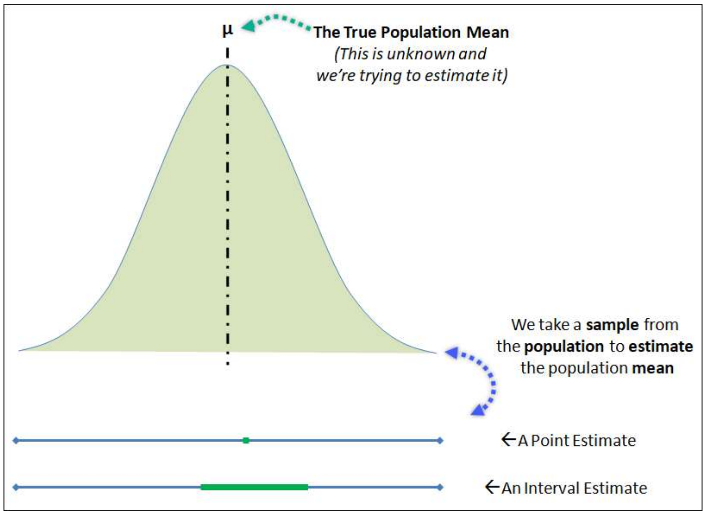
Point Estimates & Confidence Intervals

There are three kinds of lies: lies, damned lies, and statistics. - Benjamin Disraeli


μ
The True Population Mean
(This is unknown and we're trying to estimate it)
We take a sample from the population to estimate the population mean
←A Point Estimate
←An Interval Estimate

We're about to move into the fun part of statistics - Inferential Statistics, and I'm super excited!

Inferential statistics is a collection of tools & techniques that allow us to draw conclusions about populations based on information obtained from sample data.

The best place to start in **Inferential statistics** is with the basic concept of **Estimators.**

There are two types of estimators, **Point Estimates** & **Interval Estimates**, and we will discuss the similarities & differences between these two within the chapter.

From a very basic perspective these two tools allow you to estimate population parameters (mean, variance, etc) using data taken from a sample.

I'll say that again. . .

We're using **sample data** to **estimate** a **population parameter**.

This chapter is laid out in two sections, the first is dedicated to the **Point Estimate**, and the second is for the **Interval Estimate**.

Within the **Point Estimate** section of this chapter we will:

- Review the concept of Populations & Samples
- Discuss the Point Estimate for the Population Mean, Population Variance & Standard Deviation,
- Discuss the concept of an Unbiased & Efficient Estimates &
- A review of the concept of Standard Error

Within the **Interval Estimate** Section we will:

- What is a Confidence Interval
- Confidence Intervals for the Population Mean
- Confidence Intervals for Population Variance & Standard Deviation
- Confidence Intervals for Proportions

Populations, Samples & Inferential Statistics

Ok, so we've discussed the fact that ***inferential statistics** is a collection of methods that allow us to make inferences about a population based on information obtained from a sample of data.*

So let's jump into that and cover some of these key terms and look at their definitions so that you're crystal clear on what we're doing here.

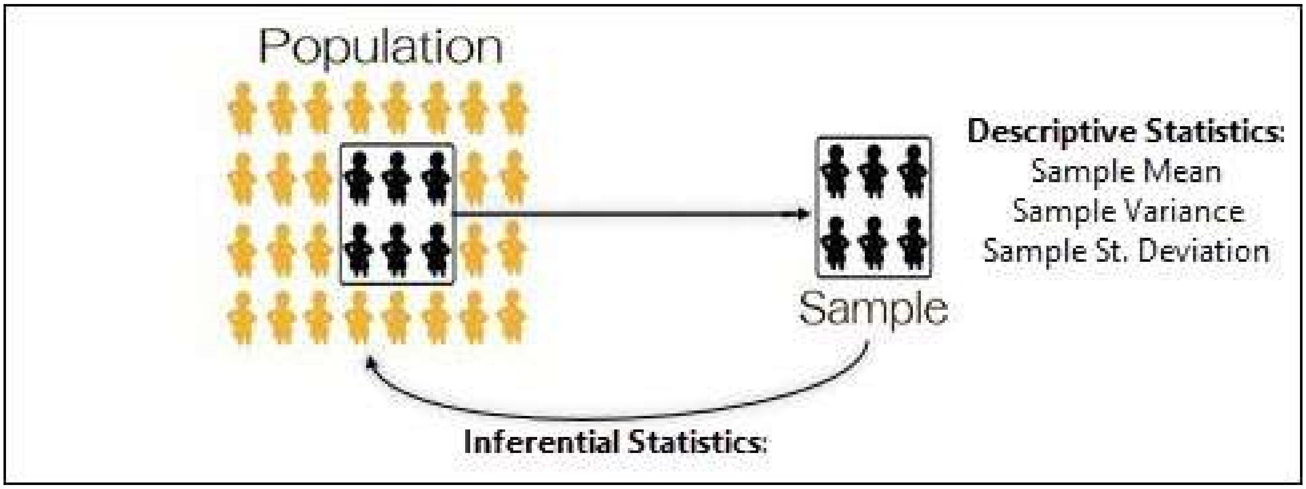
Within statistics, a **population** *is defined as a total set of objects, events or observations about which you want to study.*

In quality engineering we're often interested in knowing the mean & standard deviation associated with populations we're dealing with.

