

1.5.3

Identifying and Addressing Outliers

Now that we've worked on measures of central tendency, variance, and standard deviation, let's find the outliers of the dataset. The concept of outliers is something Louise could use to better plan her campaign by eliminating extreme data points that are not representative of the data we want. For example, there may be large funding goals in the dataset that are for theater-building proposals. Louise just wants funding for her play, not a theater, so she'll be able to eliminate those data points. Let's take a closer look at outliers and how to handle them when working with data.

In datasets, **outliers** are extreme points of data; they can be much larger than the rest of the data or much smaller. But how do we define "extreme"? We can use the tools we've just learned along with some guidelines generally accepted by statisticians. There are two main techniques for determining outliers, and each technique uses a measure of central tendency and a measure of spread. We can use either the mean and standard deviation together, or the median and interquartile range (IQR) together.

NOTE

Why don't we use variance to determine outliers? We use standard deviation because taking the square root of the variance standardizes the "units" of the variance to match the "units" of the dataset. (This is also why it's called "standard" deviation.)

If we decide to use the first method—mean and standard deviation—the guideline is that any value that is more than 3 standard deviations higher or lower than the mean is considered an outlier. If we decide to use the second method—median and IQR—two guidelines need to be followed:


- Any value greater than the upper quartile plus $1.5 \times \text{IQR}$ is considered an outlier.
- Any value less than the lower quartile minus $1.5 \times \text{IQR}$ is considered an outlier.

Score to be Submitted

Based on last attempt

Score
Time spent: 1m 46s

Review

 Retake

Which method do we use? In almost all cases, the IQR rule is preferred. Medians and quartiles are **robust statistics**, which means that they are less sensitive to outliers.

Consider a county with a small population of people making a modest living. Now, imagine a billionaire moves into the county. The median income would barely change, if at all, but the mean would catapult to a much higher value. In fact, if the county is small enough, everyone but the billionaire could end up being "below average" based on the mean.

So if the IQR rule is preferred, why is there a method that uses mean and standard deviation to determine outliers? For one thing, mean and standard deviation can be calculated more quickly. Finding percentiles requires sorting the data, which can be time-consuming with large datasets. The mean and standard deviation can be calculated without sorting data, which means that our computers won't need to work as hard to perform the calculations.


Now that we can identify outliers, what do we do with them? This is a tricky question. Changing or removing data points changes the story you're trying to tell with your data. If the identified outliers are a mistake (say, the data was entered with a typo), ideally, we would just want to correct the mistake and leave the data point in the dataset; if that's not possible, we would have to throw out the data point. However, if an outlier is a legitimate member of the dataset, it's better to leave it in and tell the full story of the data.

Score to be Submitted

Based on last attempt

Score
Time spent: 3m 4s

Review

 Retake