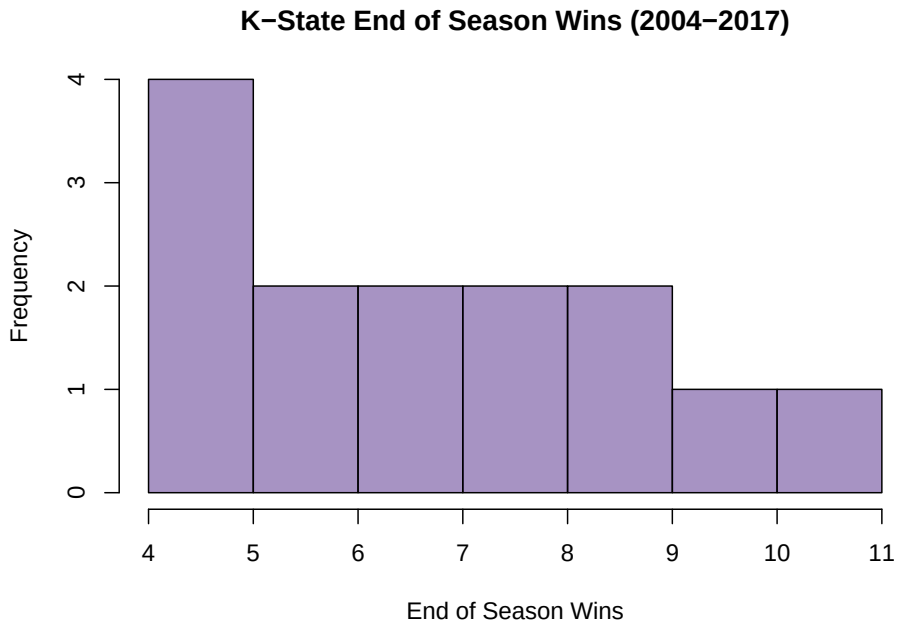
Chapter 1.4 - Measures of Spread

> "It's easy to predict what most people would do, but very hard to predict what any one person would do." - Unknown

The later chapters of this textbook will thoroughly establish what I consider to be the only philosophical truth of Statistics: The entire science of Statistics is a pointless exercise in an ideal world. The problem is that we live in an imperfect world with insufficient technology and incomplete mathematical theory. Thus, statistics is the most efficient science we [humans] can pour our lifespans into.

For example, as much as it's painful to admit K-State *does* lose football games. In a perfect world we would have a lineup of Tyler Locketts, Daren Sproles', and Jake Waters' all linked to a Bill Snyder hive mind. Since (for inexcusable reasons) they can't achieve this, their distribution of seasonal wins looks like this:



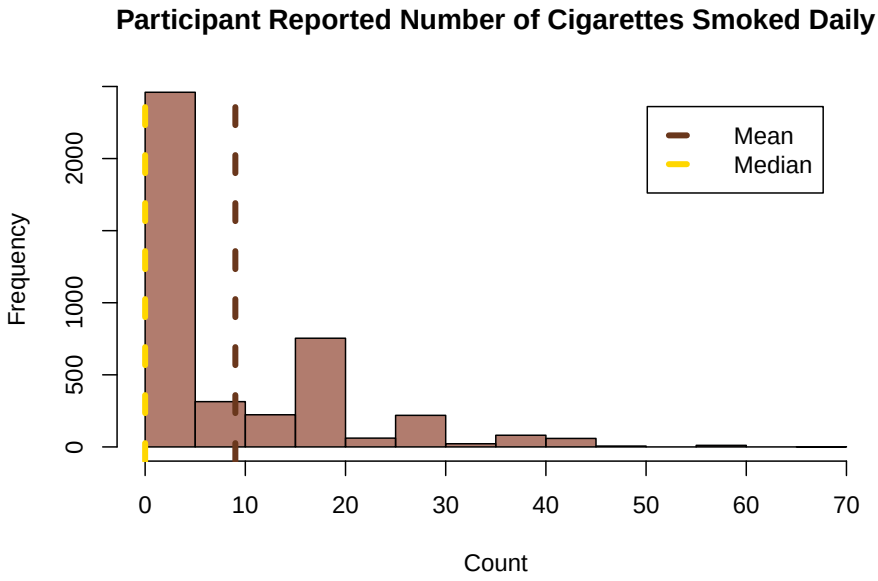
When we look at this data we can see a certain **spread** to it. Just as the center can be described numerically, so can the spread. Using both of these numeric descriptors is sometimes more than enough to describe the *entire shape* of a data set.

