This Chapter's Topics

Today, we're going to talk about three things:

- Frequency distributions
- Graphs
- Charts

Frequency distributions

Frequency distributions take raw ("ungrouped") data, and summarize it by grouping the data into predetermined categories.

For us to group data, our categories must be mutually exclusive and exhaustive.

These criteria should be familiar from our discussion of ordinal and nominal data.

Quantitative frequency distributions

A *quantitative* frequency distribution is one where we categorize quantitative data, such as ordinal, interval, and ratio variables.

One possible approach: we could create categories of people in this room based on age.

Let's look at an example.

An awesome example, dude

Suppose we have some data on how much adolescents acting like Beavis and Butthead annoys their parents; specifically, we know how many times “heh” or “huh” is used by a child before his or her parents get angry.

This data is presented in a table on page 31. We found 25 parents, and found out that it took between 2 and 29 Beavis-like utterances by their child to anger them.

That’s all well and good, but looking at 25 numbers isn’t very informative. So we need to break it down a bit, by using categories. A good way to categorize numbers is through *class intervals*.

Class Intervals

A “class interval” is a predetermined numerical category that can contain more than one possible observation in its range.

The size of class intervals is fairly arbitrary. However, with some skill, we can select reasonably-sized intervals that are helpful.

How big should our intervals be?

Size and starting point

Let's choose an interval size of 5. This way, there will be six categories that cover the data set (from 2 to 29).

It's probably best to start at one, unless your data has negative values, because it tends to make the math easier.

So, our categories are 1–5, 6–10, 11–15, 16–20, 21–25, and 26–30.

Do these categories meet our criteria?

We should make sure the categories are mutually exclusive and exhaustive.

Now that we've satisfied ourselves of that, we can find absolute and relative frequencies.

Absolute frequencies

The *absolute frequency* of a category is simply the number of observations that fit in that category. So, let's figure out the absolute frequencies:

Class Interval	Absolute Frequency (af)
1–5	2
6–10	5
11–15	6
16–20	7
21–25	3
26–30	2

Cumulative frequency

The *cumulative frequency* is simply the absolute frequency of the current category plus the absolute frequencies of the lower categories. That is, it is the running count up to, and including, the current category. So:

Class Interval	Absolute Frequency (af)	Cumulative Frequency (cf)
1–5	2	2
6–10	5	7
11–15	6	13
16–20	7	20
21–25	3	23
26–30	2	25

Getting better...

Note that the cumulative frequency after the final category and n should be the same.

These frequencies give us a fairly good idea of how the responses are distributed. Around 20, it looks like most parents are quite livid.

Now we can look at two other frequency measures: the *relative frequency* and *cumulative relative frequency*. These are particularly useful when we want to compare frequencies but have different n's.

Relative frequency

The relative frequency of any category is simply the absolute frequency of that category, divided by n. Or, in math terms:

$$\mathrm{rf} = \frac{\mathrm{af}}{n}$$

Cumulative relative frequency

In the same vein, we can look at the cumulative relative frequency for categories. We can calculate it from the cumulative relative frequency:

$$\mathrm{crf} = \frac{\mathrm{cf}}{n}$$

We can also calculate it from the relative frequencies, as a running total of the rf's. The cumulative relative frequency at the end should be 1.00 (or 100%).

Back to our example

So, let's figure out the relative and cumulative relative frequencies:

Class Interval	af	cf	rf	crf
1–5	2	2	.08	.08
6–10	5	7	.20	.28
11–15	6	13	.24	.52
16–20	7	20	.28	.80
21–25	3	23	.12	.92
26–30	2	25	.08	1.00

Going qualitative

We can also make qualitative frequency distributions based on nominal variables; for example, race, marital status, and family types.

Other than the way the groups are categorized, the process is exactly the same for qualitative frequencies. For example, let's try to group the class based on gender and Greek membership, and figure out the frequencies.

Fraternity member	?
Sorority member	?
Male GDI	?
Female GDI	?

Making pretty pictures

Another good approach to summarizing data is through visualization: producing charts and graphs. There are four types of graph that can be useful for visualizing data:

- Histograms
- Frequency polygons
- Bar charts
- Pie charts

Histograms

Most spreadsheet programs can produce these for you, but let's look at how to produce them on our own.

Histograms are used with quantitative frequency distributions. They are pretty simple to draw; the main thing to remember is that the horizontal (X) axis is *continuous*, not discrete. This means we have to set a cut point between each class interval.

For example, the cut point between the 1–5 and 6–10 intervals is $(5+6)/2 = (11)/2 = 5.5$.

End Points for Histograms

At the end points of the frequency table, we treat the "missing" adjacent groups as if they were there. So, the right side of the 26–30 interval would be $(30 + 31)/2 = (61)/2 = 30.5$, not 30.

Now we can produce the histogram.

Frequency polygons

A frequency polygon is a lot like a histogram, even though they don't appear to be the same. Rather than making big boxes between the end points of an interval, we pretend that all of the data is at the middle of the interval, and connect the dots.

We find the middle of the interval by adding the two end points together and dividing by two ($(l + h)/2$). So to find the middle of the 6–10 interval, we find $(6 + 10)/2 = (16)/2 = 8$.

So, now we can draw a frequency polygon.

The reason the histogram and frequency polygon are related is clear if you draw one on top of the other.

Bar Charts

Bar charts are a lot like histograms (indeed, many people confuse the two). The important difference is that bar charts are used for *qualitative* data.This also means that the position on the X axis is *meaningless* for bar charts.

For example, let's make a bar chart based on our Greek/non-Greek survey of the class.

Pie Charts

Pie charts can be used for either qualitative or quantitative frequencies, and the technique for drawing them is exactly the same for both types of frequency.

The important thing to know when drawing pie charts is that they are based on circles. Since a full circle is 360°, we can split up the circle based on the relative frequencies.

More Pie Charts

For example, since the relative frequency of 1–5 in the Beavis data is .08, we can figure out how big its slice needs to be by multiplying the relative frequency by 360°: $.08 \times 360 = 28.8\,\mathrm{deg}$.

While you might have to do some work with a protractor to draw one by hand, pie charts can easily be drawn in most spreadsheet programs and statistical packages.

Homework

Read the chapter summary and conclusions, starting on Page 47.

Do all parts of question 2, except you don't have to actually draw the pie chart. (You should figure out how big each slice should be based on the relative frequencies, however.)

Next time: real math!