Today's Topics

- The mode, median, and mean of frequency distributions
- Why dispersion is important
- The **range**
- The **mean (or average) deviation**
- The **variance**
- The **standard deviation**

Central tendencies and grouped data

When we have a frequency distribution, we no longer have a direct way to figure out the central tendency. It is a lot like the problem of the weighted mean.

To estimate the **mode**, we take the midpoint of the class interval with the highest absolute frequency. For example, if we look at Table 4.7, we see that the class interval that has the greatest absolute frequency is 16–20, so we estimate the mode to be 18.

Note that this is only an estimate: we don't know the real mode, because we'd need all of the data to figure it out.

Medians of grouped data

To figure out the **median**, first we have to calculate the cumulative frequencies (see Table 4.8). Now we need to find the "middle observation." Since $n = 24$, the middle observation is $i = (n+1)/2 = (24+1)/2 = 25/2 = 12.5$, which is in the class interval 11–15.

However, to get an actual number, we need to figure out what value in the interval is most likely to be the "12.5th" value.

An ugly formula

$$
\begin{aligned}
\text{Mdn} &= \text{Lower real limit} + (\frac{n/2 - \text{cf below}}{\text{af of Mdn interval}})\text{Interval size} \\
&= 10.5 + (\frac{24/2 - 7}{6})5 \\
&= 10.5 + (.833\overline{3})(5) = 14.66\overline{6}
\end{aligned}
$$

Means of grouped data

The mean of grouped data has a similar formula to that for the weighted mean. The difference is that for the "means" we use the interval midpoints, and for the "weights" we use the absolute frequencies.

Hence, the formula is:

$$\bar{x} = \frac{\sum \mathrm{af} m}{n},$$

where m is the midpoint of each interval. (Work through example.)

Dispersion (or variability)

Measures of central tendency are nice in that they tell us something about the “typical” member of a data set. However, they don’t tell us very much about how the data is distributed; we don’t know if all of the values are concentrated around a single point, or if they are widely dispersed.

To see why central tendency measures aren’t enough, look at Table 5.1. If you calculate the mean of both columns, you’ll find that it is 5; however, the second column has much greater variability than the first (in fact, the first column has no variability whatsoever).

Home on the Range

That's where measures of dispersion or variability come in. These measures give us an idea of how the data is distributed around the midpoint.

One simplistic measure of variability is called the *range*. The range is simply the difference between the highest and lowest values in the data set:

$$\text{Range} = X_{\max} - X_{\min},$$

where the subscripts denote the maximum and minimum values, respectively.

More on the Range

The range isn't normally used by statisticians, as it only tells us about the extreme values of the data, not about the data as a whole.

For example, the range of U.S. income only tells us how much more Bill Gates makes than a homeless guy. . . this is not the most helpful information.

The mean deviation

One approach that looks like it might be useful is to look at how far each observation in the data set is from the center. A simple way to do this is to find out the difference between each X and the mean:

$$\text{Mean Deviation (population)} = \sum (X - \mu)$$

This measure is called the *mean deviation* (or sometimes the "average deviation"). It looks promising, but it doesn't work; in fact, the mean deviation of any data set is *always zero*. Why?

This semester's only proof

(You can ignore this if you don't want to be a math geek.)

$$\begin{aligned}
\text{Mean Deviation} &= \sum (X - \mu) \\
&= \sum X - \sum \mu \\
&= \sum X - N\mu \\
&= \sum X - N\frac{\sum X}{N} \\
&= \sum X - \sum X = 0.
\end{aligned}$$

The variance

Because the average deviation doesn't work, mathematicians figured out that a way to make it work is to *square* each of the differences, so they are always positive.

When you square the differences, and divide by N, you have something called the *variance*. The variance is defined as:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

We divide by N so the variance doesn't automatically increase with more observations. (This wasn't a problem with the average deviation since the positives and negatives cancelled each other out.)

Uglier but easier formulas

That formula is called the *definitional* formula of the variance, because it is the most simple expression of it. However, by rearranging the terms, we can make the formula into one that is easier to calculate: the *computational* formula (if you want a proof, see me after class!):

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$$

This is the formula you should use to calculate the *population* variance. It is easier to figure out because you don't need to go through and subtract everything from the mean first.

Samples aren't populations

For a sample, we have to use a slightly different formula. The reasoning for this is fairly technical, but for our purposes all you need to do is remember that when you have a sample, you have to use the sample formula. That formula is:

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

Note that we use $n-1$, not n, on the larger fraction, and that we use s^2, not σ^2, to denote the sample variance. (Work through example.)

The standard deviation

The variance is nice, but it isn't in the same units as the data, because we squared all of the numbers in the data set. If we want the variance to be comparable to the data, we have to fix it later.

The *standard deviation* is based on the variance. Since we squared the numbers to get the variance, it makes sense to take the square root to get the variance back in the same units.

Formulas for the std. deviation

$$\sigma = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}}$$

$$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$$

Homework

You should work through the additional example starting on page 75. You don't need to turn this in. Also, you should independently look at the discussion of skewed distributions on pages 62–63.

You should also do all parts of Question 2 on page 79.

Your paper topic proposal is also due on Monday.

Monday, we will start discussing Chapter 6. I also plan to briefly review for the midterm exam. The exam will cover all material through today's class.