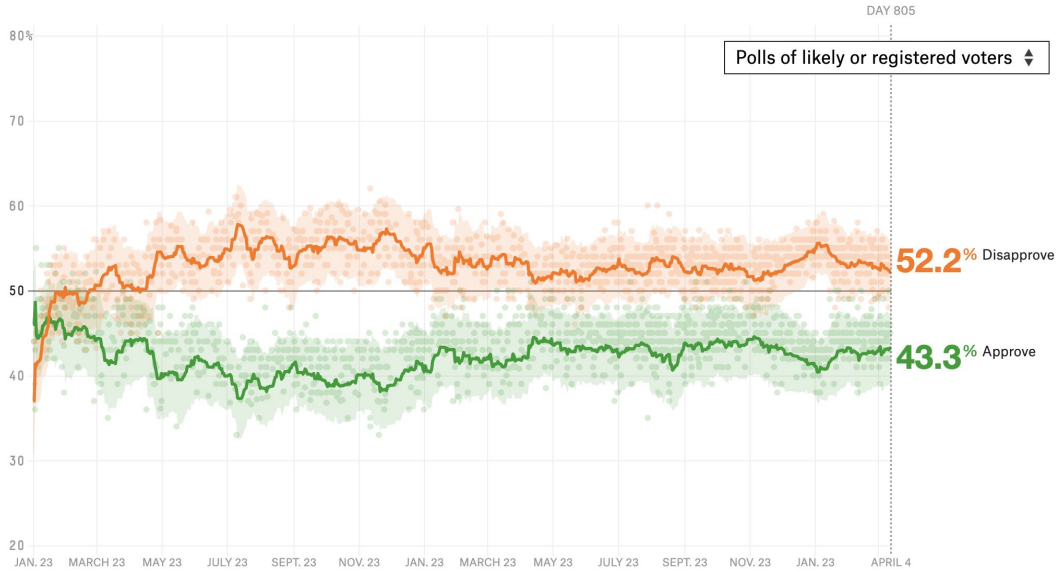
Sampling

Data Science Immersive

Guiding Questions



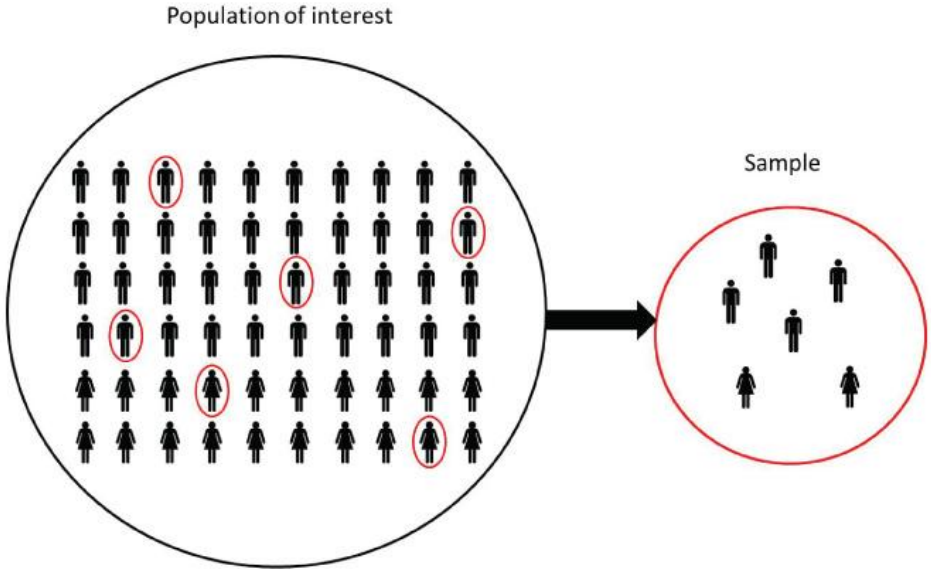
During elections companies poll voters to try and predict who will win the election. When reporting the results, you will hear the phrase “margin of error”. What is the margin of error? How and why do they calculate it?

