Essential Math for Data Analysis Using Excel Online

Module 2, Lab 3: Count Data

Learning Objectives

- Examine the nature of count data in action.

Description

Learners will look at the coffee cups variable in greater detail. We will look at how the data are discrete, and how the minimum (=MIN) is bounded at zero, whereas the max (=MAX) can be infinitely large. Learners will look at the histogram and skew (again) from that perspective.

Data set

Mod2Lab.csv

Overview

Data come in different flavors. Two of the major types of numerical data are *discrete* and *continuous*. In this lab, we'll examine the same coffee data from the previous labs and identify some of their characteristics.

What You'll Need

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

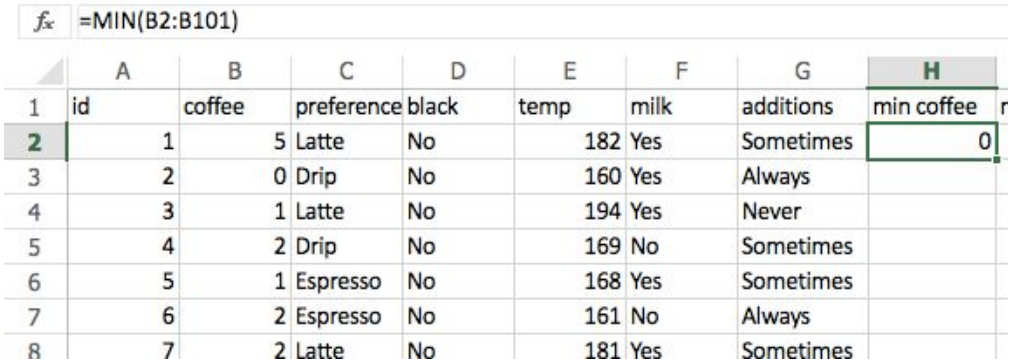
Exercise 1: Keep It Discrete

1. Open the data set in Excel. There should be 100 different rows, with column headings for various coffee preferences.

2. For starters, let's do some quick practice with Excel's MIN and MAX functions. Create two new columns for the minimum and maximum values for the coffee consumption variable (column B).

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

3. Now use the MIN function to find the lowest value from column B. The syntax is **=MIN(first cell:last cell)**.



	A	B	C	D	E	F	G	H
1	id	coffee	preference	black	temp	milk	additions	min coffee
2	1	5	Latte	No	182	Yes	Sometimes	0
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	

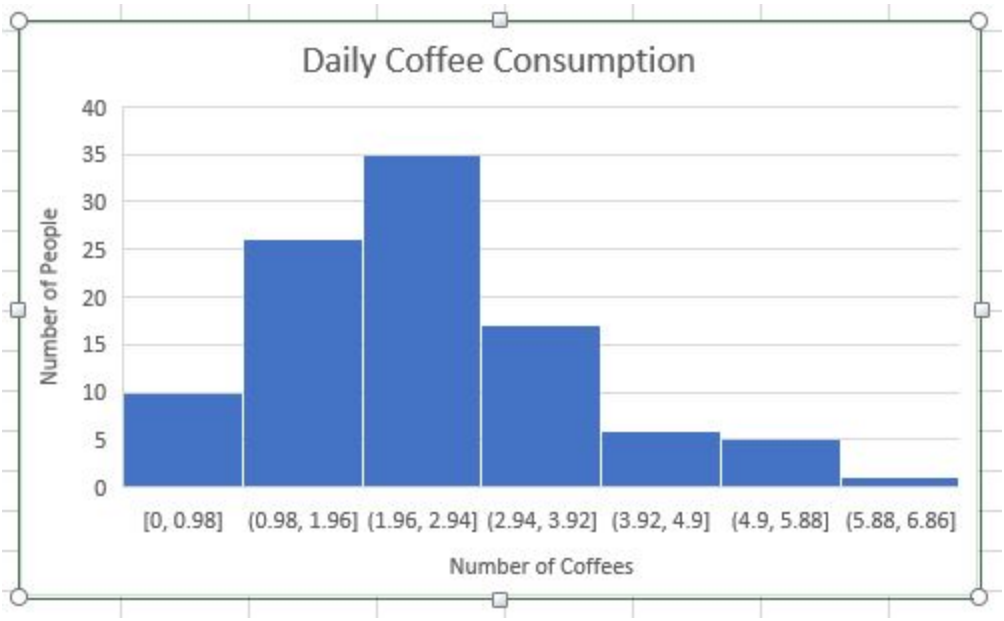
It's probably not surprising that the lowest value from column B is 0. After all, you can't drink fewer than zero cups of coffee. Therefore, the data are *bounded* at 0; they can't get any lower than that. This has nothing to do with statistics, per se, but everything to do with the type of data. In this case, we *counted* the cups of coffee, and thus there are certain values the variable can (and cannot) take.

4. Now find the maximum value. The MAX function works in the same way as the MIN function. The syntax is **=MAX(first cell:last cell)**. Once again, your range is all the data from B2 to B101.

	A	B	C	D	E	F	G	H	I
1	id	coffee	preference	black	temp	milk	additions	min coffee	max coffee
2	1	5	Latte	No	182	Yes	Sometimes	0	6
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 maximum value here is 6, which means nobody in this data set drinks more than 6 cups of coffee per day. That's probably healthy, but here's the thing: Unlike the minimum value, there's no maximum boundary for this data set. Theoretically, a person could drink as many cups of coffee in a day as they wanted. Again, this is because our variable was counted.

5. Take another look at the histogram of the coffee consumption data from a couple labs ago.



The values and the skew make a lot more sense in light of the max and min values. The data can't possibly dip below zero, which is why the "Number of Coffees" axis starts at 0. Similarly, the fact that most human beings can't handle an absurd amount of coffee explains why the graph is skewed positively (i.e. there are more values clustered around 1 and 2 cups, and fewer values at the 6-cup end.)

Notice, however, that even though the coffee data are all whole numbers (e.g. 1 cup, 2 cups, etc.), they *could* take on decimal or fraction values — it's perfectly reasonable for a person to drink half a cup of coffee per day. In business terms, though (i.e. if the data represented the number of coffees sold to or purchased by consumers), only whole cups of coffee would make sense. You can't buy 1.87 cups of coffee.

When data are limited to certain values in this way, the data are classified as *discrete*. The opposite of discrete data is *continuous* data, which means the data can take on any value. In a simplified sense, you could think of it this way: **You *count* discrete data, but you *measure* continuous data.**

As another example of discrete data, think about the other axis: the number of people. Half a person doesn't make any sense. Therefore, the number of people is also discrete in the sense that it's always a whole number.

As a general rule, if you can count something, it's discrete. Further, if you can count something, it is probably positively skewed, and it probably has a minimum value of zero. This isn't always the case, but it's important. There are summary statistics (such as the mean) that do best when the data are *not* skewed (we will get into this later, but just to give you a preview). Therefore, when choosing summary statistics, we *have* to think about the shape, type, and skew of the data.