Charts Are Like Pasta - Data Visualization Part 1: Crash Course Statistics #5

Crash Course: Statistics
https://youtube.com/watch?v=hEWY6kkBdpo
https://nerdfighteria.info/v/hEWY6kkBdpo

Hi, I'm Adrian Hill, and this is CrashCourse Statistics.

So for the last few episodes, we've discussed ways to summarize data using numbers. We used measures of central tendency, and measures of spread. But sometimes, it can be helpful to actually see your data in addition to having numbers to describe it.

Data visualizations are important to understand, because you'll see them pretty much everyday, in the news, on Facebook and magazines. Maybe I'll make an infographic of all the places we see data visualizations.

[Intro music]

There are two main types of data that we might encounter, categorical and quantitative. Quantitative data are quantities, numbers that have both order and consistent spacing. For example, how many ounces of olive oil are in each American home. If three families told you how many ounces of olive oil they have, you could put them in a meaningful order- from least to greatest, or greatest to least. This order also has consistent spacing. An increase in one ounce of olive oil is the same, whether you go from zero to one ounce, or from a hundred to a hundred and one ounces. These properties allow us to do simple math with the data, like taking the mean, or calculating the standard deviation.

Categorical data doesn't have a meaningful order or consistent spacing. For example, favourite kind of pasta. You might like penne, rotini, linguine, even angel hair. But there's no objective way to put those pastas into a meaningful order. Is penne truly better than linguine? Where does rotini fit in? It would be pasta madness to try to put them in order!

The simplest way to display categorical data is to make a frequency table. A frequency table shows you all the categories and all the numbers of data points that fall in that category. In other words, it's frequency. To change a frequency table into a relative frequency table, we just need to take each raw frequency and divide by the number of total points to get a decimal between zero and one. Some of you may be used to reading decimals as percentages, but if you're not, just multiply by a hundred to get the percentage. For linguine, we have ten divided by fifty, which is point two, or twenty percent of the group.

Relative frequency tables have the benefit of being easy to compare. No matter what we're measuring or how many data points we have, it's easy to compare percentages. If twenty percent of people like linguine, we can see that's a smaller percent than the sixty-seven percent of people who like pineapple on pizza, or greater than the ten percent of my family who thinks statistics are scary.

The relative frequency table for a favourite pasta might look like this. We can also add more than one variable to our frequency table. We could ask people to rate their favourite pasta sauce, and make a combined frequency table, or a contingency table of both pasta and sauce preference. If I were planning a party, and needed to pick some pasta for the group, my best bets would be the rotini with red sauce, and penne with red or white sauce. And because I'm planning a party, and because I'm having food, I did look it up- the chance of death by choking on food in the U.S in a given year is one in one-hundred thousand six-hundred eighty-six.

But sometimes, we don't want just numbers in our visualization. Earlier in the series, I talked about how it can be hard to wrap your head around numbers, especially when they get really big or really small. There are other more visual ways to represent categorical data.

One way to do this is with a bar chart. A bar chart uses the frequencies that we saw in our frequency table to create bars that have a height equal to the frequency. That way, we can compare the height of bars instead of looking at raw numbers. Here's a bar chart representing the pasta data we saw in our original frequency table. You can see that penne is by far the most chosen pasta, and how it compares to angel hair. Bar charts display a lot of information in a very simple graph. They can also display the frequencies of multiple variables. Let's say we want to compare each of these pasta types with either red or white sauce. We can either stack frequencies, so it gives us the same information as our contingency table, or we can have bar charts side by side.

Pie charts are another way of displaying categorical data. They use the relative frequency of categories to portion out pieces of a circle, just like a pie. The higher the relative frequency, the bigger the slice of pie a category gets. Pie charts are useful, because our eyes are pretty good at comparing slices. Our pasta data in a pie chart looks like this. Pie charts are great at visually displaying one variable, but they struggle to effectively display more than one variable, like our pasta and sauces contingency table.

Another way to display categorical data is a pictograph. Pictographs represent frequency with pictures. A picture like the ball in this basketball participation graph will represent some number of units, say a hundred kids. So if Riverdale High had five-hundred and fifty students participate in their basketball programs, then the graph would show five and a half basketballs. Sometimes pictographs represent frequencies by increasing the size of the picture instead. And it's not wrong, but it's more difficult for us to visually compare, especially for small differences, which can be misleading. Plus, at a casual glance, we don't know what the size difference means. Are we comparing the diameter of the basketballs, or are we comparing their areas?