But **why** would we care about describing the spread of data? I'd first ask whether or not it's **fair** to make inference from the center alone? Is the mean resistant? Is the median representative of all the data? Does the mode say anything about outliers?

Spread is an important metric for understanding the **differences** or **variation** in data. These differences are hidden when we only consider the center, and the trend of a sample is hidden when we only consider the spread. So while we go over the methods of describing variation in data it's important to think through what can be added to complete the picture of our data (without ever visualizing it on a graph).

Range

When looking at the mean and median for the smoking data it was clear that the median described the majority of participants much better than the mean.



But there's a fundamental problem with exclusively considering the median here, we lose almost all of the information and nuance contained in the the data. A measure of spread can help regain some of that lost information, but as statisticians are communicators first and scientists second we should look to use a measure that's *easily explained*.

The simplest measure of spread at our disposal would be to report lowest and highest values in our data. By doing this we allow the audience to make inference about the "in-between" of our data without looking at it.

Minimum	Median	Maximum
0	0	70

It seems redundant to place 0 in that table twice, right? As it turns out, it would be redundant to place the minimum and maximum in that table *regardless of their values*. We can turn them into a single value **without** losing any information by calculating the **range**.

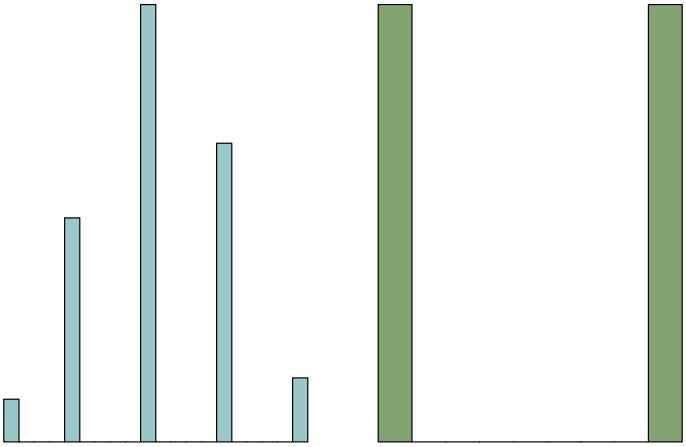
Range: Difference between the largest and smallest data value

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Our table suddenly has a blank space that we could use to give more context to our data!

