Descriptive Statistics

(And a little bit on rounding and significant digits)

Now that we know what our data look like, we'd like to be able to describe it numerically. In other words, how can we represent our data using just one or two numbers?

Actually, there are many ways to describe data numerically; we'll look only at the most common and make a few comments on other methods.

Let's start by discussing how we would best represent our data using just one number. A couple of ideas come to mind:

minimum

maximum

3^{rd} largest number

mode

mean

median

Let's discuss these. The *minimum* and *maximum* are rather interesting - many human activities center around finding the minimum or maximum. Think of the Olympics, for example, or the Guinness book of world records. However they don't really do a good job representing all of the data, only the extreme values.

The *third largest number* isn't really interesting, nor is it useful (in all fairness, who remembers who came in third in the Olympics?).

The *mode* we already discussed. It's represents the data values that are most common. Interestingly, most statistics textbooks mention the mode in a chapter on descriptive statistics and then forget all about it. Except as a descriptor for distributions it's not really all that useful (some theoretical stuff excepted).

That leaves us with the mean and median. Both are extensively used in statistics. Let's discuss the mean first.

Mean

First note that what me mean here is the *sample* mean. We'll discuss samples and populations later, but for now you probably know that when we calculate a mean we take a sample. For example, if we want to know the average weight of rabbits in Northern Virginia, we catch a sample of 20 rabbits and calculate the mean. We don't catch every single rabbit in Northern Virginia (every rabbit in Northern Virginia would be the population).

The sample mean is designated by the symbol $\bar{y}$ and is given as follows:

$$\bar{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n}$$

If you look closely at this, you should realize that what it tells us is to add up all the values in our sample, and then divide by the sample size. This is what you've been doing every time you calculate an average (although you may not have realized it was a sample average).

Just in case, let's do a simple example. We'll use the following (simulated) data for the the levels of estrogen in women during the second trimester (in pg/mL):

2670 4870 2900 1841 4233 5709 4493 2393 6159 7110

We take these data and calculate the sample mean as follows:

$$\begin{aligned}
\bar{y} &= \frac{\sum\limits_{i=1}^{n} y_i}{n} \\
&= \frac{2670 + 4870 + 2900 + 1841 + 4233 + 5709 + 4493 + 2393 + 6159 + 7110}{10} \\
&= \frac{42378}{10} = 4237.8
\end{aligned}$$

A final comment about the population mean. Some textbooks will define the population mean as follows $\mu = \sum\limits_{i=1}^{N} y_i/N$, where N is the population size (not sample size). This is technically correct, but not terribly useful since we very rarely have the luxury of measuring the entire population (see the rabbit example above). The topic of samples and populations is central to statistics and will be discussed soon.

Median

Now let's discuss the median. The median simply describes the number in the middle. If we have an odd number of data points (if n is odd), then this is very easy to calculate. Let's take the above example of estrogen levels and leave off the first number (2670). That gives us the following sample:

4870 2900 1841 4233 5709 4493 2393 6159 7110

It should be obvious that to get the number in the middle we *first need to sort our data*:

1841 2393 2900 4233 4493 4870 5709 6159 7110

And now we can see that the number in the middle is the *sample* median = 4498 pg/mL.

On the other hand, if we have an even number of data points (n is even), then we don't have a number in the middle. In this case we *average* the middle two numbers. Let's put 2670 back into our data set so we now have (sorted):

1841 2393 2670 2900 4233 4493 4870 5709 6159 7110

Now we look at the middle two numbers, 4233 and 4493, and calculate the average:

$$Sample \text{ median} = \frac{4233 + 4493}{2} = 4363$$

Let's summarize:

> If the sample size is odd, the sample median is the number in the middle.
>
> If the sample size is even, the sample median is the average of the middle two numbers.

Before we go on to look at other descriptive statistics, let's discuss the mean vs. the median. Which is better? Well that depends (don't you love a vague answer like that?).

In statistics the mean is used most often. One of the reasons for this is that it represents more of the data. Each data point is added and makes up part of the final answer. The median only uses the value of the middle (or middle two) numbers, and as such isn't as *representative* of the data. The mean includes the actual value of all the data in the calculation, the median does not.

