Plots, Outliers, and Justin Timberlake: Data Visualization Part 2: Crash Course Statistics #6

Crash Course: Statistics
https://youtube.com/watch?v=HMkllhBI91Y
https://nerdfighteria.info/v/HMkllhBI91Y

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

Last time, we left off talking about different data visualizations, the ones we encounter every single day, whether it's a chart on the subway telling us the prevalence of heart disease in different age groups, or a histogram on BuzzFeed showing us how many times people use Lyft each week. These visualizations allow us to get to know data with our eyes, and today we'll dive deeper into data visualization and make all sorts of beautiful graphs and talk about some really extreme situations, like the person who watched *Sandy Wexler* on Netflix like four hundred times, which seems high.

[Intro music]

Last episode, we looked at histograms, which used the height of a bar to show how frequently data occur. We can also use this format to make a dot plot. A dot plot takes a histogram and replaces the solid bars, which use their height to show frequency, with dots. There's one dot for each data point contained in the bar. So we can just count the number of dots to find out how many there are. The dot plot for our olive oil data looks like this, unsurprisingly similar to the histogram for that data. Or check out this dot plot of how often this sample of people called their moms this month.

This gives us a nice way to explore the general shape of our data, but we still lose information about the individual data values, just like with the histogram. Occasionally, we want that extra information. Enter the stem and leaf plot. A stem and leaf plot is a cousin of the dot plot. It also gives us information about data and their frequencies by stacking objects on top of each other. However, stem and leaf plots use values from the raw data instead of dots.

So we'll turn our olive oil dot plot into a stem and leaf plot. And no, I'm not going to explain my olive oil fixation. First, we need to split each data value into a stem and leaf. Stems are related to the "bins" or bars in our histogram or dot plot. Take our dot plot for example. Each stack of dots might represent a range of five ounces, from zero to four ounces, five to nine ounces, all the way up to a bar with all the data in the eighty to eighty-four ounce range. The stem for a bin of data is the digits that all the values in the bin have in common. For the ten to fourteen ounce range, each value has a one at the beginning of the number, so the stem is one. For the eighty to eighty-four ounce range, the data all have an eight at the beginning, so the stem would be eight. We can have larger stems too. If the data went all the way up to two thousand six ounces, we could have a stem of two-zero-zero, but that's probably too much for our olive oil example.

Now that we have all of our stems, we can add the leaves. Each stem, like in a real plant, can have multiple leaves. They're stacked on top of each other, so that the height of the stack shows you how frequently data appear in that bin, just like a dot plot. The actual leaf is the rest of the digits that are not in the stem. If one of our data points is thirteen, and the stem for that range is one, that takes care of the one, so the leaf is three. Leaves appear in numerical order, from the stem out. So leaves that are smaller digits are closer to the stem.

From a distance, stem and leaf plots look a lot like a dot plot or histogram. If you squint your eyes, the leaves almost look like bars or dots. Unsquinting them will allow you to see even more information than a histogram or a dot plot will tell you. You get to see what the individual values are, and how they're spread out within a bar. Stem and leaf plots are usually flipped on their sides, so the stems are listed vertically and the leaves extend horizontally. Here's a stem and leaf plot with the number of pieces of gum each of your extended family members has chewed in the last month.

Now let's talk about boxplots. Boxplots use some of our measures of central tendency and spread to visually display our data. A boxplot is also called a box and whiskers plot. It has two major parts: the box and the whiskers. The box is a rectangle that stretches across the inter-quartile range of our data, from Q1 to Q3. At the median, there's a line splitting the rectangle into halves. If one of those halves is larger than the other, that quartile is more spread out. Since each quartile has the same number of data points, the smaller the quartile, the less spread out that portion of the data is. Imagine the difference between fitting twenty clowns in a car and fitting twenty clowns in a regulation-sized football field. Same number of clowns, more space to make balloon animals. Attached to either end of this box are the whiskers, which help show the minimum and maximum of all the data, as long as it's within one and a half times the inter-quartile range of the median. This value sets our fences. We use one and a half times the inter-quartile range, because most of the data will be within this range, especially if your data is normally distributed. We'll get into this more in future episodes.

Most of the data will be inside the fences. Any data outside is flagged as a potential outlier. It can be tempting to think of outliers as data that's wrong somehow, but that's not always the case. Values outside the whiskers are less likely than data near the boxplot, but they're not impossible. For example, it's pretty unlikely that if you dial random numbers into your phone you'll call a Domino's Pizza, but it is possible. Rare values do happen. Keeping those rare but possible values can be important. When the local news shows you a boxplot of local rents and decides that the bottom one thousand rent values are outliers, the graph they display could be misleading. Those rents are real values that you could expect. Taking them out will make your visualization less informative, and might lead you to think that the average rent is higher than it actually is. However, some values that are flagged as outliers may not be expected in your data at all. Perhaps Neymar snuck into your amateur pick-up soccer game without you knowing. His off the charts agility scores are not representative of the population you're interested in, since he's Neymar, not an amateur. Or maybe you made a typo on your spreadsheet and wrote five hundred pounds instead of five pounds for your data on the weights of pet teacup pigs. That'd be a giant teacup pig. The problem is you may not always know the difference between a point that's valid but rare and one that's a mistake. Since we need a way to decide, it's useful to have a pre-set cutoff for when we discard the data.