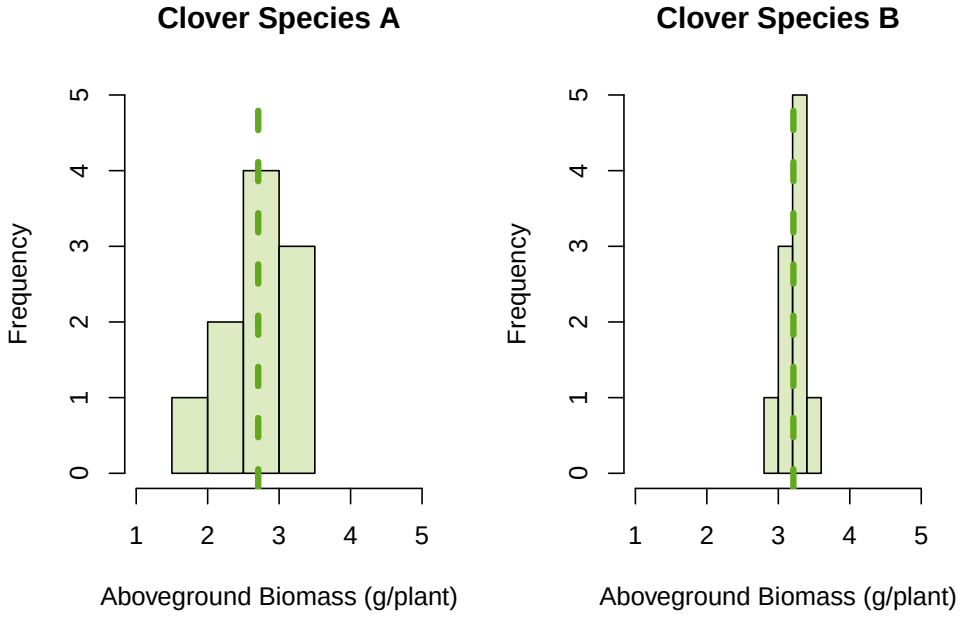
Mean	Median	Range
10	0	70

With every measure/metric there's good and bad. Range **does** let us look at spread, but its difficult to differentiate between data sets with range alone:



These two data sets could have the same range **and** mean, but does that make them the same data? It would be helpful to use a measure that can clear up these differences between data sets. How could we measure this to begin with?

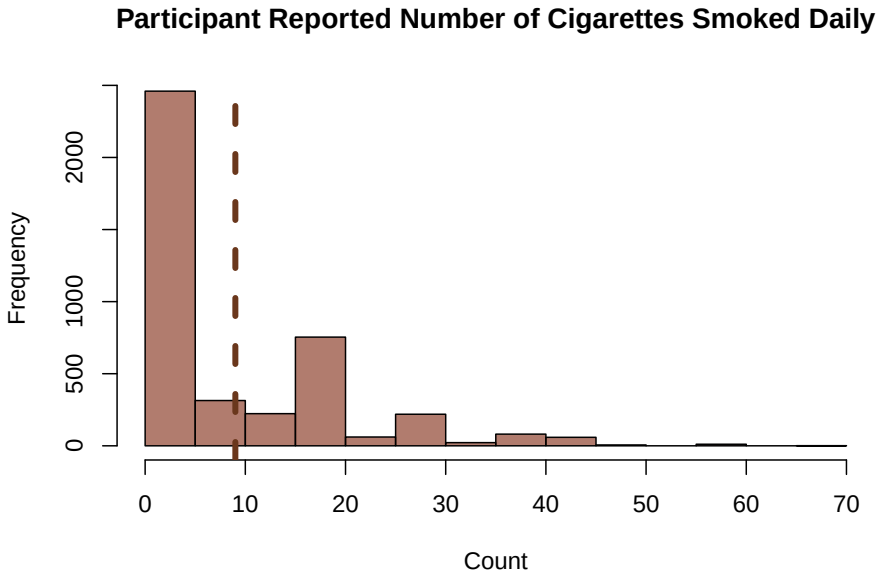
Consider a simple experiment measuring the aboveground biomass of two different clover species:



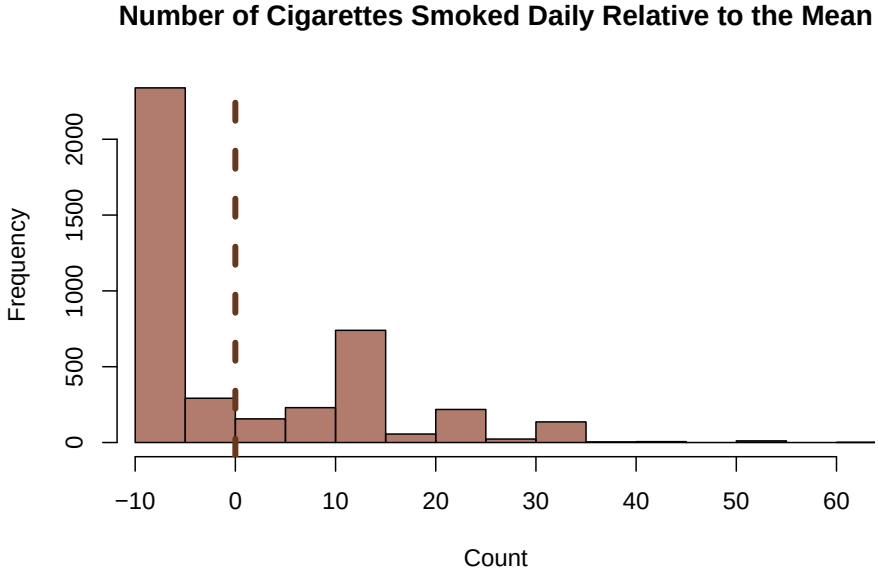
Species B has *smaller* spread, more of the data is **clustered around the mean**. Meanwhile species A has *larger* spread, more of the data is **far from the mean**. This isn't just a coincidence, it's a *distinct feature* of spread in data. What we can do to better describe the nuance of spread is use the **mean** as a **reference point**.

