The Shape of Data: Distributions: Crash Course Statistics #7

Crash Course: Statistics
https://youtube.com/watch?v=bPFNxD3Yg6U
https://nerdfighteria.info/v/bPFNxD3Yg6U

Hi, I'm Adriene Hill. Welcome to Crash Course Statistics.

We spend a lot of time talking about data visualization and different kinds of frequency plots — like dot plots and histograms — they tell us how frequently things occur in the data we actually have.

But so far in this series, the data we have talked about usually isn't all the data that exist. If I want to know about students loan debt in America, I am definitely not going to ask over 300 million Americans. I'm just lazy like that.

But maybe I can find the time to ask 2000 of them. Samples, and the shapes they give us, are shadows of what all the data would look like. We collect samples because we think they'll give us a glimpse of the bigger picture. They'll tell us something about the shape of all the data. Because it turns out we can learn almost everything we need to know about data from its shape.

[Opening music]

Picture a histogram of every single person's height. Imagine the bars getting thinner and thinner and thinner as the bins get smaller and smaller, till they are so thin that the outline of our histogram looks like a smooth line, since this is a distribution of continuous numbers and there is an infinite possibility of heights. I am 1.67642 (and on and on) meters tall. If we let our bars be infinitely small, we get a smooth curve, also known as the distribution of data. A distribution represents all possible values for a set of data, and how often those values occur.

Distributions can also be discrete, like the number of countries people have visited. That means they can only have a few set values that they can take on. These distributions look a lot more like the histograms we are used to seeing. Like a histogram, the distribution tells us about the shape and spread of data. We can think of distributions as a set of instructions for a machine that generates random numbers.

Let's say it generates the numbers of leaves on a tree. You may be wondering why we have a tree-leaf-number generating machine. The idea here is that everything can generate data, it's not just mechanical stuff. Its leaves, and animals, and even people. The distribution is what specifies how the knobs and dials on our machine are set. Once the machine is set, every time there is a new tree, the machine pops out a random number of leaves from the distribution.

It won't be the same number each time, though. Thats because it's a random selection based on the information the knobs and dials tell us about our distribution of leaves. When we look at samples of data generated by our leaf machine, we're trying to guess the shape of the distribution and how that machine's knobs and dials are set.

But remember, samples of data are not all the data, so when we compare the shapes of two samples of data, we're really asking whether the same distribution, these machine settings, could have produced these two different but sort of similar shapes.

If you got an especially expensive electricity bill last month, you might want to look at the histogram of your average daily energy consumption this month, and the same month last year. Check them out side by side. It is not that realistic to expect that you consumed energy at exactly the same rate this month as you did the year before; there are probably going to be some differences. But your question is whether there is enough difference to conclude that your energy consuming behaviors have changed.

When we think about data samples as being just some of the data made using a certain distribution shape, it helps us compare samples in a more meaningful way.

Because we know that the samples approximate some theoretical shape, we can draw connections between the sample and the theoretical machine that generated it, which is what we really care about. While data come in all sorts of shapes, let's take a look at a few of the most common, starting with the normal distribution.

We mentioned the normal distribution when we talked about the different ways to measure the center of data — since the mean, median, and mode of a normal distribution are the same. This tells us that the distribution is symmetric, meaning you could fold it in half and those halves would be the same, and that it's unimodal, meaning there's only one peak. The shape of a normal distribution is set by two familiar statistics: the mean and the standard deviation.

The mean tells us where the center of the distribution is. The standard deviation tells us how thin or squished the normal distribution is. Since the standard deviation is the average distance between any point and the mean, the smaller it is, the closer all the data will be to the mean. We'll have a skinnier normal distribution. Most of the data in the normal distribution — about 68% — is within one standard deviation of the mean on either side. Just like the quartiles in a box plot, the smaller the range that 68% of the data has to occupy, the more squished it gets.

Speaking of box plots, here's what the box plot for normally distributed data looks like. The two halves of our box are exactly the same because the normal distribution is symmetric.

You've probably seen the normal distribution in a lot of different places; it gets called a Bell Curve sometimes. Attributes like IQ and the amount of Froot Loops you get in a box are approximately normally distributed. Normal distributions come up a lot when we look at groups of things, like the total value rolled after 10 dice rolls, or birth weights. We'll talk more about why the normal distribution is so useful in the future.

As we've seen in this series, data isn't always normal or symmetric, often times it has some extreme values on one side, making it a little bit skewed. Age at death during the Middle Ages is left-skewed, 'cause lots of people died young, while the time it takes to fill out the Nerdfighteria survey was right-skewed because some people lollygagged.

