Gmail **Meredith Dodd <meredith.m.dodd@gmail.com>**

Notes

1 message

Meredith Dodd <meredith.dodd@woz-u.com> Tue, Mar 31, 2020 at 12:21 PM
To: "meredith.m.dodd@gmail.com" <meredith.m.dodd@gmail.com>

Ok, welcome all to this workshop on how to calculate measures of central tendency and distribution with MS Excel. For those of you without Microsoft products, you don't need it – you can use Google Sheets just as well. All of the same formulas apply.

I'm going to go ahead and pull up a dataset on different cat breeds. Because who doesn't love a good cute, cuddly kitten? We could all use more things that make us smile in the current health crisis!

Now the first thing I want you to take in with this dataset is that there are no spaces in data names. It uses something called "CamelCase" – meaning that you start each word off with a capital letter, like the humps on a camel. This is a best practice in working with data! Most data programs will foul up with spaces, so get used to naming things without them as early as possible.

We are going to be focusing in on the AvgKittenPrice, since measures of central tendency and distribution can only be done on continuous, or quantitative, variables. You don't have to have whole numbers, but you do need numbers that mean something, and don't just represent a category.

Another best practice is to make sure you label everything! You don't want to be scratching your head later, wondering what you did. 33 weeks goes by quickly, but you're going to be cramming so much information in that some is bound to fall out. That's ok, as long as you know where to look back and find it! So I'm going to start by labeling one of these cells with a header of the variable name, and then the next one as the mean. It doesn't matter where you label or put your information – you can choose any empty cell. I tend to do mine on the upper right, but you could easily do it and the bottom of your dataset too, or wherever makes sense to you.

Mean is just another word for average. They are the same thing. You find the mean when you add everything up, and then divide by the numbers you have. Scary, considering you have 68 rows of kitten breeds, right? Well, don't worry, because you don't have to do it by hand. Excel has a great formula for you – it's called =average.

Once you start typing, you'll see some suggestions pop up. You want the first one. You can double click on the name, or you can ago ahead and finish typing out the word average and then add an opening parentheses. Parentheses are these round guys here and come above the 9 and 0 on your standard keyboard. You'll use them a lot too as you get to programming, so it's a good idea to learn their name now!

Now you're in a holding period. Excel is ready to spring this formula into action, but it doesn't know what you want to take the average of! You have two options. If you want all the data in a particular column, then the easiest thing to do is just click on it's top letter, like this. That will highlight the entire row, and you don't even have to do any scrolling! Then just close the parentheses up and you are off and away!

The other option is to highlight the cells you want to use. If I open up the formula again, you can see the method and that you get the same result. Now I'm going to delete that, because I've already done it and we don't need a duplicate.

So that's mean! Not too difficult. And now check in with median, which is the middle of your data if you lined them all up in order from smallest to largest, or vice versa. The formula for median is =median(

How about we try for mode? Start off by labeling a cell for mode. Mode is the number that repeats the most, or the one that has the highest frequency. It is probably used the least, but still nevertheless good to know. The formula for mode is =mode(

So you may have noticed something here. While the mean and the median are pretty similar (about 15 points off), mode is waaaaay off! Well, there's a reason why mode gets used least of all. Since it is calculated by frequency, which is definitely different than the other two, it is more likely to be very different. This also illustrates another good point here, though – there is a reason why there is more than one measure of central tendency!! It can be interesting and useful to look at all three – you may end up drawing different conclusions about your data.

Let's take a moment and have another "gut check" here. Thinking critically about your data is another important skill you will develop over time. This says that the average price of a kitten is 860. It doesn't specify a currency. Datasets often won't give you all the information you need. But take a good think on this for a second. Do you think it's likely that on average, people are spending 860 dollars on a cat? I don't know about you, but that seems pretty steep to me. Think about what you paid for a cat or similar animal, or a neighbor paid, or family paid. You probably know someone who's bought some sort of small and furry pet recently. Were they pay nearly a thousand dollars for it? Probably not. This is what's called a gut check. When you see your numbers, and think something may be off – chances are it probably is! Now look and see what is in column F. I see a label for "Malaysia Popularity." Now, I don't know much about Malaysia. You may or may not! But I do know how to google! I first typed in "currency for Malaysia" and found that they use the ringgit. Doesn't that sound like a fun name for currency? But in any case, the next thing I looked up was the exchange rate for the ringgit to USD. And guess what – 860 ringgit is about 200 USD. Is that making more sense, in an average amount paid for a cat? Yes. Especially considering that the mode for paying for a cat was 100, which converted to usd is only $23 – about the price of a pet adoption from a shelter often. Combine that with an examination of a few instances of what breeds are 100, you'll find that they are generic – if they were dogs, we'd call them mutts! So, price in ringgits is looking more and more likely – and with this amount of exploration, you should feel pretty confident in your educated guess.