Variance

Let's start with the smoking data and it's mean:

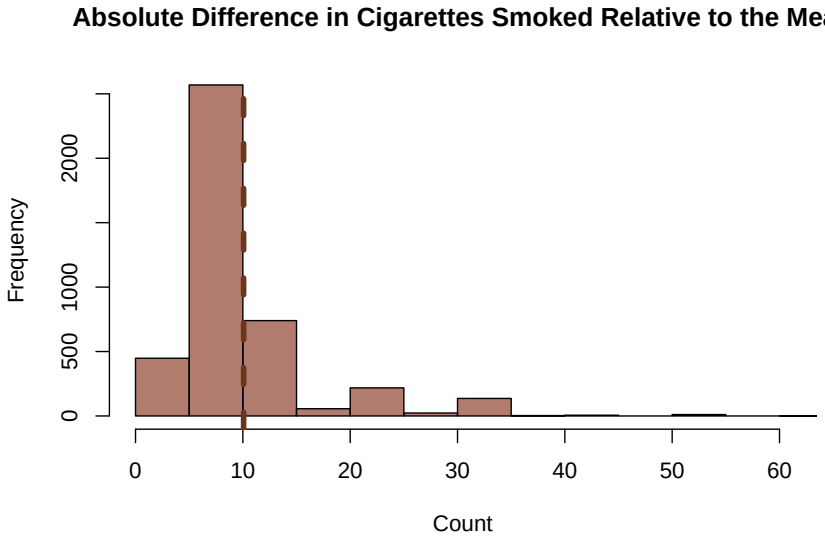


And "re-define" the mean as 0, then adjust every other data point to the new mean:



We haven't accomplished anything productive. The reason for this is because our data has only *shifted*. The value of mean, median, and mode may have changed but their **location** hasn't. More importantly the data is now complete nonsense; how can someone smoke negative numbers of cigarettes?

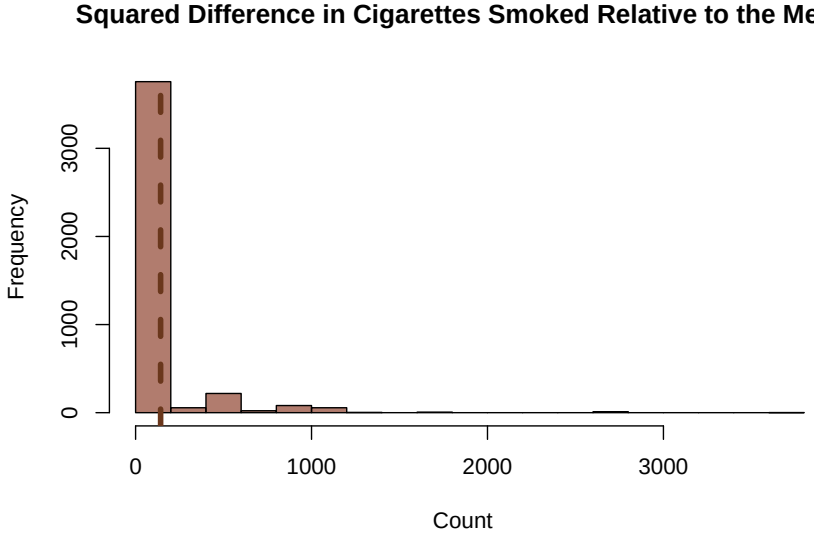
If we focused on *absolute* difference between a data point and the mean we would be able to see how different the data is from the center, which would be much more interpretable, so let's try that. All we have to do is make all of our **negative** values into **positive** ones.