However, we rarely have the time or resources to measure all values associated with our populations, which is where sampling helps us.

A **Sample** is *defined as a unique subset of a population.*

You can see that in the image below where we have a population of people, and we've taken a sample from that population.



Recall back to chapter two on Statistics (Collecting & Summarizing Data Part 2) where we discussed the difference between a **Statistic** & a **Parameter**.

When we talking about sample data and we calculate the mean or standard deviation, we are calculating a **Statistic**.

***Statistics are associated with samples*.**

When we're talking about the entire population - the population mean or population standard deviation, we're talking about a **Parameter**.

Within the table below you can see the common Population Parameters & their associates sample statistics.

	Population	**Sample**
	Parameter	***Statistic***
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Std. Deviation	σ	s
Size	N	n

Parameters, Statistics & Estimates

In inferential statistics, we take our **sample data** and we calculate our **sample statistics**.

We can then use those sample statistics to estimate the population parameter; which is often times what we're really looking to understand.

These sample statistics are used within this concept of an **estimate,** where there are two types of estimates, **Point Estimates** & **Interval Estimates**.

A **point estimate** is a type of estimation that uses a single value, oftentimes a sample statistic, to infer information about the population parameter as a single value or point.

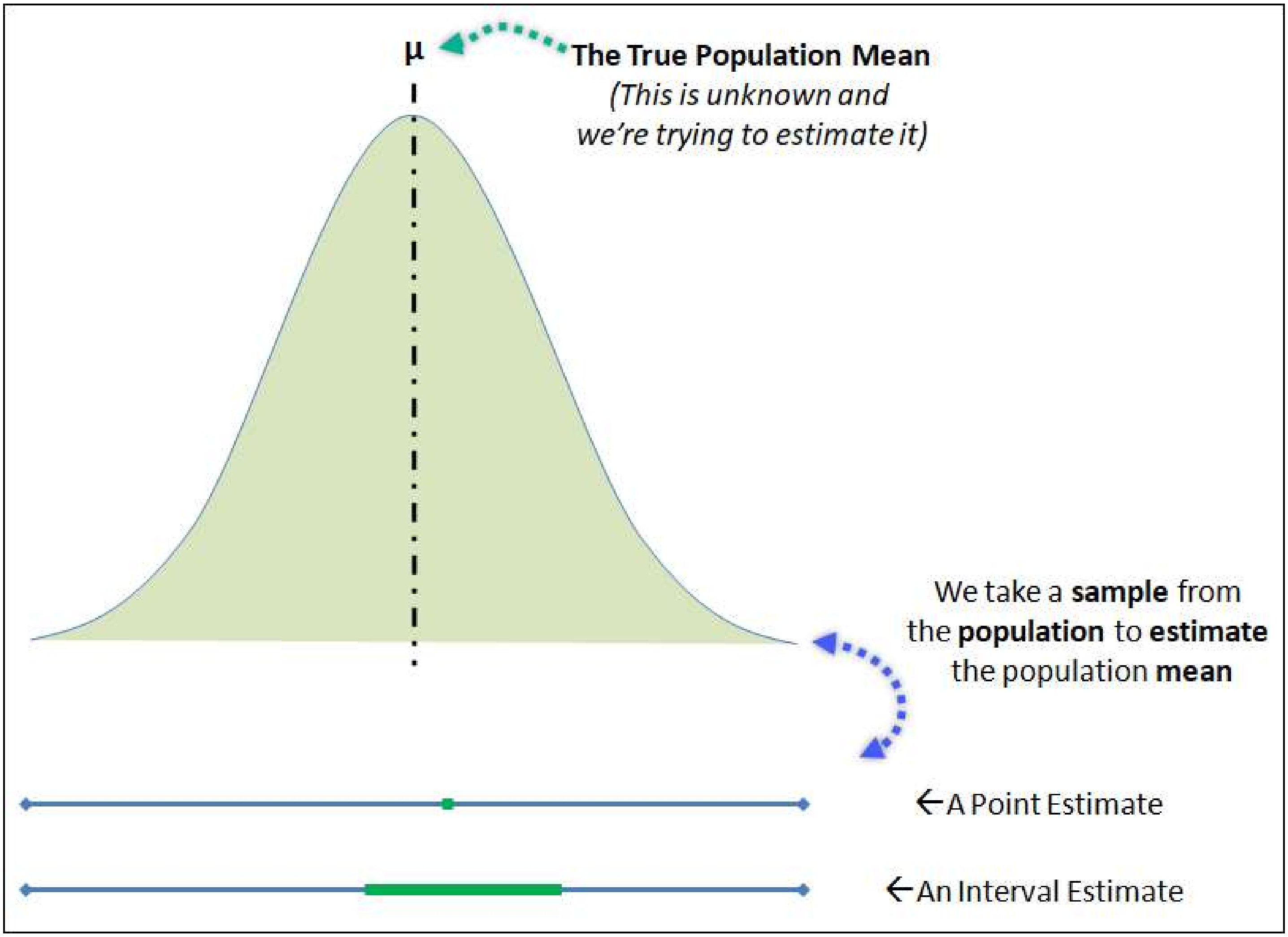
An **interval estimate** is a type of estimation that uses a range (or interval) of values, based on sampling information, to "capture" or "cover" the true population parameter being inferred.

