Essential Math for Data Analysis Using Excel Online

Module 3, Lab 1: Mean, Median, Mode

Learning Objectives

- Calculate the mean, median, and mode in Excel.
- Examine how, and when, they give different conclusions about a variable.

Description

Learners will analyze the coffee data set again, but this time with summary statistics: mean, median, and mode.

Data set

Mod2Lab.csv

Overview

The mean, median, and mode each use different logic to measure the "center" of a data set. The usefulness of each measure depends on the type of data and the level of skew. In this lab, we'll use Excel to estimate the mean, median, and mode of the coffee data we used in the last module, and we'll see which measure of center gives us the most useful information about this particular data set.

What You'll Need

To complete the lab, you will need the online version of Microsoft Excel.

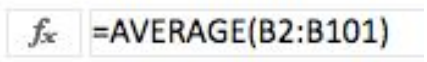
Exercise 1: Coffee Consumption Mean, Median, and Mode

1. Open the data set in Excel. There should be 100 different rows, with column headings for various coffee preferences. Column B gives the "coffee consumption" variable, or the number of cups of coffee each person drinks in a day.

2. Create three new columns for the mean, median, and mode.

	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean coffee	median coffee	mode coffee
2	1	5	Latte	No	182	Yes	Sometimes			
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			
7	6	2	Espresso	No	161	No	Always			
8	7	2	Latte	No	181	Yes	Sometimes			

3. Click into cell H2 and find the mean of all the values in column B. Remember, the mean is the mathematical average, so you can use the AVERAGE function again. Be sure to include all the data from B2 to B101.



	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean coffee	median coffee	mode coffee
2	1	5	Latte	No	182	Yes	Sometimes	2.02		
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			

The mean is 2.02, which means that on average, the participants in this survey drink 2.02 cups of coffee per day. Let's see how that stacks up with the median and mode.

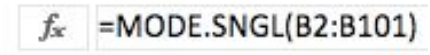
4. The median of a data set is the value in the middle when the data are arranged from least to greatest. If the data set has an odd number of values, the median is the value in the middle. If the data set has an even number of values, the median is the *mean* of the two central values.

   To find the median in Excel, click into cell I2 and use the MEDIAN function. The syntax is **=MEDIAN(first cell:last cell)**, so use B2 to B101 as the range again.

	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean coffee	median coffee	mode coffee
2	1	5	Latte	No	182	Yes	Sometimes	2.02	2	
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			
7	6	2	Espresso	No	161	No	Always			

The median number of coffees is 2, which is slightly lower than the mean of 2.02. We'll look at the reason for this discrepancy in a minute.

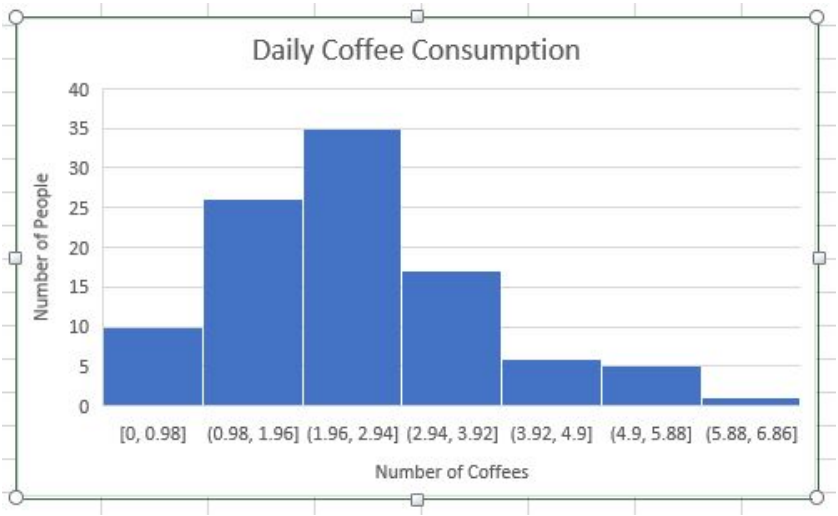
5. Now for the mode. The mode of a data set is just the value that appears the most frequently. (There can be more than one mode if multiple values show up with the same frequency, but that's not the case in our coffee data set.) Excel's MODE.SNGL function gives the most frequently-occurring number in a data set. The syntax is **=MODE.SNGL(first cell:last cell)**. Use B2 to B101 as the range again.



	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean coffee	median coffee	mode coffee
2	1	5	Latte	No	182	Yes	Sometimes	2.02	2	2
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			

The mode of this data set is also 2, which is the same as the median.

6. Which measure of center is the most useful in this case? It'll help to take another quick look at the histogram you created for this coffee data back in Module 2:



- Mean: The mean is only useful when the data are more or less symmetric, or normally distributed. The data here are skewed because those high extreme values (like the person who drinks 6 cups of coffee in a day) are pulling the data's average a bit higher than it would normally be. So in this case, the mean isn't the best measure for this data set — it's slightly biased by the outliers.

