The Normal Distribution: Crash Course Statistics #19

Crash Course: Statistics
https://youtube.com/watch?v=rBjft49MAO8
https://nerdfighteria.info/v/rBjft49MAO8

Him I'm Adriene Hill, and welcome back to CrashCourse Statistics. This is the episode you've been waiting for. The episode we designed our shelf for. The episode that you have heard a lot about.

(Normal Distribution Montage)
"The shape of a normal distribution-- from the normal distribution-- follows a normal distribution-- in a normal distribution that we mentioned earlier-- normal distributions into-- specially if your data is normally distributed"

Well, today, we'll get to see why we talk so much about the normal distribution.

[Intro Music]

Things like height, IQ, standardized test scores, and a lot of mechanically generated things, like the weight of cereal boxes, are normally distributed, but many other interesting things, from blood pressure to debt to fuel efficiency, they just aren't. One reason we talk so much about normal distributions is because distributions of means are normally distributed, even if populations aren't. The normal distribution is symmetric, which means its mean, medium, and mode are all the same value, and its most popular values are in the middle, with skinny tails to either side. In general, when we ask scientific questions, we're not comparing individual scores or values, like the weight of one blue jay or the number of kills from on League of Legends game, we're comparing groups, or samples, of them. So, we're often concerned with the distributions of the means, not the population.

In order to meaningfully compare whether two means are different, we need to know something about their distribution: the sampling distribution of sample means. Also called the sampling distribution for short. And, before we go any further, I want to say that the distribution of sample means is not something we create, we don't actually draw an infinite number of samples to plot and observe their means. This distribution, like most distributions, is a description of a process. Take income: income is skewed, so we might think the distribution of all possible mean incomes would also be skewed. But, they're actually normally distributed. In the real population, there are people that make a huge amount of money, think Oprah, Jeff Bezos, and Bill Gates. But, when we take the mean of a group of three randomly selected people, it becomes much less likely to see extreme mean incomes, because, in order to have an income that's as high as Oprah's, you'd need to randomly select 3 people with pretty high incomes, instead of just one.

Since scientific questions usually ask us to compare groups rather than individuals, this makes our lives a lot easier, because instead of an infinite amount of different distributions to keep track of, we can just keep track of one: the normal distribution. The reason that sampling distributions are almost always normal is laid out in the Central Limit Theorem. The Central Limit Theorem states that the distribution of sample means for an independent, random variable, will get closer and closer to a normal distribution as the size of the sample gets bigger and bigger, even if the original population distribution isn't normal itself.

As we get further into inferential statistics and making models to describe our data, this will become more useful. Many inferential techniques in statistics rely on the assumption that the distribution of sample means is normal, and the Central Limit Theorem allows us to claim that they usually are.

Let's look at a simulation of the Central Limit Theorem in action. For our first example, imagine a discrete, uniform distribution, like dice rolls. The distribution of values for a single dice roll looks like this. With a sample size of one, the regular distribution of dice values, there's only one way to get a one, one way to get a two, one way to get a three, and so on. But, we want to look at the mean of say, two dice rolls, meaning our sample size is two. With two dice, let's first look at all the sums of the dice rolls we can get: 2,3,4,5,6,7,8,9,10,11,12. There's only one way to get a 2 and 12, either two ones or two sixes, but there's six ways to get a 7: one and six, two and five, three and four or six and one, five and two, and four and three. Which lends significance to the number 7, which is the number you'll roll most often.

But back to means, we have the possible sums, but we want the mean, so we'll divide each total value by two, giving us this distribution. Even though our population distribution is uniform, the distribution of sample means is looking more normal, even with a sample size of 2. As our sample size gets bigger and bigger, the middle values get more common, and the tail values are less and less common.

We can use the multiplication rule from probability to see why that happens. If you roll a die one time, the probability of getting a 1, the lowest value, is $1/6$. When you increase the number of rolls to two, the probability of getting a mean of one is now 1/36, or 1/6 times 1/6 since you have to get two ones to have a mean of 1. Getting a mean value of 2 is a little bit easier since you can have a mean roll of 2 both by rolling two 2's, but also by rolling a 3 and a 1, or a 1 and a 3. So, the probability is 3 times 1/36.

If we had the patience to roll a die 20 times, the probability of getting a mean roll value of 1 would be 1/6 to the 20th, since the only way to get a mean of 1 on 20 dice rolls is to roll a one every single time. So, you can see that even with a sample size of only 20, the means of our dice rolls will look pretty close to normal. The mean of the distribution of sample means is 3.5, the same as the mean of our original uniform distribution of dice rolls,a nd this is always true about sampling distributions: their mean is always the same as the population they're derived from. So with large samples, the sample means will be a pretty good estimate of the true population mean.