Agenda

- Population v. Sample
- Sampling Statistics
- Central Limit Theorem
- Confidence Intervals
- T-Distribution
    - Degrees of freedom

Flatiron Polling Services

The real power of statistics comes from applying the concepts of probability to situations where you have data. The results, called statistical inference, give you probability statements about the population of interest based on that set of data.



Introduction to Inferences

There are two types of statistical inferences:

1. Estimation - Use information from the sample to estimate (or predict) the parameter of interest.

Using the result of a poll about the president's current approval rating to estimate (or predict) his or her true current approval rating nationwide.

2. Statistical Test - Use information from the sample to determine whether a certain statement about the parameter of interest is true.

The president's job approval rating is above 50%.

Gathering a Sample

You work for a Democratic presidential nominee and the campaign wants to know how they will do in the upcoming primary. So they task you with giving them an estimate of their support.

How would you go about acquiring your data to create your estimate?

We are trying to minimize the **bias** of our sample while simultaneously minimizing our **cost.**

Assumptions of a Simple Random Sample

- Independence: Each sampled value must be independent from one another
    - What does this imply about whether we should sample with or without replacement?
- Randomized: Each individual selected for the sample should be randomly selected
- Sample Size: Must be sufficiently large for your desired effect size

Population v. Sample Terminology

Term	Population = Parameter	Sample = Statistic
Count of items	N	n
Mean	μ	$\bar{x}$
Median	$\tilde{\mu}$	$\tilde{x}$
Standard Dev.	σ	S
Estimators =	$\hat{\mu}$	$\bar{x}$
	$\hat{\sigma}$	s

When we are observing something from a sample, it is considered a **point estimate** of population parameters

How can we make informed judgements from a given sample?

Sample Means

Imagine we have a group of small dogs and that have the following weights:
19, 14, 15, 9, 10, 17 so the population mean = 14lbs

Let's start with obtaining all of the possible samples of size $n=2$ from the populations, sampling without replacement. The table below show all the possible samples, the weights for the chosen puppies, the sample mean and the probability of obtaining each sample.

Sample Means

Below are the possible means of the samples we could draw, along with the probability of drawing them.

$\bar{x}$	9.5	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.5	16.0	16.5	17.0	18.0
Probability	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$

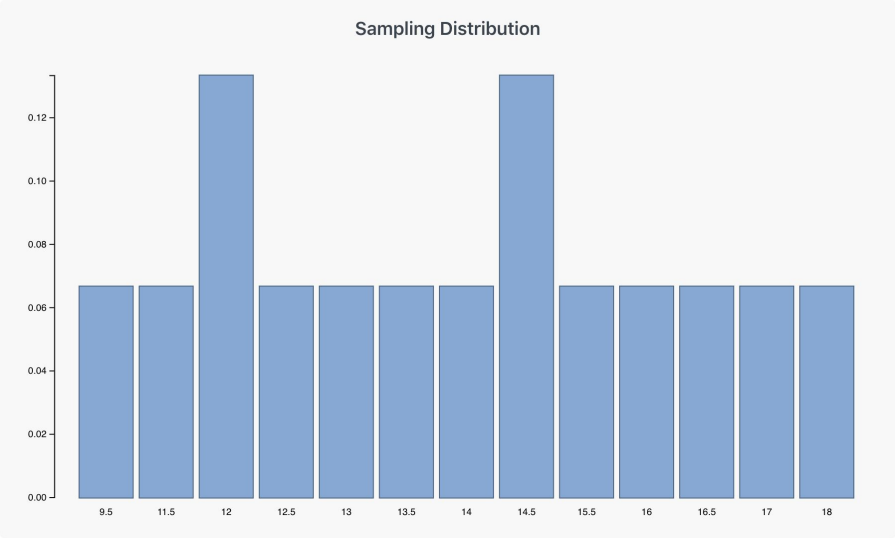
The mean of the sample means is

$$\mu_{\bar{x}} = \sum \bar{x}_i f(\bar{x}_i) = 9.5\left(\frac{1}{15}\right) + 11.5\left(\frac{1}{15}\right) + 12\left(\frac{2}{15}\right)$$
$$+ 12.5\left(\frac{1}{15}\right) + 13\left(\frac{1}{15}\right) + 13.5\left(\frac{1}{15}\right) + 14\left(\frac{1}{15}\right)$$
$$+ 14.5\left(\frac{2}{15}\right) + 15.5\left(\frac{1}{15}\right) + 16\left(\frac{1}{15}\right) + 16.5\left(\frac{1}{15}\right)$$
$$+ 17\left(\frac{1}{15}\right) + 18\left(\frac{1}{15}\right) = 14$$

Sampling Distribution

Sampling Distribution

The sampling distribution of a statistic is a probability distribution based on a large number of samples of size *n* from a given population.