The likelihood that the interval estimate contains the true population parameter is given by the **confidence level**.

You can see both of these estimates below.

The top of the image is the **population distribution**, with the **true population mean** shown, which is the population parameter that we're attempting to **estimate**.

This population mean can be estimated by a single point estimate, or as an interval estimate.



Point Estimates

Ok, let's quickly jump into the first type of estimate, the **Point Estimate**.

A **point estimate** is a type of estimation that uses a single value, oftentimes **a sample statistic**, to **infer** information about the **population parameter**.

Let's go through some of the major point estimates which include point estimates for the population **mean**, the population **variance** and the population **standard deviation**.

Point Estimate for the Population Mean

So, let's say we've recently purchased 5,000 widgets to be consumed in our next manufacturing order, and we require that the average length of the widget of the 5,000 widgets is 2 inches.

Instead of measuring all 5,000 units, which would be extremely time consuming and costly, and in other cases possibly destructive, we can take a sample from that population and measure the average length of the sample.

As you know, the sample mean can be calculated by simply summing up the individual values and dividing by the number of samples measured.

$$\boldsymbol{Sample\ Mean}:\ \bar{X} = \frac{\sum x}{n}$$

Example of Sample Mean Calculation

Calculate the sample mean value of the following 5 length measurements for our lot of widgets: **16.5, 17.2, 14.5, 15.3, 16.1**

$$\boldsymbol{Sample\ Mean}:\ \bar{X} = \frac{\sum x}{n} = \frac{16.5 + 17.2 + 14.5 + 15.3 + 16.1}{5} = 15.9$$

Point Estimate for the Population Variance & Standard Deviation

Similar to this example, you might want to estimate the variance or standard deviation associated with a population of product.

The point estimate of the population variance & standard deviation is simply the sample variance & sample standard deviation:

$$\boldsymbol{Sample\ Variance}:\ \boldsymbol{s^2} = \frac{\boldsymbol{\sum(x_i - \bar{x})^2}}{\boldsymbol{n-1}} \qquad \& \qquad Sample\ Standard\ Deviation:\ s = \sqrt{\frac{\boldsymbol{\sum(x_i - \bar{x})^2}}{\boldsymbol{n-1}}}$$

Example of Sample Standard Deviation

Let's find the sample standard deviation for the same data set we used above: **16.5, 17.2, 14.5, 15.3, 16.1**

x*i*	$(\boldsymbol{x_i - \bar{x}})$	$(\boldsymbol{x_i - \bar{x}})^2$
16.5	(16.5 - 15.9) = 0.6	0.36
17.2	(17.2 - 15.9) =1.3	1.69
14.5	(14.5 - 15.9) =-1.4	1.96
15.3	(15.3 - 15.9) =-0.6	0.36
16.1	(16.1 - 15.9) =0.2	0.04
		4.41

$$\boldsymbol{Sample\ Standard\ Deviation}:\ s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{4.41}{5-1}} = 1.05$$

Make sense? Now let's switch on to a few important topics before jumping into the confidence interval section.

Unbiased & Efficient Estimators

Anytime we're using an **estimator to infer a population parameter**, you will naturally **incur some risk** (or likelihood) of **inferring incorrectly**.

Then entire field of **Inferential statistics**, by nature, involves a certain element of risk, which we will talk a lot about over the next few chapters.

So, to minimize the **risk** associated with **estimators**, we desire two characteristics of a **high quality estimator**, that they are **unbiased & efficient**.

Unbiased

*An **unbiased estimator** is one who's expected value is equal to the population parameter being estimated.*

