Common and Rare

Thomas Kinzeler & Daniel Kaplan

8/12/2020

Orientation

English has many words to describe what statistician's call *frequency*: common, unusual, rare, infrequent, uncommon, and so on. There's no precise, numerical meaning to these words; they are used to create an impression.

In statistics, it's helpful to have standard ways to refer to frequency. The standard deviation provides a widely accepted measure of commonality or rareness. It's not just statisticians who use this measure. They are used in psychology and social science, in physical science (where the standard deviation is often called "sigma"). In medicine, the standard deviation underlies a surprising number of diagnostic criteria. In criminal cases in court, the usual standard for evidence is "beyond a reasonable doubt." But in civil discrimination cases, for instance employment discrimination or jury selection, the US Supreme Court has described compelling evidence as lying outside "two or three standard deviations."

In this lesson, you are going to explore the use of the standard deviation as a kind of ruler for expressing frequency. For simplicity of speech, we'll adopt the English words "common," "uncommon," and "rare" to refer to specific intervals:

- common: within 2 standard deviations of the mean. For a variable with a normal distribution, 95% of cases are common.
- rare: beyond 3 standard deviations of the mean. For a variable with a normal distribution, a quarter of one percent (that is, 0.27%) of cases are beyond 3 standard deviations.
- uncommon: not common but not rare. For a variable with a normal distribution, uncommon covers about 5% of cases.

Since it's so long-winded to say "within 2 standard deviations of the mean," statisticians have adopted a scale called the *z-score*. In the language of the z-score, "within 2 standard deviations of the mean" is written $|\ z\ | < 2$. Similarly, "rare on the left size of the mean" is $z < -3$.

Activity

Open up the Little App Density Little App. (See footnote[1]). In the Data tab in the top tool bar, set the Source Package to Little Apps, and the Data set to NHANES2. Set the response variable to `height_adults`.

1. The graphic shows a traditional plot of the distribution of the response variable, called a *density plot*. If you're familiar with a histogram, you might like to think about a *density plot* as a kind of smoothed histogram without the jagged, abrupt changes from bar to bar.
2. Using your everyday experience, write down a range of human heights that you think of as "common." Similarly, write down a tall height that's rare and a short height that's rare.

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

- Within the app, click on the Graph tab in the toll bar at the top. You will see a density plot with grey bars. In the middle is the bar representing the mean. On either side of that are bars that represent values of common, uncommon, and rare on either side of the mean. Slide the corrsponding bars to the values you picked as common, uncommon,and rare. Just click on the bar and drag your mouse either left or right. The bar will move to the new value. According to the app, what fraction of cases fall into the ranges common, uncommon, and rare. Note that there are two ranges for uncommon: one on the left side of the mean and one on the right side. That's also true for the ranges for rare.

3. In the Data tab, switch the Data set to `Natality_2014`. Use `mager`, the age of the mother when she gave birth, as the response variable. Go back to the Graph tab, Drag the bars so that about 10% of the distribution is to the left of common and 10% is to the right.
    - *At what age is the leftmost boundary of 'common'?* . . .
    - *At what age is the rightmost boundary of 'common'?* . . .
4. As set in (3), common covers about 80% of the distribution. If you place the cursor on the lower common bar and leave it there for a few seconds, the z-score will be displayed above the graph, next to the bookmark icon in the tool bar. You can do this at any point on the curve.
    - *At what standard deviation measure is the leftmost boundary of 'common'?* . . .
    - *At what standard deviation measure is the rightmost boundary of 'common'?* . . .
5. Consider blood pressure. A high systolic blood pressure is generally defined to be at or above 130 mmHg. Switch back to the `NHANES2` data set and select `systolic` as the response variable.
    - *What fraction of the people in NHANES2 have a systolic pressure above 130 mmHg? . . .