Sample Means - Larger *N*

Now imagine we increased the size of our sample from 2 to 5.

Sample	Weights	$\bar{x}$	Probability
A, B, C, D, E	19, 14, 15, 9, 10	13.4	1/6
A, B, C, D, F	19, 14, 15, 9, 17	14.8	1/6
A, B, C, E, F	19, 14, 15, 10, 17	15.0	1/6
A, B, D, E, F	19, 14, 9, 10, 17	13.8	1/6
A, C, D, E, F	19, 15, 9, 10, 17	14.0	1/6
B, C, D, E, F	14, 15, 9, 10, 17	13.0	1/6

Standard Error of the Mean

The standard error of the mean is the standard deviation of the sampling distribution.

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

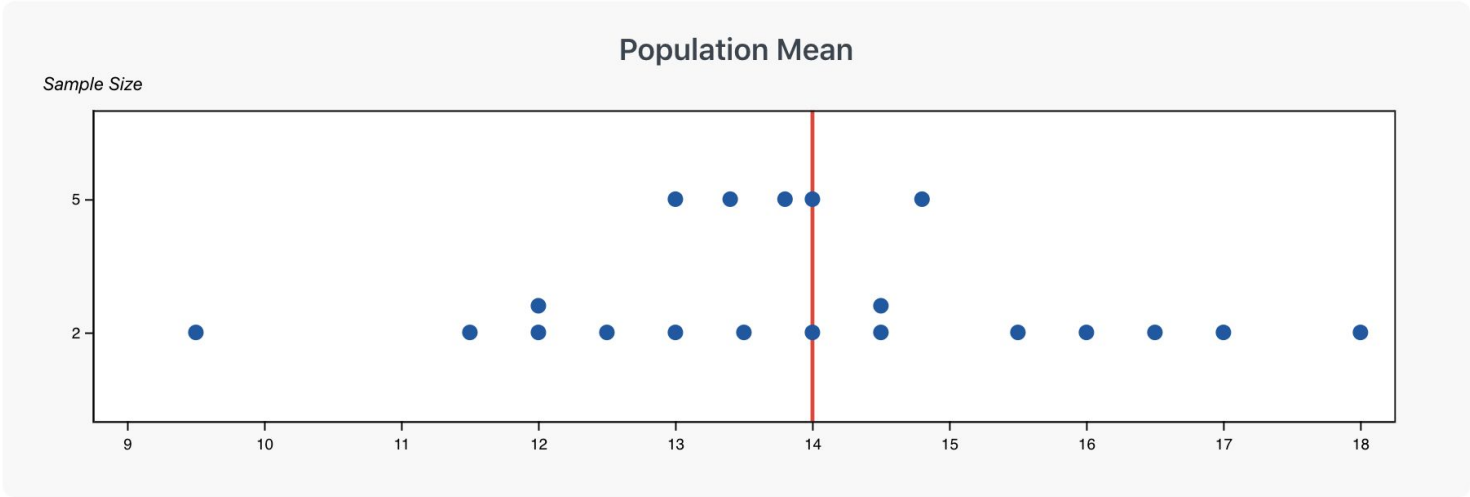
- σ_x = standard error of $\bar{x}$
- σ = standard deviation of population

If we do not know the population standard deviation, we can approximate for it by using the sample standard deviation.

