Print

Measures of Central Tendency and Location

In the previous videos, you learned how to use histograms to describe the centering, shape, and spread of a distribution.

You can also describe these features, along with many other characteristics of the distribution, using summary statistics.

In this table, you see many descriptive statistics for continuous variables in the Impurity data set. You can see that there are 100 observations, and that none of the variables are missing any values.

Most of the descriptive statistics for continuous variables are measures of central tendency and spread. The most commonly used measures of centering are the mean and the median.

Measures of spread include the range, standard deviation, quartiles (listed here as quantiles), and interquartile range. The minimum and maximum values, which are used to compute the range, also provide an indication of spread.

You learn about the standard error in the next lesson.

In this video, you learn about measures of central tendency of the data. You also learn about quartiles, quantiles, and percentiles.

Perhaps the most well-known measure is the arithmetic average, or mean. When we calculate the mean for a variable with n observations, we use the term sample mean, and we denote the sample mean as x-bar ($\bar{x}$).

Let's say we measure impurity for a small sample of five batches. The sample mean is 5.96. Of course, we don't need to compute sample statistics by hand.

The sample mean for the sample of 100 batches of polymer is 6.12. The sample mean is the center of gravity, or the balancing point, of the distribution. What happens to the sample mean if there is an extreme value?

Let's see what happens to the mean if the largest value is 8.5 instead of 7.

To keep the distribution in balance, the mean shifts toward this higher value. If there are extreme values, or if the distribution is highly skewed, the mean might not be the best way to describe the central tendency of the distribution.

Another popular measure of central tendency, which is not sensitive to extreme values, is the median. The sample median is written x~ (x tilde). The median is the middle value of a distribution that divides the data into two equal parts. Half of the data points are below the median, and the other half are above the median.

To find the median for our five values, we sort the data from low to high. The median is the middle value, or 6.1.

If we have an even number of observations, the median is the average of the two middle values. For example, the median for these observations is the average of 6.1 and 6.3, which is 6.2.

For the Impurity data, the sample median is 5.86. The distribution is slightly right-skewed, so the mean is slightly higher than the median.

If you have data that are highly skewed, the median can be a more representative measure of the central tendency of the distribution. For example, take salaries at a company. The distribution of salaries is usually right-skewed, with the salaries for the top executives much higher than the salaries for most of the employees. These higher salaries inflate the mean. So, instead of reporting the mean salary to represent the "typical" salary at the company, the median salary is often reported.

The median divides the data into two parts, but we can divide the data into more than two parts.

For example, we can divide the data into four equal parts, or quarters. The values that separate the quarters are called quartiles. Twenty-five percent of the values fall below the first quartile, or Q1.

Fifty percent of the values fall below the second quartile, or Q2. This is the median, and 75% of the values fall below the third quartile, or Q3.

Note that, because the data values might not fall exactly at the quartiles, interpolation is used to compute these values.

For information on how JMP computes quartiles, see the Read About It for this module.

For the Impurity data, the first quartile is 4.95 and the third quartile is 7.07. The quartiles are used to create a graph called a box plot.

An outlier box plot is displayed above the histogram in JMP, along with other information that helps you interpret the centering, shape, and spread of the distribution.

Let's take a quick look at the box plot (without this other information).

The left side of the box is the first quartile, the right side of the box is the third quartile, and the vertical line drawn in the box is the median, or the second quartile.

You learn more about how to use and interpret box plots in upcoming videos.

In addition to dividing the data into halves and quarters, we can divide the data into 100ths, or percentiles.

For example, 90% of the Impurity measurements fall below 8.23. Note that percentiles are also referred to as quantiles. Percentiles are often reported for test scores and physical measurements.

For example, you might score in the 95th percentile on an exam or be in the 50th percentile for height.

You also see percentiles reported for highly skewed data, as a way of trimming off the long tail of the distribution.

Consider the time to deliver a product, which is right-skewed. In this example, 90% of the product was delivered in less than 20.9 days.

This value might be more meaningful to your organization, and to your customers, than reporting the median or the mean.

Statistical Thinking for Industrial Problem Solving

Close