- Median: The median is a good backup measure when the mean is biased in some way (see above). It's based on both frequency and order, and it isn't easily biased by skew. Think about why: The median ignores *all* values except the middle one or two. Thus, skew and extreme scores have little impact on it. We still *prefer* the mean if we can use it (after all, it takes information from every data point), but the median is an excellent backup.

- Mode: The mode can be very useful with categorical data because it shows which category was the most popular. But it isn't super useful with numerical data like these coffee preferences.

So even though the median and mode are identical in this data set, the most useful measure here is the median.

Exercise 2: Coffee Temperature Mean, Median, and Mode

1. Run through those same steps for the coffee temperature variable (column E). Start by deleting those three new columns you created in Exercise 1, and replace them with columns for the temperature mean, median, and mode.

	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean temp	median temp	mode temp
2	1	5	Latte	No	182	Yes	Sometimes			
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			
7	6	2	Espresso	No	161	No	Always			

2. In cell H2, use the AVERAGE function to find the mean. This time, the range of data runs from E2 to E101 (that's the temperature column).

	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean temp	median temp	mode temp
2	1	5	Latte	No	182	Yes	Sometimes	172.69		
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			

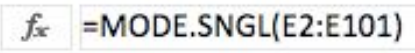
The mean temperature preference is 172.69 degrees.

3. In cell I2, use the MEDIAN function on the data from E2 to E101.

	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean temp	median temp	mode temp
2	1	5	Latte	No	182	Yes	Sometimes	172.69	171	
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			

The median or middle temperature was 171 degrees.

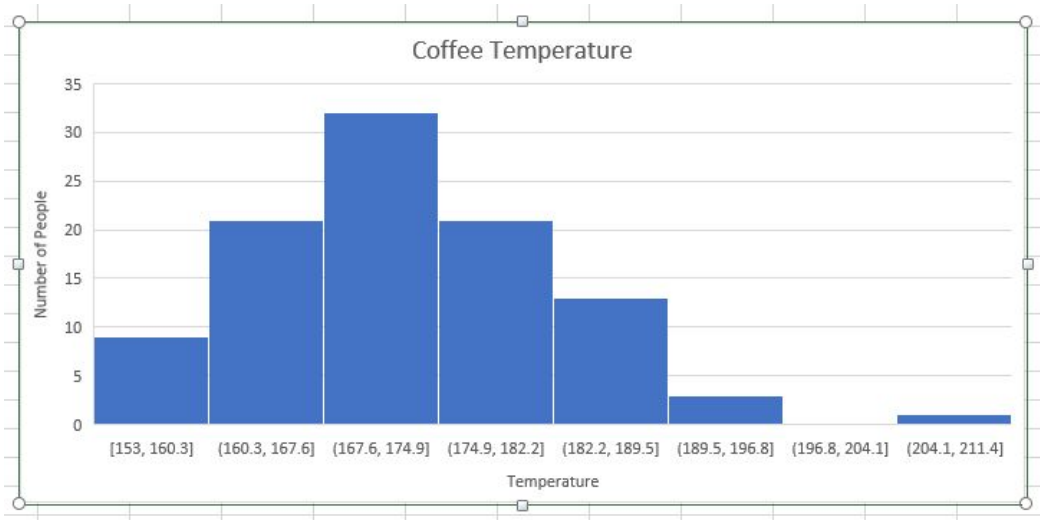
4. In cell J2, find the mode using the MODE.SNGL function. Once again, be sure to use the data from the temperature variable (cells E2 to E101).



	A	B	C	D	E	F	G	H	I	J
1	id	coffee	preference	black	temp	milk	additions	mean temp	median temp	mode temp
2	1	5	Latte	No	182	Yes	Sometimes	172.69	171	170
3	2	0	Drip	No	160	Yes	Always			
4	3	1	Latte	No	194	Yes	Never			
5	4	2	Drip	No	169	No	Sometimes			
6	5	1	Espresso	No	168	Yes	Sometimes			
7	6	2	Espresso	No	161	No	Always			

The mode, or the most frequently-occurring value, was 170 degrees. All three measures of center were different this time around.

5. Which measure is the most useful here? Once again, taking a look at the histogram of this data is helpful to see the skew.



- Mean: These data are also skewed by that extreme value on the far right, which represents a person who likes their coffee piping hot. Therefore, that extreme value is pulling the mean higher than the other measures of center, so the mean isn't the most useful measure here.

- Median: Smack-dab in the middle, the median is often the most useful measure when the data aren't normally distributed (i.e. when there's some skew involved). The median is once again our best bet.

- Mode: It's helpful to know which value shows up the most, but with numerical data like these temperatures, the mode isn't the *most* useful measure in terms of how the data are centered.

Note: In cases where the data have a symmetrical, normal distribution, the mean, median, and mode would all have the same value.