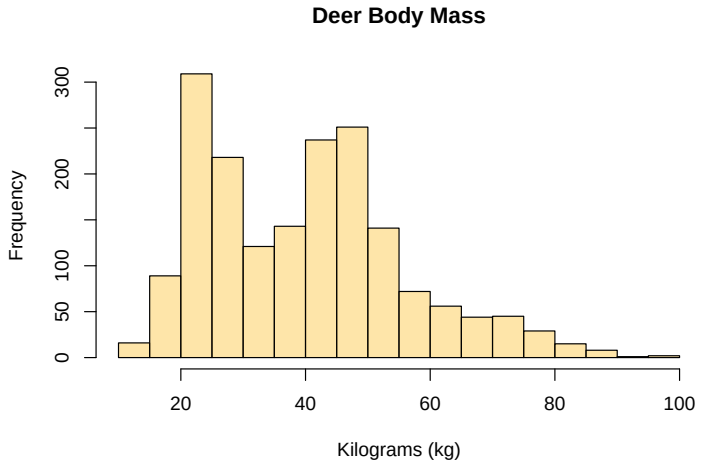
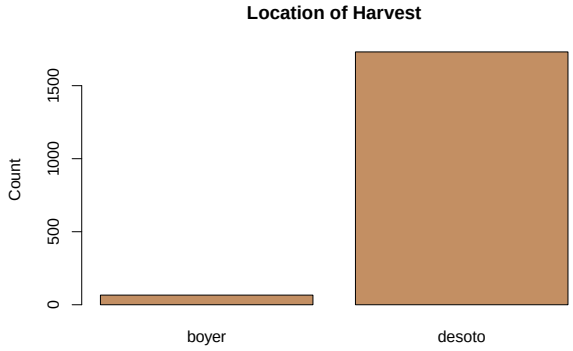
Chapter 1.3 - Measures of Center

> "Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable." - Bobby Bragan

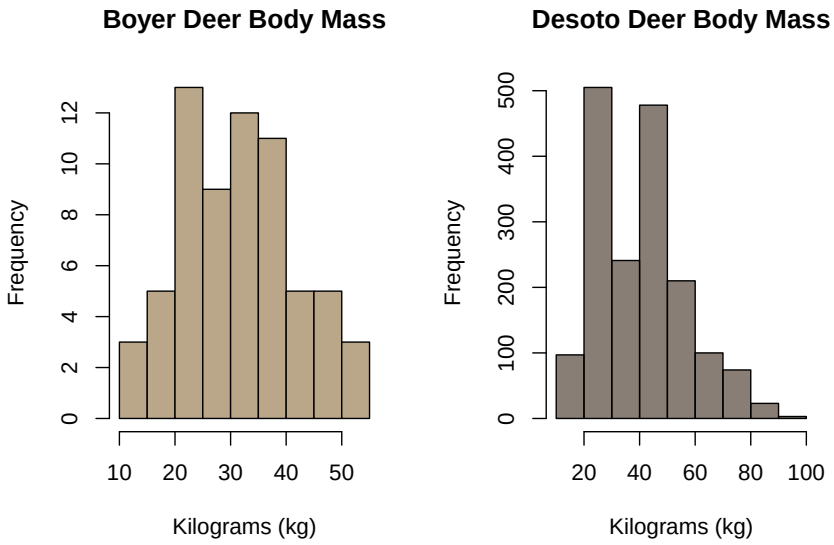
Let's look at the weights of our white-tailed deer once more.



When we consider the features of our sampled deer like body weight it's *probably important* to consider where the samples came from. You'd might weigh less if the place you lived had no food, we should be able to apply the same concept to any living thing.