There are two separate distributions we're talking about. There's the original population distribution that generating each individual die roll, and there's a distribution of sample means that tells you the frequency of all the possible sample means you could get by drawing a sample of a certain size, here 20, from that original population distribution. Again, population distribution. And sampling distribution of sample means.

But, while the mean of the distribution of sample means is the same as the populaation's, it's standard deviation is not, which might be intuitive since we saw how larger sample sizes render extreme values, like a mean roll value of 1 or 6--very unlikely, while making values close to the mean more and more likely. And it doesn't just work for uniform population distributions. Normal population distributions also give normal distributions of sample means, as do skewed distributions, and this weird looking guy. In fact, with a large sample, any distribution with finite variance will have a distribution of sample means that is approximately normal. And, this is incredibly useful. We can use the nice, symmetric and mathematically pleasant normal distribution to calculate things like percentiles, as well as how weird or rare a difference between two sample means actually is.

The standard deviation of a distribution of sample means is still related to the original standard deviation. But as we saw, the bigger the sample size, the closer your sample means are to the true population mean, so we need to adjust the original population standard deviation somehow to reflect this. The way we do it mathematically is to divide by the square root of n, our sample size. Since we divide by the square root of n, as n gets big, the standard deviation, or sigma, gets smaller. Which, we can see in these

simulations of sampling distributions of size 20, 50, and 100. The larger the sample size, the skinnier the distribution of sample means.

For example, say you grab 5 boxes of strawberries at your local grocery store, you're making the pies for a pie eating contest, and weigh them when you get home. The mean weight of a box of strawberries from your grocery store is 15 ounces. But, that means that you don't have quite enough strawberries. You thought that the boxes were about 16 ounces, and you wonder if the grocery store got a new supplier that gives you a little less. You do a quick Google search and find a small, strawberry company's blog. They package boxes of strawberries for a local grocery store, they list the mean weight of their boxes, 16 ounces, and the standard deviation, 1.25 ounces.

That's all the information we need to calculate the distribution of sample means for a sample of 5 boxes. Part of the mathematical pleasantness of the normal distribution is that if you know the mean and standard deviation, you know the exact shape of the distribution. So, you grab your computer and pull up a stats program to plot the distribution of sample means with a mean of 16 ounces and a standard deviation of 1.25 divided by the square root of 5, the sample size. We call the standard deviation of a sampling distribution the standard error, so that we don't get it confused with the population standard deviation. It's still a standard deviation, just of a different distribution.

Our distribution of sample means for a sample of 5 boxes looks like this. And, now that we know what it looks like, we can see how different the mean strawberry box weight of 15 ounces really is. When we graph it over the distribution of sample means, we can see that it's not too close to the mean of 16 ounces, but it's not too far away either. We need a more concrete way to decide whether the 15 ounces is really that far away from the mean of 16 ounces.

It might help is we had a measure of how different we expect one sample mean to be from the true mean, and luckily we do: the standard error, which tells us the average distance between a sample mean and the true mean of 16 ounces. This is where personal judgment comes in. We could decide, for example, that if a sample mean was more than 2 standard errors away from the mean, we'd be suspicious. If that was the case, then maybe there was some systematic reduction in strawberries, because it's unlikely our sample mean was randomly that different from the true mean. In this case, our standard error would be 0.56. If we decided 2 standard errors was too far away, we wouldn't have that much to be suspicious about. Maybe we should hold off on posting a nasty comment on the strawberry farmers blog.

Looking at the distribution of sample means helped us compare two means, but we can also use sampling distributions to compare other parameters like proportions, regression coefficients, or standard deviations, which also follow the Central Limit Theorem. The Central Limit Theorem allows us to use the same tools, like distributions, with all kinds of different questions. You may be interested in whether your favorite baseball team has better batting averages, and your friend may care about whether Tylenol cures her headache faster than ibuprofen. Thanks to the Central Limit Theorem, you can both use the same tools to find your answers.

But, when you look at things on a group level instead of the individual level, all these diverse shapes and the populations that make them converge to one common distribution: the normal distribution. Voila! And the simplicity of the normal distribution allows us to make meaningful comparisons between groups, like whether hiring managers hire fewer single mothers or whether male chefs make more money. These comparisons help us know where things fit in the world.

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

CrashCourse Statistics is filmed in the Chad and Stacy Emigholz Studio in Indianapolis, Indiana, and it's made with the help of all these nice people. Our animation team is Thought Cafe. If you'ld like to keep CrashCourse 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.

CrashCourse 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.

[Outro Music]