Week 2 Lecture 1:

Displaying and Describing Categorical Data

Slide 2:

Review categorical data – Categorical data describes a category. For our purposes, we’ll picture *columns* that describe a category. They are interpreted as text, even if they are numeric. We cannot do math with categorical variables.

If we can’t do math with categorical variables, what *can* we do? One thing we can do is count them.

Slide 3:

Frequency Tables - Counting categorical variables in a dataset is sometimes done using a frequency table – in Excel we use a Pivot Table to do this.

\* Excel demo with 311 data. Create counts, then create percentages.

This a dataset from NYCOpenData, which is an online repository of publicly available data related to NYC government agencies. This is a fantastic data resource, and I highly recommend it when you need to find data for projects later in the course.

This particular dataset is from 311, NYC’s non-emergency hotline. Each record represents a 311 call with a unique ID, date, agency, etc. that came in between June 1st and June 12th of this year. Looking at this dataset, which columns are categorical? All of these can be considered categorical. Created\_Date is not a great category, but if you removed the time it could be useful depending on what you’re looking for.

Now let’s ask a question: How many calls went to each agency during this time period? Unlike our sample dataset from the slides, there are far too many rows to count manually (95K+). To do this, we use a Pivot Table to do the counting.

\*Click on dataset – Insert – Pivot Table on far left, insert in new worksheet.

Now we have to define what we want our Pivot Table to do. Remember, we want to know how many calls each agency received. We have a list of our variables (columns) at the top, and we have to decide which ones we are interested in, and whether we want them to be in the rows or columns of our pivot table. We select agency\_name, since that is what we are interested in. This automatically gets put in the rows of our table.

Now we have to select our “values”. We know we want to count something, so let’s select something we can count for each agency. How about unique\_key? Since each record represents a call, if we count the unique keys, we should answer the question. When we select it, Excel assumes this is what we want as our values, but what happened? It automatically sums our unique keys, but what do we know about categorical variables? We cannot do math with them, so this doesn’t make any sense.

Select the drop down arrow, and value field settings, and change the Sum to Count. Now congratulations - we’ve answered our original question – we see the number of calls each department got in the time period of our dataset. This series of counts is sometimes called a *frequency distribution*.

But let’s think about what we did here. We counted the number of unique keys in each category What if we selected something other than unique key to count? How about descriptor? Why are the counts different? Because some records are missing a descriptor, and therefore are not being counted.

\*Demonstrate filter on original dataset to find records with null descriptor.

It’s always best to count a unique key or id record, since we know there will not be missing values.

So, we answered our question, but are these numbers that useful to us? Sometimes instead of raw counts, we want to see them as a percentage of a total. The Pivot Table can do that for us.

\*Select drop down menu, value field settings, show values as, % of Grand total

Doing this allows us to easily compare values across different categories.

Slide 4:

Data visualization –

Aka, charts, graphs, maps, etc. We use data visualization techniques for the same reason we use frequency tables – to summarize large amounts of data.

Data visualization is great when you’re exploring data for the first time. Things that might not be obvious at first glance in a frequency table may quickly jump out on a chart. For this same reason, charts can also be far more impactful for *communicating* findings than a table or text.

\*Covid example – During the covid pandemic, were news outlets showing tables each night with state names in rows and counts/percentages of cases for each? No – they used maps with varying color gradients so viewers could quickly find hotspots and draw conclusions about the state of the pandemic. They also used line charts to show rolling averages of infections and hospitalizations.

Slide 5:

Bar charts

Bar charts are charts that display counts are percentages across categories – the same thing our frequency tables do. You have your categories on one axis and your values on the other. They are extremely simple, and are a great place to start when exploring data for the first time.

Fortunately for us, Excel can create charts from pivot tables quickly and easily. Let’s say we want to make a bar chart to show what we found in our previous pivot table.

\*Back to Excel, click on agency pivot table, insert, pivot chart, column chart.

Excel calls vertical bar charts “column” charts, and horizontal bar charts “bar charts”. In reality, these are both called bar charts. As a quick note, when I have long category names, horizontal bar charts usually look better.

Bar charts are as simple as data visualization gets, but they are often incredibly effective. Something cool about bar charts is that you can have counts or percentages as your values and it doesn’t change the graph (show example).

What is something we notice now that we have this data in visual form? Personally, I am now seeing that calls to the NYPD are accounting for way more of the total calls than I noticed in the table. We certainly could’ve seen this in the table, but the visual makes it much more obvious.

There are entire classes on data visualization techniques, so we aren’t going to dive into that too heavily in this course aside from a few big rules of thumb. Rule number 1 of data visuzualization: simplify your charts as much as possible, but no more. No need to add colors or text that don’t add to a viewers understanding of what’s going on.

In our chart, it would be helpful to have a title. How about “311 Reports by Agency

June 1-12, 2025”? This way people can see what the bars represent.

