Slide 1 – Title Slide

Hello and welcome to Week 5, Part 4 of EGM101: Frequency. In this lesson, we'll talk about how we can group our datasets and start to look at the ways that we can start to summarize our data.

Slide 2 – Grouping Data

When we're working with large datasets, we typically want or need to group, or categorize, our data. Exactly how we do this depends on the kind of data that we're working with.

For example, with nominal data, we can count the number of occurrences of each value of the dataset. In this example, which summarizes the different types of responses to an airline survey on passenger complaints, we see the nature of the complaint in the first column of the table, and the number of complaints corresponding to each type of complaint in the second column. From this table, we can see that most of the passenger complaints have to do with uncomfortable seats or inadequate leg room, but we also see that there are also complaints about narrow aisles, not enough toilets on the flight, or a "catch-all" category of other.

All that we've done here is taken a large number of complaints, categorized them according to what the complaint is about, and then counted the number of complaints that fall into each category.

If we have discrete data, we can do something similar – here's an example for a different survey question, asking respondents how often they have visited their local zoo in the past year. Each row of the table tells us how many people responded with a given value, and we see that most of the respondents, 90, haven't been to the zoo at all in the past year, while only one respondent has been to the zoo more than 6 times in the past year.

In each of these examples, we're grouping the data according to different categories – either the type of complaint in the nominal example, or the individual values, or a range of values, in the discrete example.

With continuous data, though, we can't do this – we need some other way of grouping the data.

Slide 3 – Tally charts

Before we cover how we can group continuous data, though, we'll introduce the idea of a tally chart. In the example dataset shown here, we have a total of 100 values ranging from 0 to 10. Maybe this represents the number of visits a total of 100 people have made to a particular supermarket in the past month.

Rather than displaying this as an unorganized table, we can instead summarize these data and display them as a tally chart – here, each row corresponds to a distinct value for number of visits to the supermarket, and the second column will show the number of times that value appears in the data.

For each distinct value that we see in the dataset, we then add a tally mark to the table. So, starting here in the upper left-hand corner of the table with “5”, we add a tally mark to the row corresponding to 5. Moving to the next value, 6, we add a tally mark to the row corresponding to 6.

And now, hopefully, you get the idea. In the final tally chart here, you see that I’ve grouped values by 5s, which is a conventional way to help us count the number of tallies more easily.

So here, we can see that most people surveyed visited this particular supermarket between 4 and 6 times in the past month, though some people visited as many as 10 times, while others didn’t visit at all. The point of doing this is that this kind of summary helps us understand the data more quickly than just looking at the raw numbers shown in the table up above – we can more easily see that values like 4, 5, and 6 appear more often in the data compared to others.

Now, this leads us to a question: what do we do with continuous data or discrete data with a large range? For example, if we had numbers ranging from 0 to 100, we wouldn’t want to go through and have a row for each individual value – we would instead want to group the values together somehow. It turns out, this is exactly how we would deal with continuous data, as well.

Slide 4 – Fish sampling

Let’s say that we’ve gone out and taken a sample of 25 salmon from a local river, and we’ve measured their lengths.

This table shows the results – here, we use the variable *n* to denote the number of fish that we’ve sampled, with the values in the table showing the length of the fish in millimeters.

Like with the previous example, though, this is a bit messy. From this, it’s not easy to tell what the largest or smallest lengths are. We don’t see very many repeat values, either, which means that we don’t want to just put together a tally chart.

Slide 5 – Stem-and-Leaf plots

Instead, we can make something called a stem-and-leaf plot, which is an effective way of organizing numerical data, especially if it’s continuous or covers a large range – for example, the fish lengths that we measured.

On the left side of the stem-and-leaf plot, we have the “stem” – this is all but the last digit of our numbers. So, since we have values starting from 574, going all the way up to 635, we would have stems of 57, 58, 59, and so on, up to 63. Effectively, we’re grouping our values into groups of 10, and the stems are the whole number values when we divide the data by 10.

Next, we have the “leaf” part of the “stem-and-leaf” plot, on the right side of the plot. As you might be able to guess, this is where we record the final digit of the value. So, starting again at the upper left-hand corner of the table with the value of 601, we would put a “1” in the row corresponding to “60”, like so. With the value “623”, we put a “3” in the row corresponding to “62”, and so on.

