This Chapter's Topics

Today, we're going to talk about multiple measures of *central tendency*:

- the **mode**
- the **median**
- the **arithmetic mean**
- the **weighted mean**

What is central tendency?

When we have a large amount of data, it is often helpful to summarize it.

While frequency distributions and charts can be useful, sometimes we want to summarize it more succinctly.

A measure of *central tendency* gives us an idea of what the "middle point" of a data set is.

The Mode

Perhaps the simplest measure of central tendency is the **mode**.

The mode is the value that appears most frequently in a given data set. That is, it is the *most common* value.

The symbol for the mode of a *sample* is $\mathrm{Mo}_{\bar{x}}$.

The mode of a *population* is shown as Mo_{μ}.

À la mode

For example, if we look at Table 4.1, we can see that the most common hair color in Calvin's class is **brown**.

Because we have a population ($N = 50$), $\mathrm{Mo}_{\mu} = \mathrm{Brown}$.

Note that we can use the mode even with nominal data.

A sample mode

We can also use the mode with ordinal data. Table 4.2 presents ordinal data on a sample of students' attitudes toward the statement, "Taking a statistics class should be a requirement for all students to graduate."

As we can see from the table, the most common response is *strongly agree (1)*, so that is the modal value.

Since we have a sample ($n = 25$), $\text{Mo}_{\bar{x}} = 1$.

Limitations of the mode

The mode has a couple of limitations that are worth mentioning:

- What do we do if there is a tie? There's no standard answer. Schacht says there isn't a mode in that case. Other statisticians say that you can have multiple modes.

- The mode may be the *most common* answer, but it isn't necessarily *representative*. For example, in table 4.3, the most common number is 20, but it is at the low end of the range.

Because of these limitations, we often use other measures.

The median

The *median* of a data set is the value that is in the middle; like Schacht says, it's like the median of a divided highway.

At the median value, half of the observations are above it, and half of the observations are below it.

IMPORTANT: The median is *meaningless* with nominal measures. Without some ordering constraint, we can't have a median.

The median works best with interval or ratio data, but can also be used with ordinal data.

Finding the median

To find the median, first we need to arrange the data in *rank order*. That is, we need to sort it from highest to lowest (or from lowest to highest; it doesn't matter). Then the median is simply the middle value.

Like the mode, there are special symbols for the median.

The symbol for the median of a *sample* is $\mathrm{Mdn}_{\bar{x}}$.

The median of a *population* is shown as Mdn_{μ}.

What if n is even?

To find what observation is the median, we take $i = (n+1)/2$ (for a sample) or $i = (N+1)/2$ (for a population). So the median of X will be X_i.

If we have an even number of data points, we take the two observations closest to the middle and average them. For example, if $N = 10$, we look at observations 5 and 6, because $(N+1)/2 = (10+1)/2 = 11/2 = 5.5$.

For example, the median of the data in Table 3.24 (page 52) is $(45+47)/2 = 46$.

Don't be mean

Both the mode and the median only look at some of the values in a data set. However, we may want a summary that looks at *all* of the data.

The summary that we most often use is called the *arithmetic mean* (often simply called the mean, as in Schacht). When most non-statisticians use the word "average," what they are talking about is the arithmetic mean.

We use the symbol $\bar{x}$ for the sample mean, and μ for the population mean.

Calculating the mean

The arithmetic mean of a sample or population is simply the sum of all of the values in the data set, divided by n or N.

In math terms:

$\bar{X} = \frac{\sum X}{n}$ and $\mu_X = \frac{\sum X}{N}$

(For now, the subscript on μ isn't important. But when you have multiple variables, it is better to keep track of things that way.)

An example

So, if we look at the data on the top of page 58, we can figure out the sample mean:

$$\sum X = 8 + 2 + 7 + \cdots + 4 + 2 = 37$$

$$\bar{X} = (\sum X)/n = (37)/10 = 3.7.$$

Notes about the mean

Like the mode (but unlike the median), the mean can be unrepresentative. If there are large gaps in the data (extreme values), they can skew the mean. For example, mean family sizes are skewed because a relatively small number of people have large families.

The mean can only be used with interval or ratio data. The mean of an ordinal measure is meaningless, and the mean of a nominal measure would depend on the values assigned to each category—which are arbitrary!

Hence, while we can say that the average American is of Caucasian descent (the mode), we can't say that the "average American" is 0.70 white, 0.16 black, and 0.14 other.

This is heavy

When we have summarized data, or we have already taken means of subgroups, we can't find the mean directly. We have to use something called the *weighted mean*.

The weighted mean is called that because it *weighs* the values taken from more observations more highly. We calculate the weighted mean as follows:

$$\bar{X}_w = \frac{\sum_i W_i X_i}{\sum_i W_i}$$

That is, we multiply the weight by the group value for each group, add these up, and divide by the total of the weights.

Weighted mean example

Let's look at Table 4.4. The weighted mean is calculated as follows:

$$\bar{X}_w = \frac{\sum_i W_i X_i}{\sum_i W_i} \quad (1)$$

$$= \frac{15 \times 40 + 20 \times 38 + 25 \times 35 + 30 \times 33 + 35 \times 30}{15 + 20 + 25 + 30 + 35} \quad (2)$$

$$= \frac{4275}{125} \quad (3)$$

$$= 34.2 \quad (4)$$

Homework

From chapter 4:

- Question 2
- Question 4

Next time: we'll finish Chapter 4 (skewed data and frequency distributions), and start Chapter 5 (range, variance, and standard deviation).