If in this analysis we are only trying to show that a huge chunk of 311 calls goes to the NYPD, do we even need numbers on the horizontal axis?

\*Remove axis labels by clicking + sign, axis, primary horizontal

This cleans up our chart area and still gets our point across. We’ll stop here.

Slide 6:

Pie charts show proportions of categories in a dataset. Excel can create them from Pivot Tables just as easily as bar charts, but very often then are useless. Look at our example from the 311 dataset. We can still see that the NYPD gets the biggest proportion of calls, but all other agencies are completely lost.

Second rule of data viz – almost never use a pie chart…I say “almost” because there are a few cases where they are helpful. People find pie charts very intuitive, and when you have less than 4 categories, they can sometimes work. However, even with only a few categories, it can be hard to detect differences if the proportions are close. When categories exceed 4 or 5, they quickly become cluttered.

Slide 7:

Two variables –

So far we have just explored a single variable in our 311 dataset – agency name, but things get much more interesting when we add a second. We can use the familiar concept of a frequency table to display a second variable. Since we put our first variable in rows, we can put a second in columns.

This structure is sometimes called a “contingency table” because it shows the frequency distribution of a variable *contingent on* the value of another variable. In our example here, we can see the frequency distribution of female participants across our eye color variable, and our male participants separately. In other words, we’re looking at the distribution of eye color *contingent on* gender. Similarly, we can look at the distribution of gender *contingent on* eye color. This is called a conditional distribution.

\*Back to Excel – start on single variable frequency table and delete pie chart. Return values to raw counts again?

Let’s think about our original frequency table, and see if we can ask a question that involves another variable. What if we want to know the distribution of agencies contingent on the borough the call came from? Remember, one of our columns tells us the borough that the call came from, so we should be able to add this variable to the table to view the conditional distributions.

\*select borough from variable list and add to columns

Does everyone see what I just did? We can now see the distribution of agencies contingent on borough, as well as the distribution of boroughs contingent on agencies. By adding a second variable, we have significantly increased the number of questions we can answer. For example:

What % of reports came from Brooklyn? 29404/95854=30.68%

We can answer this in Excel by viewing our counts as percentages

\*value field settings – show values as - % of grand total OR % of row total. Switch back to counts afterwords.

What % of reports from Brooklyn were for the Sanitation Department? 3177/29404=10.8%

\*value field settings – show values as - % of column total (let students pick between grand total, column total, or row total)

The homework will have you answer a few questions like this using a contingency table.

Quick but important note:

\*switch from % of column total to % of row total

Lets look at the calls for the department of housing preservation and development. Notice how many come from the Bronx – nearly a third. Now look at the parks and recreation department – only 9% of calls come from the Bronx. The numbers vary pretty widely for each department, actually. In statistics, we might say that these two variables are DEPENDENT – meaning that the agency the call goes to DEPENDS on the borough the call came from. Conversely, if we saw similar numbers across each of the boroughs for all agencies, we might say these variables are INDEPENDENT – meaning one has nothing to do with the other. So if I’m in the parks and recreation department and a call comes to me from 311, there is an equal probability of it being from any of the 5 boroughs.

There are robust tests for this that we’ll cover later in the semester, but just be aware that this is a big concept in stats and analytics. For now, just try to get a sense for the concept.

Now that we have our two variables, lets again try to visualize the data

\*switch back to raw counts before trying grouped bar chart.

Let’s make a column chart from our pivot table. Notice we see our “row” variable along the horizontal axis, and they are broken up by our “column” variable in different colors.

This is called a grouped bar chart, or “clustered column” chart in Excel, which was our default option. Excel provides a few more options for this type of chart depending on how you want to see it. Let’s look at a stacked bar chart. Can everyone see what happened? We basically just stacked all the bars on top of each other rather than having them side by side. Generally speaking, this is not a good way to display data. Why? Let’s compare the number of calls to the dept of housing preservation to the number calls to the sanitation department that came from Brooklyn. I’ll wait. If we hover over the bars we can see the numbers. Roughly 3700 compared to 3100 – a difference of about 600 calls. That is a lot of calls, and we can barely detect it because the bars don’t start at 0. Let’s switch back to our grouped bar chart to see it again. This is why grouped bars are generally superior.

In your homework you’ll have to make a chart, so feel free to explore some of the options you have in Excel. You cam make the chart 3d. Be careful – these may look cool, but remember you are trying to convey a message or answer a question. Don’t let your visual distract from that mission.

Simpsons paradox – use sales example. Who is the better salesperson? Think critically.

College admission example.

Homework questions: use another 311 dataset

1. What % of all reports came from Staten Island? (2141/47255=4.53%)
2. What % of all reports were for the Department of Transportation?
3. What % of reports from Manhattan were for the NYPD?
4. Come up with some question about the data and try to answer it with a pivot table and display it on a chart. Feel free to use filters to explore something you find interesting. Use your best judgement on the chart and explain your choices.