Common, uncommon, and rare

StatPREP Class Activity

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 rare-and-common Little App. (See footnote[1]). Set the response variable to `height_adults` and leave the explanatory variable at `.none`.

1. The graphic shows a traditional blot 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.

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

- There are sliders in the app to let you define numerically what range of values are common, uncommon, or rare.

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.

- Within the app, set the sliders to correspond to what you wrote down for common and rare. 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. Turn on `sex` as an explanatory variable.

   What fraction of women are marked as either uncommon or rare on the tall side? What fraction of men? . . . . . .

   Similarly, what fraction of women are marked as uncommon or rare on the short side? Of men? . . . . . .

   - Adjust the sliders so that about 2.5% of women are marked as uncommon or rare on the tall side, and another 2.5% on the short side. Consequently, about 95% of women will be marked with a "common" height. Using those slider settings, answer these questions for men:

   What fraction lie outside the common range on the short and on the tall side? . . .

   For the women, you set the slider so that the short and tall fractions are roughly equal. Are those fractions roughly the same for men? . . .

4. Switch to the `Births_2014` data frame with `age_mother` as the response variable. Set the explanatory variable to `.none`. Arrange the sliders 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'?* . . .

5. As set in (4), common covers about 80% of the distribution. On the top of the graph is a scale marked in terms of standard deviations from the mean.

   - *At what standard deviation measure is the leftmost boundary of 'common'?* . . .

- *At what standard deviation measure is the rightmost boundary of 'common'?* . . .

6. Look for an explanatory variable that will split up the data into groups such that one group has a much higher fraction of uncommonly young births than another group.

   What explanatory variable did you find that satisfies the criterion? . . .

7. Consider blood pressure. A high systolic blood pressure is generally defined to be at or above 130 mmHg. Switch back to the `NHANES` data frame and select `systolic` as the response variable with `.none.` as the explanatory variable.

   - *What fraction of the people in NHANES have a systolic pressure above 130 mmHg? . . .

   Set the explanatory variable to be `diabetes`. Is there a difference in the fraction of people with diabetes who have high blood pressure compared to the fraction of people without diabetes with high blood pressure?

   - *What is the fraction of people with diabetes who have high blood pressure?* . . .

   - *What is the fraction of people who do not have diabetes who have high blood pressure?* . . .

Version 0.2, 2019-05-29, Thomas Kinzeler and Danny Kaplan, Word version