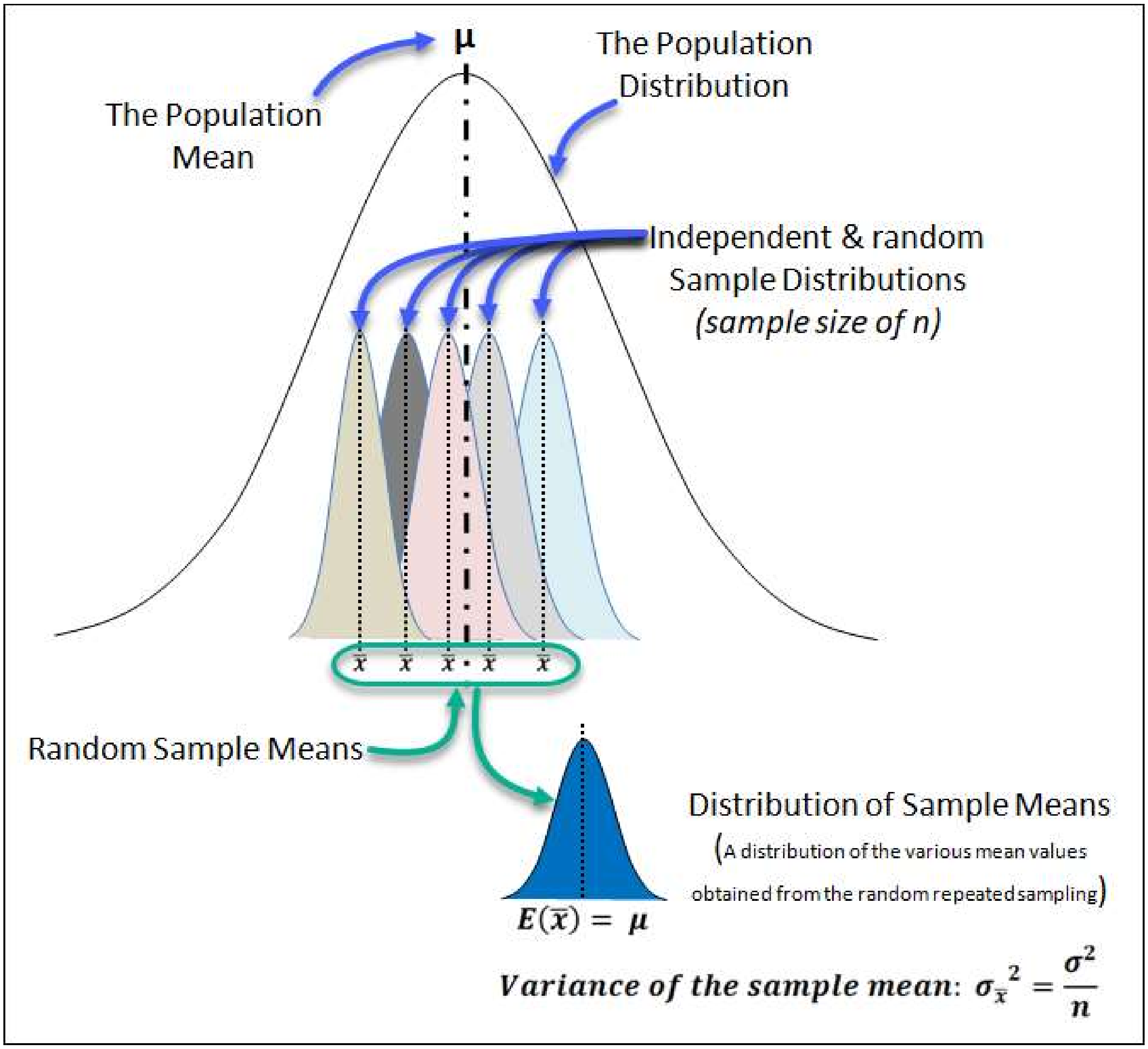
Consider the situation where we are **repeatedly sampling (sample size = n) from a population distribution.**

Each sample would have its own distribution of values, which are all shown under the main population distribution.

Let's say you sampled 100 units from a population of 1,000 and you calculated the sample mean.

Based on the random nature of sampling, you'd expect each sample taken to likely have a different sample mean.

Now let's say you repeated this sampling 30 times; and you plotted the **distribution of sample means**.


μ
The Population Distribution
The Population Mean
Independent & random Sample Distributions (sample size of n)
$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$
Random Sample Means
Distribution of Sample Means (A distribution of the various mean values obtained from the random repeated sampling)
$E(\bar{x}) = \mu$
$Variance\ of\ the\ sample\ mean:\ \sigma_{\bar{x}}^{\ 2} = \frac{\sigma^2}{n}$

This new distribution of sample means has its own variance & expected value (mean value).

$$E(\bar{x}) = \mu$$

A point estimate (the sample mean, in this example) is considered unbiased if its expected value is equal to the parameter that it is estimating.

Bias & Variance

This same thing can be said for the sample variance (S^2), in that the expected value of the sample variance can be shown to be equal to the population variance (σ^2).

A quick caveat about that here, if you compare the equation for **Population Variance** against the **Sample Variance**, you'll notice they have **different denominators**.

$$\boldsymbol{Population\,Variance:\;\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}} \qquad \boldsymbol{Sample\,Variance:\;s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}}$$

The population variance is divided by **N**, while the sample variance is divided by **n-1**, this is to **account for bias**.

If you calculated the sample variance using only N, you'd find that the estimate tends to be too low and therefore biased.

This is why we divide by n-1 as this has been shown to be an unbiased estimate of the population variance.

Efficiency

The second characteristic of a high-quality estimator is one that is **efficient** and it is a reflection of the **sampling variability of a statistic**.

When it comes to estimating a population parameter like the population mean, there can often be many ways to estimate that population mean.

A **more efficient estimator** is one that has the **lowest variance** (sampling variability) associated with it.

For example, to estimate the population mean you could use two different estimators.

The first estimator could be the **sample mean**, while the second sample could simply be **a single observation** from the population as an estimate.

To pick the more efficient estimator between these two options (sample mean v. a single observation) we must understand the variance associated with each estimator.

The variance of a single observation is simple, it's equal to the population variance σ^2.

The variance of the sample mean distribution is: $Variance\ of\ sample\ mean: V(\bar{x}) = \frac{\sigma^2}{n}$

If we compare these together we can see that the variance of the sample mean is smaller than the variance of a single observation.

$$\frac{\sigma^2}{n} < \sigma^2$$

Therefore, we can conclude that the sample mean is a more efficient estimator of the population mean than a single observation because its variance is lower.

