Measures of Spread: Variance and Standard Deviation

The sample variance, s^2, is a measure of the variability of your data around the mean.

The variance is the average of the squared differences between each observation and the sample mean.

Let's take a look at this formula using the small sample with five impurity values.

To compute the sample variance, you take the difference between each observation and the sample mean, you square each of these differences, you add up all of the squared differences, and then you compute the average by dividing by the sample size minus 1.

You'll learn why we divide by n-1 instead of n in a few moments.

For this example, the variance is 0.763. Of course, you don't need to do this by hand, but it's important to know how the variance is computed and what the variance measures.

The variance is computed by squaring the deviations from the mean, so the unit of measure for the variance is the square of the unit of measure for the variable.

Remember that the unit of measure for the Impurity data is percent, so the variance is 0.763% squared.

What if you are studying Temperature in degrees Celsius? The variance will be in degrees Celsius squared. How would you interpret this value?

Instead of using the sample variance, it can be easier to interpret the sample standard deviation, s.

This is calculated by taking the square root of the sample variance. As a result, the sample standard deviation is in the same unit of measure as the variable of interest. This makes the standard deviation much easier to interpret.

The standard deviation for our small sample is 0.873% impurity. The sample standard deviation gives us a measure of how far, on average, the individual values are from the sample mean. For our small sample, the values are 0.873 units away from the mean, on average.

Let's look at the larger sample, with 100 batches of polymer. The standard deviation of impurity for this sample is 1.514.

A natural question is, "Why do you take the square root of the variance?" Why don't you just calculate the average distance between the values and the mean directly? Remember that the sample mean is the balancing point for the data.

Some of the deviations from the mean will be positive, and some will be negative. The sum of the deviations is zero.

Using the sum of the squared deviations and computing the variance gets around this issue. In fact, this idea of squaring the deviations is used in many of the methods you'll see throughout this course.

Let's take another look at the formula for the variance. Instead of dividing by n, the number of observations in the sample, we divide by n-1.

This value, n-1, is called the degrees of freedom. Intuitively, degrees of freedom are the number of freely varying observations in a sample-based calculation.

Let's explore what is meant by the term "freely varying." Consider our small sample. The sample mean is used to compute the deviations, and we know that the deviations sum to zero.

After we calculate the first four deviations, we know the value of the fifth deviation without having to calculate it. The fifth deviation is fixed, because it's based on the other four deviations.

Degrees of freedom are attached to many sample-based calculations. You'll see this term again in future modules, but it won't be emphasized.

What's important is that you understand, at least conceptually, why degrees of freedom are reported.

In these videos, you learned how to describe the shape, centering, and spread of your data.

In upcoming videos, you learn additional tools for graphing and exploring continuous data. You also learn how to describe categorical data using numerical summaries and graphics.

Statistical Thinking for Industrial Problem Solving

Close