2.4: Measures of Central Tendency (Sameness)

We've seen that we can get a sense of data by plotting dots in a graph, and by making a histogram. These tools show us what the numbers look like, approximately how big and small they are, and how similar and different they are from another. It is good to get a feeling about the numbers in this way. But, these visual sensitudes are not very precise. In addition to summarizing numbers with graphs, we can summarize numbers using numbers (NO, please not more numbers, we promise numbers can be your friend).

From many numbers to one

Measures of central have one important summary goal: to reduce a pile of numbers to a single number that we can look at. We already know that looking at thousands of numbers is hopeless. Wouldn't it be nice if we could just look at one number instead? We think so. It turns out there are lots of ways to do this. Then, if your friend ever asks the frightening question, "hey, what are all these numbers like?". You can say they are like this one number right here.

But, just like in Indiana Jones and the Last Crusade (highly recommended movie), you must choose your measure of central tendency wisely.

Mode

The mode is the most frequently occurring number in your measurement. That is it. How do you find it? You have to count the number of times each number appears in your measure, then whichever one occurs the most, is the mode.

> *Example: 1 1 1 2 3 4 5 6*

The mode of the above set is 1, which occurs three times. Every other number only occurs once.

OK fine. What happens here:

> *Example: 1 1 1 2 2 2 3 4 5 6*

Hmm, now 1 and 2 both occur three times each. What do we do? We say there are two modes, and they are 1 and 2.

Why is the mode a measure of central tendency? Well, when we ask, "what are my numbers like", we can say, "most of the number are, like a 1 (or whatever the mode is)".

Is the mode a good measure of central tendency? That depends on your numbers. For example, consider these numbers

> *1 1 2 3 4 5 6 7 8 9*

Here, the mode is 1 again, because there are two 1s, and all of the other numbers occur once. But, are most of the numbers like, a 1. No, they are mostly not 1s.

"Argh, so should I or should I not use the mode? I thought this class was supposed to tell me what to do?". There is no telling you what to do. Every time you use a tool in statistics you have to think about what you are doing and justify why what you are doing makes sense. Sorry.

Median

The median is the exact middle of the data. After all, we are asking about central tendency, so why not go to the center of the data and see where we are. What do you mean middle of the data? Let's look at these numbers:

> *1 5 4 3 6 7 9*

Umm, OK. So, three is in the middle? Isn't that kind of arbitrary. Yes. Before we can compute the median, we need to order the numbers from smallest to largest.

> *1 3 4 5 6 7 9*

Now, 5 is in the middle. And, by middle we mean in the middle. There are three numbers to the left of 5, and three numbers to the right. So, five is definitely in the middle.

OK fine, but what happens when there aren't an even number of numbers? Then the middle will be missing right? Let's see:

> *1 2 3 4 5 6*

There is no number between 3 and 4 in the data, the middle is empty. In this case, we compute the median by figuring out the number in between 3 and 4. So, the median would be 3.5.

Is the median a good measure of central tendency? Sure, it is often very useful. One property of the median is that it stays in the middle even when some of the other numbers get really weird. For example, consider these numbers:

> *1 2 3 4 4 4 5 6 6 6 7 7 1000*

Most of these numbers are smallish, but the 1000 is a big old weird number, very different from the rest. The median is still 5, because it is in the middle of these ordered numbers. We can also see that five is pretty similar to most of the numbers (except for

1000). So, the median does a pretty good job of representing most of the numbers in the set, and it does so even if one or two of the numbers are very different from the others.

Finally, outlier is a term will we use to describe numbers that appear in data that are very different from the rest. 1000 is an outlier, because it lies way out there on the number line compared to the other numbers. What to do with outliers is another topic we discuss sometimes throughout this course.

Mean

Have you noticed this is a textbook about statistics that hasn't used a formula yet? That is about to change, but for those of you with formula anxiety, don't worry, we will do our best to explain them.

The mean is also called the average. And, we're guessing you might already now what the average of a bunch of numbers is? It's the sum of the numbers, divided by the number of number right? How do we express that idea in a formula? Just like this:

$$ \text{Mean} = \bar{X} = \frac{\sum_{i=1}^{n} x_{i}}{N} \label{mean} $$

"That looks like Greek to me". Yup. The $\sum$ symbol is called sigma, and it stands for the operation of summing. The little "i" on the bottom, and the little "n" on the top refers to all of the numbers in the set, from the first number "i" to the last number "n". The letters are just arbitrary labels, called variables that we use for descriptive purposes. The x_{i} refers to individual numbers in the set. We sum up all of the numbers, then divide the sum by N, which is the total number of numbers. Sometimes you will see $\bar{X}$ to refer to the mean of all of the numbers.

In plain English, the formula looks like:

$$\text{Mean} = \dfrac{\text{Sum of my numbers}}{\text{Count of my numbers}} \nonumber $$

"Well, why didn't you just say that?". We just did in Equation \ref{mean}.