$$\sigma_x \approx \frac{s}{\sqrt{n}}$$

- s = sample standard deviation

Sampling Error



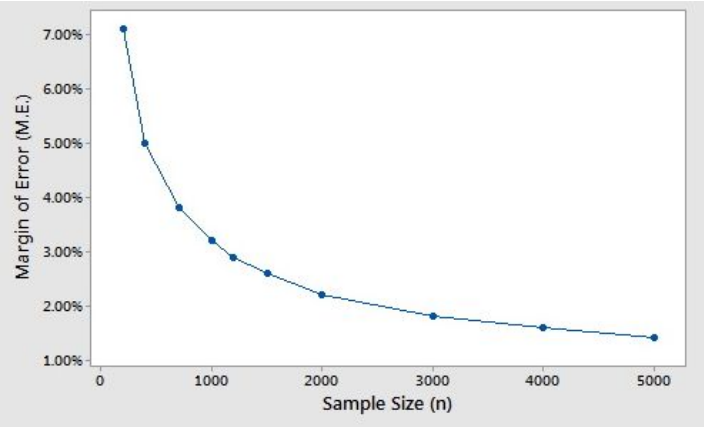
Sampling Error

The error resulting from using a sample characteristic to estimate a population characteristic.

Sample size and sampling error: As the dotplots above show, the possible sample means cluster more closely around the population mean as the sample size increases. Thus, the possible sampling error decreases as sample size increases.

Standard Error of the Mean

How should sample size influence standard error of the mean?



$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_x \approx \frac{s}{\sqrt{n}}$$

Important implication: The Standard Error of the mean remains the same as long as the population standard deviation is known and sample size remains the same.

Sampling Distribution of the Mean

If the population is normally distributed with mean μ and standard deviation σ, then the sampling distribution of the sample means is also normally distributed no matter what the sample size is.

When we know the sample mean is normal or approximately normal, and we know the population mean, μ, and population standard deviation, σ, then we can calculate a z-score for the sample mean and determine probabilities for it where:

$$Z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}$$

Sampling Distribution of the Mean Applied

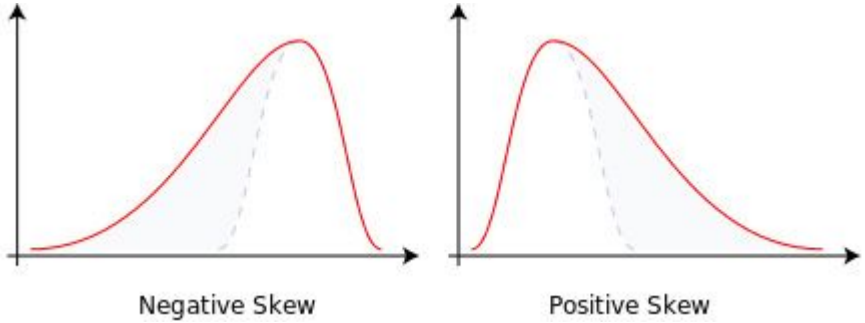
Pablo Escobar is preparing to sell a ton of cocaine to a buyer. He knows that his employees have packed the cocaine into bricks that have an average weight of 1KG and a standard deviation of .03 kg.

A potential buyer plans on taking a sample of 5 bricks and weighing them. He will not purchase the shipment of coke if the sample mean is below 0.95 kg. What is the probability the buyer will back out of the deal? $P(\bar{y} < .95) = ?$

$$Z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$$

Central Limit Theorem

What happens when the sample comes from a population that is not normally distributed? This is where the Central Limit Theorem comes in.



For a large sample size (we will explain this later), $\bar{x}$ is approximately normally distributed, regardless of the distribution of the population one samples from. If the population has mean μ and standard deviation σ, then $\bar{x}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Central Limit Theorem

Demonstration: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Notes on the CLT for this demonstration:

- If the population is skewed and sample size small, then the sample mean won't be normal.
- When doing a simulation, one replicates the process many times. Using 10,000 replications is good..
- If the population is normal, then the distribution of sample mean looks normal even if $n=2$
- If the population is skewed, then the distribution of sample mean looks more and more normal when n gets larger.
- Note that in all cases, the mean of the sample mean is close to the population mean and the standard error of the sample mean is close to

Central Limit Theorem

Sampling Distribution of the Sample Mean

The sampling distribution of the sample mean will have:

- the same mean as the population mean, μ
- Standard deviation [standard error] of $\sigma/\sqrt{n}$

It will be Normal (or approximately Normal) if either of these conditions is satisfied

- The population distribution is Normal
- The sample size is large (greater than 30).

