1.4: Numerical Descriptions of Data I: Measures of the Center

Oddly enough, there are several measures of central tendency, as ways to define the middle of a dataset are called. There is different work to be done to calculate each of them, and they have different uses, strengths, and weaknesses.

For this whole section we will assume we have collected n numerical values, the values of our quantitative variable for the sample we were able to study. When we write formulæ with these values, we can't give them variable names that look like $a, b, c, \ldots$, because we don't know where to stop (and what would we do if n were more than 26?). Instead, we'll use the variables $x_1, x_2, \ldots, x_n$ to represent the data values.

One more very convenient bit of notation, once we have started writing an unknown number (n) of numbers $x_1, x_2, \ldots, x_n$, is a way of writing their sum:

[def:summation] If we have n numbers which we write $x_1, \ldots, x_n$, then we use the shorthand **summation notation** $\sum x_i$ to represent the sum $\sum x_i = x_1 + \cdots + x_n$.[2]

[eg:subscriptssums] If our dataset were $\{1, 2, 17, -3.1415, 3/4\}$then n would be 5 and the variables $x_1, \ldots, x_5$ would be defined with values $x_1 = 1$, $x_2 = 2$, $x_3 = 17$, $x_4 = -3.1415$, and $x_5 = 3/4$.

In addition[3], we would have $\sum x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 1 + 2 + 17 - 3.1415 + 3/4 = 17.6085$.

Mode

Let's first discuss probably the simplest measure of central tendency, and in fact one which was foreshadowed by terms like "unimodal."

[def:mode] A **mode** of a dataset $x_1, \ldots, x_n$ of n numbers is one of the values x_i which occurs at least as often in the dataset as any other value.

It would be nice to say this in a simpler way, something like "the mode is the value which occurs the most often in the dataset," but there may not be a single such number.

EXAMPLE 1.4.4. Continuing with the data from Example 1.3.1, it is easy to see, looking at the stem-and-leaf plot, that both 73 and 90 are modes.

Note that in some of the histograms we made using these data and different bin widths, the bins containing 73 and 90 were of the same height, while in others they were of different heights. This is an example of how it can be quite hard to see on a histogram where the mode is... or where the mode**s are**.

Mean

The next measure of central tendency, and certainly the one heard most often in the press, is simply the average. However, in statistics, this is given a different name.

[def:mean] The **mean** of a dataset $x_1, \ldots, x_n$ of n numbers is given by the formula $\left(\sum x_i\right)/n$.

If the data come from a sample, we use the notation $\overline{x}$ for the **sample mean**.

If $\{x_1, \ldots, x_n\}$ is all of the data from an entire population, we use the notation μX [this is the Greek letter "mu," pronounced "mew," to rhyme with "new."] for the **population mean**.

EXAMPLE 1.4.6. Since we've already computed the sum of the data in Example 1.4.2 to be 17.6085 and there were 5 values in the dataset, the mean is $\overline{x} = 17.6085/5 = 3.5217$.

EXAMPLE 1.4.7. Again using the data from Example 1.3.1, we can calculate the mean $\overline{x} = \left(\sum x_i\right)/n = 2246/30 = 74.8667$.

Notice that the mean in the two examples above was not one of the data values. This is true quite often. What that means is that the phrase "the average *whatever*," as in "the average American family has X" or "the average student does Y," is not talking about any particular family, and we should not expect any particular family or student to have or do that thing. Someone with a statistical education should mentally edit every phrase like that they hear to be instead something like "the mean of the variable X on the population of all American families is ...," or "the mean of the variable Y on the population of all students is ...," or whatever.

Median

Our third measure of central tendency is not the result of arithmetic, but instead of putting the data values in increasing order.

DEFINITION 1.4.8. Imagine that we have put the values of a dataset $\{x_1, \ldots, x_n\}$ of n numbers in increasing (or at least non-decreasing) order, so that $x_1 \leq x_2 \leq \cdots \leq x_n$. Then if n is odd, the **median** of the dataset is the middle value, $x_{(n+1)/2}$, while if n is even, the median is the mean of the two middle numbers, $\frac{x_{n/2}+x_{(n/2)+1}}{2}$.

EXAMPLE 1.4.9. Working with the data in Example 1.4.2, we must first put them in order, as $\{-3.1415, 3/4, 1, 2, 17\}$so the median of this dataset is the middle value, 1.

EXAMPLE 1.4.10. Now let us find the median of the data from Example 1.3.1. Fortunately, in that example, we made a stem-and-leaf plot and even put the leaves in order, so that starting at the bottom and going along the rows of leaves and then up to the next row, will give us all the values in order! Since there are 30 values, we count up to the 15^{th} and 16^{th} values, being 76 and 77, and from this we find that the median of the dataset is $\frac{76+77}{2} = 76.5$.

Strengths and Weaknesses of These Measures of Central Tendency

The weakest of the three measures above is the mode. Yes, it is nice to know which value happened most often in a dataset (or which values all happened equally often and more often then all other values). But this often does not necessarily tell us much about the over-all structure of the data.

