Sampling

Data Science Immersive

FLATIRON SCHOOL

Outline

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

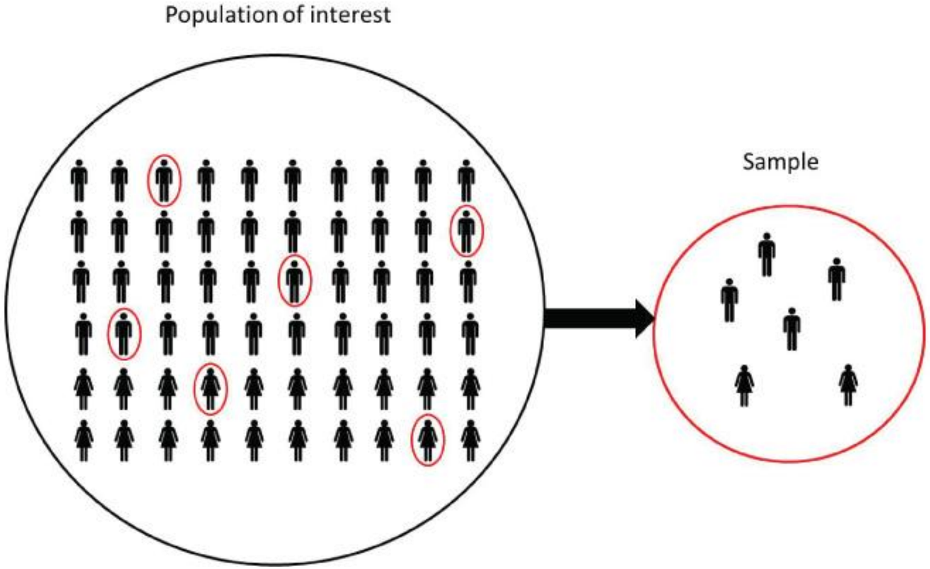
Flatiron Commission for Public Safety

The mayor's office has hired Flatiron Data Science Immersive students to determine a way to fix traffic congestion and prevent pedestrian-car collisions. To get started, we decide to determine percentage of New Yorkers own a car.

How should we go about doing this?

Flatiron Commission for Public Safety

We need to gather a sample!



Gathering a Sample

What are some approaches to acquiring a sample?

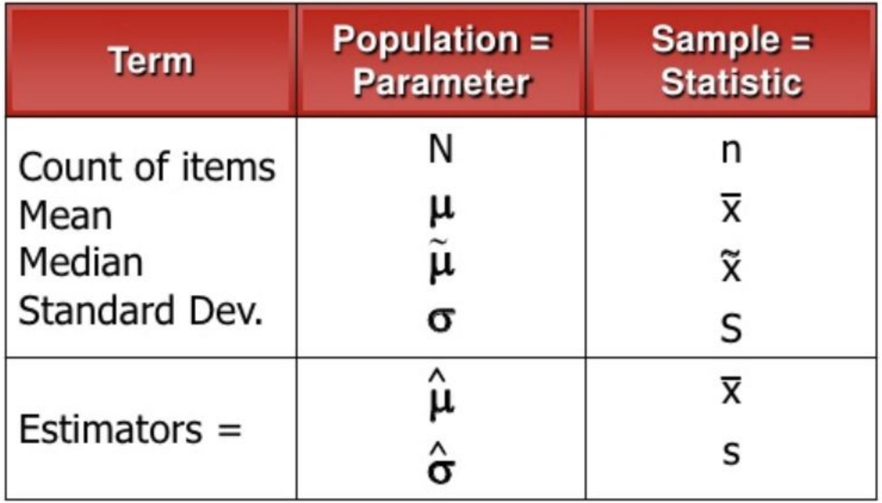
- Stand outside Flatiron and ask random people until *n* responses
- Go to a randomly assigned street corner at a random time and ask *n* people if they own a car
- Go to multiple randomly assigned street corners at random times and *n* people if they own a car

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 =	$\hat{\mu}$ $\hat{\sigma}$	$\bar{x}$ 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?