Sampling Part 2

Data Science Immersive

Confidence Intervals

Confidence Interval
An interval of values computed from sample data that is likely to cover the true parameter of interest.

General form of a confidence interval

$$sample\ statistic \pm margin\ of\ error$$

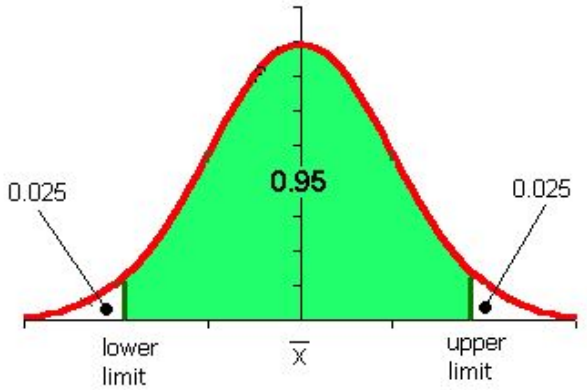
Interpretation of a Confidence Interval

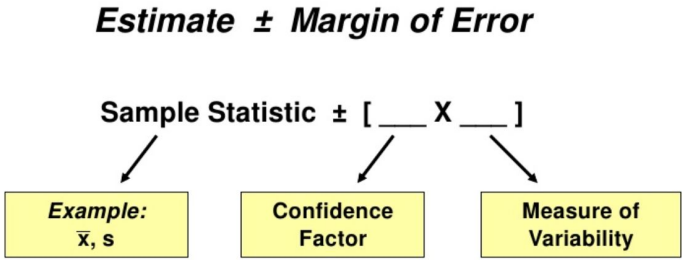
The interpretation of a confidence interval has the basic template of: "We are 'some level of percent confident' that the 'population of interest' is from 'lower bound to upper bound'. The phrases in single quotes are replaced with the specific language of the problem.

Confidence Intervals

Our level of confidence that if we obtained a sample of equal size through the same process, our sample would contain the population mean.

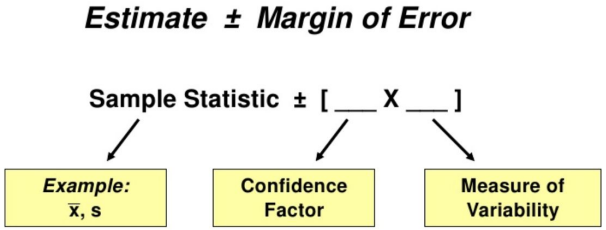
IT IS NOT: The % chance the population mean lies within our sample interval. (Many people will say this!)





Confidence Intervals

Assuming a 95% confidence interval….



If we know population variance

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

If we do not know population variance

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

If we have a small sample size (generally n < 100)

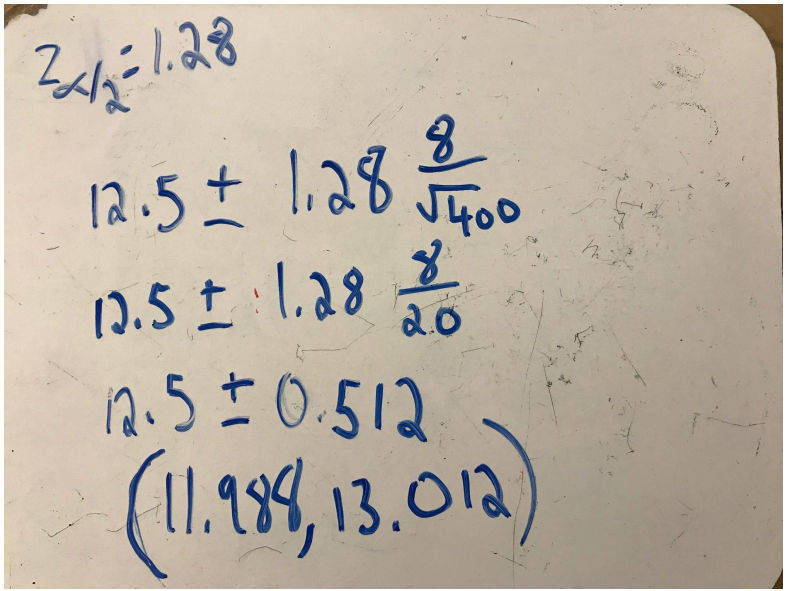
$$\bar{x} \pm t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