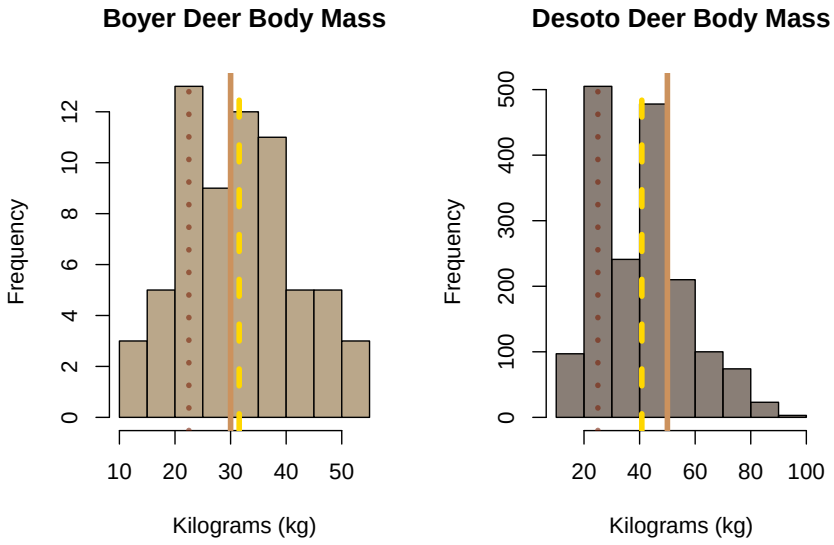


It's apparent that the majority of the sample is sourced from Desoto, so there may be merit in splitting up our sample by location and looking at those body weights to figure out if location has a real impact on body weights.



So far we've learned enough language to describe how these two histograms differ **in shape**. But how could we describe their differences numerically, without showing the histograms? We'd want some way of representing the entirety of our samples, numerically, without using a graph, and without lazily showing all of the data in a table.

Intuition might lead us towards the center, but what would we consider the center? Do we use the dead center of the histogram? Should we consider the frequency of our data to be relevant and use the center of the data instead? What about the highest point of our histogram (a.k.a. the most frequent observation)?



This is the moment in our language learning where we discover the horrors synonyms, those pesky words that all have roughly the same meaning but are completely different spellings. Fortunately we'll be cracking into our statistical thesaurus and resolving as much confusing as possible.

Mean

When a statistician talks about the "mean" value of something, they're referring to the *average* value. Most people know what an average is. The average height of students in a classroom, average SAT scores, average temperature for the day. There's a lot of power behind this sort of publicity— if we describe something using the average we can rely on our audience understanding us without needing much (if any) additional context.

The recipe for calculating a mean (or average) is rather direct. You take all of the values, add them up, and divide them by the number of values you added. We even call this method "averaging".

7	3	12	3	5

$$\text{Mean} = \frac{7+3+12+3+5}{5} = \frac{30}{5} = 6$$

We can combine our previously learned phrases with this new vocabulary:

If the data we calculated a mean for comes from a **sample**, we call it a **sample mean**. If the data we calculated a mean for comes from a **population**, we call it a **population mean**.

As stated, the best part about the mean is that it's fairly ubiquitious, in that almost everyone's heard of it. One of the reasons for this is that it's very easy to interpret: The average salary for a position can be interpreted as the expected salary for someone with standard qualifications.

The problem with the mean is that it's very sensitive to things called *outliers*. Consider the following:

> You walk into a restaurant and sit at the bar. You notice that the man to your left is Tom Brady and the only other person sitting at the bar is Payton Manning. Thus the average number of Super Bowl rings owned by the people sitting at that bar is 3. That's an absolutely correct statement, but it's very misleading.
>
> As of 2021 the top 1% of households in the United States hold 32.3% of the country's wealth, while the bottom 50% hold 2.6%. Describing the average household salary in the U.S. isn't incredibly informative because there's a small number of data points that are over-represented by the mean.

We refer to a measurement as *resistant* if it isn't susceptible to outliers. We find resistance to be a very attractive feature when the data we're working with has *strong* outliers, not necessarily when it has **any** outliers.