The data has been fundamentally changed, as the new title suggests. The mean of this data can now be interpreted as the average difference in cigarettes smoked between participants *in the original data set*, what a convenient statistic!

Unfortunately, we can't use this specific method. While absolute values are graphically simple and intuitive in explanation, they're absolute demons in calculation. Statistics has only recently enjoyed the advantages of computers and numerical calculus, which means most of the standard practices in statistics revolve around methods that are *analytically tractable*. That is to say, they can be calculated to a final solution by hand.

Fortunately, the fix for this problem is quite simple. Instead of taking the **absolute value** of the differences, we'll just use the elementary concept of multiplying two negatives and **square** the differences:



While it looks much more disasterous on a graph than the absolute differences, squared differences are *significantly* nicer to calculate by hand. Another important note is that the mean needs a *slight* nudge to fix a problem called "bias". A detailed explanation and proof would be a topic for a higher level, mathematical statistics textbook. For the purposes of this text, we simply need to remember that in place of dividing the summation of these squared differences by the total sample size, n, we divide by $n-1$.

What we just did was a rudimentary "proof" (to use the term very loosely) of the measure of center known as **variance**.

Variance: The average squared difference of data from the mean

By definition, variance should **never be negative**. It is bounded between 0 and ∞. **Larger** variance means **more variability**. As variance shrinks to 0 our data set becomes a table of the exact same value.

Given:

$$\text{Sample Mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample variance (denoted s^2) is defined as:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

Again, that $n-1$ component seems sneaky and insignificant but it's **vital** to ensure that our estimate of sample variance is sound.

Similarly:

$$\text{Population mean} = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Population variance (denoted σ^2) is defined as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

We can get away with using population size (N) as is, without subtracting one, due to the nature of "bias" only concerning itself with the difference between our estimate and the truth. We consider any parameter calculated directly from a population to be "the truth".

Remember that statistics is about making inferences about a population parameter using sample statistics

In practice we **almost never** directly calculate population variance, rather we use sample variance to **estimate** population variance

---

Standard Deviation

When we made the swap from **absolute** difference to **squared** differences we gained analytical simplicity in exchange for messy interpretation. The units for our smoking example are "Cigarettes^2", which *should be* confusing. What *shouldn't be* confusing is the solution to this problem:

$$\sqrt{\text{Cigarettes}^2} = \text{Cigarettes}$$

We've successfully derived the last measure of center discussed in this chapter, **standard deviation**.

Standard Deviation: The average difference between the data and the mean.

The formula for standard deviation is extremely direct when we consider variance to be a complete variable (rather than a separate, larger formula):

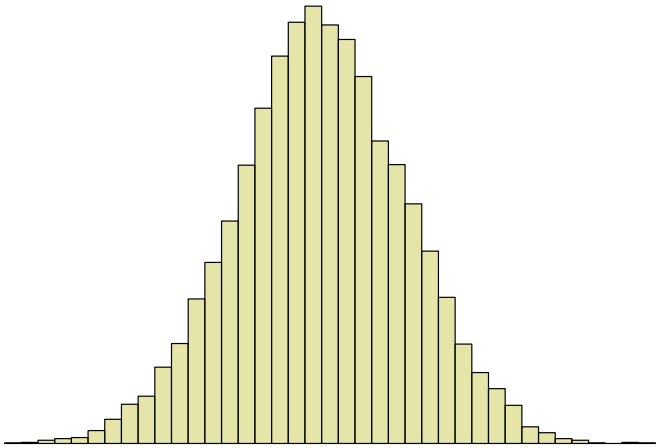
$$\sqrt{\sigma^2} = \sigma \rightarrow \text{Population Standard Deviation}$$

