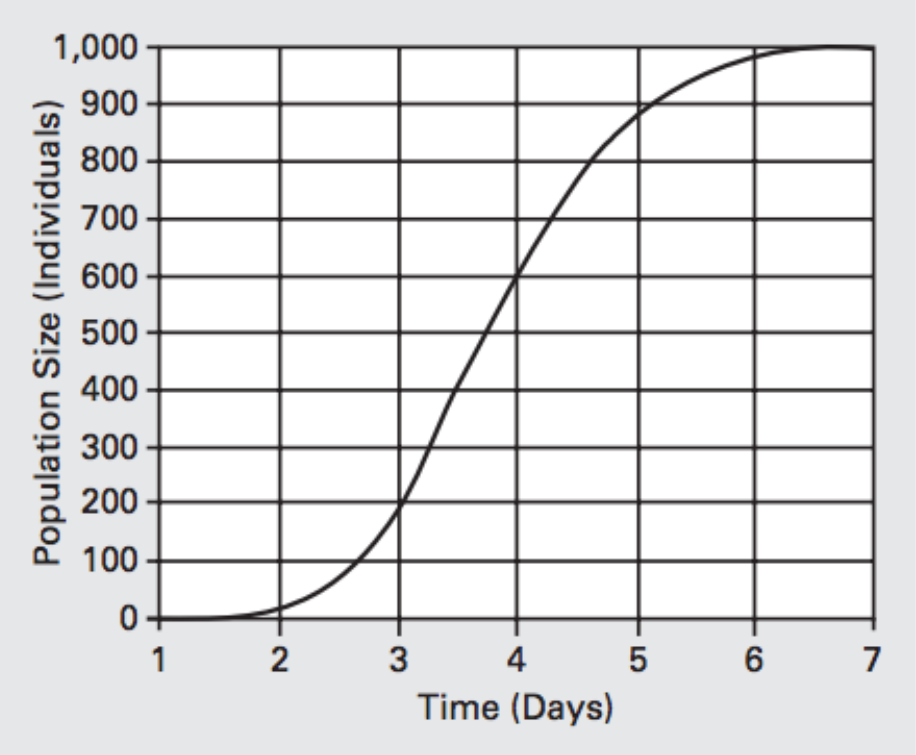
Review of Stats for Bio

CALCULATING THE SLOPE

Use the equation: $M = (y_2 - y_1) / (x_2 - x_1)$



Use the graph above to calculate the mean rate of population growth (individuals per day) between day 3 and day 5. Give your answer to the nearest whole number.

Answer: ~ 340-360 individuals per day (**UNITS MATTER**)

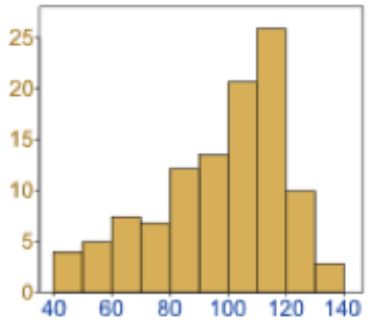
Answer Calculation: M = [875-175] / [5-3] = 350

When Data is Normally Distributed (parametric data) – you can calculate mean, SD, SE.

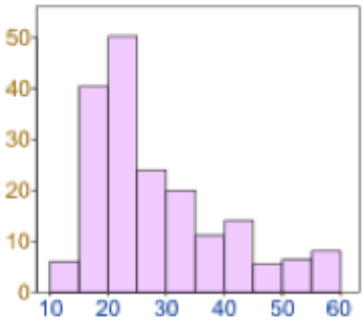
But… what does it mean when data is Normally Distributed?

Data can be "distributed" (spread out) in different ways.

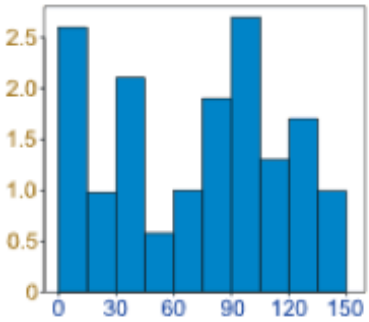
It can be spread out more on the left



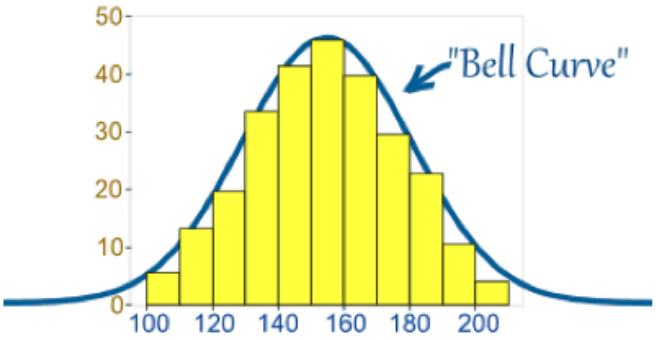
Or more on the right



Or it can be all jumbled up



But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



A Normal Distribution

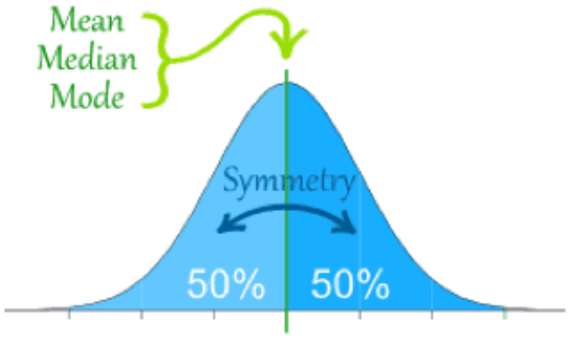
The "Bell Curve" is a Normal Distribution.
And the yellow histogram shows some data that follows it closely, but not perfectly (which is usual).

It is often called a "Bell Curve" because it looks like a bell.

Many things closely follow a Normal Distribution:

- heights of people
- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test

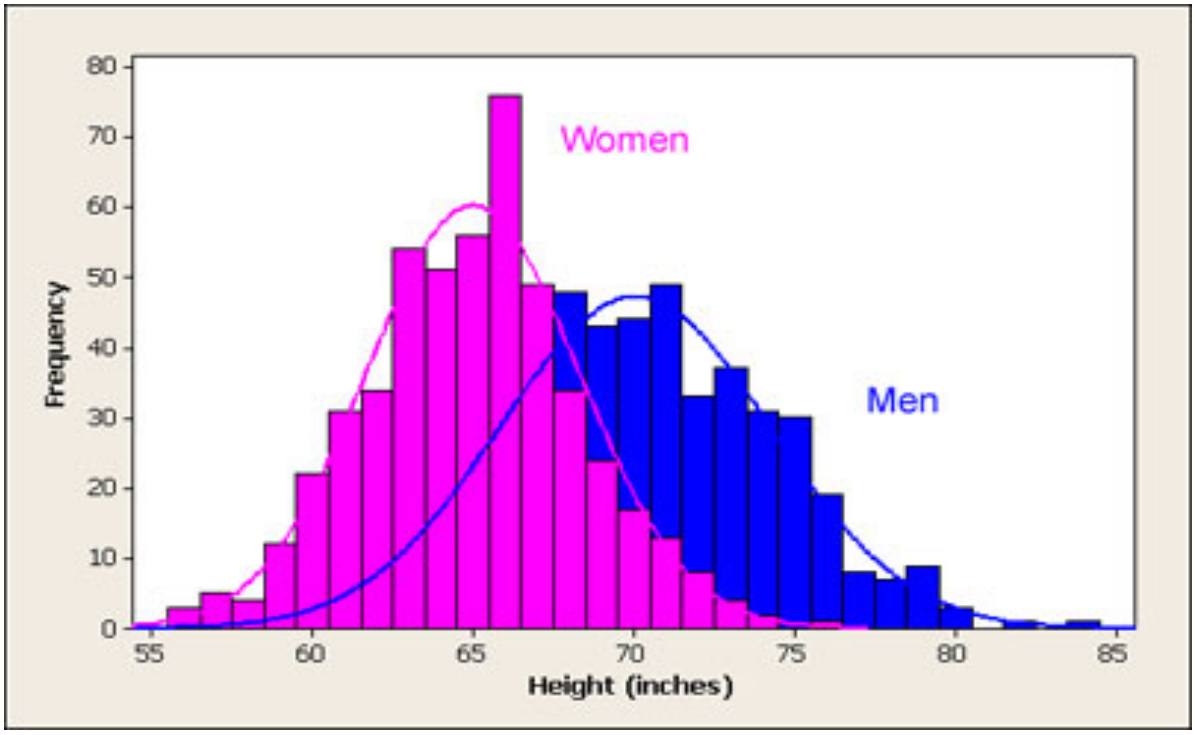
We say the data is "normally distributed":