EXAMPLE 1.4.11. Suppose we had the data

$$\begin{array}{cccccccccc} 86 & 80 & 25 & 77 & 73 & 76 & 100 & 90 & 67 & 93 \\ 94 & 83 & 72 & 75 & 79 & 70 & 91 & 82 & 71 & 95 \\ 40 & 58 & 68 & 69 & 100 & 78 & 87 & 25 & 92 & 74 \end{array} \tag{1.4.1}$$

with corresponding stem-and-leaf plot

Stem										
10	0									
9	0	1	2	3	4	5				
8	0	2	3	6	7	8				
7	0	1	2	3	4	5	6	7	8	9
6	7	8	9							
5	8									
4	0									
3										
2	5	5								

This would have a histogram with bins of width 10 that looks exactly like the one in Example 1.3.2 – so the center of the histogram would seem, visually, still to be around the bar over the 80s – but now there is a unique mode of 25.

What this example shows is that a small change in some of the data values, small enough not to change the histogram at all, can change the mode(s) drastically. It also shows that the location of the mode says very little about the data in general or its shape, the mode is based entirely on a possibly accidental coincidence of some values in the dataset, no matter if those values are in the "center" of the histogram or not.

The mean has a similar problem: a small change in the data, in the sense of adding only one new data value, but one which is very far away from the others, can change the mean quite a bit. Here is an example.

EXAMPLE 1.4.12. Suppose we take the data from Example 1.3.1 but change only one value – such as by changing the 100 to a 1000, perhaps by a simple typo of the data entry. Then if we calculate the mean, we get $\overline{x} = \left(\sum x_i\right)/n = 3146/30 = 104.8667$,

which is quite different from the mean of original dataset.

A data value which seems to be quite different from all (or the great majority of) the rest is called an *outlier*[4] What we have just seen is that **the mean is very sensitive to outliers**. This is a serious defect, although otherwise it is easy to compute, to work with, and to prove theorems about.

Finally, the median is somewhat tedious to compute, because the first step is to put all the data values in order, which can be very time-consuming. But, once that is done, throwing in an outlier tends to move the median only a little bit. Here is an example.

EXAMPLE 1.4.13. If we do as in Example 1.4.12 and change the data value of 100 in the dataset of Example 1.3.1 to 1000, but leave all of the other data values unchanged, it does not change the median at all since the 1000 is the new largest value, and that does not change the two middle values at all.

If instead we take the data of Example 1.3.1 and simply add another value, 1000, with- out taking away the 100, that does change the media: there are now an odd number of data values, so the median is the middle one after they are put in order, which is 78. So the median has changed by only half a point, from 77.5 to 78. And his would even be true if the value we were adding to the dataset were 1000000 and not just 1000!

In other words, **the median is very insensitive to outliers**. Since, in practice, it is very easy for datasets to have a few random, bad values (typos, mechanical errors, *etc.*), which are often outliers, it is usually smarter to use the median than the mean.

As one final point, note that as we mentioned in §4.2, the word “average,” the unsophisticated version of “mean,” is often incorrectly used as a modifier of the individuals in some population being studied (as in “the average American ...”), rather than as a modifier of the variable in the study (“the average income...”), indicating a fundamental misunderstanding of what the mean *means*. If you look a little harder at this misunderstanding, though, perhaps it is based on the idea that we are looking for the center, the “typical” value of the variable.

The mode might seem like a good way – it’s the most frequently occurring value. But we have seen how that is somewhat flawed.

The mean might also seem like a good way – it’s the “average,” literally. But we’ve also seen problems with the mean.

In fact, the median is probably closest to the intuitive idea of “the center of the data.” It is, after all, a value with the property that both above and below that value lie half of the data values.

One last example to underline this idea:

EXAMPLE 1.4.14. The period of economic difficulty for world markets in the late 2000s and early 2010s is sometimes called the **Great Recession**. Suppose a politician says that we have come out of that time of troubles, and gives as proof the fact that the average family income has increased from the low value it had during the Great Recession back to the values it had before then, and perhaps is even higher than it was in 2005.

It is possible that in fact people are better off, as the increase in this average – mean – seems to imply. But it is also possible that while the mean income has gone up, the *median* income is still low. This would happen if the histogram of incomes recently still has most of the tall bars down where the variable (family income) is low, but has a few, very high outliers. In short, if the super-rich have gotten even super-richer, that will make the mean (average) go up, even if most of the population has experienced stagnant or decreasing wages – but the median will tell what is happening to most of the population.

So when a politician uses the evidence of the average (mean) as suggested here, it is possible they are trying to hide from the pubic the reality of what is happening to the rich and the not-so-rich. It is also possible that this politician is simply poorly educated in statistics and doesn’t realize what is going on. You be the judge ... but pay attention so you know what to ask about.

The last thing we need to say about the strengths and weaknesses of our different measures of central tendency is a way to use the weaknesses of the mean and median to our advantage. That is, since the mean is sensitive to outliers, and pulled in the direction of those outliers, while the median is not, we can use the difference between the two to tell us which way a histogram is skewed.

FACT 1.4.15. If the mean of a dataset is larger than the median, then histograms of that dataset will be right-skewed. Similarly, if the mean is less than the median, histograms will be left-skewed.