Let's compute the mean for these five numbers:

> *3 7 9 2 6*

Add em up:

> *3+7+9+2+6 = 27*

Count em up:

> *i_{1} = 3, i_{2} = 7, i_{3} = 9, i_{4} = 2, i_{5} = 6; N=5, because i went from 1 to 5*

Divide em:

> *mean = 27 / 5 = 5.4*

Or, to put the numbers in the formula, it looks like this:

$$ \text{Mean} = \bar{X} = \frac{\sum_{i=1}^{n} x_{i}}{N} = \frac{3+7+9+2+6}{5} = \frac{27}{5} = 5.4 \nonumber $$

OK fine, that is how to compute the mean. But, like we imagined, you probably already knew that, and if you didn't that's OK, now you do. What's next?

Is the mean a good measure of central tendency? By now, you should know: it depends.

What does the mean mean?

It is not enough to know the formula for the mean, or to be able to use the formula to compute a mean for a set of numbers. We believe in your ability to add and divide numbers. What you really need to know is what the mean really "means". This requires that you know what the mean does, and not just how to do it. Puzzled? Let's explain.

Can you answer this question: What happens when you divide a sum of numbers by the number of numbers? What are the consequences of doing this? What is the formula doing? What kind of properties does the result give us? FYI, the answer is not that we compute the mean.

OK, so what happens when you divide any number by another number? Of course, the key word here is divide. We literally carve the number up top in the numerator into pieces. How many times do we split the top number? That depends on the bottom number in the denominator. Watch:

$$ \frac{12}{3} = 4 \nonumber $$

So, we know the answer is 4. But, what is really going on here is that we are slicing and dicing up 12 aren't we. Yes, and we slicing 12 into three parts. It turns out the size of those three parts is 4. So, now we are thinking of 12 as three different pieces $12 = 4 + 4 + 4$. I know this will be obvious, but what kind of properties do our pieces have? You mean the fours? Yup. Well, obviously they are all fours. Yes. The pieces are all the same size. They are all equal. So, division equalizes the numerator by the denominator…

"Umm, I think I learned this in elementary school, what does this have to do with the mean?". The number on top of the formula for the mean is just another numerator being divided by a denominator isn't it. In this case, the numerator is a sum of all the values in your data. What if it was the sum of all of the 500 happiness ratings? The sum of all of them would just be a single number adding up all the different ratings. If we split the sum up into equal parts representing one part for each person's happiness what would we get? We would get 500 identical and equal numbers for each person. It would be like taking all of the happiness in the world, then dividing it up equally, then to be fair, giving back the same equal amount of happiness to everyone in the world. This would make some people more happy than they were before, and some people less happy right. Of course, that's because it would be equalizing the distribution of happiness for everybody. This process of equalization by dividing something into equal parts is what the mean does. See, it's more than just a formula. It's an idea. This is just the beginning of thinking about these kinds of ideas. We will come back to this idea about the mean, and other ideas, in later chapters.

> *Pro tip: The mean is the one and only number that can take the place of every number in the data, such that when you add up all the equal parts, you get back the original sum of the data.*

All together now

Just to remind ourselves of the mode, median, and mean, take a look at the next histogram. We have overlaid the location of the mean (red), median (green), and mode (blue). For this dataset, the three measures of central tendency all give different answers. The mean is the largest because it is influenced by large numbers, even if they occur rarely. The mode and median are insensitive to large numbers that occur infrequently, so they have smaller values.

```r
library(ggplot2)
my_mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
exp_num<-round(rexp(1000,.15),digits=1)
qplot(exp_num, col=I("grey"), fill=I("white"), bins=30)+
  geom_vline(xintercept=mean(exp_num), color="red", size=1.5)+
  geom_vline(xintercept=median(exp_num), color="green", size=1.5)+
  geom_vline(xintercept=my_mode(exp_num), color="blue", size=1.5)+
  theme_classic()
```

run restart restart & run all

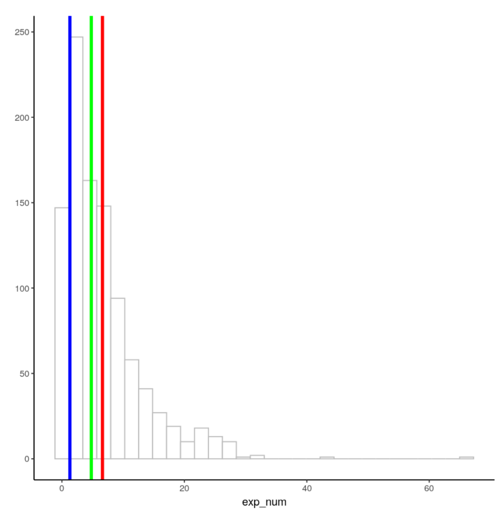

Figure \(\PageIndex{1}\): A histogram with the mean (red), the median (green), and the mode (blue)