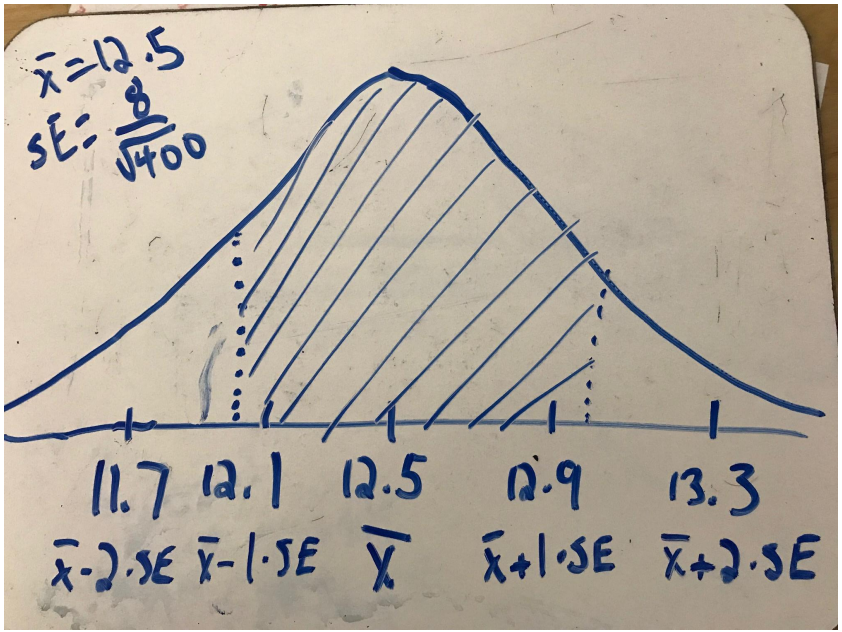
Confidence Interval Citi Bike Example

Imagine you take a sample of 400 Citi Bike cyclists and determine that their average time is 12.5 minutes with a standard deviation of 8 minutes. What is the 80% confidence interval for this sample?

Confidence Interval Citi Bike Example

Imagine you take a sample of 400 Citi Bike cyclists and determine that they average time is 12.5 minutes with a standard deviation of 8 minutes. What is the 80% confidence interval for this sample?





Confidence Interval Practice Problem

Suppose we want to estimate the average weight of an adult male in Dekalb County, Georgia. We draw a random sample of 1,000 men from a population of 1,000,000 men and weigh them. We find that the average man in our sample weighs 180 pounds, and the standard deviation of the sample is 30 pounds. What is the 95% confidence interval.

When Normal Distribution Breaks Down

When performing an experiment, we can assume that the central theorem holds and therefore can assume a normal distribution of your sample *if*

- The population standard deviation is known
- The sample size is greater than 100

If these conditions do not hold……..

You can use the T-Distribution!!

“Student’s” T-Distribution

- William Sealy Gosset, a statistician at Guinness Brewing Company, was running experiments to determine the highest yielding strains of barley
- Published a paper detailing the t-distribution under the pseudonym “Student” because Guinness had a policy that its employees could not publish research

T-Distribution

When performing an experiment, we can assume that the central theorem holds and therefore can assume a normal distribution of your sample *if*

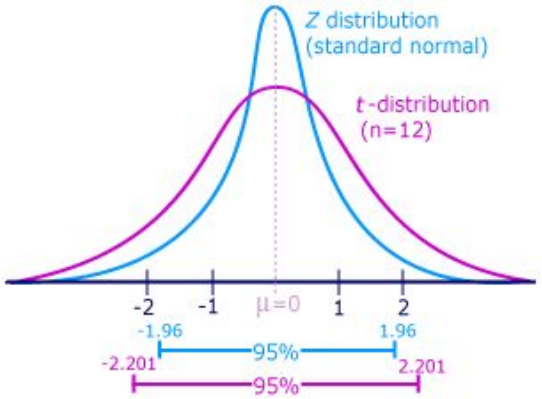
- The population standard deviation is known
- The sample size is greater than 100