What we end up with is a visual representation of the “shape” of the data – we can see that most of our fish have lengths between 590 and 610 mm, with a few that are much larger than that, and a few that are smaller than that. Like the tally plot, this is an effective way to summarize our data, and it works for both continuous values as well as datasets with a very large range of values.

Slide 6 – Grouping Continuous Data

So, the answer to the question “what do we do with continuous data?” is: group it! It’s the same answer that we saw for discrete values with a large range.

But, this doesn’t answer all of our questions. How do we choose the number or size of the groups that we use? If we choose too few groups, then we end up oversimplifying the data. If we choose too many groups, we end up with way too much detail, and we lose the benefits of grouping/summarizing the data.

So, how do we find the Goldilocks “just right” number of groups? In general, we want to use equally-spaced intervals, and pick between 5 and 20 groups, or classes. We also want to choose “easy” boundaries – in the stem-and-leaf plot example, we grouped values into groups of 10. In the different textbooks recommended in this module, there are a number of other rules or even methods for calculating the “ideal” group size that you’ll find, but these aren’t necessarily hard-and-fast rules that you always have to follow. Stick to the basic rules we’ve discussed here, and you should be fine.

Slide 7 – Frequency and Relative Frequency

There are a few additional definitions that we’ll need to learn when we’re talking about grouping data like this. Keeping to our fish length example, the first term that we’ll define is frequency, which is just the number of times that a particular value, or group of values, appears in the data. Looking back at the fish lengths, we see that we had one value between 570 and 579.99 mm, 5 values between 580 and 589.99, and so on. Because we’re counting the number of times something happens, frequency is always a whole number.

We can also define the relative frequency, which is the proportion of times that a value, or range of values, occurs. It’s equal to the frequency divided by the total number of values present in the dataset – since we have 25 lengths to consider, in this example the relative frequency is the frequency divided by 25. We can express the relative frequency as a fraction, like 1/25, or as a decimal: 0.04; you may also see it expressed as a percent. One thing to remember is that the relative frequency is always between 0 and 1, because it makes up some proportion of a whole.

The last term we’ll introduce here is the cumulative relative frequency, which is the sum of all of the previous relative frequencies. In our fish example, we can see that nearly half of our fish have a length less than 600 mm – the cumulative relative frequency for the 590 to 599.99 mm group is 0.48, or 48% of the data. Just like the relative frequency has to be between 0 and 1, the cumulative relative frequency has to add up to 1 – if it doesn’t, we need to go back and check our math.

Slide 8 – Histograms

The last topic we'll cover in this lesson is histograms – these are a graphical representation of the frequency distribution, and they look like a bar chart of frequency or relative frequency values. In a histogram, the bars should touch but not overlap, just like the groups that they represent.

Histograms give us a quick way of seeing the "shape", center, and spread of our data. In this example, we see that most of our values are somewhere around 600 mm or between 580 and 610 mm, and there are more values in the lower part of the range than in the upper part.

You may also see this information plotted as a line graph, known as a "frequency polygon." In a frequency polygon, the markers or dots are plotted at the center of the intervals, like you can see here.

Slide 9 – Summary

In this lesson, we saw that one of our goals in quantitative skills is summarizing our data, to make it easier to manage and make it easier to spot patterns in the data.

With nominal or discrete data, this is easy: we just count the number of occurrences of each category or value.

If we have continuous data, or discrete data with a large spread of values, though, we need to group the data into intervals, to help make sure that what we have is a more concise summary.

Continuing on our theme of displaying data, we also saw how we can display this information, using things like tally charts, stem-and-leaf plots, or histograms.

Slide 10 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapters 1.3 and 2.1-2.2; Caswell, Chapter 4; or Weiss, Chapter 2.3. I didn't discuss it here, but one of the reasons that we choose evenly-spaced groups for estimating frequency is to provide an honest summary of the data – the video linked here, which is part of the course that the Bergstrom and West book grew out of, discusses this idea in more detail. That's all for this lesson – I hope you found it interesting, and you have any questions, please don't hesitate to e-mail me or post in the discussion forum on blackboard. Bye!