The fact that the median doesn't use the actual value of all the numbers is, however, one reason it sometimes does much better in representing a set of numbers than the mean. Let's use a non-biological example.

What's the average income in your neighborhood? In Fairfax County, it's probably pretty decent and comes in at about $50,000. Notice that this is per person, not per household (that figure's much higher).

Now suppose Bill Gates decided to move into your neighborhood. What happens to the average in just your neighborhood?

Let's figure it out:

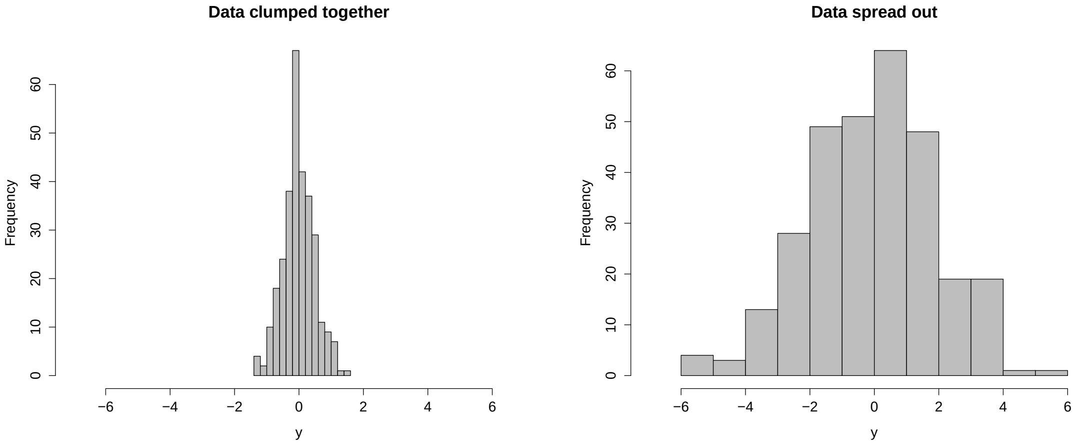
> Let's assume 25 single family homes in your neighborhood, each with two people making about $50,000. The mean is (obviously) $50,000 (i.e., $50,000 \times 50 = 2,500,000$ then $2,500,000/50 = 50,000$).
>
> Now Bill Gates moves in. Last year his income was about $11,000,000,000. The new average is $(2,500,000 + 11,000,000,000)/51 = 11,002,500,000/51 = 215,735,294$.
>
> In other words, the average income in our little neighborhood is now $215,735,294.

Does this really *represent* our neighborhood? No, it doesn't. The median, on the other hand, is hardly affected. It moves up half a spot. It doesn't care how big the largest value is - the largest value could be *anything* and the median would be affected in exactly the same way (if Bill Gates made 10 times as much money, it wouldn't make a difference to the median).

In this example, the median does a much better job in representing the neighborhood. Incidentally, you'll notice that for income the median is often used instead of the mean for exactly this reason.

So now we have an idea how to measure the center of our distribution. What else might we be interested in? The next most important thing is to figure out how spread out our data are. Are all the observations sort of similar, or are the rather different from each other?

Here's an example:



Here we also have some candidates:

range

average absolute deviation

variance

standard deviation

As we did above, let's go through these. Well start with the *range*. The range is defined as follows:

$$\textit{range} \;=\; \textit{maximum value} \;-\; \textit{minimum value}$$

The range is interesting (often for the same reasons the maximum and minimum are interesting, and it does tell us something about the spread of our data. But it's too sensitive to *outliers* or extreme values (think of what Bill Gates would do the the range of salaries).

Really, what we want is some kind of *average* distance that each observation is from the mean. Something like an *average deviation*. This isn't on the list above, but let's try it anyway.