Notation

Teachers tend to know other teachers, it's a bizarre phenomenon. One of my colleagues in education refers to mathematical notation as "letter math" and regularly shares her woes about how mathematics made sense until the alphabet stole the show. Ironically, this person teaches math.

It's at least anecdotally true that mathematical notation is a significant source of strife for students as they make the leap from simple arithmetic to real maths. At this point in time I still haven't figured out a way to make notation so simple and intuitive that students don't suffer through it. The best I can do is try to convince you of it's usefulness so that hopefully you're more motivated to learn it.

Let's start at the beginning of our "math alphabet". We can denote data values as $x_1, x_2, x_3, ...$, thus x_1 refers to the observed value of the **variable** x from **individual** 1.

It's important to recognize the difference between convention and law when it comes to notation. It's **convention** to denote data values as letters towards the end of the alphabet: x, y, z. You don't have to do this, you can use different letters or even full words— $Deer_1$ can refer to the first deer in a table of deer body weights.

It's **law** to denote them consistently. If you refer to the first fish in your data set as F_1 you can't suddenly change that to d_4 without *at least* explaining it.

Most of what you'll see with notation is convention. You **need** to learn convention so that you can understand the equations you'll be working with. You **should** use convention so that others can understand your work. But you don't have to as long as you explain what you've changed.

Sample size (the number of individuals in the sample) is denoted with a lower case n.

Population size is denoted with a capital N.

Summation refers to the summ, or addition, of everything contained in the expression. We denote this with the Greek capital "Sigma": Σ.

We'll often see a variable like i or j put underneath the summation symbol, set equal to a number. This is referring the some "index", or starting place, for a variable that's acting as multiple values. If we put a number or variable at the top of our summation symbol then we're denoting a "stopping point" for the summation.

$$\sum_{i=1}^{n}$$

With this notation we can easily describe "The summation of x_i to the n^{th} term, starting from 1".

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$$

So if we have a data set of 10 values and we want to sum the first half we could write:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

Despite the fact that our data set, x_i, contains values up to x_{10}.

Using "sigma notation", as we'll refer to is moving forward, we can express the sample and population mean formulas:

- Sample mean (denoted $\bar{x}$):

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

- Population mean (denoted μ):

$$\frac{1}{N}\sum_{i=1}^{N} x_i$$

Greek letters typically refer to **population parameters**, while lower-case English (Latin) letters represent **sample statistics**.

In practice the usage of the formula is far less painful than its appearance suggests:

Student	1	2	3	4	5	6	7	8	9	10
Abscences	2	6	1	2	4	0	1	3	0	2

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{1}{10}(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10})$$

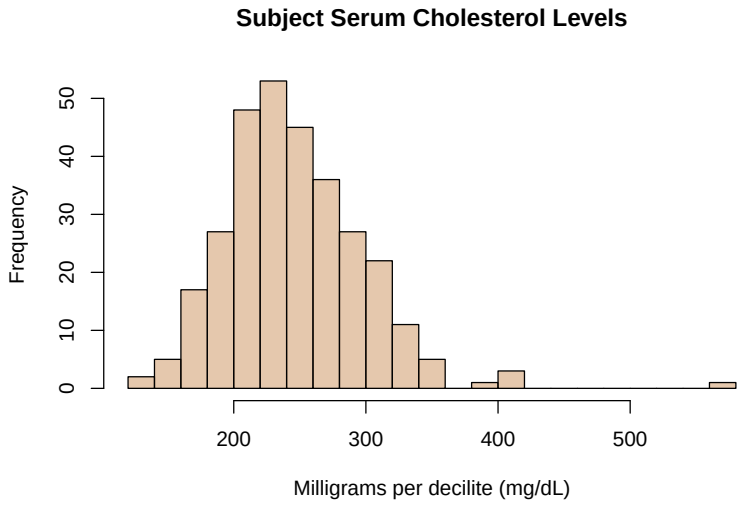
$$\frac{1}{10}(2 + 6 + 1 + 2 + 4 + 0 + 1 + 3 + 0 + 2)$$

$$\frac{1}{10} * 21 = \frac{21}{10} = 2.1$$

The wonderful gift that notation gives us is the ability to condense larger algorithms (mathematical recipes) into single lines. Without sigma notation we would have to write out a sentence every time we wanted to describe how to calculate a mean. With sigma notation we can fit the entire process onto postage stamp.