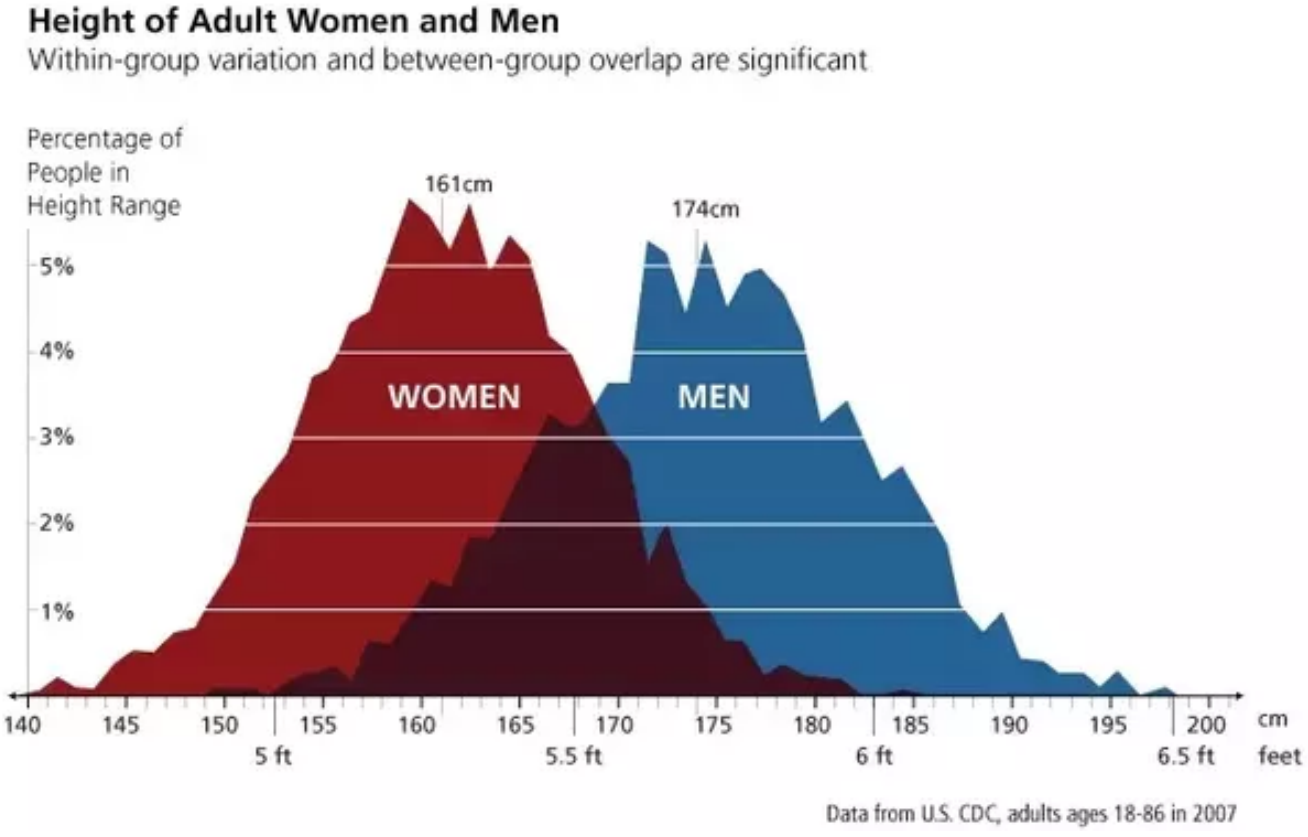


The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

Below are two examples of normally distributed curves for the height of women and men.





TRUE POPULATION VS SAMPLE POPULATIONS

Calculations like mean, standard deviation, standard error can describe the entire or true **population (N)** that you are studying, but collecting the data from **EVERY SINGLE SUBJECT** in a population to compute these statistics is often ***not*** possible. Therefore, we usually collect data for a subset that hopefully represents accurately the true population. This **sample population (n)** is used to calculate sample means, sample standard deviation of the mean, and sample standard error of the mean.

Sample size *[the # of individuals included in a subset]* is important when students try to estimate how confident they can be that the sample set they are trying to analyze accurately represents the entire population. **The larger the sample size the better!!!**

- *We want usually, at minimum, **30 subjects** in a sample to feel more confident that the sample is representative of the true population the sample is a part of.*

MEAN [of a Sample Population]

The **mean of the sample [Sample Mean], or X bar, is the sample average** (the sum of the numbers in the sample divided by the total number in the sample).

Mean

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

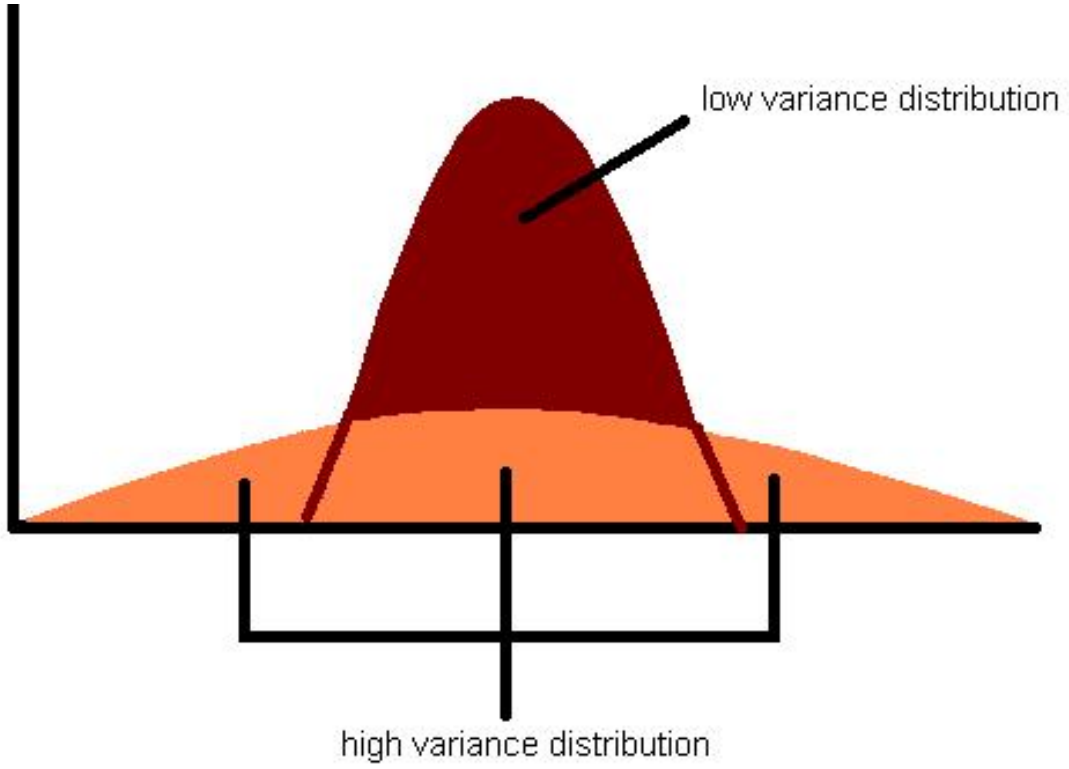
The mean summarizes the entire sample and hopefully provides an estimate of what the **entire population's true mean [μ]** is since, remember, it may be impossible to actually calculate the mean of the entire population your sample is derived from.

STANDARD DEVIATION (SD)

The **sample standard deviation (SD)** measures **how spread the data is in the sample** population, which thus provides an **estimate of the variation in the sample set**.

A large sample standard deviation indicates that the data has a lot of variability.

A small sample standard deviation indicates that the data is clustered close to the sample mean – *that all subject values are not too different from the mean of the sample population.*

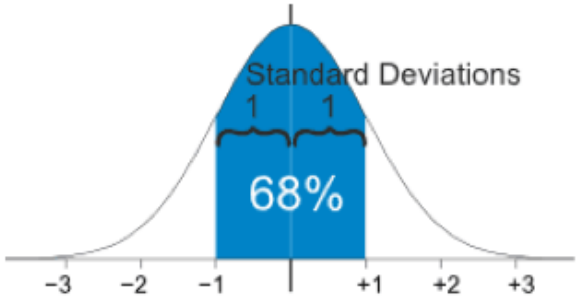


Standard Deviation*

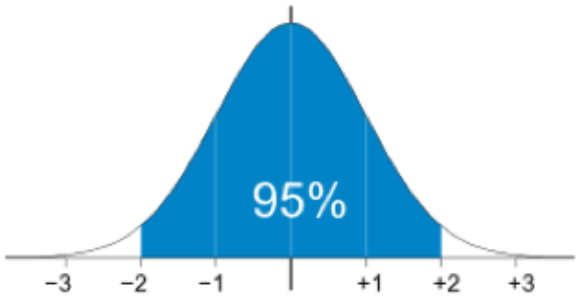
$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

The number of **standard deviations from the mean** is also called the "Standard Score", "sigma" or "z-score". Get used to those words!

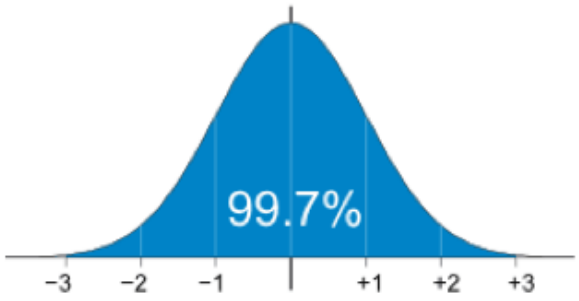
When we calculate the standard deviation we find that (generally):



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean

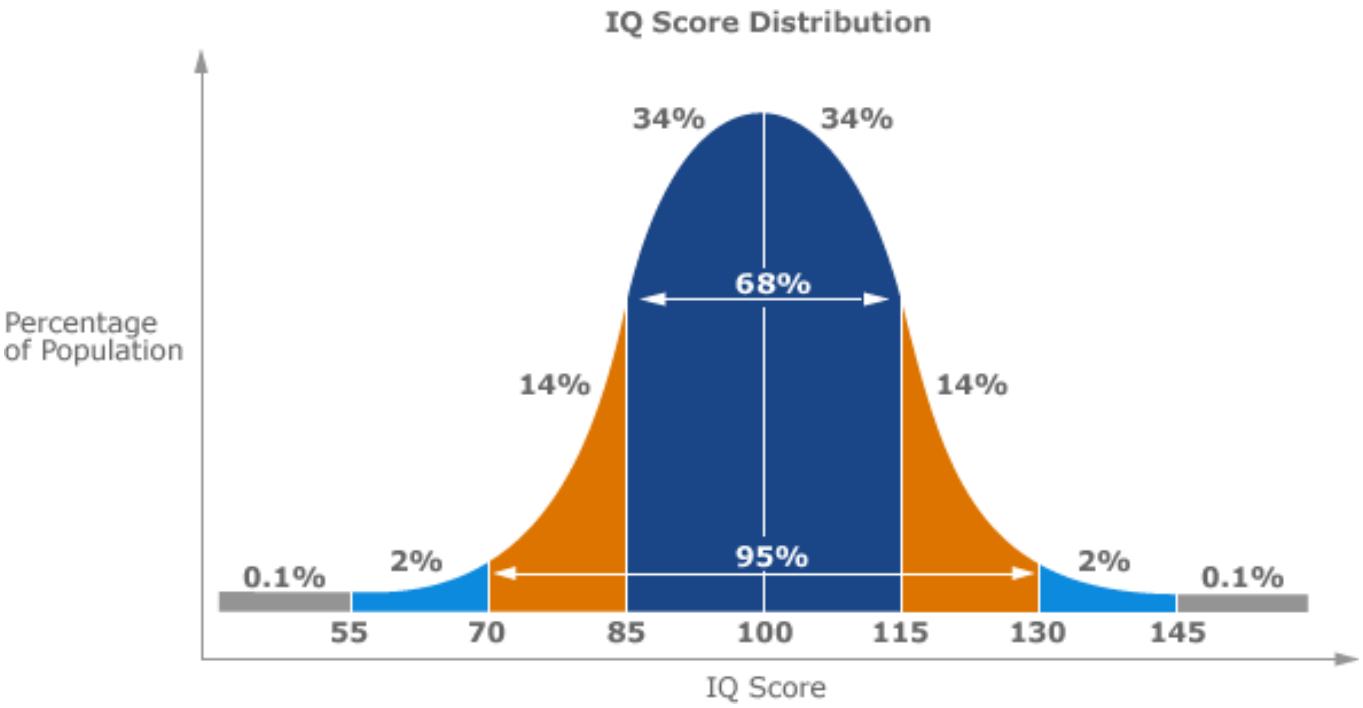


99.7% of values are within
3 standard deviations of the mean

As you can see from the graphs above, **a little more than two-thirds of the data points [68%], when the data is NORNALLY DISTRIBUTED, will fall between +1 standard deviation and -1 standard deviation from the sample mean.** *[34% of the data is expected to fall below the mean and 34% is expected to fall above the mean].*

More than 95% of the data falls between ±2 standard deviations from the sample mean. [Again, 47.5% of the data points expected to fall below the mean and 47.5% expected above the mean].

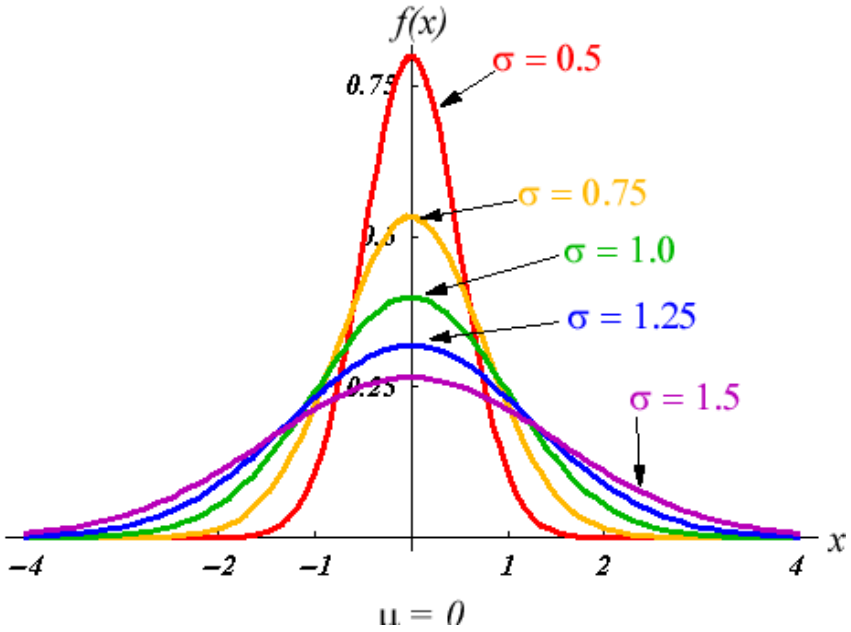
Example:



Sample standard deviation can be referred to by SSD, S, or σ.

You can be asked how many standard deviations a data point is away form the mean…1SD, 2SD, 3SD? This is known as the **z-score**. Your data can have a z-score of + or – 1, + or – 2, and + or – 3 depending on if it falls 1 SD above the mean or 1 SD below, 2 SD's above the mean or 2 SD's below the mean etc…

The larger the standard deviation, the more spread your data is around the mean.



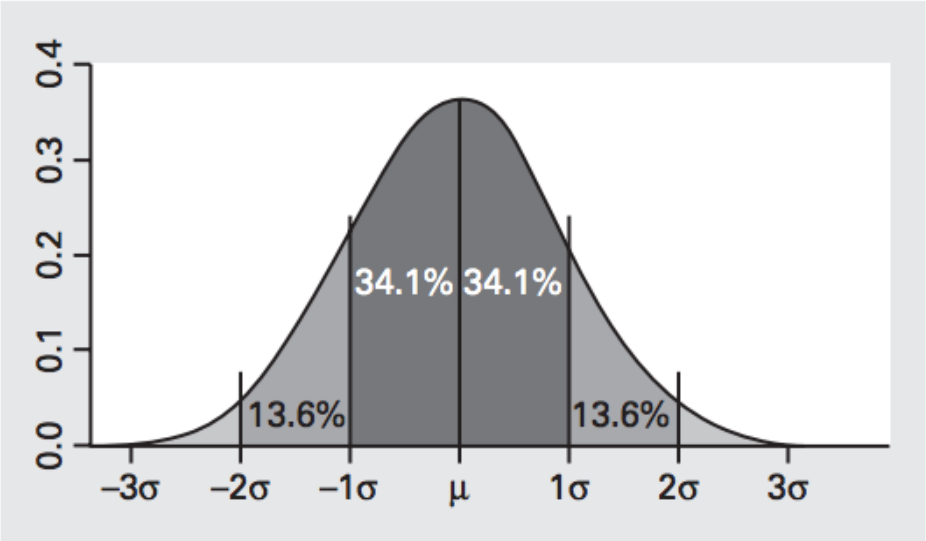


Figure 6. Standard Deviations and a Normal Distribution

A little more than two-thirds of the data points will fall between +1 standard deviation and −1 standard deviation from the sample mean. More than 95% of the data falls between ±2 standard deviations from the sample mean.

Standard deviations [SD] for a population [N] or, most often, SD a sample population [n] can be displayed visually on a graph. *You will be asked to interpret &/or draw such graphs!*

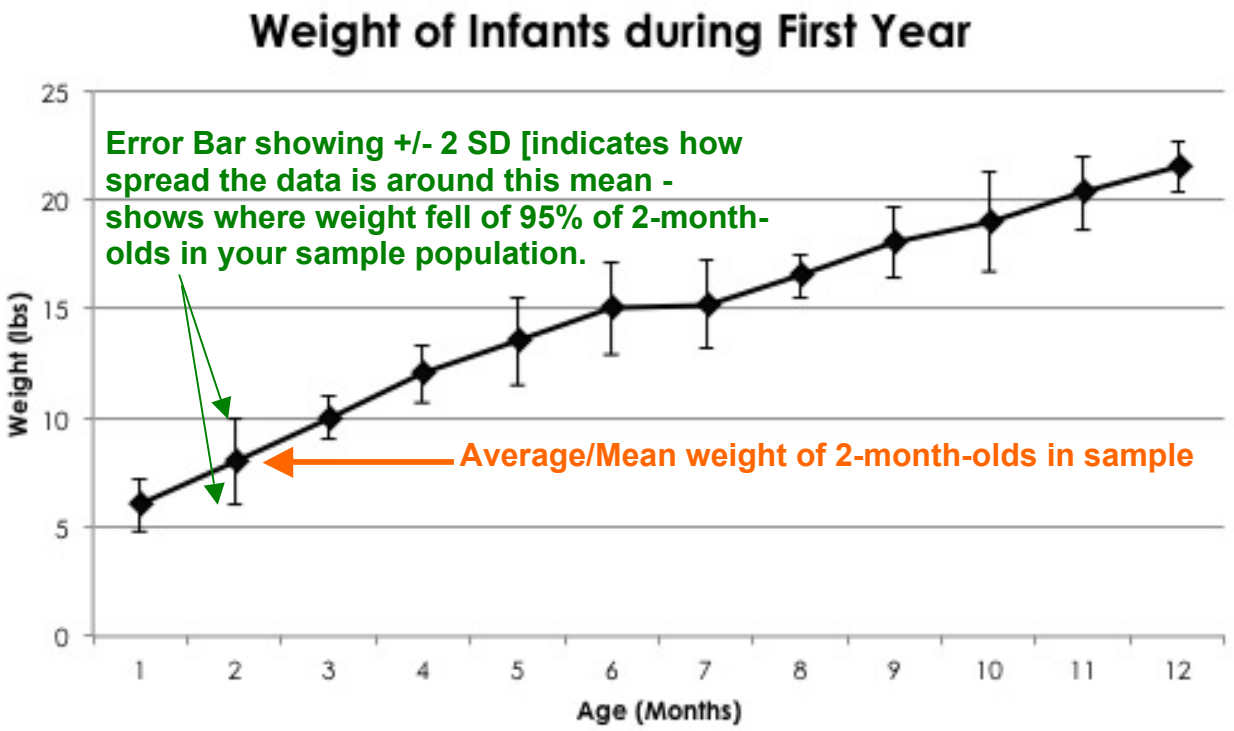
How to Add Standard Deviation Bars on Graphs?

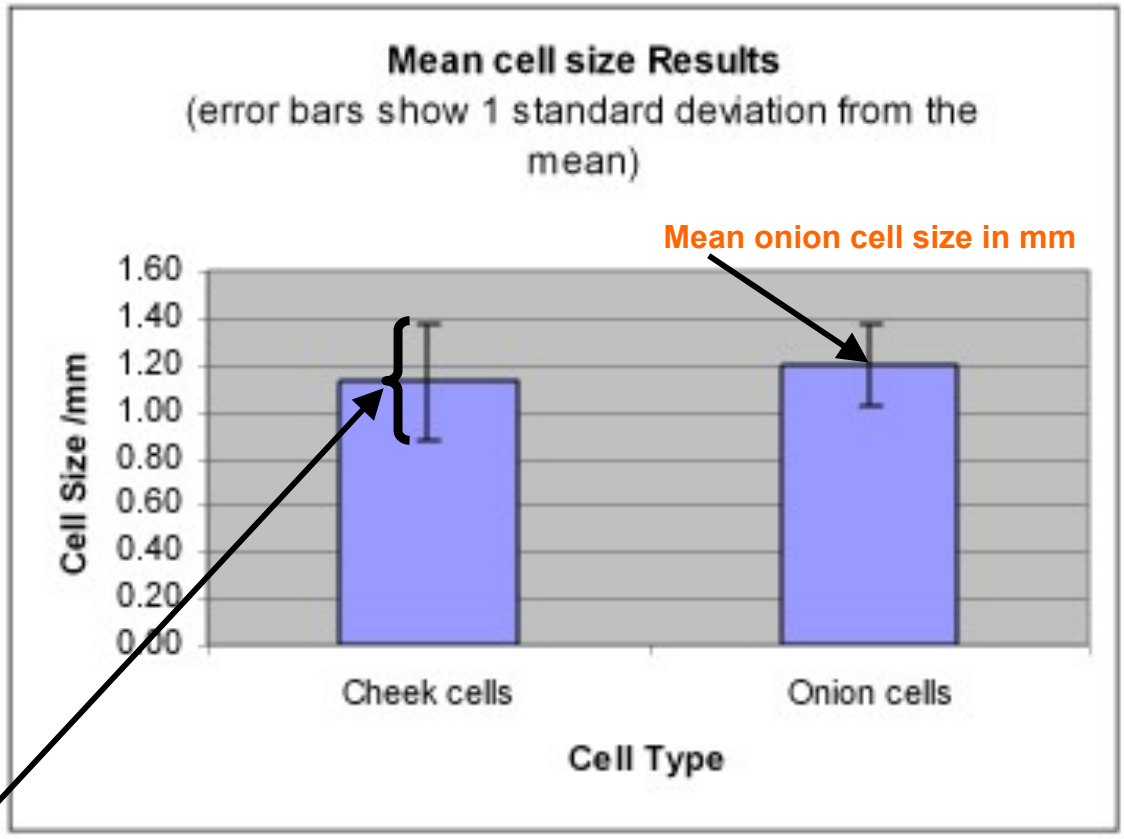
1. Find out the SD and the sample Mean.

2. You can then find out where 68% of the data falls by calculation **Mean ± 1SD** or where 95% of the data falls by calculating **Mean ± 2SD**.

3. Draw a normal graph, plotting your data points of your sample population as you are accustomed to doing.

4. **Add error bars to show the spread within your sample population around the means.** [The length you draw the error bars around your mean depend on if you are asked to make error bars that show +/- 1 SD or that show +/- 2 SD or +/- 3 SD]





Error Bars showing +/- 1 SD [highlights what cheek cell size 68% of check cells size in your sample population have]

SAMPLE STANDARD ERRORS (SE or SE of the Mean)

Remember, we would love to collect data on every single subject in an entire population **[N]**. This is often not physically possible. We often collect data from population samples **[n]** – subsets of the true population.

- Could you really measure the length of every single fire ant in Florida, one by one, in order to calculate the true Florida fire ant population mean length? No.

- You may have to pick and survey a patch of land in Florida, measure the length of all fire ants in this sample population, and calculate the sample mean length of fire ants in this sample patch in order to get an estimate of what the average fire ant length is in all of Florida.

- **Ideally, the average length of the fire ants in your sample population [n] matches exactly the average length of all fire ants in the entire state of Florida [N], which then allows you to use your data from your sample to make an accurate conclusion about the average length of fire ants in Florida, even if you didn't get to measure every single fire ant in all of Florida.**

- It is important then to ask "how well does the mean length of ants in your sample patch actually match the true mean of the entire Florida fire ant population?" "Is your sample mean an accurate representation of the mean of the entire Florida fire ant population?" ***In reality, the sample mean may or may not always match perfectly to the true population mean.***

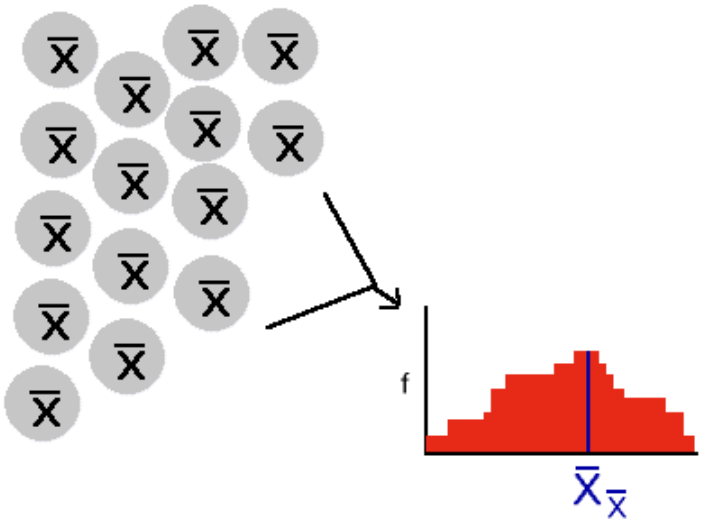
- For this reason, **we do not rely on data from just one sample population either when estimating the mean of the true population.** A researcher would collect data on fire ant length from **MANY [at least 30 sample]** patches of land spread throughout Florida, calculating the sample ant length means in *each* of these sample patches.

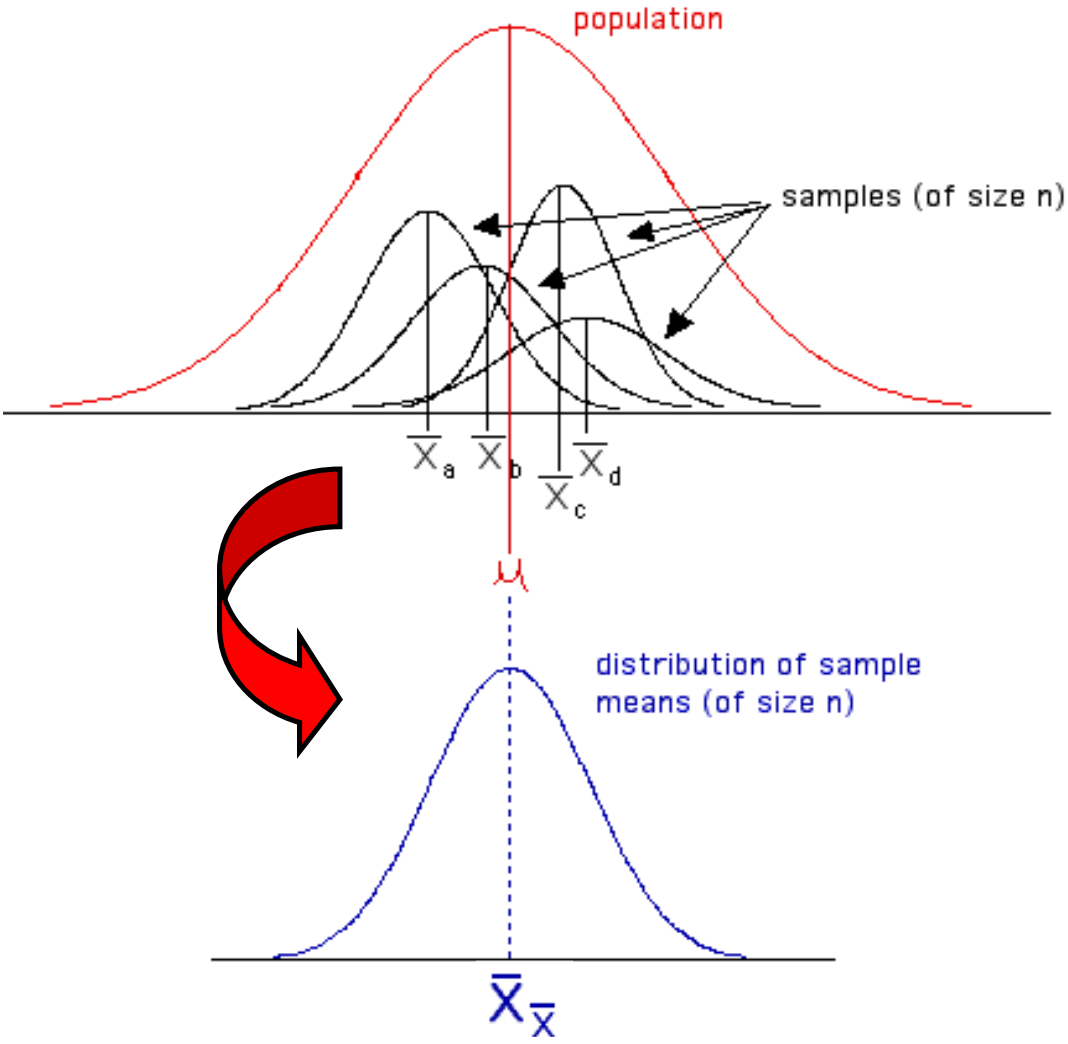
- **The true population mean is expected to fall somewhere among or around the means of all the different sample populations** [which themselves are expected to vary a little from each other], but how confident is the researcher in concluding that the true population mean value would fall among the sample mean values?

Sample standard error (SE) is a statistic that allows a person to make an inference about how well the sample mean matches up to the true population mean.

> *Recall, Standard Deviation is a measure of the amount of **variance of the data around the sample mean**.*
>
> *Standard Error is a measure of **how close the sample mean matches up to the true population (N) mean**.*

If one were to take a very large number of samples [multiple **n**'s] (**at least 30 different samples**) from a given population (**N**), the **means** for **each** sample can be plotted on a graph and they would form an approximately normal distribution too—*a normal distribution this time of sample means.*





The distribution of sample means, then, is a theoretical construct that helps us define our boundaries of confidence in our samples – how confident are we that the sample means are indicative of my true population mean or, stated differently, ***how confident are we that my true population mean [μ] falls within my sample mean calculations?***

Again, **Sample standard error (SE) or standard error of the mean (SEM)** is a statistic that allows a person to make an inference about how well the sample mean matches up to the true population mean.

> *Recall, Standard Deviation is a measure of the amount of* ***variance of the subject data around the sample's own mean*** *[= a comparison of individual data to the sample group's mean]*
>
> *Whereas, Standard Error [of the Mean] is a measure of* ***how close the sample mean matches up to the true population (N) mean***

Standard error is calculated from the following formula:

Standard Error of the Mean*

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

S = the sample standard deviation [SD]
n = the sample size.

A sample mean of ±1 SE describes the range of mean values about which an investigator can have approximately 68% confidence that the range includes the true population mean.

Even better, a sample with a ±2 SE defines a range of mean values with approximately a 95% certainty that the true population mean would be included within that range. In other words, if the sampling, for example, were repeated 20 times with the same sample [n] size each time, the confidence limits, defined by ±2 SE, would include the true population mean approximately 19 times on average. Meaning, ***you are confident that 95% of the time the true population mean falls within the range of sample mean values that are 2 Standard Errors [of the Mean] above and below the mean of all sample means.***

This is the inference: it is a statistic that allows investigators to gauge just how good their estimate of the true population mean actually is form their sample mean data.

Standard Error [of the Mean] Graphs

Ex: Let's say you are researching the difference in width of English ivy leaves based on the amount of sunlight they are exposed to. You ask…do shady English ivy leaves have a larger maximum diameter than sunny English ivy leaves?

It would not be practical to collect all of the leaves growing in an area to measure their width at the widest point of each leaf. Instead, as in most biological investigations, it is advisable to choose smaller samples to inform an investigation. Samples chosen should be as random and as unbiased as possible.

The student collected and measured the maximum width, in centimeters, of 30 leaves from each of two habitats: Shady Areas and Sunny Areas. The data is shown in the table below.

Table 2. A Student's Leaf Measurement Data

Shady Leaves (in cm)	Sunny Leaves (in cm)
3.7	3.2
5.2	3.5
5.4	4.1
5.7	4.3
5.8	4.4
5.8	4.6
6.0	5.0
6.1	5.0
6.5	5.2
6.5	5.2
6.6	5.3
6.8	5.4
7.0	5.6
7.3	5.7
7.3	5.7
7.4	5.8
7.7	6.0
7.9	6.0
8.0	6.4
8.1	6.5
8.1	6.7
8.2	6.7
8.3	7.1
8.9	7.1
9.0	7.1
9.4	7.3
9.9	7.5
9.9	7.9
9.9	8.0
10.4	8.2

We can describe the nature of data statistically. The number of leaves in each sample is 30 **[n = 30]**. We can calculate the mean width of the shady leaves in the one sample and the mean width of the sunny leaves in the other sample. The mean for shady leaves is 7.43 cm.

We can calculate how much variation in width exists compared to the average shady leaf width. Are there a lot of different widths of shady leaves or are they all very similar in width. Standard Deviation tells us what fraction of shady leaf widths fall within a certain distance from the mean shady leaf width. We can calculate SD for the sunny leaves as well.

Standard Deviation For Shady Leaves:

After using the SD formula, the Standard Deviation for shady leaf data is 1.63 cm. So… +/- 1 SD = 7.43 cm +/- 1.63 cm

Conclusion: Because the data is normal [normally distributed], statistically, **68% of the widths of shady leaves are between 5.8 cm and 9.06 cm.**

The nature of the sample's distribution can be described by combining the sample mean with the sample standard deviation. Can you tell me, with 68% confidence, where the true population mean of all shady leaves falls based on the sample shady leaf width data you collected?

Standard Error For Shady Leaves:

How does the mean of the sample of shady leaves compare to the true population mean of shady leaves? **There is around 68% probability that the true population mean lies within the boundaries of the sample mean ±1 sample standard error.**

After using the SEM formula, the Standard Error of the Mean for shady leaf data is 0.30cm. So…. +/- 1 SE = 7.42cm +/- 0.30 cm [for a 68% confidence].

Conclusion: Because the data is normal [normally distributed], statistically, there is a **68% probability [we are 68% confident] that the true population's shady leaf width mean falls between 7.72 cm and 7.012 cm** (a calculated CONFIDENCE INTERVAL).

The researcher can calculate the same statistical inference data for the sunny leaves.

Table 3. Descriptive Statistics

	Shady Leaves	Sunny Leaves
Mean	7.43	5.88
Standard Deviation	1.63	1.32
n	30	30
Standard Error	0.30	0.24

The researcher then produces the bar chart shown in Figure 8 (below) to compare the means visually. *Notice that **she includes error bars of ±1 SEM**, which you should be able to both interpret with accurate wording & draw too!!!*

> Remember, the error bars on the first column in this graph tell us that we are 68% confident that the shady leaf width of the true population of shady leaves falls within this range of shady leaf width sample means. Similarly, the error bars on the second column in this graph tell us that we are 68% confident that the sunny leaf width of the true population of sunny leaves falls within this range of sunny leaf width sample means.

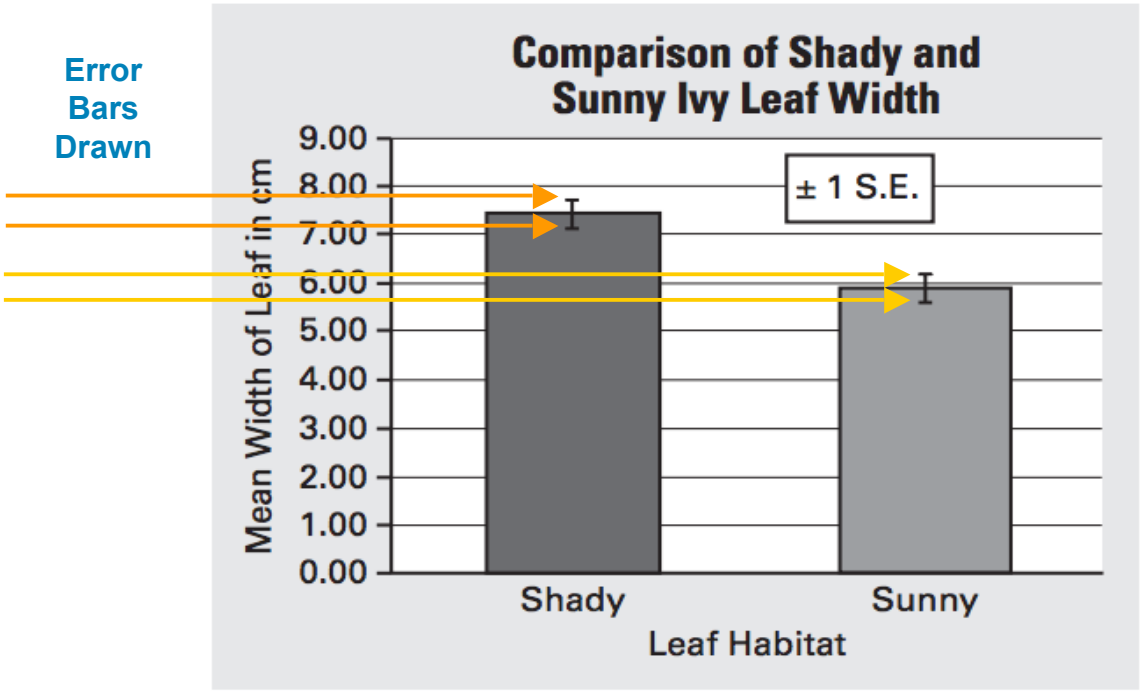


Figure 8. Bar Chart with Standard Error Bars Comparing the Means for Shady and Sunny Ivy Leaf Width

Can the researcher conclude that the leaves growing in shade have on average a larger width compared to those growing in the sun? You may feel like concluding "yes" because the columns reach different heights. **But we do not reach these conclusions without statistical input and analyzing the SEMs.**

What if she happened to sample shady leaves that are larger in width **by chance**, missing other shady leaves that actually are smaller in widths resulting in her mean sample width of shady leaves not being an exact representation of your true population shady leaf mean? And what if, when she selected sunny leaves to measure, **just by chance**, she accidentally selected a sample where more were shorter and her mean doesn't actually represent the true population width of all sunny leaves? ***How confident is she that the 2 POPULATIONS' MEAN LEAF WIDTH really differ? If the true mean of the shady leaves is smaller than her sample average and her true mean of the sunny leaves is larger than your sample average, then perhaps the two populations have actually similar mean leaf widths!!!***

It is very important to note thus whether the error bars of the shady vs sunny leaf lengths **overlap**.

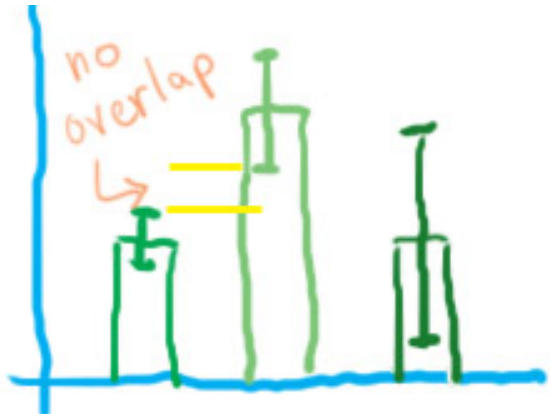
It is evident from the graph that the error bars for the shady leaf mean **do NOT overlap** with the error bars for the sunny leaf mean. In fact, had the student chosen to plot ± 2 SE error bars, they, too, would not have overlapped.

This non-overlap strongly suggests that there is a high probability [68% confidence] that the two populations [the two population leaf width means] are indeed statistically different from each other and that the variation in width is not due to chance alone.

So…

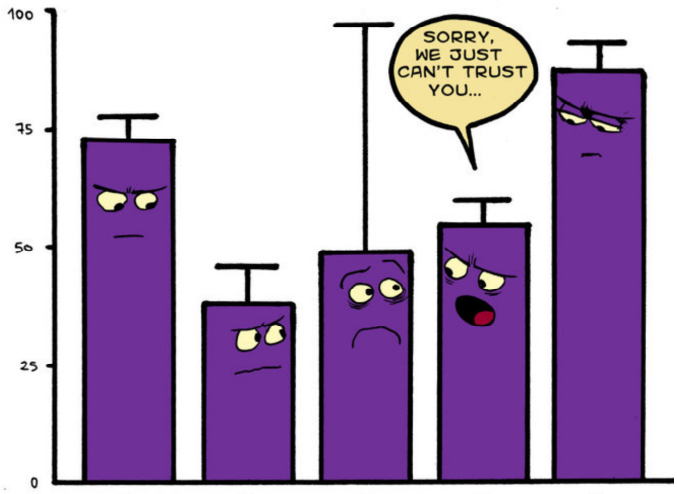
When Error Bars **DO NOT OVERLAP**, we can confidently state [at a 68% or 95% confidence level depending on if we calculate +/- 1 or +/- 2 SEM] that the two population means [based on the sample means] **are** statistically different.

- **The real population means DO differ statistically.**



When Error Bars **DO OVERLAP**, we can confidently state [at a 68% or 95% confidence level depending on if we calculate +/- 1 or +/- 2 SEM] that the two population means [based on the sample means] **are not** statistically different.

- **The real population means DO NOT differ statistically [even if the sample means are different]**



Just like we can graph Standard Error bars on a graph around a mean data point, we can also graph Standard Deviation bars on a graph around a data point.

Below see an example graphs showing either SD or SE bars. *Would you be able to describe with accurate wording & conceptual understanding what these two graphs and their error bars show?* ***Give it a try without looking at the answers below the graph.***

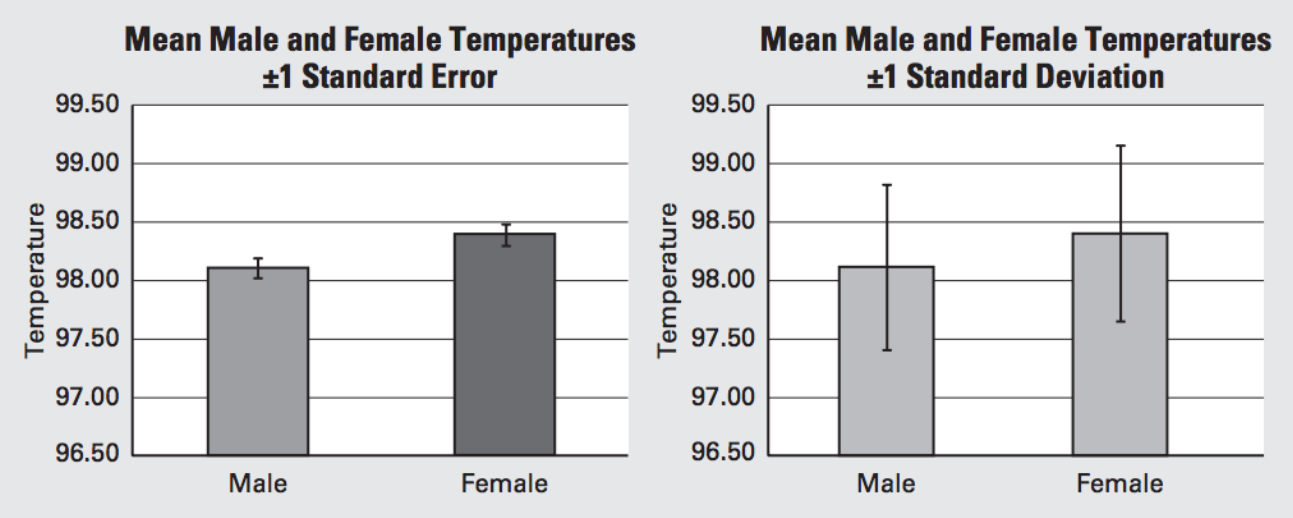


Figure 14. A Comparison of Standard Error and Standard Deviation

.
.
.
.
.
.
.
.
.

The 1st graph's error bars compare sample means to true population means [Standard Error of the Mean].

- We are 68% confident that the mean male body temperature in the whole population of males falls within the error bar range calculated & drawn [on the column showing the mean male body temperature of our sample of males measured].

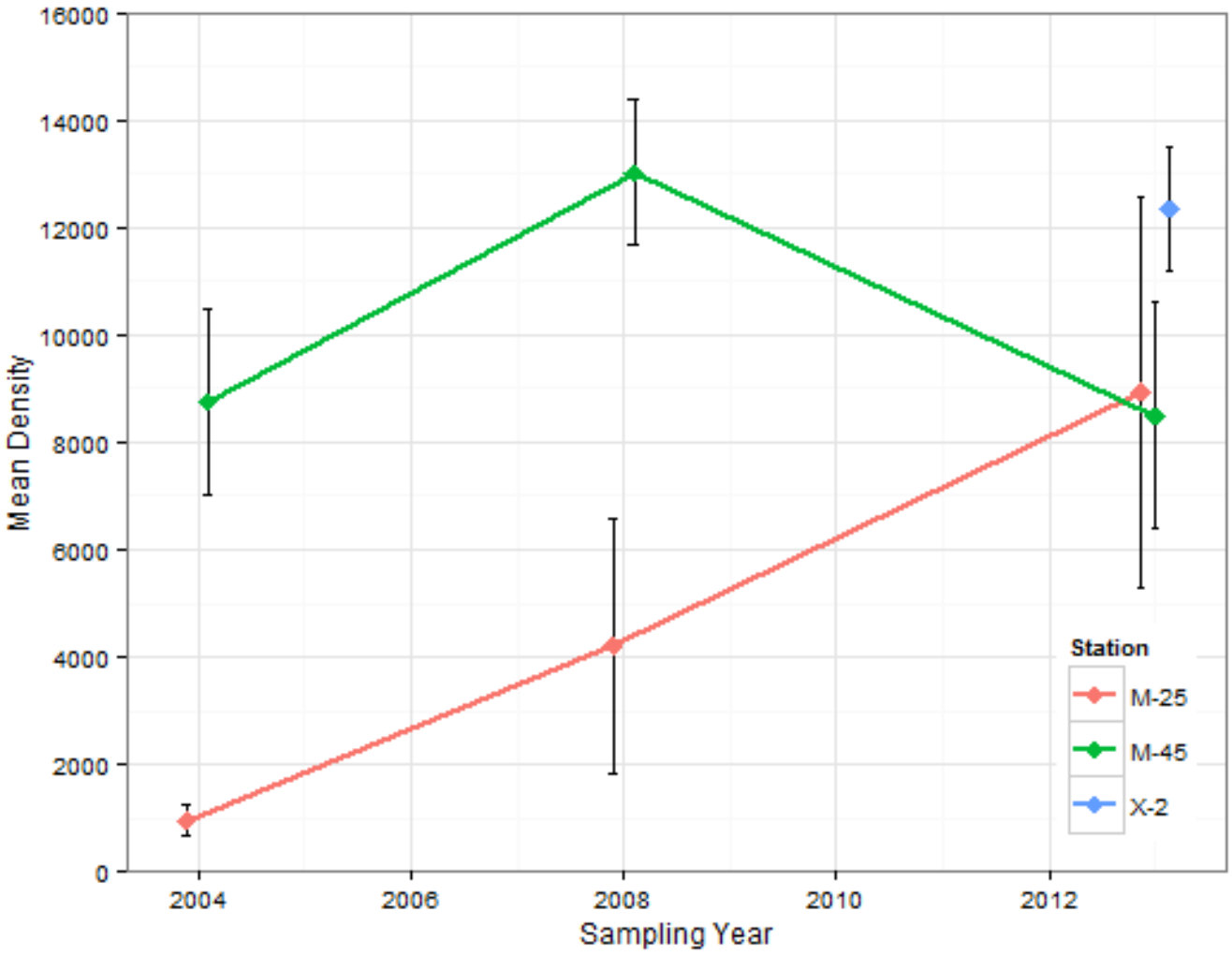
The 2nd graph's error bars indicate how much variation there exists among the male body temperatures in the sample of males tested [Standard Deviation].

- 68% of sampled males have a body temperature that falls within the error bar range calculated & drawn [on the column showing the mean male body temperature of the sample of males measured].

Let's look at a line graph with error bars.

Scenario #1: Researchers collected data on the number of lady bugs in five 1 km2 regions near each of 3 different field stations in a tropical rain forest over several years. **If the bars are Standard Deviation Error Bars, what does the graph tell you when comparing the data collected from the three field stations?**

Scenario #2: Researchers collected data on the a 1 km2 surrounding each of 3 different field stations in a tropical rain forest over several years. **If the bars are Standard Error of the Means Bars, what does the graph tell you when comparing the data collected from the three field stations?**



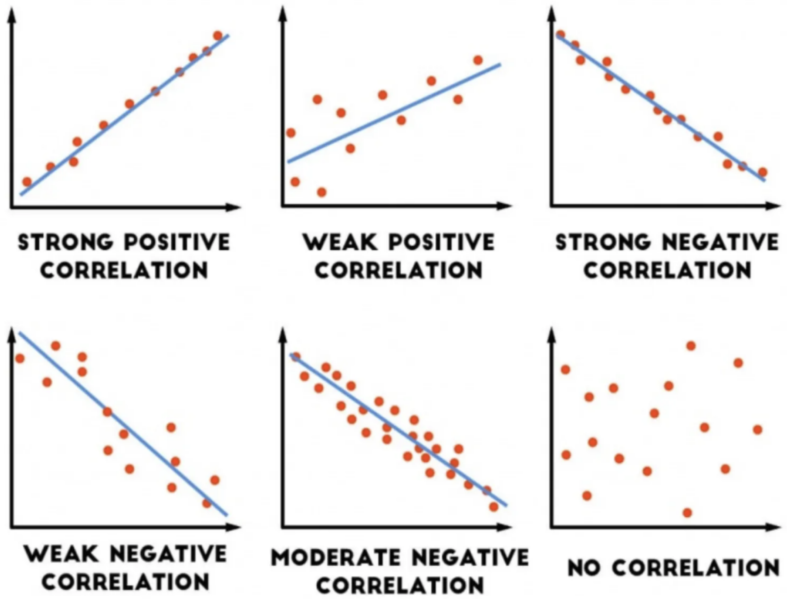
SCATTER PLOTS AND REGRESSION LINES

A scatter plot (a.k.a. scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point.

Scatter plots' are used to identify and/or show if a relationship exists between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Scatter plots can be used to graph or identify a correlational relationship between two variables. These graphs do **NOT** automatically indicate a cause-and-effect relationship exists between two variables. Meaning, they do **NOT** state that a change in the value of one variable is the **cause** of any change in value of another variable. An experiment would have to be run to **test** if one variable actually is the cause of a change in the other variable as opposed to two variables both changing in a certain, predictable way due to another cause. If evidence supports that there actually is a cause and effect relationship between two variables, instead of merely correlational relationship between the two, then the one variable will be called the independent variable and the other the dependent variable. *(Ex: When ice cream sales increase, the rate of homicides also increases, a positive **correlation** between these two variables, but not, as it turns out, a cause-and-effect relationship. People eat more ice cream in the summer when it is warm out. People also gather and party more when it is warmer out, leading to more conflicts and confrontations, that sometimes lead to murder. Eating ice cream does not cause people to kill.)*

Scatter plots allow you to predict, given a particular horizontal value, what the vertical value may be. Relationships between variables can be described in many ways: **positive or negative, strong or weak, linear or nonlinear,** depending on the trend seen in the data and the distance of the data points from the regression line or line of best fit.



When a scatter plot is used to predict or identify a correlational relationship between variables, it is common to add a trend line to the plot. The trend line is a best fit line or line of best fit for the data points—it summarizes all the points in a single linear equation. Line of best fit is a line drawn through a scatter plot of data points that best represent their distribution by **minimizing the distances between the line and these points**.

Regression lines results from statistical calculations known as regression analysis and serves to illustrate the relationship among the data. **Regression Lines** highlight **trends** in data.

Ex: Is heart rate in humans related to body temperature? Does one increase or decrease as the other increases or decreases?

To answer this, a scatter plot should be made first with the"DEPENDENT" variable on the Y-AXIS and the "INDEPENDENT" variable on the X-AXIS.

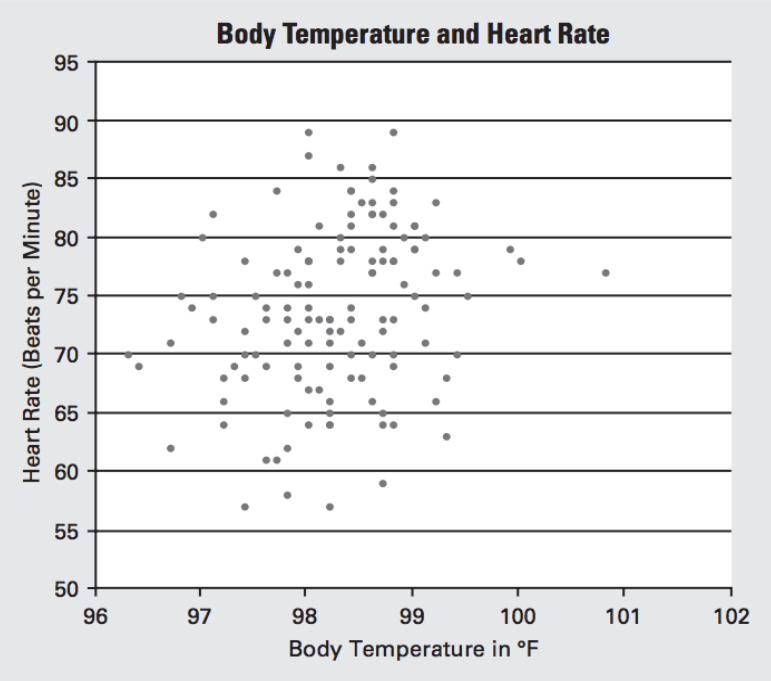


Figure 16. Scatterplot of Heart Rate and Body Temperature

But, is there a general trend? If so, what sort of relationship might exist between these variables?

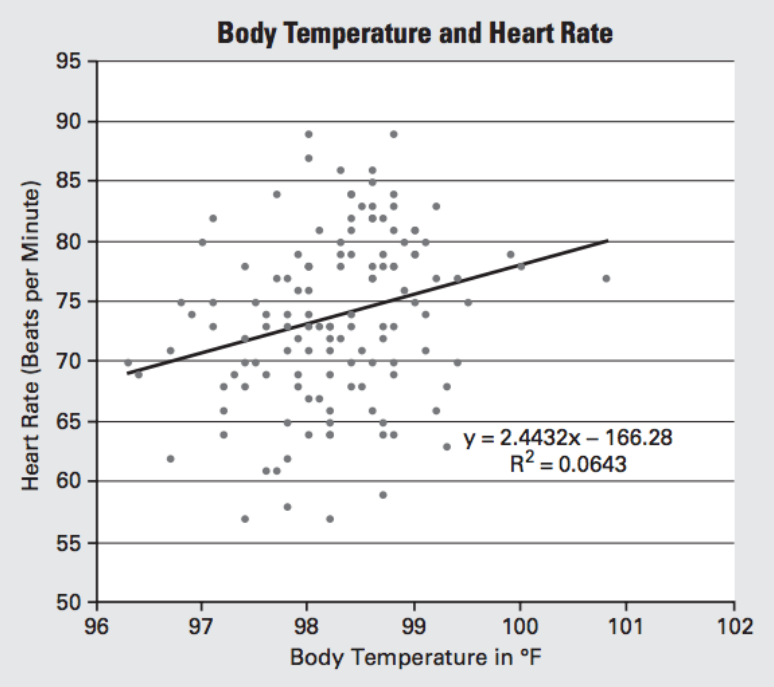


Figure 3. A Scatterplot with a Linear Regression Line

Regression analysis provides a measure of how the two variables are related. **The resulting regression line minimizes the distance of the actual scores from the predicted scores.**

There appears to be a positive correlation between body temperature and heart rate—that is, when body temperature goes up, so does the heart rate, and vice versa.

When Data is NOT Normally Distributed - you will calculate median, modes, quartiles, with Box-and-Whisker Plots (instead of the mean, SD, SE)

Some measurement data will not be normally distributed. The data distribution may be skewed or have large or small outliers (nonparametric data).

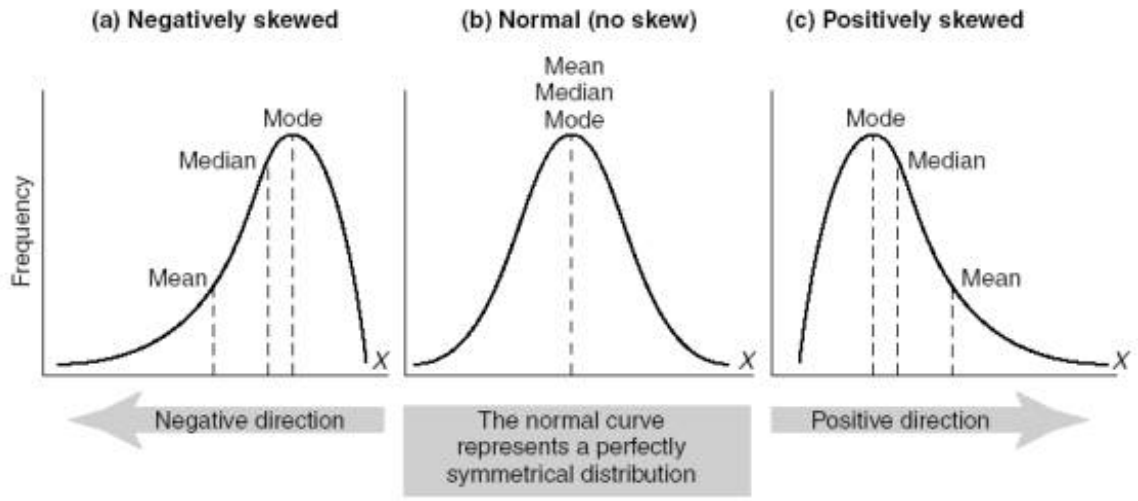


FIGURE 15.6 Examples of normal and skewed distributions

Example of non-normal distributed data:

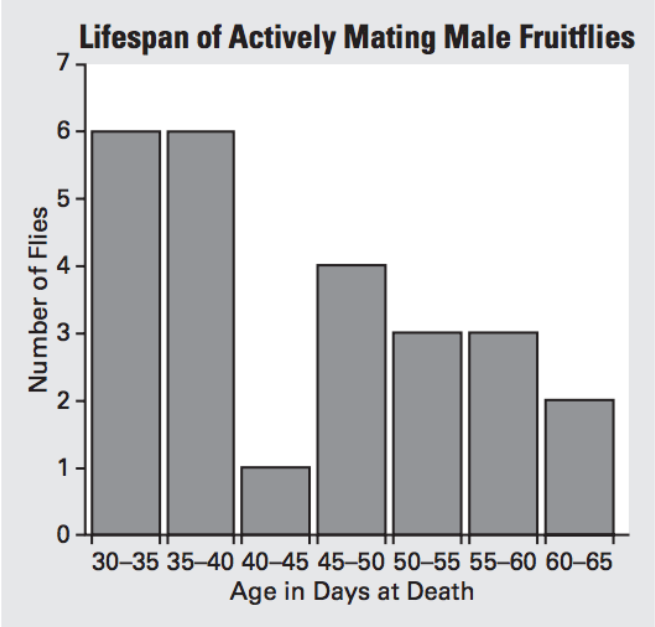


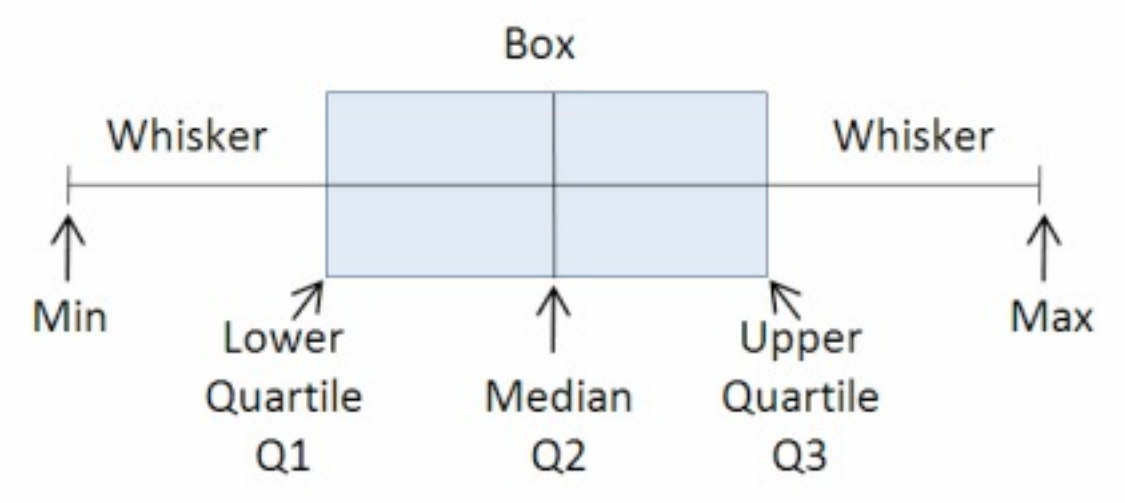
Figure 5. Histogram Showing Nonparametric Data [*Source: Hanley, James A., and Stanley H. Shapiro. "Sexual Activity and the Lifespan of Male Fruitflies: A Dataset That Gets Attention."* Journal of Statistics Education *2, no. 1 (1994).*]

BOX-AND-WHISKER PLOTS (WHISKER PLOTS)

For nonparametric, non-normally distributed data (where the mean is not equal to the median and the mode), the appropriate graph is a **box-and-whisker plot (a.k.a. a box plot or a whisker plot).**

In the graph, the ticks at the tops and bottoms of the vertical lines show the highest and lowest values in the dataset, respectively.

The data is broken down into quartiles.

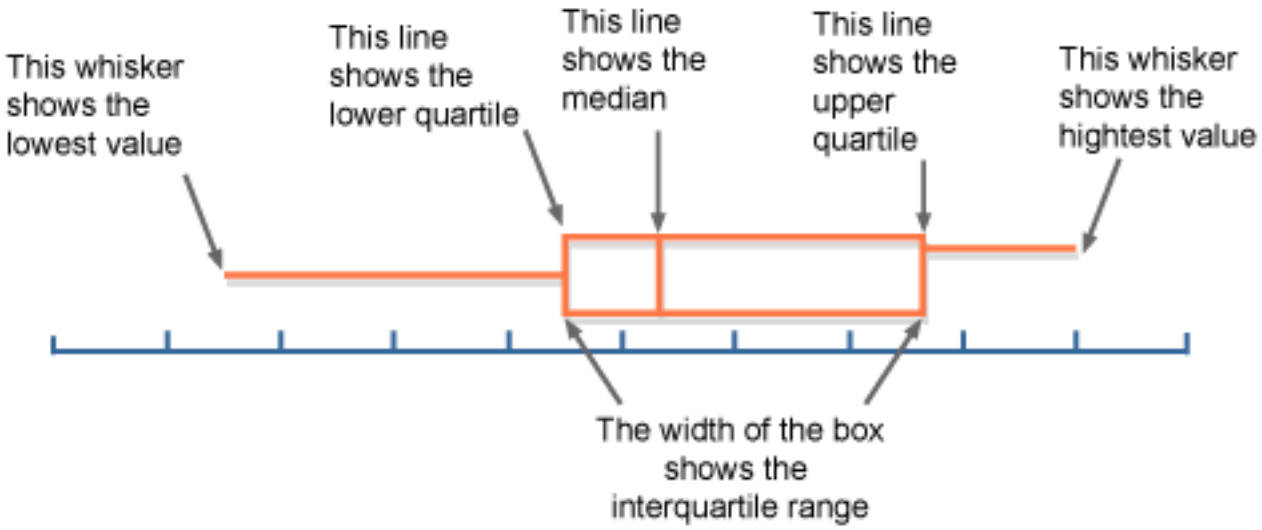


The right whisker represents the range in which the **top 25% of the data points fall**.

The left whisker represents the range in which the **bottom 25% of the data points fall.**

The box itself represents where the **middle 50% of the data falls.**

The horizontal line represents the **median**.



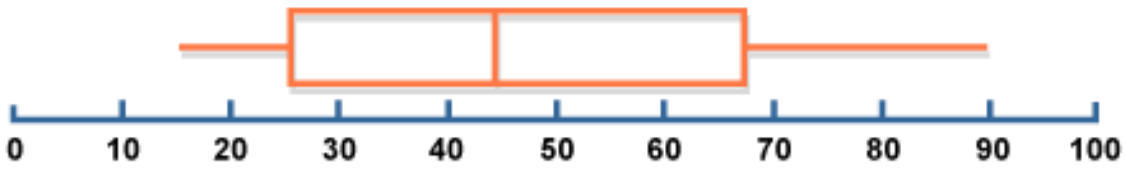
Example

The oldest person in Mathsminster is 90. The youngest person is 15.

The median age of the residents is 44, the lower quartile is 25, and the upper quartile is 67.

Represent this information with a box-and-whisker plot.

Solution



The graph below allows the investigator to determine at a glance, in this case, that the ash leaves appear to decay the fastest and the beech leaves take longer to decay.

Bag	% Decay		
Number	Ash	Sycamore	Beech
1	51	40	34
2	63	33	15
3	44		26
4		52	21
5	48	48	
6	32	35	11
7	70	44	19
8	48	63	32
9	57	40	

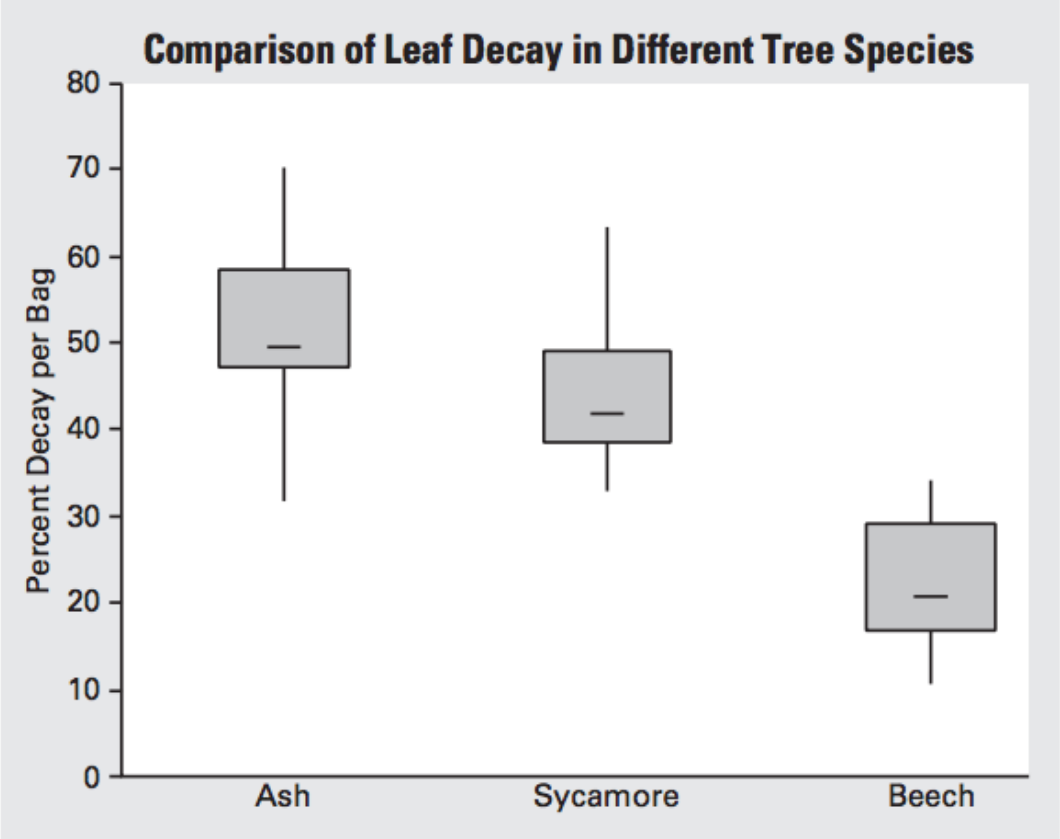


Figure 4. Nonparametric Data and Their Representation in a Box-and-Whisker Plot
[Source: Redrawn from "Merlin_examplels.xls," available as part of a download at: http://www.heckmondwikegrammar.net/index.php?highlight=introduction&p=10310]