Measures of Spread: Range and Interquartile Range

Although it's important to understand the central tendency of a distribution, understanding the spread or dispersion of a variable can be just as important.

The most common measures of spread are the range, interquartile range, sample variance, and sample standard deviation.

The easiest way to quantify the spread of the distribution is to compute the range. The range is simply the difference between the largest value and the smallest value.

For the Impurity data, the smallest value, reported as the minimum, is 3.078, and the largest, or maximum value, is 10.535, so the range is 7.457. That is, the Impurity values span a range of 7.457 units.

The interquartile range is another measure that is straightforward to calculate. The interquartile range is the range of the middle 50% of the values. This is the difference between the third quartile and the first quartile.

For the Impurity data, the third quartile is 7.069, and the first quartile is 4.951, so the interquartile range, or IQR, is 2.118. This means that the middle 50% of the data spans a range of 2.118 units. This is displayed as the width of the box in the box plot.

The interquartile range can also help identify potential outliers. Do you notice the whiskers that are drawn on either side of the box? Each whisker is drawn to the furthest point within 1.5 times the interquartile range from the corresponding quartile.

If your data are approximately normally distributed, individual points that are plotted beyond these whiskers are potential outliers. Here, we see one potential outlier. Notice that we're using the term potential outliers. You don't necessarily want to exclude or delete potential outliers from your analysis.

Let me explain this with an example. Here are a histogram and a box plot of 1000 observations that were simulated using a normal distribution.

Note that 7 of the 1000 observations are plotted outside the whiskers. Is there anything unique or special about these 7 observations? Should we get rid of them?

No, they just happen to be extreme by random chance alone. They're valid observations that should be included in the analysis.

You might have an outlier for a number of reasons. An outlier might occur by chance alone, as we see here. There might be a measurement error or a data entry error. The observation might be a real value, or an outlier might be the result of the shape of the distribution.

For a skewed distribution, disconnected points are not necessarily outliers.

Here's an example. The first histogram and box plot are for a variable that is right-skewed, and the second is for a variable that is normally distributed. Notice the difference in the box plots?

For a highly skewed distribution, there will generally be some points that fall beyond one of the whiskers, simply because the distribution is skewed. These points might not be outliers.

If you determine that an observation should be removed from analysis, you should have a good reason for doing this. That is, there should be an assignable cause.

In the next video, you see how to hide and exclude observations from an analysis in JMP. Note that there are statistical procedures for identifying and handling outliers, but that discussion is beyond the scope of this course.

Let's return to the discussion of box plots. The lengths of the whiskers also tell you about the shape of the distribution.

If the whiskers are approximately the same length, and if the median is centered within the box, then your distribution is likely symmetric. However, if one whisker is much longer than the other, the distribution is probably skewed.

For our skewed data, you can see that the lower whisker in this box plot is shorter than the upper whisker, and the median is not centered within the box.

Let's look at one final example. Here, the data are highly skewed, with most of the values close to zero. There are 1000 observations. The histogram alone is not very informative.

For example, can you tell how many values fall beyond 500? A box plot adds complementary information to the histogram. You can't see the box plot itself, but you can now see the observations in the tail of the distribution, and this might be of greater interest to you.

So far in this lesson we've talked about the range and the interquartile range as measures of spread. A box plot is an efficient graph of these measures. Perhaps the two most important and popular measures of variability are the variance and the standard deviation. We discuss these measures in an upcoming video.

Close