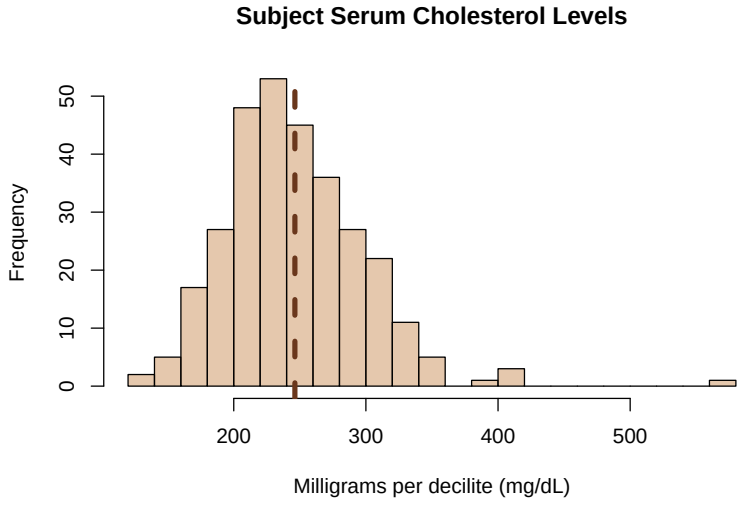
Median

Outliers are "extreme" values that fall outside of the denser (more connected) regions of our data set.

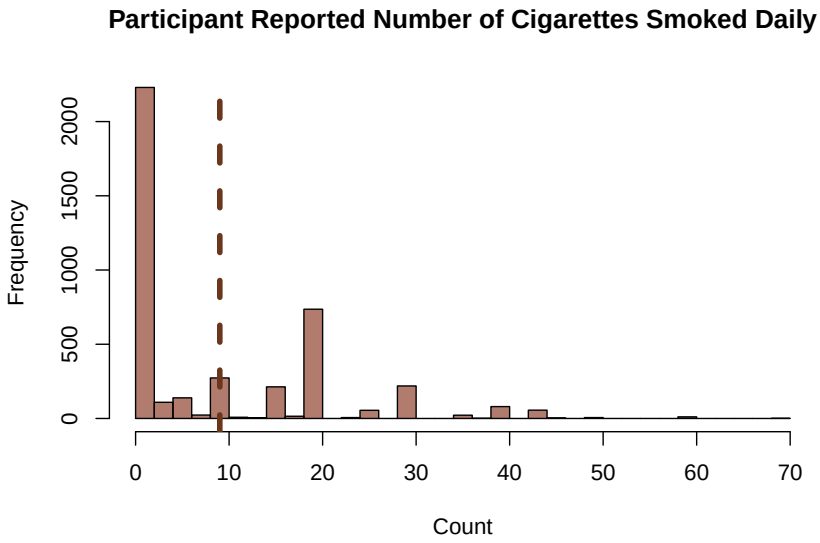


It's not inherently negative to have outliers in a data set. Usually they're a true measurement that's *supposed* to happen but is just *rare* or infrequent. It's important to never throw out data and we'll see why in later chapters. For now though, we have to figure out how to work with the outliers we have.

The influence of extreme values is sometimes quite minimal. If we take the average of these cholesterol levels and represent them with a horizontal line:



We can see that the mean is still contained within the densest region of the data. We might not consider these to be impactful outliers, so using the mean as a measure of center is easily justified. Sometimes it's not so defensible; consider the study below tracking daily cigarette consumption of participants.