For the population mean, we measure the **efficiency (sampling variability)** of our sample using the **Standard Error**.

Standard Error

As discussed above, whenever we create a distribution of sample means, that distribution of sample means will also have a certain amount of variability to it.

This is a reflection of the **efficiency (sampling variability)** associated with our sampling (size of n). The smaller the standard error, the less sampling variability.

Within the context of this chapter, we're only going to discuss the standard error associated with the sample mean distribution.

With some statistics work that's out of scope of this text, we can prove that the variance of the sample mean distribution:

$$\boldsymbol{Variance\ of\ sample\ mean}: \boldsymbol{V(\overline{x})} = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}$$

The **Standard Error** of this sample mean distribution is analogous to the Standard Deviation in that it is a reflection of the dispersion or spread of sample mean values around the population mean.

Similar to the standard distribution, the **Standard Error** is the square root of the sample mean distributions variance.

$$\boldsymbol{Standard\ Error\ of\ The\ Sample\ Mean}: \boldsymbol{S.E.} = \sqrt{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}} = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

The Standard Error is equal to the population standard deviation, divided by the square root of n.

We will use this concept of the **standard error** in the next section when we discuss the **confidence interval**.

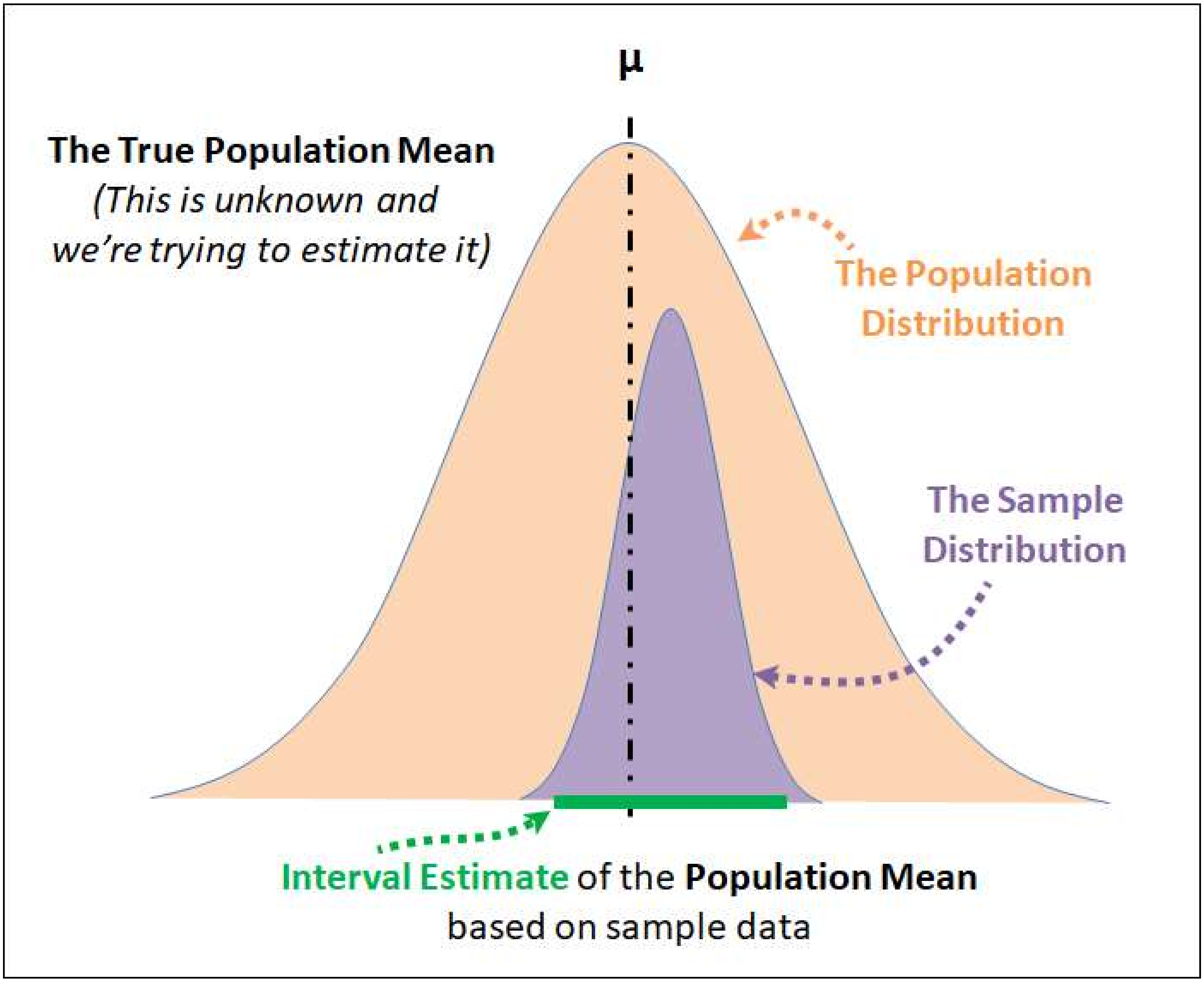
Confidence Interval (The Interval Estimate)

An **interval estimate** is a type of estimation that uses a range (or interval) of values, based on sampling information, to "capture" or "cover" the true population parameter being inferred / estimated.

