Measures of Spread: Crash Course Statistics #4

Crash Course: Statistics
https://youtube.com/watch?v=R4yfNi_8Kqw
https://nerdfighteria.info/v/R4yfNi_8Kqw

Hi, I'm Adriene Hill, and welcome to Crash Course Statistics!

So in the last episode, we talked about the middle set of data. What statisticians call the central tendency. Today, we're heading to the data of both sides of that the middle - what statisticians call measures of spread. Not confused to be with the gauge of the quality my peanut butter and jelly sandwich. Measures of spread, right, get it? Cause you spread the peanut butter and jelly?

Anyway, statistical measures of spread or dispersion tell us how dated a spread around the middle. That let's us know how well the mean or median represents the data and how much we trust conclusions based on the mean and median.

And I can hear you saying to the screen, "C'mn, Adrian, when would anyone use this in real life?" I say or may or not have said that too at some point. I have no official comment.

But measures of spread are all around. From test scores from when you found that you scored in the 99th percentile on the LSAT. Economists use measures of spread to study income inequality. Investors use them to try to identify price bubbles. Bubbles they might want to try to avoid. Gamblers use them to try to figure out how much they might win or lose. Posters use measures of spread of help calculate margins of error.

So yeah, they come up in real life and heads up - there's some math coming your way. We're not going to spend a lot of this series doing calculations, but for this one, it's important.

[Opening music]

Let's do a thought experiment to compare measures of spread. And since you're probably watching this on YouTube on right, we'll talk about YouTube viewers and their ages.

You're a Youtuber with big dreams and amazing content. But as a growing channel, you need to know more about your audience. YouTube will give you some information about this, usually in the form of a fancy chart. One of the pieces or information you can calculate in the range of your audience's age. Range takes the largest number of our data set and subtracts the smallest number in the set to give us the distance between these two extremes. The largest the distance, the more spread out our data is.

With the range, we're able to quantify the distance between our most extreme points. We can often sense the groups are different, and our ranges confirm it. If we looked at the range of your audience's age, we get a better idea of the full spectrum of people who watch.

If you have 13-year-olds watching, you might want to limit the adult content, but if you have people over 40 watching, you may still need to explain some of the slang you're using. #lit, #fam. Am I doing that correctly?

But the range won't tell you about your core audience. These are the people who you appeal to most. This might be better summarized by the interquartile range or IQR which doesn't consider extreme values. The IQR looks at the spread of the middle 50% of your data.

So in this example, the ages of your audience, the IQR will give you a better idea of who is the primary group watching you. A lifestyle guru like Bethany Mota might have an IQR of or 13-25, whereas I may guess somebody like John Oliver has an IQR that older - maybe in the range of 22-40. Their overall range can be similar; I'm there are 13-year-olds and 60-year-olds watching both of the channels, but the IQR gives us a better idea of their core audience.

So let's introduce numbers and do a little math: let's say 10 basketball players have scored the following number of points in the first part of the game: 1, 3, 3, 4 , 5, 6 , 6, 7, 8 , and 8.

The median is 5.5. That divides that data in two halves. To divide it further in quarters, we find the median of each of those halves which are 3 and 7. Q1 and Q3 respectively. The 4 quartiles here are from 1-3, 3-5.5, 5.5-7, and 7-8. The IQR is the difference between Q3 and Q1, or in this case 7 minus 3 which is 4.

If the median is closer to one of the ends of the IQR, it means that quartile has a smaller range. Since each quartile has the same number of data points, it means that for that quartile, the same amount of points are closer to each other. But we're still losing a lot of information about how to spread out all the data is, since only two of the data points are being used to calculate the range and the interquartile range.

There are measures of spread that include all of our data, just like the mean. Take the variance which can give us a better sense of how spread out the whole data set is. Let's take a scatter plot of all our data points and draw a straight line across the graph at the mean. Then draw lines from each point straight down to the mean line. Those lines represent the deviation, or difference from each point to the mean. Now imagine a square with sides the length of the deviation line. The area of all the squares for every point, divided by the number of data points, is the variance.

But it turns out the if you use the same formula to calculate the variance of a sample, it would be "biased". This is, the sample variance would consistently be a little smaller than the real variance of the population. We divide the number of samples minus 1 in order to get the sample variance to be unbiased - or a better guess for the population variance.

For example, say that the Mets, the Yankees, the Angels, the Dodgers, and the Astros have two 2, 3, 5, 8, and 8 wins each. The mean number of wins of the group of teams is 5, 25 divided by 5. To calculate the variance, we take each number and subtract the mean, square this difference, then add all of these squared differences together, and divide by the number of data points minus one. The variance of this set of baseball is 9 plus 9 plus 0 plus 9 plus 9 all divided by 4, which equals 9 squared wins. And yes, I know 9 squared wins doesn't mean anything, but when we square our numbers, we're also squaring our units right along with them.