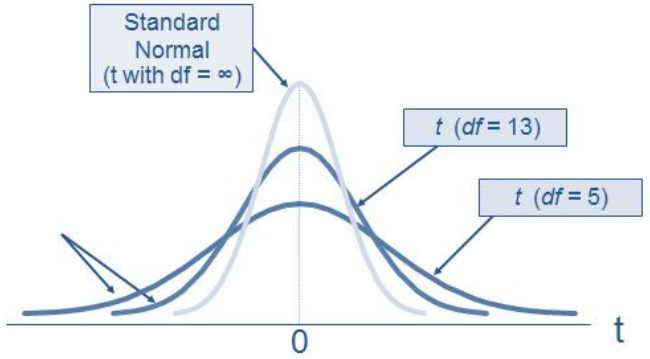
However, if neither of these conditions hold true, we need to account for the greater uncertainty, by using the t-distribution family

DF = degrees of freedom = *n-1*

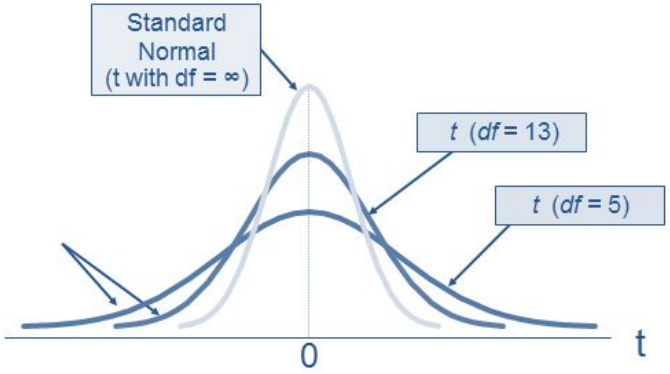
[Interactive T-Distribution]

What happens to the shape of our t-distribution as our sample size increases?





T-Distribution



PDF:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Parameters: : $\nu > 0$ where ν is degrees of freedom (n-1)

T-Score v. Z-Score

95% DISTRIBUTION COMPARISON

Z-distribution, ± 1.96

Student's t-distribution

n	df	Interval
10	9	± 2.262
30	29	± 2.045
75	74	± 1.993
100	99	± 1.984

Brief Deviation

When calculating for the sample standard deviation or variance, you must divide by the degrees of freedom instead of N.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

DF = N - 1

This is due to something called [Bessel's Correction]

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

Confidence Intervals

Confidence Interval
An interval of values computed from sample data that is likely to cover the true parameter of interest.

General form of a confidence interval

$$sample\ statistic \pm margin\ of\ error$$

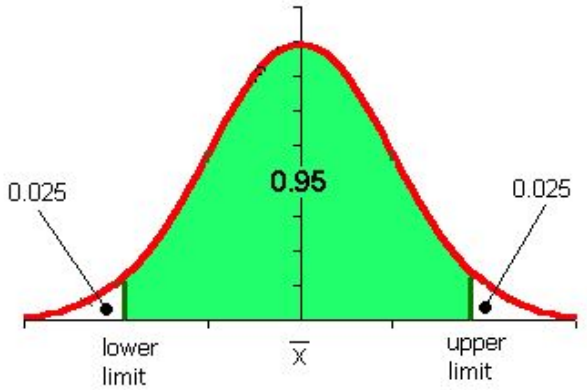
Interpretation of a Confidence Interval

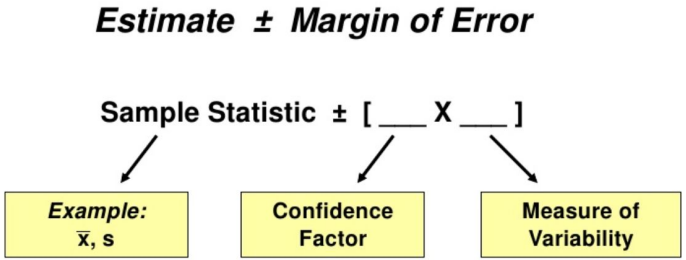
The interpretation of a confidence interval has the basic template of: "We are 'some level of percent confident' that the 'population of interest' is from 'lower bound to upper bound'. The phrases in single quotes are replaced with the specific language of the problem.