$$\sqrt{s^2} = s \rightarrow \text{Sample Standard Deviation}$$

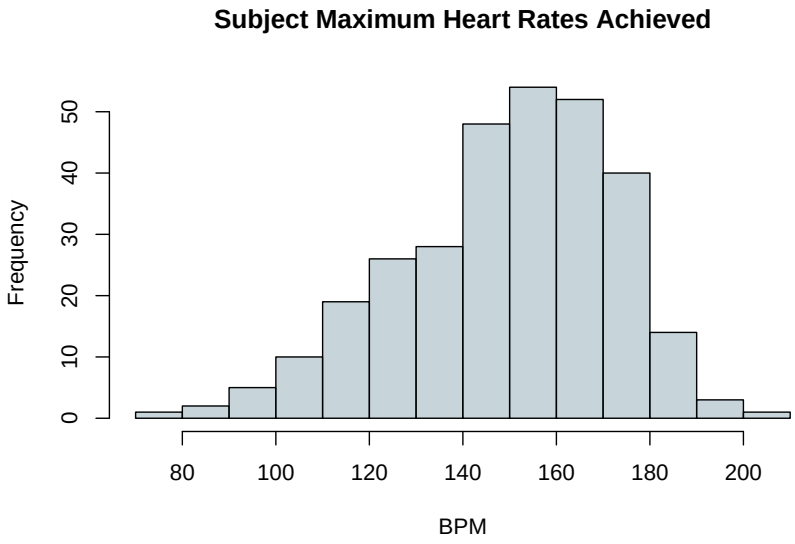
The purpose of standard deviation is primarily to restore the original units of the data that we calculated variance with. However standard deviation is still used throughout the majority of equations in statistical inference, so don't fool yourself into brushing it aside as nothing more than a convenient tool for communication..

Empirical Rule

Some shapes of data are common enough that we attach names and discover generalized trends with them. One of the more (if not most) famous shapes is **bell-shaped**. Approximately symmetric with a modal peak in the center.



The data doesn't need to be *flawlessly bell-shaped* to be described as bell-shaped, for example this data on heart rates could be considered **approximately** bell-shaped:



One of the many reasons statisticians are so keen on bell-shaped data is because of a fun phenomenon known as the **empirical rule** (also known as the 68-95-99.7 rule).

The Empirical Rule

For a population that has an *approximately* **bell-shaped** distribution:

Approximately 68% of the data is within **ONE** standard deviation of the mean

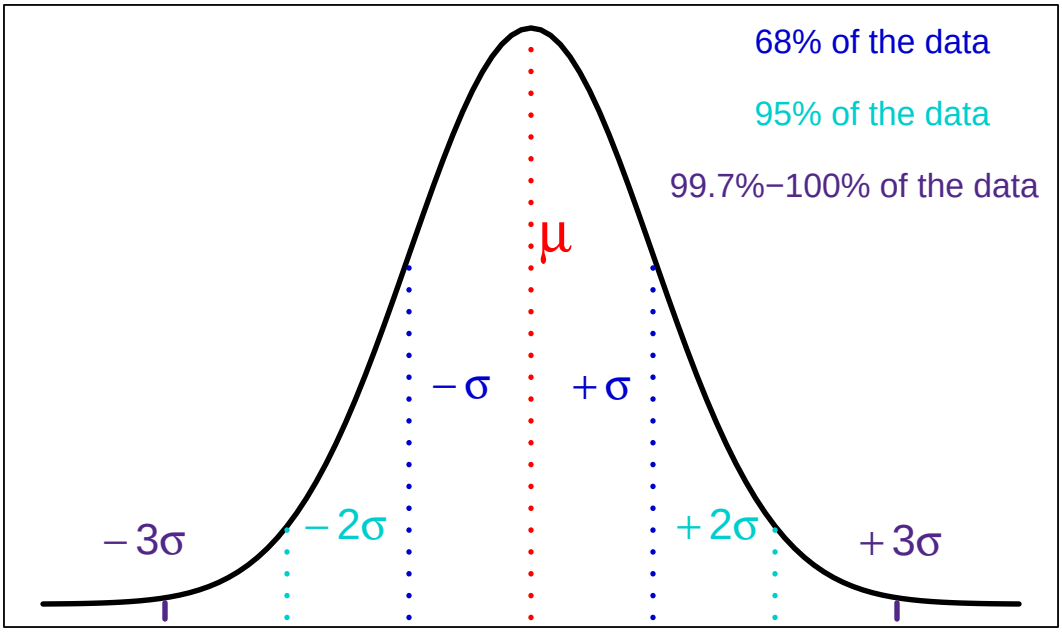
$$\approx 68\% = \begin{cases} \mu - \sigma \\ \mu + \sigma \end{cases}$$