In a box plot of data from a skewed distribution, the median will not usually split the box into two even pieces. Instead the side with the skewed tail will tend to be stretched out, and often, we'll see a lot of outliers on that side, just like the box plot of the Nerdfighteria survey times. When we see those features in our sample of data, it suggests that the distribution that generated our data also has some kind of skewed tail.

Skew can be a useful way to compare data. For example, teachers often look at the distribution of scores on a test to see how difficult the test was. Really hard tests tend to generate skewed scores, with most students doing pretty poorly and a few who still ace it.

Say we flash pictures of 20 Pokemon and asked people to name them. Here are their grades. Or another sample from a test asking people to list all 195 countries. We can compare the shapes and centers of these two groups of tests, as well as any other notable features.

First of all, these two samples look pretty similar. Both have a right skew. Both have a pretty low center, but the second test has a more extreme skew.

Bigger skewed tails usually mean that the data — and therefore the distribution — has both a larger range, and a bigger standard deviation than data with a smaller tail. The standard deviation is higher because not only are extreme data further away from the mean, they drag the mean toward them, making most of the other points just a little further from the mean too. While the direction of the skew tells you where most of the data is — always on the opposite side of the skewed tail — the extremeness of the skew can help you mentally compare the approximate measures of spread, like range and standard deviation.

But we compare the shapes of two samples in order to ask whether the shape of the distributions that generated them are different, or whether one shape could have randomly created both samples. In terms of our machine analogy, we ask whether one machine with its knob settings could have spit out two sets of scores, one that looks like test A, and one that looks like test B. Answering that question gets complicated, but we'll get there.

Now that we've examined the tails, let's look at the middle of some distributions. Almost all the distributions we've seen so far are unimodal — they only have one peak. But there are many times when data might have two or more peaks. We call it bimodal, or multimodal data. And it looks like the back of a camel, or maybe like two of our unimodal distributions pasted side by side. And, that's probably what's happening — the unimodal stuff, not the camel thing.

Often when you see multimodal data in the world it's because there are two different machines with two different distributions that are both generating data that is being — for some reason or another — measured together.

One possible example of this is the length in minutes that the geyser Old Faithful erupts. Most eruptions last either about 2 minutes or about 4 minutes, with few eruptions around the 3-minute mark, giving us a bi-modal distribution. It's entirely possible that there are two different mechanisms behind the data, even though they're being measured together.

For example, one set of conditions may lead to an eruption that's about 2 minutes long, and another — maybe a different temperature or latency — leads to a different kind of eruption, which lasts on average 4 minutes. Since these two potentially different types of eruptions are being measured together, the data looks like they come from one distribution with two bumps, but it's likely that there's two unimodal distributions being measured at the same time.

Another example that you don't need to be a geologist to understand is the race times for some marathons. While this data may look like it comes from a unimodal distribution, in reality there are two big groups of people who run a marathon: those who are competing, and those that do it to prove they can do it. There's usually one peak around the time that all the professional runners cross the finish line, and another when the amateurs do.

While we don't know for sure that bi-modal data is secretly two distributions disguised as one, it's a good reason to look at things more closely.

We'll finish today with uniform distribution. Even though we haven't mentioned uniform distributions yet, you've probably come across them in your everyday life. Each value in a uniform distribution has the same frequency, just like each number on a die has exactly the same chance of being rolled. When you need to decide something fairly, like which of your six roommates has to do dishes tonight, or which friend to take to the Jay-Z concert - the best thing you can do is use something, like a die, that has a uniform distribution. That gives everyone an equal chance of being picked.

And you can have uniform distributions with any number of outcomes. There are 20-sided dice. When you're in Vegas playing a round of roulette, the ball is equally likely to land in any of 38 slots.

There's a difference between the shape of all the data, and the shape of a sample of the data. When we talk about uniform distribution, we're talking about the settings of that data-generating machine, it doesn't mean that every sample — or even most samples — of our data will have exactly the same frequency for each outcome. It's entirely possible that rolling a die 60 times results in a sample shaped like this, even if we know the theoretical distribution looks like this.

Using statistics allow us to take the shape of samples that has some randomness and uncertainty, and make a guess about the true distribution that created that sample of data. Statistics is all about making decisions when we're not sure. It allows us to look at the shape of 60 dice rolls and figure out whether we believe the die is fair, or whether the die is loaded, or whether we need to keep rolling. Whether it's finding the true distribution of eruption times at Old Faithful, or showing evidence that a company is discriminating based on age, gender, or race. The shape of data gives us a glimpse into the true nature of what is happening in the world.

Thanks for watching and DFTBA...Q. I'll see you next time.

Crash Course Statistics is filmed in the Chad and Stacey Emigholz Studio in Indianapolis, Indiana, and it's made with the help of 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.