Now onto measures of distribution! As the name suggests, these help you see how your data points are distributed. Is your data all in one place, centered around the mean? Do you have two major sections of data? Is data all over the place, with lots of extreme values? Measures of distribution help you get to the bottom of some of those mysteries.

The first measures of distribution you will explore are min and max. Min, or minimum, is the very smallest value you have, and max, or maximum, is the very largest value you have. The formula for min is =min(. The formula for max is =max(. So now you know that the smallest price was 100 ringgits, and the largest was 2000 ringgits.

There is no formula for range. If you were a very dramatic group, now would be a nice time for a theatrical gasp! Huh?! But don't worry, you can easily do simple math as well in excel, using calculated fields (i.e. things generated from Excel's auto formulas). So the equation for range is simply the max minus the min. Just like the formulas, you activate excel's awesome powers with =. But this time, you can just click on the numbers you want to use and provide excel with the appropriate operand (addition, subtraction, multiplication, division). It will look something like this. Excel conveniently highlights the numbers in different colors, so even if your spreadsheet is large, you can see what numbers your equation was based off of later. Then hit enter, and you have the range, which is one way to measure the spread of the data. The larger the range, the more sprawled out your data is. 1900 is a pretty big spread!

Quartiles mean breaking up your data into four parts. The root of quartile, quart, means 4 in latin. Latin's influenced most of our modern languages today, from the English quartiles in statistics to the Spanish word for four, quatro. In any case, if you're examining quartiles, you'll be looking at the data in fours! If you're breaking your data into fourths, then you end up with five dividing lines – a start, a first box line, a second line, a third line, and an end. It will look something like this: The start is actually your min. It thus follows that the end is your max! The second line is also known as your median. So, it turns out, you've already done more work than not in creating your quartiles! All you need is the first line, called your first quartile maker, and your third line, called your third quartile marker. As always, start by labeling those bad boys, so you know what goes where.

The formula for quartiles is the same, whether you are hitting up the first or the third, which also makes this the most complicated formula you have encountered before, because you now have more than just the option of selecting data. Let's give it a shake. It is =quartile(, and you want the one at the bottom. Start by highlighting the data, and then you get the options! Type in 1 for first quartile, or 3 for third quartile. We are going to start with the first.

Now onto the third! Do the exact same thing as before, only type a three instead of a one for the second set of options. Incidentally, with the quartile ranges, min, max, and median, you now have all the information you need to create a boxplot. Also sometimes called a box and whisker plot – how fitting for today's subject matter, meow!

There is a special type of range called inter-quartile range. Do not be afraid, fair citizen! As a range, it still is all about the difference between two things. But instead of the difference between min and max, it is the difference between quartile 1 and quartile 3. You can do it the same way you would range, by using an equals sign and then clicking on the cells you want to do math with. And there you are! Because the first and third quartiles are not as extreme values as the min and max, you see that the IQR is smaller than the regular range.

Hang in there- you only have two more to go! You have variance, and standard deviation. Both are measures of how far out your data is spread, with the bigger the number, the more varied and widespread your data is. Variation and standard deviation are related: take your standard deviation and square it, and you have variance! Or, if you'd like it in the reverse as well, take the square root of your variance, and voila! Magic! Standard deviation!

Excel has different formulas for the population and sample standard deviation and variances. Basically it's the difference between having .s or .p at the end. You can go ahead and use the .s for everything, because in this course we will almost never be dealing with the population. Remember that the population is ALL of something that you're measuring…and

guess what? It is super hard to get ALL of anything in one place and actually get information about every single thing. So, that's one reason why data scientists need to know about samples – it is very hard to get everyone together! It's for convenience.

But back to calculating standard deviation! The formula is =stdev.s(. And the formula for variance is =var.s(. Notice that the variance is quite a bit larger than the standard deviation – so you can see proof that variance is std squared.

Alright! You should now be set up to use Microsoft Excel or Google Sheets to calculate measures of central tendency and distribution. Isn't that the cat's meow?

Meredith Dodd, Ph.D. | Data Science Program Chair and Instructor

meredith.dodd@woz-u.com

o: 480-291-8068

powered by WOZ U

https://woz-u.com