Interval estimates are created using a **confidence level**, which is the probability that your interval truly captures the population parameter being estimated.

Because we use a confidence level, we often call these interval estimates a **confidence interval**.

You can see an example of the confidence interval below.



The image starts with the population distribution in orange, and this distribution has an unknown population mean, which we're attempting to estimate.

Then we take a sample (of size n) from that population, and that sample distribution is shown in purple.

We can then create our confidence interval based on that sample data.

Let's talk more about this confidence level before jumping into the interval calculations.

What does a Confidence Level Mean?

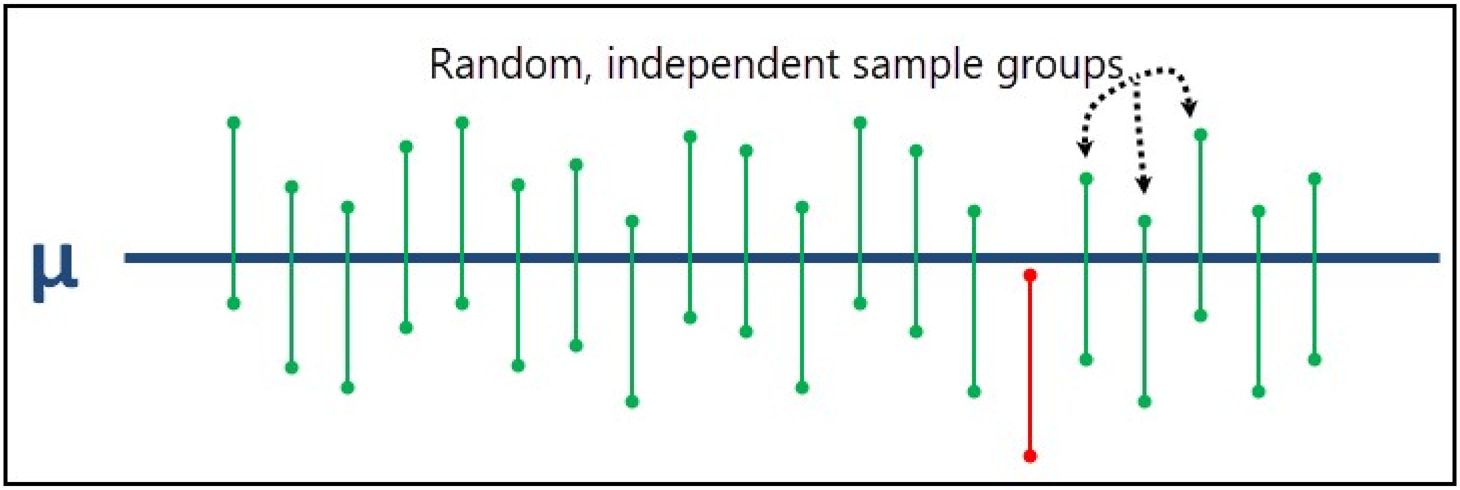
The confidence level is an often miss-understood concept.

When we say that we have 95% confidence in our interval estimate, we do not mean that *95% of the overall population falls within the confidence interval.*

The confidence level is the probability that your confidence interval truly captures the population parameter being estimated.

So if we have a 95% confidence level, we can be confident that 95% of the time (19 out of 20), our interval estimate will accurately captures the true parameter being estimated.

If you look at the graph below, the true population parameter (μ in this case) is shown as the solid blue line down the middle.



In 19 of the 20 intervals created, the true population mean is captured within those 19 intervals. There is only 1 interval that does not capture the true population mean and it's shown in **red**.

Up until this point I've only used the 95% confidence level, but your confidence level can vary.

You can be 80% confident, 90%, 99% ,etc.

The confidence level you choose is based on risk - specifically your **alpha risk** (α).

Alpha risk is also called your **significance level** *and it is the risk that you will not accurately capture the true population parameter.*

Your Confidence Level then is equal to 100% minus your significance level (α).

Confidence Level = 100% - Significance Level (α)

So if your significance level is 0.05 (5%), then your confidence level is 95%.

The Confidence Interval Equation

Ok, similar to the confidence level, I wanted to start by talking generically about the confidence interval equation, so that you understand the different components.