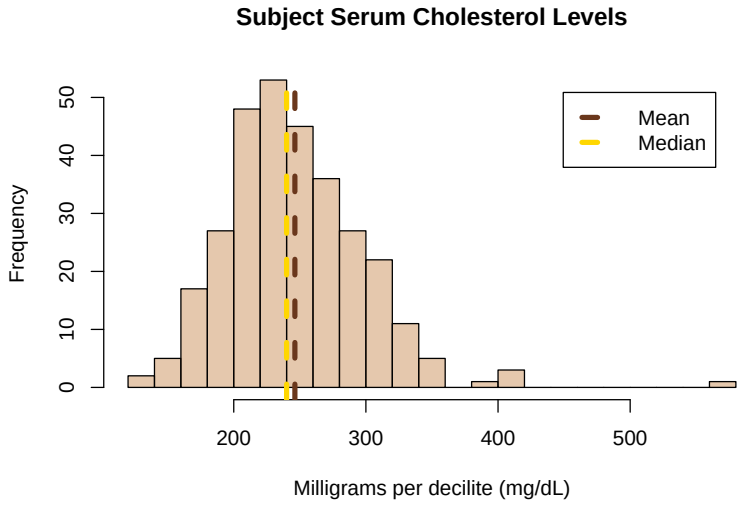
We can see that the *vast majority* of these participants don't smoke at all, yet our outliers affect the mean so much that we're led to believe that our participants are smoking quite a bit ($\bar{x} = 9$). This is because the mean considers magnitude **and** frequency. So, despite our data being made up of mostly zeros, the mean is considering all of those "extreme" values to be just as important.

The easiest work around for this would be to reduce our measurements reliance on magnitude. That is to say, if we place more weight on how frequently an observation shows up in our data set then we should reduce the impact of outliers.

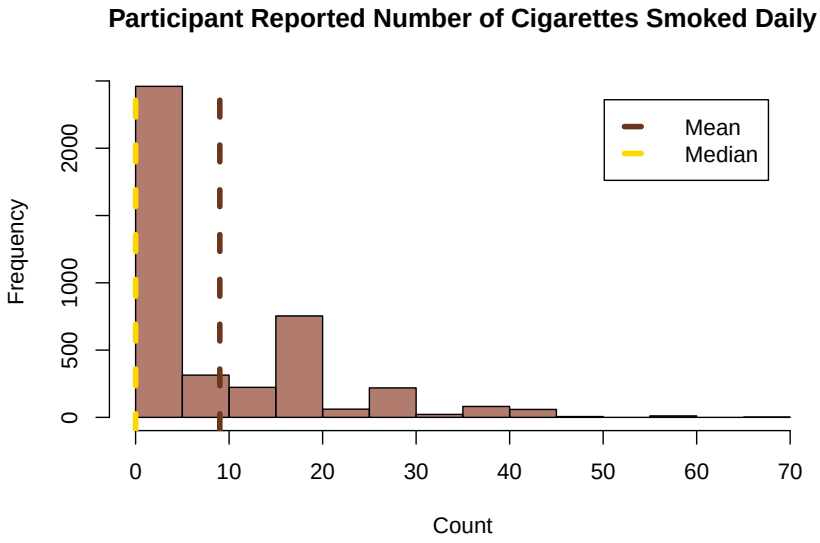
One of the ways we do this is by placing our data into a line from "smallest" to "largest" and picking the dead center. We call this the **median**.

Median: The middle value, where half the data are below and half are above.

Occasionally calculating the median produces a result *similar* to the mean:



But in the case of strong outliers, we can see that it gives us a better representation of the center:



We'll look at two common techniques for calculating the median. The first (elimination) is useful for small data sets ($n \lessapprox 20$) as it's very direct and quick to implement. The second (derivation) is reliable for any size of data set and less prone to error, but relies on a formula which early learners may find cumbersome.

To calculate the median for the data below using **elimination**:

7	3	12	3	5

1. Sort the data in **increasing** order (low to high).

3	3	5	7	12

2. Mark or remove values evenly from each end until you hit the center value.

	3	5	7	

		5		

In this case the median is **5**.

Using **derivation**:

If n is **odd**: Choose position $\frac{(n+1)}{2}$ in the ordered data set

$$\frac{(n+1)}{2} = \frac{(5+1)}{2} = 3$$

We would pick the 3^{rd} data point after sorting, which is still 5 in the original data set.

If n is **even**: Pick $\frac{n}{2}$ and $\frac{n}{2}+1$ and average the two data points. Note that the averaging of the two central data points is necessary regardless of the method used to find the median.

We should still order the data:

7	3	12	3	5	8

3	3	5	7	8	12

$$\frac{6}{2} = 3 \ , \ \frac{6}{2} + 1 = 4$$

3	3	**5**	**7**	8	12

$$\frac{5+7}{2} = 6$$

In this case the median is 6. Notice that the value for the median doesn't appear in the data at all. This is an intended result.

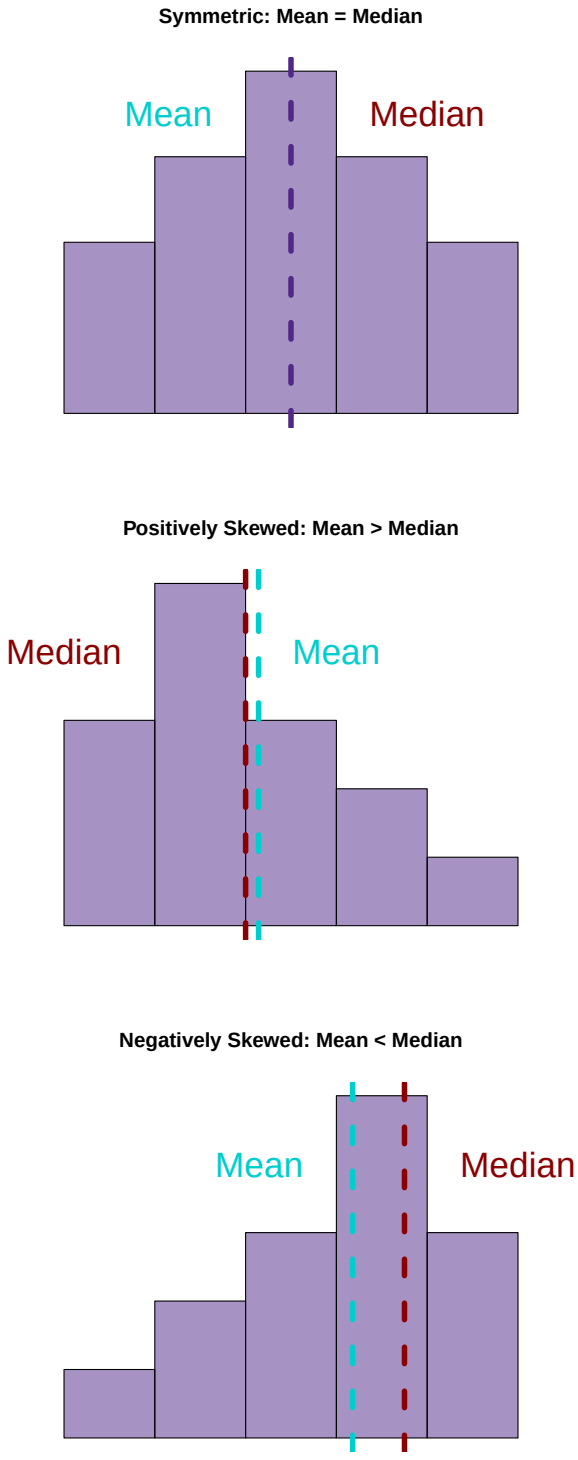
The median is a very useful measure of center because it doesn't make direct use of all of the data. That said, it doesn't disregard the shape of the data entirely, it still incorporates some consideration of *magnitude* without putting *significant weight* to it.

