Mean, Median, and Mode: Measures of Central Tendency: Crash Course Statistics #3

Crash Course: Statistics
https://youtube.com/watch?v=kn83BA7cRNM
https://nerdfighteria.info/v/kn83BA7cRNM

Hi I'm Adriene Hill, and welcome to Crash Course Statistics. In the last video, we tried to make sense of ginormous numbers. And teeny-tiny numbers.

Today we're going to talk about less showy numbers. The numbers stuck in the middle. The averages. The medians. The modes. They may not seem as mind-blowing. Or all that flashy. But, turns out they are really, really important.

Those middle numbers are often the ones that get ALL the press attention. Get tossed back in forth in political debate. And, they give us this little fun bit of trivia: What's the average, or mean, number of feet people have?

It's not 2. Turns out the average number of feet people has is a little less than 2. Cause the average takes into account the small number of people out there with fewer than 2 feet.

So, if you have two feet, you have more than the average number of feet. And with that, let's get what into "measures of central tendency" are and why they're useful.

[Opening music]

If your boss asks you for a report on this quarter's sales numbers but is rushing to a meeting and only has time to listen to one piece of information about the data, that piece of information you give her should probably be a measure of central tendency.

The center of a bunch of data points is usually a good example or summary of the type of data we can expect from the group as a whole. One common measure of the middle is the mean. You've likely heard it called the average - though all of these measures are sometimes called averages.

Some people call it the expectation of a set of data. The mean, or average, takes the sum of all the numbers in a data set, and divides by the number of data points. So, if 10 pregnant dogs give birth to 50 total puppies--the average litter size is 5 puppies. Each data point, in this case each litter of puppies, contributes equally to the calculation. Awwwww.

Here's another example. Say you you have ten dollars and your best friend has 20 dollars, the mean amount of cash you two have is 15 dollars. Ten-plus-twenty-divided by two. But saying that the mean is fifteen dollars, doesn't mean you each can buy that 12-dollar BFF necklace that you've been eyeing - you know, the one with the half-a-heart that fits together. You personally only have ten dollars in your pocket. The average of a set of data points tells us something about the data as a whole, but it doesn't tell us about individual data points.

The mean is good at measuring things that are relatively normally distributed. Normal means a distribution of data that has roughly the same amount of data on either side of the middle, and has its most common values around the middle of the data. Data that are distributed normally will have a symmetrical bell shape that you've probably seen before.

A distribution shows us how often each value occurs in our data set, which is also known as their frequency. Imagine you are trying to impress your new college dorm mates by guessing how many times they've each seen *Harry Potter and the Sorcerer's Stone*. Your mom is in the entertainment industry and you overheard, at her last dinner party, that 18 year olds, on average, had seen the movie five times each. That's a lot of Quidditch.

So you should guess your new friends have seen the movie five times each. (Unless you can clearly see Slytherin tattoos.) You won't be right each time, but it's your best guess. It might not be the best way to impress them though. It's not a great party trick.

Sometimes, the mean is misleading. For instance: life expectancy in the Middle Ages. As we explored in Crash Course World History, there was an incredibly high rate of infant mortality in the days before modern medicine, but the people who made it to adulthood lived relatively long lives. Because of the high rate of infant and child mortality, the average life expectancy was about thirty years. But things weren't nearly as dire as all that. Not if you actually made it to 30. In the 13th century a male who lived to 30 was likely to make it into his fifties.

To give unusually large or small values, also called outliers, less influence on our measure of where the center of our data is, we can use the median.
Unlike the mean, the median doesn't use the value of every data point in it's calculation. The median is the middle number if we lined up our data from smallest to largest.

For example, if you have two cats, Julian has one cat, and Erik has three cats, the median number of cats in your little cat-loving group would be two. When we put the number of cats in order from least to most cats, two is in the middle. But what if there's no middle number? You invite Will to join your cat group. He has an impressive - or is it excessive - total of fourteen cats. Now there are four cat owners. There is no one middle number; both two and three are in the middle.

In this case there are differing opinions on how to calculate the median, but most often we take the mean of the two middle numbers, so our median would be 2.5 cats. Meow. Meow Me-

Let's go to the Thought Bubble.

Imagine ten artists have been working for years, together, to come up with a new, fresh way to tie macrame knots. The standard square knot just wasn't inspiring them the way it used to. And finally, they do.

Viola! The abracadabra-doolittle knot!

So these 10 artists go out to celebrate. And they go to a relatively modestly priced restaurant cause macrame artists don't make all that much money.
Each of them pulls in about $20,000 a year. So the average...or mean... income in around the table is $20,000. And the median income is also $20,000.