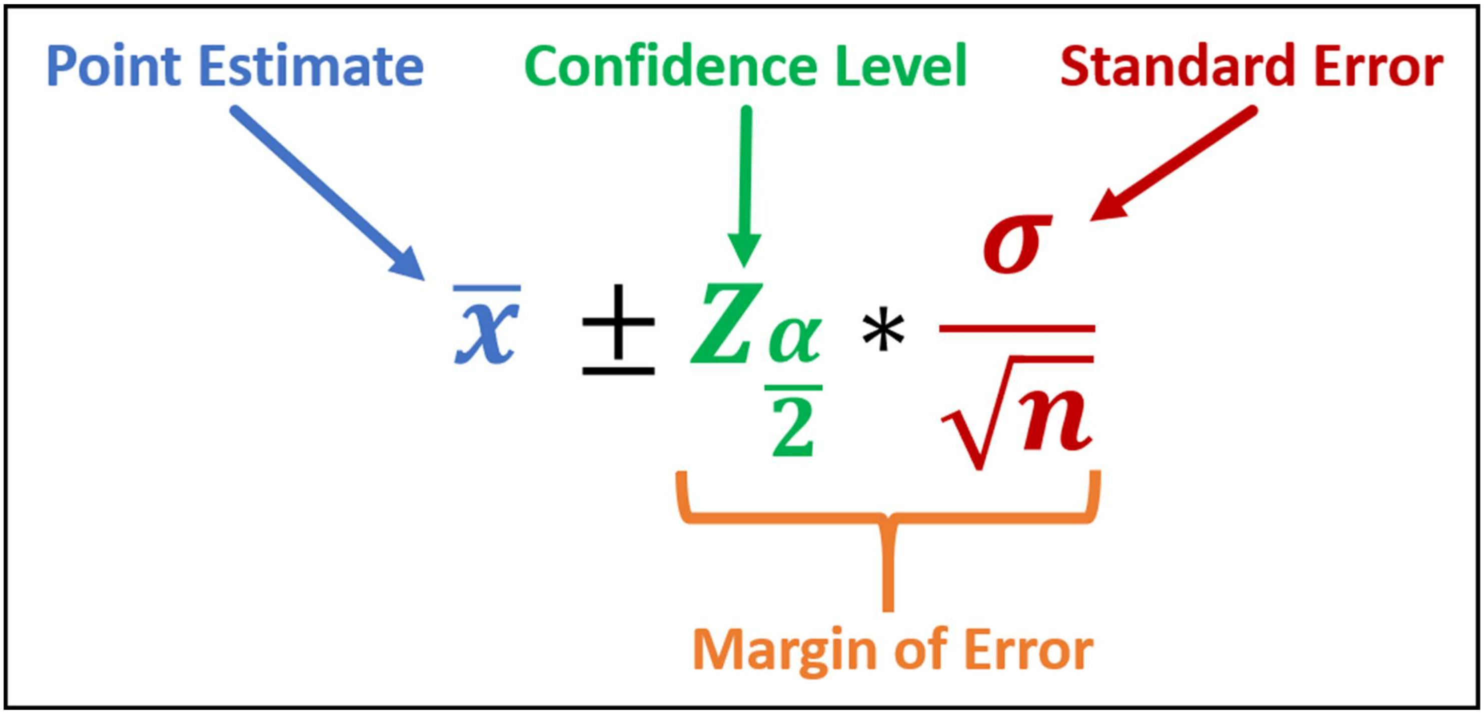
Once you get this part, the various situations (Mean, variance, proportion), is just an adaptation of this equation.

The confidence interval equation is comprised of 3 parts: **a point estimate (also sometimes called a sample statistic)**, **a confidence level**, and a **margin of error**.

The **point estimate**, or statistic, is the most likely value of the population parameter and the **confidence level & margin of error** represents the amount of uncertainty associated with the sample taken.

Confidence Interval = Point Estimate ± Confidence Level * Margin of Error

When dealing with the population mean, the confidence interval looks like this:


Point Estimate
Confidence Level
Standard Error
$\bar{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$
Margin of Error

The standard error can be impacted by the sample size (n) associated with your sample, where a larger sample means a smaller margin of error. The margin of error is also impacted by the standard deviation.