Even though squared wins isn't an understandable unit to us, the variance is still a really useful number to have, because it tells use how much variability is in our data. In our baseball example, it tells us roughly how far each team's win record is from the mean. We'll see it pop up quite often once we get to the inferential statistics.

For now, let's go to the Thought Bubble.

Professor Hooch has hired you to analyze student's broom speeds for the Howart's Quidditch team. There are 15 Gryffindors, so you measure how long it takes them to fly around the field twice. And here's the plot of the times, in seconds, that it takes each student to complete the trip.

Looks like a few of the Muggleborn students who didn't grow up using magic took a lot longer than their classmates who grew up in wizarding families. Our mean of all students is 36.47 seconds, but if we take out the Muggleborn students, the mean is down to 29.67 seconds. Means are very easily changed by extreme values.

But the median doesn't change as much: it only goes from 30 seconds to 29.5 seconds when we pull out the Muggleborn. The range changes greatly, going from 46 seconds to 20 seconds

because the extreme values determine the high number in our range calculation.

The variance is also greatly affected since those slow Muggleborn students in inflate our mean. If we take our those Muggleborn times, the rest of our data is quite close together, reflected in the variance of about 36 seconds squared. But once we add those times back in, the variance shoots up to about 228 seconds squared, which matches our intuition that the group is more spread out.

You can see that the distance between points and the new mean are much larger than before we put the Muggleborn times back in. These Muggleborn times change our measures of spread and center.

But that doesn't necessarily mean these data are bad. We need to think about whether unusual points belong in our data or not. And we'll talk more about unusual point or outliers a little later in the series.

Thanks, Thought Bubble.

Remember that the units of variance are squared units, like seconds squared for our flying broomstick times, or baseball wins squared for our baseball example. And yes, variance is valuable, but sometime we need something with units that make just a little more sense.

Enter standard deviation. The standard deviation is the square root of the variance, which gives back the unit that we're comfortable with: seconds or baseball wins. The standard deviations of our Quidditch data would be able 6 seconds without the Muggleborn data and about 15 seconds with it.

You can think of the standard deviation as the average amount we expect a point to differ or deviate from the mean. That means that on average, we expect students to deviate from the mean time by 6 seconds. When the Muggleborn students raise our mean, our standard deviation goes up as well. In part this happens because the other points are further from the mean, since the mean became larger.

Just like the mean, the standard deviation and variance are heavily affected by unusually small or large values. So you should always look out for extreme values in your data and be aware of the influence they can have. If you see someone reporting a mean number in an article or on TV, you can use the standard deviation - if they're thoughtful enough to give it to you - to get a better understanding of how well the mean represents the data.

If the mean number of murders per state in 2015 was 307 - which it was - then a standard deviation of 10 murders shows us that 307 is a pretty good guess for the number of murders in any individual state. But if the standard deviation was 353 murders (which it was), that guess wouldn't be nearly as accurate. And this makes some sense. You woulnd't expect Montata to have nearly as many murders as a heavily populated state like New York or California.

So let's go back to our YouTube channel. So now you have a better idea of who's watching you. And you're getting more and more viewers everyday. If you want to grow more, you realize you need to diversify your audience. So you look at the standard deviation of the ages of your audience. This will give you a better idea of whether your audience have similar ages, or whether you're appealing to many age groups.

You keep adding new content and collaborating with other YouTubers to try to reach a wider audience, and it's working! Your standard deviation is getting larger, which means you're attracting a more diverse - or more spread out - audience. And congratulations, looks like you just hit one million subscribers.

As our YouTube thought experiment showed us, the different measures of spread each give us different information about our data, but they all tell us something about how spread out the data are. Sure, you can use measures of spread to grow your YouTube channel and they're important for statisticians, but they're also valuable for us non-statisticians to ponder.

And I'm going to go a little deep here, and try not to veer into the cheesy, but here's my big take away from this episode. We all have a tendency to compare ourselves to the average. We compare our income to the average income. We compare our rent to the average rent, ur intelligence to the average intelligence. We compare our weight to the average weight of someone our age. And on and on.

From these measures of spread I take away the idea that the average whatever, on it's own, can be deeply misleading. Comparing ourselves to that single statistic can give us a false sense of failure or success, depending on how the data is spread out. So maybe stop comparing yourself to the average, or if you're really insistent on ranking yourself against everybody else, go calculate the standard deviation too.

Thanks for watching. I'll se you next time.

Crash Course Statistics is filmed in the Chad and Stacy Emigholz Studio in Indianapolis, Indiana, and it's made by all of these nice people. Our animation team is Thought Cafe.

If you'd like to keep Crash Course free for everyone, forever, you can support the series at Patreon, a crowdfunding platform that allows you to support the content you love. Thank you to all our patrons for your continued support.

Crash Course is a production of Complexly. If you like content designed to get you thinking, check out some of our other channels at complexly.com. Thanks for watching.