



Essential Math for Data Analysis Using Excel Online

Module 3, Lab 2: SD and IQR

Learning Objectives

- Calculate the standard deviation (SD) and interquartile range (IQR) in Excel.
- Examine how, and when, they give different conclusions about a variable.

Description

This lab will examine the same two variables as the prior lab, but we will now examine the SD and IQR in Excel. We will see how the IQR is more useful for variable data.

Data set

Mod2Lab.csv

Overview

There are a couple different ways to measure the variation or “spread” in a data set: the standard deviation (which we touched on in Module 1) and the interquartile range. In the same way that the different measures of center are more or less useful depending on the type of data, the standard deviation and the interquartile range can each be more or less effective with different data types. In this lab, we’ll use Excel formulas to calculate both, and we’ll explore why the IQR is more useful with skewed data.

What You’ll Need

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

Exercise 1: Coffee Consumption SD and IQR

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 a new column for the standard deviation (SD) of this variable.

	A	B	C	D	E	F	G	H
1	id	coffee	preference	black	temp	milk	additions	coffee SD
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	

3. We already talked through the ins and outs of the SD formula back in Module 1, Lab 4. Instead of working through all the math, use Excel's STDEV function. The syntax is **=STDEV(first value:last value)**. Our variable is in column B, so use B2 to B101 as the range of values. Be sure to use a colon between the cells (not a comma).

 =STDEV(B2:B101)

	A	B	C	D	E	F	G	H
1	id	coffee	preference	black	temp	milk	additions	coffee SD
2	1	5	Latte	No	182	Yes	Sometimes	1.3025228
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	

Boom. The standard deviation of this variable is about 1.3025, which means that the data points are, on average, about 1.3025 cups away from the mean. It's a bit like the margin of error — you can expect most of the data to be within 1.3025 cups of the mean.

Note that this is the *sample standard deviation*. That is, it uses the equation:

$$\hat{\sigma} = \sqrt{\frac{\sum \left((x - \bar{x})^2 \right)}{n - 1}}$$

As we discussed in the videos, this is the best equation to use when working with a sample (e.g. a small set of data from which you want to draw larger conclusions, such as future trends, inferences about other stores or markets, etc.). If you are working with a very large population data set (e.g. you have *every* observation of interest and only want to describe the data you have), then you can use the equation for a population shown in the videos, which can be done with =STDEV.P(). We won't show this here, but know that it's an option.

Here's the thing, though: Because the SD uses the mean, it suffers from the same drawbacks as the mean when the data are skewed. In the previous lab, you saw how the coffee consumption

data were skewed, and the mean wasn't the best measure of center. Similarly, the SD isn't the best measure of variability for this data set because the data are skewed.

That's where the interquartile range (IQR) comes to the rescue. The IQR uses the median instead of the mean in its formula, so it's better suited for skewed data.

- Find the IQR of the coffee variable. Sadly, there's no Excel formula to calculate the IQR directly, but there is a function we can use to find the IQR indirectly. We can use the QUARTILE.EXC function to track down the first and third quartiles, then use those to find the IQR. Here's the general formula for finding the interquartile range (IQR), where Q1 is the first quartile and Q3 is the third quartile.

$$\text{IQR} = Q3 - Q1$$

So what you'll do is use Excel functions to calculate Q3 and Q1, then find the difference between those two values.

- Set up three new columns in your spreadsheet for Q1, Q3, and the IQR.

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	coffee SD	coffee Q1	coffee Q3	coffee IQR
2	1	5	Latte	No	182	Yes	Sometimes	1.3025228			
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				

- In cell I2, use the QUARTILE.EXC function to find Q1 (quartile 1). The syntax is **=QUARTILE.EXC(range of cells, # of quartile)**. You're looking at the coffee variable, so the range is B2:B101. You want the first quartile, so enter 1 after the comma.

f_x =QUARTILE.EXC(B2:B101, 1)

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	coffee SD	coffee Q1	coffee Q3	coffee IQR
2	1	5	Latte	No	182	Yes	Sometimes	2.12132034	1		
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				

Q1 is 1. In other words, a value of 1 represents the point where about 25% of the sample is below you.

Now for Q3.


7. In cell J2, use QUARTILE.EXC again to find Q3 (quartile 3). Enter exactly the same thing you did in the previous step, but enter 3 after the comma instead of 1.

 =QUARTILE.EXC(B2:B101, 3)

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	coffee SD	coffee Q1	coffee Q3	coffee IQR
2	1	5	Latte	No	182	Yes	Sometimes	2.12132034	1	3	
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 third quartile is 3. In other words, 75% of the sample is below a 3.

8. Find the IQR by subtracting Q1 from Q3. In cell K, enter =J2-I2. Or just do the math in your head.

 =J2-I2

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	coffee SD	coffee Q1	coffee Q3	coffee IQR
2	1	5	Latte	No	182	Yes	Sometimes	2.12132034	1	3	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				

9. The interquartile range of the coffee consumption data is 2. Because the data are skewed, the IQR is a more useful measure of the data's variation. The reason is that the IQR only measures the middle 50% of data points, so it leaves out any extreme values.

Exercise 2: Coffee Temperature SD and IQR

Now let's run through the same exercise for the "temperature" variable in the spreadsheet.

1. Start by deleting the new columns you used in the last exercise. Replace them with columns for the temperature SD, Q1, Q3, and IQR.

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	temp SD	temp Q1	temp Q3	temp IQR
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				

- In cell H2, use the STDEV function again to find the standard deviation. This time, your range of data runs from E2 to E101.

f_x =STDEV(E2:E101)

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	temp SD	temp Q1	temp Q3	temp IQR
2	1	5	Latte	No	182	Yes	Sometimes	9.70441428			
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				

For the “temperature” variable, the standard deviation is about 9.704. Like we saw in earlier labs, though, the temperature data are skewed. That means the IQR will be a better measure of variability.

- In cell I2, use the QUARTILE.EXC function to find Q1 (quartile 1). The range of data is E2 to E101, and the quartile number is 1.

f_x =QUARTILE.EXC(E2:E101, 1)

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	temp SD	temp Q1	temp Q3	temp IQR
2	1	5	Latte	No	182	Yes	Sometimes	9.70441428	166		
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 first quartile is 166, which means about 25% of the data is below 166 degrees.

- Click into cell J2 and use QUARTILE.EXC again to find Q3 (quartile 3). The range of cells is still E2 to E101, but the quartile number is 3 this time.

f_x =QUARTILE.EXC(E2:E101, 3)

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	temp SD	temp Q1	temp Q3	temp IQR
2	1	5	Latte	No	182	Yes	Sometimes	9.70441428	166	179	
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 third quartile is 179, which means about 75% of the data is below 179 degrees.

- Find the IQR by subtracting the value in I2 from the value in J2.

$$f_x = J2 - I2$$

	A	B	C	D	E	F	G	H	I	J	K
1	id	coffee	preference	black	temp	milk	additions	temp SD	temp Q1	temp Q3	temp IQR
2	1	5	Latte	No	182	Yes	Sometimes	9.70441428	166	179	13
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				

- This time, the interquartile range is 13. Notice how different this is from the standard deviation of 9.704. Again, because we're dealing with skewed data, the IQR is the better measure of variability for this particular data set.