The Confidence Level & Z-Score

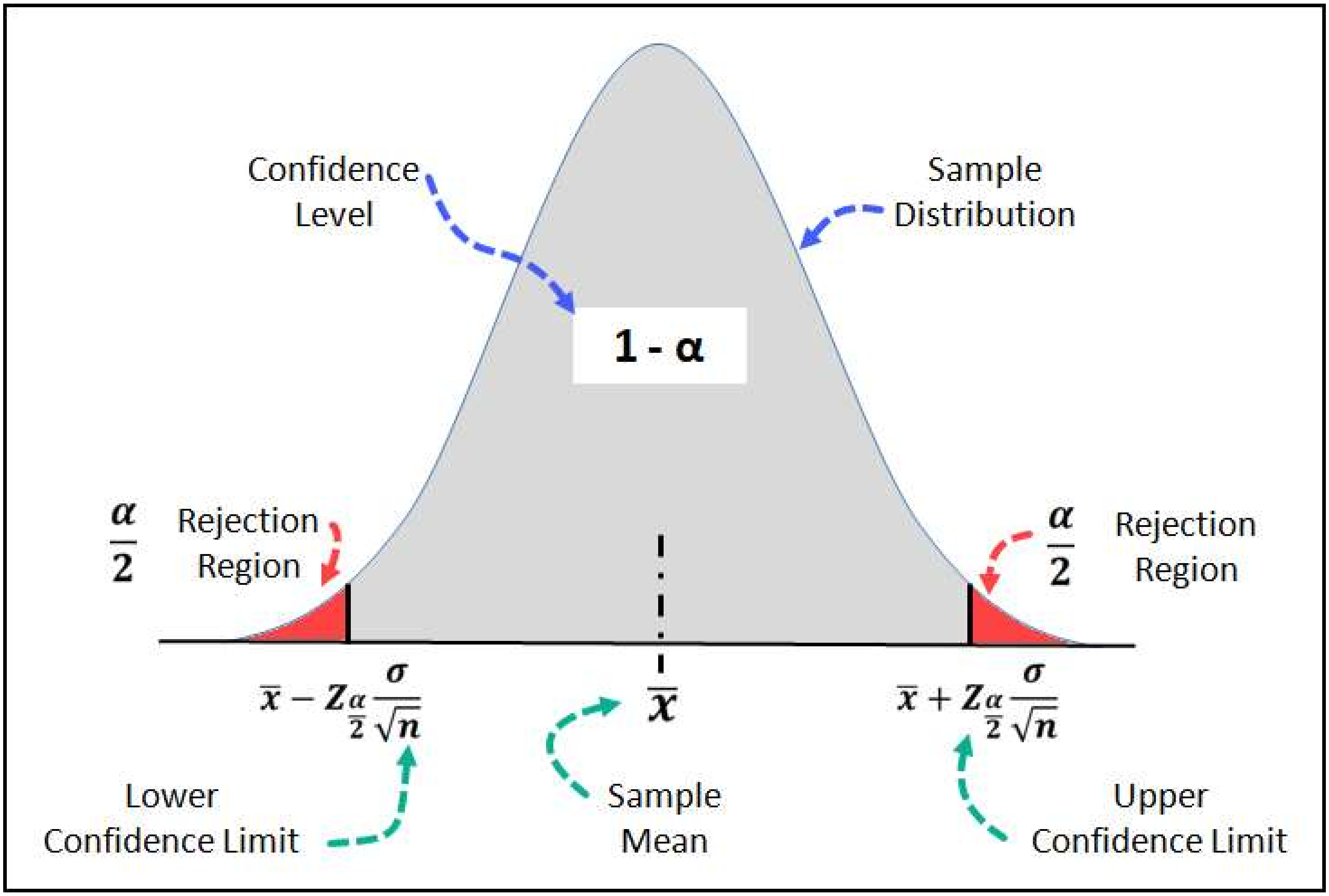
You'll notice in the example above; the confidence level is expressed as a Z-score that also references the **alpha risk** (**Significance level**).

In this situation, we're using the Z-score because the distribution of sample means is normally distributed.

Essentially what this confidence level is doing is capturing a certain proportion (95% for example) of the sample mean distribution when creating the interval estimate.

In the various other situations, where we're creating a confidence interval for variance & standard deviation, that data follows the chi-squared distribution.

So those confidence intervals use the Chi-squared value instead of the Z-score.


Confidence Level
Sample Distribution
1 - α
$\frac{\alpha}{2}$ Rejection Region
$\frac{\alpha}{2}$ Rejection Region
$\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
$\bar{x} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
Lower Confidence Limit
Sample Mean
Upper Confidence Limit

On this graph you can see that the region shaded in gray in the middle captures a certain portion of the distribution, for example if your confidence level is 95%, it would capture 95% of the distribution.

The region in red is often called the **rejection region**, but I've shown it that way to demonstration the **alpha risk**.

The alpha risk, 5% in this example, is **split in half** between the left and right tail of the sample distribution which is why the equation and the image show it as **α/2**.

In this example the sample distribution is the normal distribution, but in other interval estimates we might be using the t-distribution or the chi-squared distribution.

Confidence Interval for the Population Mean

Ok, time to jump into the meat of the confidence interval discussion and show you the actual equations.

When it comes to creating an interval estimate for the Population Mean, there are two possible equations.

These two possible equations are based on whether or not the population variance is known or unknown.

When the population variance is known, you use the normal distribution (z-score) and the population variance to create your interval estimate using the following equation:

$$\textit{Interval Estimate of Population Mean (known variance)} : \bar{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

When the population variance is unknown, you use the t-distribution (t-score) and the sample variance to create your interval estimate using the following equation:

$$\textit{Interval Estimate of Population Mean (unknown variance)} : \bar{x} \pm t_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

I've highlighted the difference between the z-score and t-score above in **red**, and I've highlighted the difference between the sample variance & population variance in **blue**.

Another way to decide between the equations above is based on the sample size **n**.

Generally, if the sample size is less than 30, the t-distribution should be used. If the sample size is greater than 30, then the normal distribution can be used.

Let's do an example of each.

Example of Interval Estimate of Population Mean with Known Variance

You've sampled 40 units from the latest production lot to measure the weight of the product, and the sample mean is 10.40 lbs. If the population standard deviation is known to be 0.60 lbs, calculate the 95% confidence interval.

Ok, let's see what we know after reading the question:

n = 40, $\bar{x} = 10.4$ lbs, σ = 0.60 lbs, α = 0.05.

Before we can plug this into our equation we need to find the Z-score associated with the 95% confidence interval.

If we look that up in the Normal Probability Table, we find Z = 1.96.

The Z-score of 1.96 is associated with an area under the curve of 0.475. This is because the normal distribution is two-sided and the alpha risk associated with one side of the distribution is 0.500 - 0.025 = 0.475.

$$\textit{Interval Estimate of Population Mean (known variance)} : \bar{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

$$\textit{Interval Estimate} : 10.4 \pm 1.96 * \frac{0.60}{\sqrt{40}}$$

$$\textit{Interval Estimate} : 10.4 \pm 0.186$$

$$95\% \textit{ Confidence Interval} : 10.21 - 10.59$$