1/30/2019

We talked about categorical data in contrast to numerical data. Numerical data is basically numbers, but sometimes we represent categories as numbers called "dummy variables", so not all numbers represent numerical variables.

An example of a number that is not numerical is a jersey number. Other examples are social security numbers or credit card numbers.

The easiest test to make is to ask yourself: does it make sense to the take average of these numbers? If not, it's really a categorical variable (usually a nominal variable) in disguise. What if I take the average of everyone's phone number? Or zip code? Not much. What about your ages or height? That makes sense so it really is a numerical variable.

Numerical variables count or measure things. They include things like height, number of children living at home, commuting distance to work, age, test scores (not letter grades), or square footage of a house.

Among numerical or quantitative data there are two types called interval and ratio. Ratio data can be described as having a "true zero", but I find the most reliable test is a ratio test. If I divide two values in the list of numbers, does the ratio make sense and does it make sense consistently? If yes, then it's ratio. If not, then it's interval.

Consider temperature. If we compare a day that is 30-degrees with a day that is 60-degrees, does it make sense to say that the 60-degree day is "twice as warm" as the 30-degree day? Not really, so this is interval data.

On the other hand, if a child is 36-inches tall, and an adult is 72-inches tall, is the adult "twice as tall" as the child? Yes, this makes sense. If I "stack" two 36-inch-tall children on top of each other, they will be the same height as the adult. So, this is ratio data.

The difference between interval and ratio data won't affect our statistics or graphs too much in this course, but sometimes it will matter when we interpret data, and whether finding percent change makes sense.

Percentiles

Percentiles apply to numerical data. (It must at least be ordered.) Often, on standardized test scores you'll get a percentile score. What does this mean?

If you score in the $80^{th}$ percentile, this means that 80% of those taking the test scored the same as you or less well.

Put another way, 20% did better.

We can apply this concept to any sort of numerical data.

If we do this by hand, we order the data from smallest to largest. We can find where the percentiles are by counting the data.

Suppose I have a data set that has 300 numbers in it. The 10th percentile is at $10\% \times 300 = 30$. So the 30th values in the (ordered) data set marks the 10th percentile. The 55th percentile is $55\% \times 300 = 165$, so the 165th term in the ordered list is the 55th percentile. You would report the value you find at that position, not the position.

If we know that a value appears in the 28th position after sorting the list in a dataset of 175, we can divide $28/175 = 0.16$ or 16th percentile. 16% of the data is less or equal to that value (whatever value it is), and so 84% is more.

In Excel, there is a function called PERCENTILE that let's you find the percentile value from a data set without sorting. Just type =PERCENTILE( highlight the column where the data is, then after a comma, specify the percentile needed with a % sign or in decimal form.

You may also see references to "deciles" which go by 10 percents, so the 2nd decile is the same as the 20th percentile. You will also see references to quartiles. The first quartile is the same as the 25th percentile, and the 3rd quartile is the same as the 75th percentile (since ¾=0.75). The median is the 50th percentile.

Video on percentiles here: https://youtu.be/TWFUNQiXbVM

Line Graphs display what is called time-series data. This is data that is gathered sequentially. And it's typically paired with a time measurement like years, months, days of the week/month, quarters, etc. in order.

In Line Graphs be sure that you have a descriptive title and axis labels. While line graphs can plot more than one data set at the same time, generally, you don't want it to be too busy or difficult to read. How much data can go on one graph depends on the kind of data being graphed. As with the pie chart, it's best to avoid 3D effects (shadows are okay but not perspective) because it can make the graph harder to read, or misleading—pretty, but still misleading.

You can watch this video: https://youtu.be/wovEy1iR9lU