Now, let's imagine that Elon Musk gets wind of this macrame milestone. Turns out, he's a huge macrame fan himself. Beauty in design. He couldn't miss up a chance to celebrate, so he decides to show up at the restaurant. Musk's total annual compensation, including his salary and stock options, is reportedly in the neighborhood of 100 million dollars.

As soon as Musk walks in the door, the average income in the room skyrockets to a little over 9 million dollars. But nobody else in the room is actually richer. Nobody feels any richer. The median income of the macrame artists and Musk is still $20,000 because most of our
group is still making $20,000.

And this isn't just the stuff of make-believe macrame world, it happens in real life too - the average is distorted by outliers.

Thanks Thought Bubble!

Alright, now say there's a controversial book on Amazon

called *Pineapple Belongs on Pizza*, with 400 reviews: 200 five-star reviews, and 200 one-star reviews. The mean number of stars given was 3, but no one in our sample actually gave the book 3 stars, just like no one could actually have the median of 2.5 cats.

In both of these situations, it can be useful to look at the mode. The word mode comes from the Latin word *modus*, which means "manner, fashion, or style" and gives us the French expression a la mode, meaning fashionable. Just like the most popular and fashionable trends, the mode is the most popular value. But we don't mean popular like Despacito-popular.

When we refer to the "mode" of our data, we mean the value that appears most in our data set. For our Amazon book review of Pineapple Belongs on Pizza the modes are both 5 and 1, which give us a better understanding of how people feel about the book.

These reviews are called "bimodal" because there are two values that are most common. Bimodal data is an example of multimodal data which has many values that are similarly common. Usually multimodal data results from two or more underlying groups all being measured together. In the case of our book, the two groups were the "love it" five-star group, and the
"hate it" one-star group.

Or for another example, if we made a graph of the times customers went to In-N-Out, we'd probably see two peaks because there's two groups of people: one around lunch time, and one around dinnertime.

The mode is useful here because it's an actual value that occurs in our data set, unlike the median and mean which can give us numbers that wouldn't actually occur and don't describe our data very well.

The mean time people come into In-N-Out may very well be 3:30pm, but that doesn't suggest you should expect an overflowing restaurant in the middle of the afternoon. You should be able to get your animal style burger and fries, without too much of a wait.

The mode is most useful when you have a relatively large sample so that you have a large number of the popular values. One other benefit of the mode is that it can be used with data that isn't numeric. Like, if I ask everyone their favorite color, I could have a mode of blue. There's no such thing as a mean or average favorite color.

The relationship between the mean, median, and mode can tell us a lot about the distribution of the data. In normal distribution that we mentioned earlier, they're all the same. We know that the middle value of the data (the median) is also the most common (the mode) and is the peak of the distribution.

The fact that the median and mean are the same tells us that the distribution is symmetric: there's equal amounts of data on either side of the median, and equal amounts on either side of the mean.

Statisticians say the normal distribution has zero skew, since the mean and median are the same. When the median and mean are different, a distribution is skewed, which is a way of saying that there are some unusually extreme values on one side of our distribution, either
large or small in our data set.

With a skewed distribution, the mode will still be the highest point on the distribution, and the median will stay in the middle, but the mean will be pulled towards the unusual values.

So, if the mean is a lot higher than the median and mode, that tells you that there's a value (or values) that are relatively large in your data set. And a mean that's a lot lower than your median and mode tells you that there's a value (or values) that are relatively small in your data set.

Let's go to the News Desk.

The average income of a US family grew 4% between 2010 and 2013. Those average paychecks expanded from $84,000 to over $87,000.

But not everyone is cheering. The median income fell five percent during those same years. Median family income dropped from $49,000 to just over 46 and a half thousand dollars.

This really happened, back in the years after the financial crisis. The mean income rose at the same time the median income fell. That's because families at the tip-top of the income distribution were making more money and pushing the mean up, while many other families were making less.

And even though unscrupulous politicians could accurately claim average incomes are rising and pat themselves on the back - it would be misleading. For most Americans during that stretch incomes were flat or falling.

This points to another really important point about statistics, a point we'll come back to time and time again during this series. Statistics can be simultaneously true and deceptive. And an important part of statistics is understanding which questions you are trying to answer. And whether or not the information you have is answering those questions.

Statistics can help us make decisions. But we've all gotta use our common sense. And a little skepticism.

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

Crash Course Statistics is filmed in the Chad and Stacy Emigholz Studio here in Indianapolis, Indiana, and it's made by all of these lovely 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 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.