samples. We only have one measuring stick in the app, so use your finger widths to record that typical length. (E.g. the interval is 3 fingers wide, or whatever.)
c. Now increase the sample size to n = 1000. Press New Sample several times and keep track of how often the max-to-min interval is shorter than the one shown by the measuring stick for n = 200, and how often it is longer.
d. Similarly compare the length of the 95% summary interval when n = 1000 to the length of the interval you recorded with your fingers in step (b).
e. Do the same with a sample size of n = 5000.

- Which is more consistent across sample sizes, the 95% interval or the max-to-min interval?

Version 0.1, 2019-04-15, Daniel Kaplan