We'll use some (made up) data on blood sugar levels in six diabetic patients after fasting (in mg/dL):

117 122 133 104 84 101

Let's calculate the sample mean:

$$\frac{117 + 122 + 133 + 104 + 84 + 101}{6} = 110.17$$

Now let's get the average deviation:

$$\begin{aligned}
117 - 110.17 &= 6.83 \\
122 - 110.17 &= 11.83 \\
133 - 110.17 &= 22.83 \\
104 - 110.17 &= -6.17 \\
84 - 110.17 &= -26.17 \\
101 - 110.17 &= \underline{-9.17} \\
\textbf{Sum:} \qquad & \qquad \mathbf{0.00}
\end{aligned}$$

So that doesn't work (we can stop here, dividing by 6 is obviously pointless). It can be shown fairly easily that the sum of the average deviations will always add up to 0. So this idea doesn't work.

But suppose we're not interested in the *direction* of the deviation, just in the magnitude. In other words, we don't care if it's a positive or negative difference, just in the total deviation (how far are we from the mean, regardless of direction).

This give us the *average absolute deviation*, which we can calculate simply by taking the absolute value of all the differences. We'll stick with the same example:

$$\begin{aligned}
|117 - 110.17| &= |6.83| &&= 6.83 \\
|122 - 110.17| &= |11.83| &&= 11.83 \\
|133 - 110.17| &= |22.83| &&= 22.83 \\
|104 - 110.17| &= |-6.17| &&= 6.17 \\
|84 - 110.17| &= |-26.17| &&= 26.17 \\
|101 - 110.17| &= |-9.17| &&= \underline{9.17} \\
\textbf{Sum:} \qquad & && \mathbf{83.00}
\end{aligned}$$

Now we can divide this by 6 and we have:

$$\text{Average absolute deviation} = 83.00/6 = 13.83$$

The average absolute deviation is used in statistics, but for a number of reasons (theoretical and practical) it is difficult to work with. So even though we now have a nice measure of spread, it doesn't do us much good for more complicated analyses. We'll have to use something else.

The problem with the average deviation is that if we add up the differences, we'll get 0. To overcome this, we used the absolute value. But another way we can make negative differences positive is by squaring them. It turns out this is the approach that is preferred in statistics (at least in most introductory statistics).

So let's talk about the *variance*. Let's use our example and calculate the variance. Now we square all the differences:

$$\begin{array}{rrr}
(117-110.17)^2 = & (6.83)^2 = & 46.69 \\
(122-110.17)^2 = & (11.83)^2 = & 140.03 \\
(133-110.17)^2 = & (22.83)^2 = & 521.36 \\
(104-110.17)^2 = & (-6.17)^2 = & 38.03 \\
(84-110.17)^2 = & (-26.17)^2 = & 684.69 \\
(101-110.17)^2 = & (-9.17)^2 = & \underline{\quad 84.03} \\
\textbf{Sum:} & & \textbf{1514.83}
\end{array}$$

The sum above (1514.83) is actually a rather important quantity - it is called the *Sum of Squares* and is abbreviated by the letters SS. In other words:

$$SS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 1514.83$$

The SS is used a lot in various types of analyses, although we won't see it much until later in the semester.

Let's finally calculate the variance. To do this, we take the SS and divide by $n-1$:

$$Sample\ variance = s^2 = \frac{SS}{n-1} = \frac{1514.83}{6-1} = \frac{1514.83}{5} = 302.97$$

We use s^2 as the abbreviation for the sample variance.

Why do we divide by $n-1$ instead of n? We'll explain that a little further down. For now just take it as a given.

So what we wind up with is a value for the sample variance.

The sample variance is used a lot by statisticians. Many statisticians don't even bother with standard deviations. Here's the formula for the variance:

$$s^2 = \frac{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

(You should make sure you understand this formula).

But the variance has one problem. It's the average of the squared deviations. In other words, the units are all squared.

Blood sugar is measured in mg/dL. The sample mean has units expressed in mg/dL, so does the median and, for that matter, the absolute average deviation. But the variance has units measured in (mg/dL)^2. This is also why the variance is so much bigger than the absolute average deviation. We need to get back to the correct units.

This leads us to the *standard deviation*. The standard deviation is the square root of the variance:

$$Standard\ deviation = s = \sqrt{s^2}$$

For our blood sugar example it becomes:

$$s = \sqrt{s^2} = \sqrt{302.97} = 17.41$$

This is at least on the same order of magnitude as our average absolute deviation.