Approximately 95% of the data is within **TWO** standard deviations of the mean

$$\approx 95\% = \begin{cases} \mu - 2\sigma \\ \mu + 2\sigma \end{cases}$$

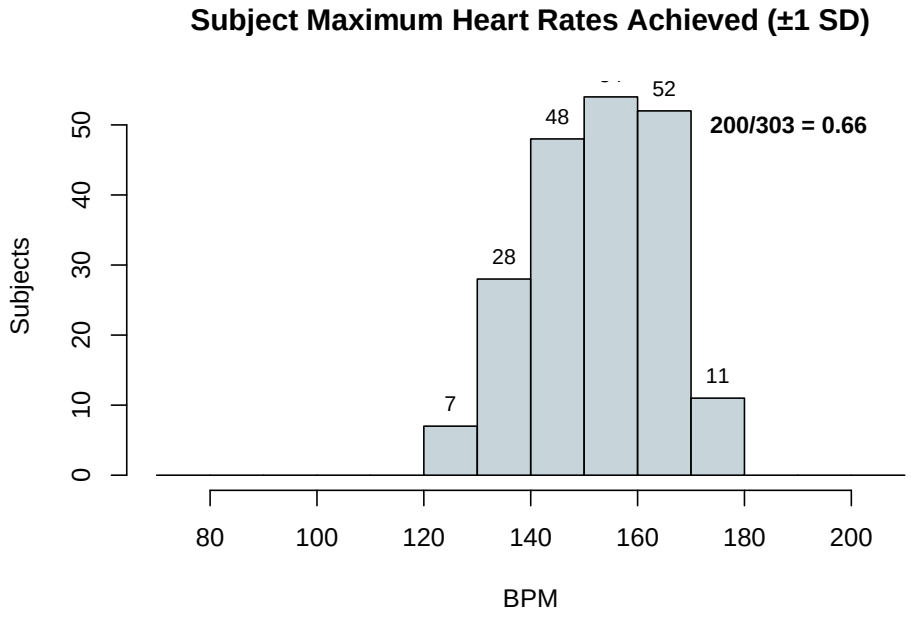
Approximately **All** or *almost all* of the data is within **THREE** standard deviations of the mean

$$\approx 99.7 - 100\% = \begin{cases} \mu - 3\sigma \\ \mu + 3\sigma \end{cases}$$