By definition the median is a **resistant** measure of center— outliers have little to no effect on the value of the median. This makes it a particularly good measurement for things like household income:

$$\text{Median Household Income (Kansas)} : 57,422 \text{ USD}$$

$$\text{Average Household Income (Kansas)} : 77,509 \text{ USD}$$

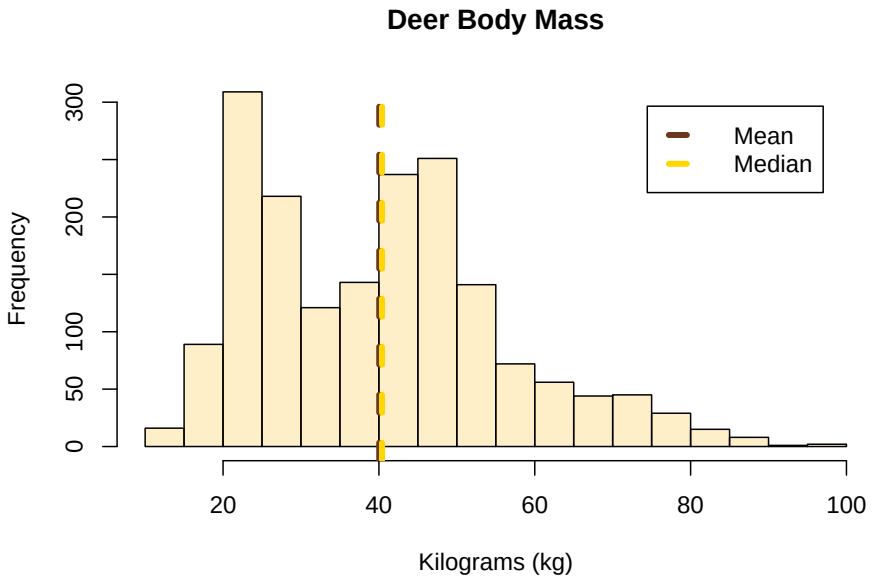
The difference between median and mean depend on *skew* of the histogram (shape of the data):



The more egregious the skew the more separated the median and mean will become, so for our toy data sets in the above histograms we see only a slight difference. Household income is a perfect example because the vast majority of families in the U.S. make less than $\$1,000,000$ annually, but the ones who make more than that tend to be *multiple orders of magnitude* above it.

Mode

On (somewhat rare) occaison, we may find that the mean and median don't suffice as measures of center for our data.



It's almost always **context** that dictates our choice of measure. It's not unusual that scientists are more worried about what value they're most likely to encounter rather than the value that's most representative of the total data set. In this case, the mode is a very useful tool.

The **mode** is the *most frequent* observation in the data, or the value that comes up the most. This is usually very useful in qualitative analyses whereas its use cases are limited in quantitative analyses.

> "Which species of *Salmonella* is most commonly growing in my flour?" is a question with real scientific significance. We can expand on that answer to figure out what methods might be needed to control the dominant species.

> "What's the most common weight of cattle on our research farms?" isn't an inherently useful question to answer. Do we care that the majority of our cattle are exactly 300 lbs.? Possibly. But we're likely more concerned with knowing that the average weight is 220 lbs. since that probably means our cattle are sick or underfed (or our sample is poorly stratified by age).

A data set can have any number of modes $(0, 1, 2, ...)$. Computing the mode is an intuitive task: count the number of times each observation occurs and pick the one(s) with the highest frequency.

3	3	5	7	12

The most frequently observed value in this data set is 3, thus the mode is 3.