The sample standard deviation (abbreviated by s) is also used a lot in statistics. We'll see it a lot as we go through the semester.

Here's the formula for the standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{SS}{n-1}}$$

Finally, let's talk about why we use $n - 1$ in the denominator for the variance. We'll go with an intuitive explanation (there's also a theoretical explanation that's based on the concept of *expected* values and how we derive the formula (theoretically) for the sample variance).

Let's take a sample of size 1. What is the variance? Using our formula for the variance we get:

$$\frac{0}{0}$$

This quantity is undefined. In other words, if we have a sample of size one, we can *not* calculate the variance. This actually makes sense because a sample of size one has no information about the variability in our population (we need a minimum sample size of 2 to say anything about variability in our population).

Are we ever justified in using n instead of $n - 1$? Confusingly, the answer is yes, but only very rarely. If we (somehow) manage to measure everything in our population, then we can use n (actually N). Many statistic textbooks make a big deal out of this and go on to give the following formula:

$$\sigma^2 = \frac{\sum\limits_{i=1}^{N}(y_i - \bar{y})^2}{N}$$

Here, the N indicates that everything in the population has been measured (σ^2 is the *population* variance, which we'll get back to later). This formula is not terribly useful as we can virtually never measure the entire population, so we'll make two comments and then ignore this formula:

1. If we have a population of size 1 (unrealistic) the formula does yield the correct answer. This time we get 0 instead of 0/0. This again makes sense because it tells us that a single individual has no variation.
2. Calculators will give you a choice as to which denominator to use (n or $n - 1$). You ought to make sure that you *always* use $n - 1$ in the denominator.

Optional material (an alternative formula for the variance):

This is mostly of historical interest, although some textbooks will still give you the following alternative formula for calculating the variance:

$$s^2 = \frac{\sum\limits_{i=1}^{n} y_i^2 - \frac{\left(\sum\limits_{i=1}^{n} y_i\right)^2}{n}}{n-1}$$

In the days before calculators (and computers) could easily calculate the variance, this was an "easier" formula to use since you didn't have to calculate all the squared differences. For computational reasons having to do with the way computers and calculators work, this formula is not really preferred these days. But if you should ever need to calculate a variance without a calculator that has statistical functions you might find it really is a bit easier to use.

(Incidentally, it's not difficult to prove that this is equivalent to the formula we used for the variance above.)

Some information on significant digits and rounding.

This is really the first time we've had to do some real math, so this is a good point to discuss information on rounding and significant digits.

> *A personal comment: I'm not very strict on this - I much prefer you give me too many digits than not enough. You will not get marked off on anything if you give me a few too many digits, but you might get marked off if you don't give me enough.*

Significant digits:

What is a significant digit?

4.5960	means the 0 is significant.
0.3425	usually means the 0 is not significant.
23,000	usually means the 0's are not significant.

So how do we use significant digits? Technically, your final answer should not have more digits than the number with the *least* number of significant digits in your original data. If we have the following three numbers, for example:

8.0 6.2 4.34

and we want to calculate the average, we should get $\bar{y} = 6.2$ (not 6.18). *But see my personal comment above!*

Rounding:

Everyone probably knows the rules of rounding - if the last digit is a 5 or above, we round up. If not, we round down. For example, rounding 6.15 to two significant digits we get 6.2. Rounding 6.14 to two significant digits we get 6.1.

If you've heard of something called statistical rounding, we will *not* be using this in this class.

Finally, an important comment about rounding and significant digits:

Never round anything until you are finished with all your calculations!

In other words, in any of your calculations keep all the digits you can in your calculations. Never round anything until you're all done. Then feel free to round to the appropriate number of significant digits. Why?

Consider the following number:

$$2/3 = 0.7\,??$$

Or how about:

$$2/3 = 0.667\,??$$

What is 2/3 really equal to? It can't be represented on a calculator (or computer), and the more you round, the less accurate your answer will be.

This is just a simple example, but the topic of mistakes or errors made by computers and calculators is actually *very* important in computer science (and in what is called the *computational sciences*, which is not the same as computer science). The best advice is as given above - don't round anything until you're all done!