Confidence Intervals

Our level of confidence that if we obtained a sample of equal size through the same process, our sample would contain the population mean.

IT IS NOT: The % chance the population mean lies within our sample interval. (Many people will say this!)





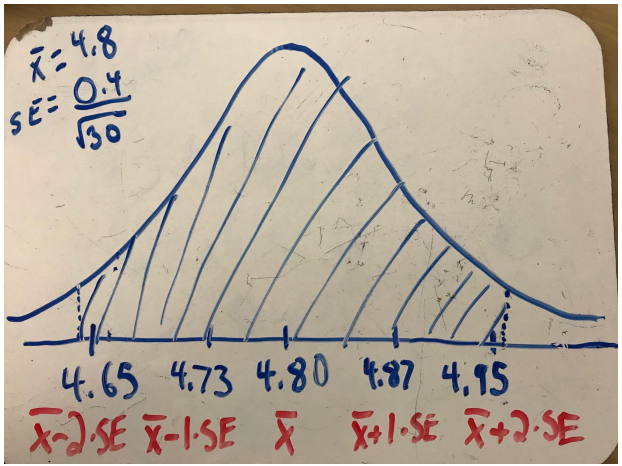
Confidence Interval Factory Example

You are inspecting a hardware factory and want to construct a 95% confidence interval of acceptable screw lengths. You draw a sample of 30 screws and calculate their mean length as 4.8 centimeters and the standard deviation as 0.4 centimeters. Calculate the bounds of your confidence interval?

\

Confidence Interval Factory Example

You are inspecting a hardware factory and want to construct a 90% confidence interval of acceptable screw lengths. You draw a sample of 30 screws and calculate their mean length as 4.8 centimeters and the standard deviation as 0.4 centimeters. Calculate the bounds of your confidence interval?



Confidence Interval Factory Example

You are inspecting a hardware factory and want to construct a 90% confidence interval of acceptable screw lengths. You draw a sample of 30 screws and calculate their mean length as 4.8 centimeters and the standard deviation as 0.4 centimeters. Caculate the bounds of your confidence interval?

```
1 import scipy.stats as scs
2 n = 30
3 mean = 4.8
4 t_value = scs.t.ppf(0.95,n-1)
5 margin_error = t_value* 0.4/(n**0.5)
6 confidence_interval = (mean - margin_error, mean + margin_error)
```

```
In [2]: 1 confidence_interval

Out[2]: (4.6759133066001235, 4.924086693399876)
```

Sampling Distribution of the Sample Proportion

What if we don't have continuous data?

Sample proportion distribution approximates a normal distribution under following 2 conditions:

$$np \geq 10$$
$$n(1-p) \geq 10$$

Then, if this set of two conditions are satisfied, sample proportion is approximately normal, with mean = p and a standard error of

$$\sqrt{\frac{p(1-p)}{n}}$$

We can then convert a sample proportion to a z-score using the following formula:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Sampling Distribution of the Sample Proportion

Trump declares that his polling numbers are above 50%, so a polling firm conducts a survey to try and see if this is true. The firm polled 75 people and found that 46% of them approved of Trump. What is the probability that Trump's true approval rating is above 50%?

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Confidence Intervals for Population Proportion

- Find the *Zα/2* multiplier for the level of confidence.
- For 95% confidence, the alpha value is 5% or 0.05 The multiplier would be a z-value with α/2, or 0.025 area to the right of it.
- Examining the standard normal table, we find that this corresponds to a z-value of 1.96

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

Confidence Intervals for Population Proportion

A random sample of 1500 U.S. adults is taken. They are asked whether they approve or disapprove of the current president's performance so far (i.e. an approval rating). Of the 1500 surveyed, 660 respond with "approve". Calculate a 95% confidence interval for the overall approval rating of the the president.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} =$$

Confidence Intervals for Population Proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

A random sample of Flatiron students is taken. They are asked whether they approve or disapprove of the NYC mayor's performance so far (i.e. an approval rating). Of the 200 surveyed, 90 respond with "approve". Calculate a 90% confidence interval for the overall approval rating of the mayor.

Any Questions?