To see how boxplots can help us look for these outliers and compare data from two samples, let's jump to the Thought Bubble. Justin Timberlake has a new album. This American-born singer and songwriter has had quite the career. I mean, he did bring sexy back. Our writer Chelsea wanted to know how going solo affected the songs he wrote. Specifically, she wanted to know the number of unique words he used per song. To satisfy her curiosity, she made a boxplot for a sample of Justin Timberlake songs once he'd gone solo, and one for a sample of songs he sang with NSYNC. The first thing we might notice is that the medians are pretty different. The median number of unique words in a Justin Timberlake song is higher than the median number of unique words in an NSYNC song. JT has a median of a hundred and twenty-nine words, versus a median of eighty-nine back in his NSYNC days. Guess we shouldn't be surprised, coming from a band that had a song titled "Bye Bye Bye." So it seems like JT may have developed a larger lyrical vocabulary when he went solo. Maybe Lance Bass was holding him back. Anyway, you might also notice that the box part of the NSYNC boxplot is a lot smaller. The squished nature of the box plot shows us that NSYNC songs have a relatively similar amount of unique words. The box plot also shows you some potential outliers to look at, shown by the points that are outside the fences of our box plot. Let's look at a song that's marked as a potential outlier in the Justin Timberlake box plot. The song is "Chop Me Up", and it has two hundred and fifty-seven unique

words, which is a lot, since the median number of unique words for a JT song is a hundred and twenty-nine. It's definitely outside the fences.

Thanks, Thought Bubble. We don't want to throw out data just because it's extreme, and "Chop Me Up" isn't part of some super-experimental Christmas album, so it's hard to tell if this is a valid data point. To get around this uncertainty, we apply our preset rule. There isn't one set rule for handling these extreme values, there are many. For now, we'll use our boxplot method, and get rid of the "Chop Me Up" data because it's outside the fences. Remember, statistics is all about uncertainty. I'm not sure if the number of unique words in "Chop Me Up" is just a rare value, or whether it's the lyrical equivalent of Neymar in a pick-up soccer game. We still have to make decisions.

For all the nerdfighters out there, you may have heard of Hank's annual Nerdfighteria census, and while you're interested in taking it, you might wonder how long it takes to fill out. You don't have all day. So you use your new data viz skills to create a boxplot of the data, and wah-wah. I can't even see the box or even the whiskers through all those extreme values. It looks like some nerdfighters were very thorough, or were very distracted by other things. Eight thousand minutes is a hundred and thirty-three hours. The plot isn't wrong, per-se, but it's not very informative, since we can't get much useful information from it. We don't have any better idea of how long it's gonna take to fill out Hank's survey. When you make or see a data visualization, it's important to remember that its job is to actually give you information. If it doesn't do that, it's not worthwhile.

Now let's go back to frequency plots and talk about one last method for visualizing quantitative data, the cumulative frequency plot. Cumulative frequency plots are like histograms, but instead of the height of a bar telling you how much data is in that specific bin, it tells you how much data is in that bin, and all previous bins. That's why it's called cumulative, it's the frequency of all the points we've accumulated up to this point. It's like a small fish getting eaten by a bigger fish, which gets eaten by an even bigger fish, and on and on. Each fish is now full of the fish it ate, and the fish that that fish ate. And side note, your odds of being killed by a shark, a noted fish eater, are about one in three point seven million. Back to our cumulative frequency plots. These plots have their moment to shine when we want to answer a question like how many Justin Timberlake songs have a hundred and sixty unique words or fewer? The cumulative frequency plot looks like this, and here's the bar that answers our question. We could also get this information by counting all the songs in the bars that are a hundred and sixty or less on our histogram, but that's more work.

Now that we've seen some good graphs and some bad, we can apply our newfound knowledge anytime we see data visualizations, which will be all the time. This, I promise you. I mean, like, until the end of time. On the bus, in your health app, or during your boss's annual company-wide meeting, you'll know that graphs are only as good as the information they communicate. If you see a bad graph out there, say something. Ask questions. Be skeptical. I'm coining a new DFTBA today: DFTBAQ — Don't Forget to Be Asking Questions. It's another way of being awesome. I want it that way. The world wants it that way. And remember, it's not just gonna be you, it's gonna be me too.

All right, I'm gone. See you next time. And yeah, I know, "I Want It That Way" was Backstreet Boys.

Crash Course Statistics is filmed in the Chad & 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 crowd funding 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.