This is channel two news. Looks like all you students out there are really hitting the books. Data from the US department of education shows the graduation rate has been climbing. So way to go, everybody! You're passing the test of life with flying colours. Let's push that stack of books even higher.

So that last pictograph, not at all to scale. See how the stacks of the books are not proportionate? It shows a difference of five percent from seventy-five percent to eighty percent with a stack of books that is over double the height of the seventy-five percent stack. This makes the difference seem huge, because the axis doesn't start at zero. And yet, an increase of eighty to eighty-one percent is shown by two stacks that are barely different in height, even though the five percent difference looks huge. Always keep an eye on those axes.

Let's loop back to quantitative data, which, as you'll remember, have a meaningful order and consistent spacing. Frequency tables can also be used to display quantitative data like age or height or ounces of olive oil in your house. We just have to create categories out of our quantitative data first. We do that with a process called binning. Binning takes a quantitative variable and bins it into categories that are either pre-existing or made up. For example, I can say that zero to fifteen ounces of olive oil is very little, sixteen to thirty-two ounce is average, thirty-three to forty-nine ounces is a lot, and fifty plus ounces is excessive, like suspiciously excessive, like Will's fourteen cats excessive. Why do you need so much olive oil?

Anyway, once I've binned my data, I can create a frequency table or relative frequency table, just like with our pasta example. It might look something like this. Binning is most useful when there's pre-existing bins for our data. Like you can divide age in years into the bins- child, teen, adult, and older adult, because those are pre-existing categories. We can also take a score on a depression test

and create two bins- clinically depressed and not clinically depressed. You can see from this example that bins don't have to be equally spaced, but if you see quantitative data that has been binned, make sure that the way it's divided up was appropriate for the situation. Unequally spaced bins can be misleading, unless there's a real world distinction to back it up.

Say politician X wants to make himself look popular, but it seems like people in their thirties really hate him, probably because he said the reason they can't afford a house is their brunch habit. Politician X wants to hide the fact that over eighty percent of people in their thirties say they won't vote for him, so he does some rebinning. Traditionally, the data are binned roughly by decade- eighteen years old to twenty-nine years old, thirty years old to thirty-nine years old, forty to forty-nine, you get the point. But Mr.X needs to hide those hateful thirty-somethings in the data. The old chart looked like this. But politician X decided to split up the thirty-somethings to make his numbers look better. He moved the data around to hide the glaring group of thirty year old dissenters. Instead of showing the truth that thirty-somethings despise him, we see a more positive view of his popularity. By splitting the thirty-somethings and putting them into two other larger groups, he can obscure their political dissatisfaction. Looking at his new table, he'd win the popularity vote in each of the five new bins. If I don't show you the number of voters per bin, it seems legit.

Another categorical graphing method we can apply to quantitative data is bar charts. When we use bar charts for quantitative data, we squish the bars together, so that their touching, and we call them histograms. The bars are squished together because the data are continuous, which means the values in one bar flow into the next bar. There's no separation like in our categorical bar charts. In histograms, like bar charts, the height of the bars tell us how frequently data in a certain range occur. A histogram also gives us information about how the data is distributed. We can estimate where the mean, median, and mode of our data are, as well as how spread out the data is. Look at this histogram for our olive oil data. For this histogram, we can see that the range of data is approximately eighty-five, since it covers the value from zero to eighty-five ounces, and that it's right-skewed, the tail is to the right, and it's centre is around twenty-five ounces. The histogram gives us more information about the data than a frequency table does. But they're still obscuring what the specific data values are.

If you read the news or watch the news, you will see these representations over and over and over. You will likely see far more of these charts and graphs than you will create. The big takeaway here as a consumer of these things is to look closely at what the visualization is actually telling you, or maybe trying to hide from you. These charts and graphs give us another way to comprehend numbers, and to see the big picture.

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

CrashCourse Statistics is filmed in the Chad & Stacey 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'd like to keep CrashCourse 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. 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.