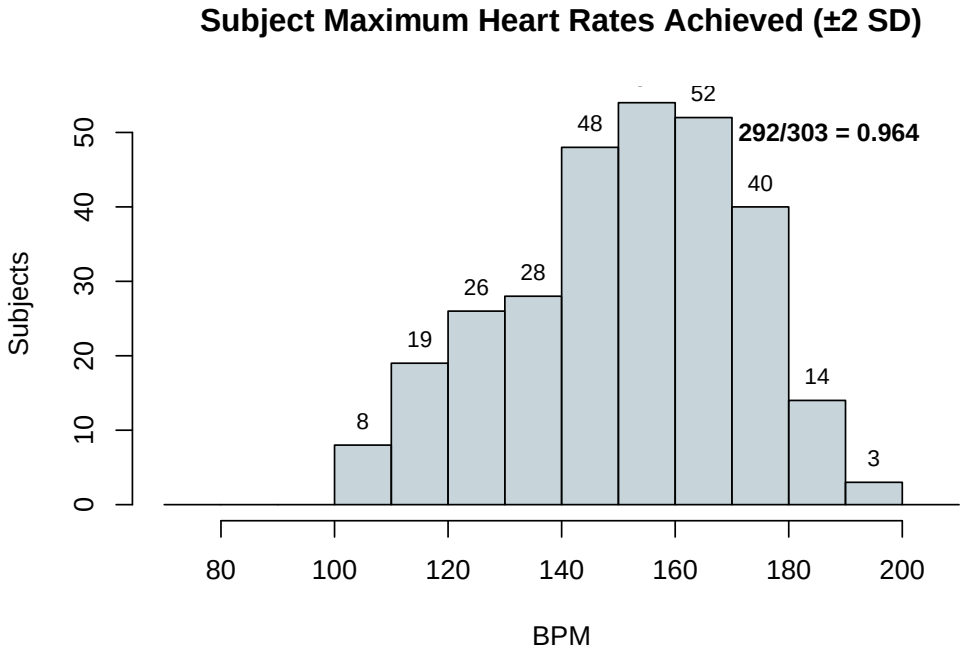


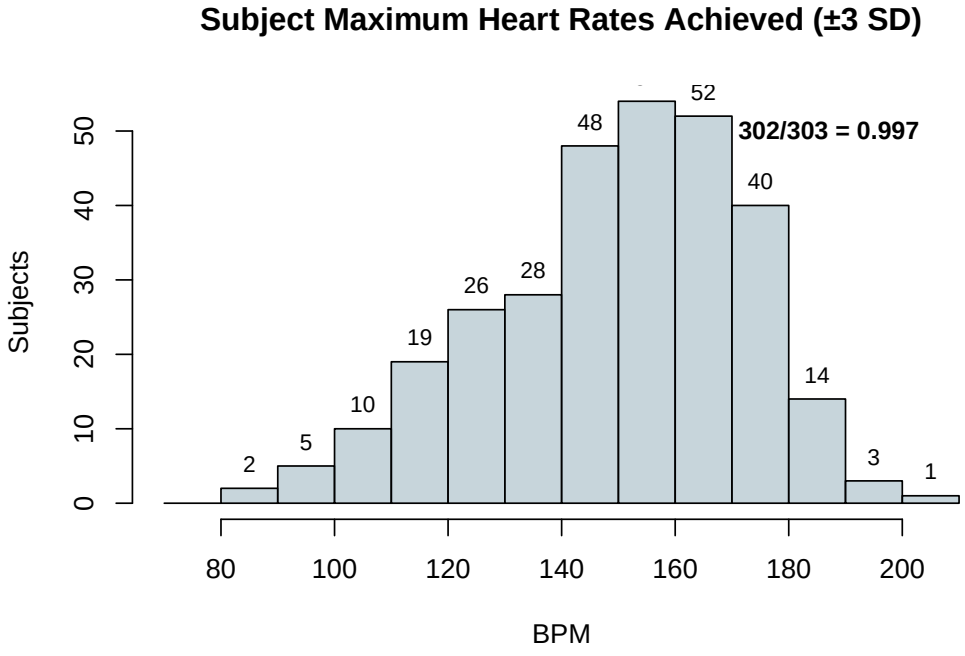
It's convenient to see *how* this rule comes about, let's look back at the heart rate data.

Using some light programming (since $n = 303$ is a bit tedious to work with by hand) we can find the mean and standard deviation of the data to be $\bar{x} = 149.65$ and $s = 22.91$ respectively. If we count up and show the data points that are between 126.74 and 172.56 ($\bar{x} \pm s$):



Since the data isn't a perfect bell-shape we'll have some error between our calculated proportion and the empirical rule, but if we go ahead and check 2 and 3 standard deviations:





We can see the empirical rule is a very good *approximation* of reality. It's also easy to see how quickly the rule falls apart when the data doesn't meet that bell-shaped standard: