Confidence Intervals: confidence interval is a range of values where we believe the true value lies.

We commonly use a 95% or 99% confidence interval.

- This is calculated using the mean of our data and the standard deviation. We can use this to find a range that the true mean most likely lies in.

- Still, it may not lie in this range. What we are saying is 95% of experiments will include the true mean, and 5% will not. This means there is a 5% chance our confidence interval does not include the true mean.

To calculate a confidence interval, we need to calculate a z score. A z score tells you how many standard deviations from the mean your score is.

$$z = (x - \mu) / \sigma$$

If our test score is 100, our mean is 80, and the standard deviation is 30, the z score would be:

$$z = (100 - 80)/30$$
$$z = 20/30$$
$$z = 0.67$$

When multiple samples are involved, we will need to account for the stand error.

$$z = (x - \mu) / (\sigma / \sqrt{n})$$

This show how many standard errors are between the sample and population mean.

Example problem: In general, the mean height of women is 65″ with a standard deviation of 3.5″. What is the probability of finding a random sample of 50 women with a mean height of 70″, assuming the heights are normally distributed?

$$z = (x - \mu) / (\sigma / \sqrt{n})$$
$$z = (70 - 65) / (3.5/\sqrt{50}) = 5 / 0.495 = 10.1$$

Calculating the confidence interval, the formula is:

$$X \pm z * s\sqrt{n}$$

X - the mean
z - the z score
s - the standard deviation
n - the number of observations

Example: We have an apple orchard and want to estimate the true mean of their size.

We get a sample of 50 random apples.

We get a mean weight of 5.7 oz with a standard deviation of 2.3.

Using this data, try to calculate the 95% confidence interval yourself.

(Remember, for a 95% confidence interval, the z score is 1.96.)

$x = 5.7$
$z = 1.96$
$s = 2.3$
$n = 50$

$5.7 \pm 1.96(2.3/50)$
$5.7 \pm 1.96 * 0.046$
5.7 ± 0.09016

The true mean of the weight of the apples is likely to lie between 5.60984 and 5.79016.

Z-intervals are not the only option out there! In our previous example, we did not use the population's data, just what we had from our sample.

A t-interval can be more useful when there is an unknown population mean. We rely on our sample standard deviation to calculate the margin of error.

$$x \pm t(s/\sqrt{n})$$

t a critical value from the t-distribution, so is the sample standard deviation, and n is the sample size.

To get the value of t, we need to first calculate degrees of freedom. This is equal to our sample size minus 1.

$$df = n - 1$$

We then use this number in a t-table.

Example with a t interval:

We get a sample of 20 random apples.

We get a mean weight of 5.2 oz with a standard deviation of 3.5.

Using this data, try to calculate the 95% confidence interval yourself.

Standard Error seems to be similar to the standard deviation except it is for sample size not the population size. If you have a small Standard error then it is telling you the sample mean is more accurate to the actual population.

Mean is the sum of values or number of values. It is calculated by totaling the sum of the numbers (in a series, collection, etc.) and dividing it by the total number. Example: 1+2+3+4 = 10/4= 2.5.

Range is the difference between the lowest and highest values. For ex. in this hamburger price the smalles pice is 5.5 and the largest is 14.

so 14-5.5=8.5

Kurtosis is a statistical measure of the outliers. If the kurtosis is greater than 3, then the dataset has heavier tails than a normal distribution (more in the tails). If the kurtosis is less than 3, then the dataset has lighter tails than a normal distribution (less in the tails).

Mode value that appears most often. The value that is most likely to be sampled.

Skewed left data is the opposite. It has a left tail and the "hump" is on the right.
Skewed right data has a long tail that extends to the right of the data set (the smaller numbers have the hump)

We will need to get the significance level, the mean, the standard deviation, and the sample size to get the confidence value.

Using our data from the cells in the previous slide, we can easily calculate these values.

For the significance level, we need 1 - confidence. So for a 95% confidence interval, we have a significance level of 0.05

=1 - confidence level
To calculate the mean, we use the average function and pass in the range of cells we want to get the mean of:

=AVERAGE(range of cells)
=AVERAGE(B3:B12)

For standard deviation, we do the same thing with the standard deviation function!

=STDEV.P(range of cells) or standard deviation, we do the same thing with the standard deviation function!

=STDEV.P(range of cells)

To get the count, call the COUNT function on the range.

=COUNT(range of cells)

Now that we have that data, we can pass it into the CONFIDENCE function. This will give us the number to add and subtract from the mean to get the confidence interval.

=CONFIDENCE(Alpha, Standard_dev, Size)

What is Standard Deviation?

A number that represents how one group differs from the mean value of entire group or data set.

What is 68-95-99.7 Rule?

The Percentages or Standard Deviation from the mean

68% of values are within 1 Standard Deviations

95% of values are within 2 Standard Deviations

99.7% of values are within 3 Standard Deviations

Example (If Mean = 100, Standard Deviation = 15)

1 Standard Deviation = 85 - 115 = 68%

2 Standard Deviations = 70-85 and 115-130 = 95%

3 Standard Deviations = 55-70 and 130-145 = 99.7%

What is the Variance?

Measures how far the set of (random) numbers are spread out from mean (their average value)

This tells us the measure of numbers from our mean

Two Types of Variance

Population - Variance from all data

Sample - Variance from a sample of the data: measures how far the set of random numbers are spread out from mean (average value) of the sample data. This variance is calculated from sample data.

Average of squared deviations about the (sample) mean, by dividing # of.

Samples - Variance could be higher because it is compensating without all data.

Variance of zero tell us that all the data values are identical.

High variance – indicates that the data points are very spread out from the mean, and from one another.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

Low variance means the range is close to the mean while high variance is the spread out from the mean.
Central Limit – belief that as the sample size get larger, the distribution will become normalized regardless of the population (especially for sample sizes over 30). It is foundational to the concept of confidence intervals and margins of error in frequentist statistics.

Law of Large Numbers – if you have a population and take large random samples from the population, it is believed and accepted as law in statistics that the sample means will become a normal distribution.

Probability – the mathematical possibility and/or likelihood of event happening.

Central Limit Theorem and Confidence Interval

- Utilize Confidence Interval (also known as accuracy): means that with a large number of repeated samples, 95% of such calculated **confidence intervals** would include the true value of the parameter.

The **mean** of the sampling distribution will cluster around the population mean.

The **standard deviation** of the sampling distribution is called the standard error.

Margin of error (also known as precision) - tells you how many percentage points your results will differ from the real population value.
Precision is one of the most important metrics for analyzing Supervised learning prediction and probability models
Example: a 95% confidence interval with a 2 percent margin of error tells us that your statistic will be within 2 percentage points of the population value 95% of the time.

Frequentist Statistics - *Utilize Confidence interval*
Means that with a large number of repeated samples, 95% of such calculated confidence intervals would include the true value of the parameter.

Bayesian Statistics - They believe in testing with a combination of prior knowledge.

Goal of Maximum – to find the optimal way to fit a distribution to the data.
Types of Distribution:
Normal

- Expect most of the meausrements (mouse weights) to be close.
- Relatively symmetrical around the mean.
- Various shapes & sizes.

Exponentia
Gamma

Normal distribution is close to the mean on all data points and is relatively symmetrical.

Once you find the correct type of distribution, you can fit that to experiments of the same type. This saves time and allows you to understand the data easier.

Normal distribution is the bell curve distribution.

A skinny bell curve means the average has a low variance.

Likelihood refers to how often you will observe the data on the curve.

Max likelihood of the mean (average) is the curve that covers most of the data.

The likelihood of the standard deviation is how likely you will see the data in a certain standard deviation. The max likelihood is the standard deviation that covers the most data points.