Central Limit Theorem

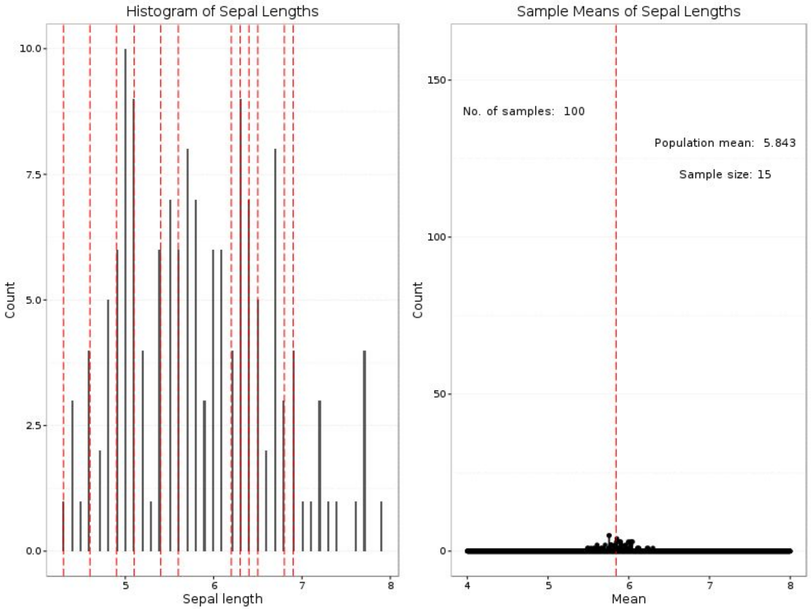
If we take repeated samples of a population, the sampling distribution of sample means will approximate to a normal distribution!

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

as n → “large”

Let’s look at an example using real data….

https://github.com/learn-co-students/nyc-mhtn-ds-102218-lectures/blob/master/week-6/2-sampling.ipynb



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.

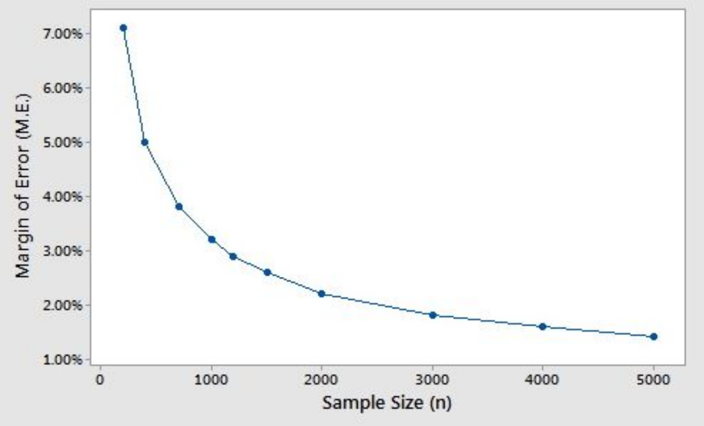
$$\sigma_x \approx \frac{s}{\sqrt{n}}$$

- s = sample standard deviation

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.

Brief Deviation

When calculating for the sample standard deviation or variance, you must divide by

N - 1 rather than N.

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

This is due to something called Bessel's Correction

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{N-1}$$

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

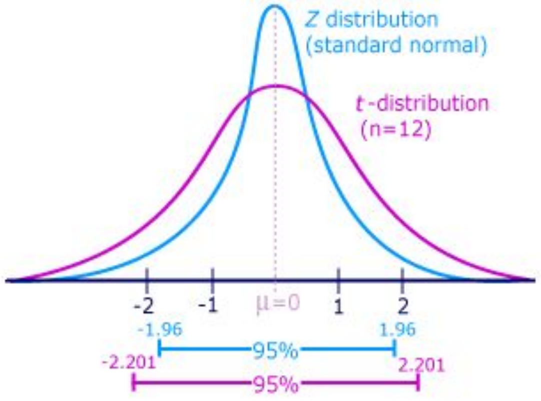
T-Distribution

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

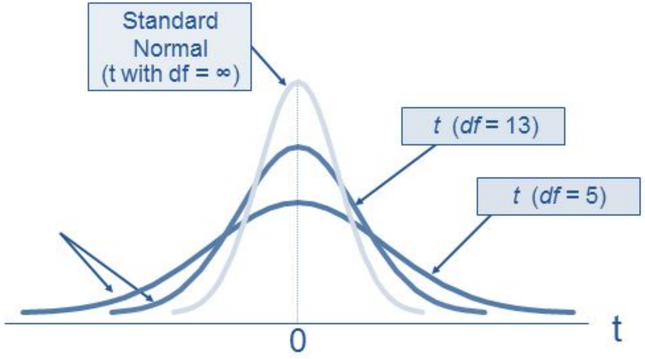
- The population standard deviation is known
- The sample size is greater than 100

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

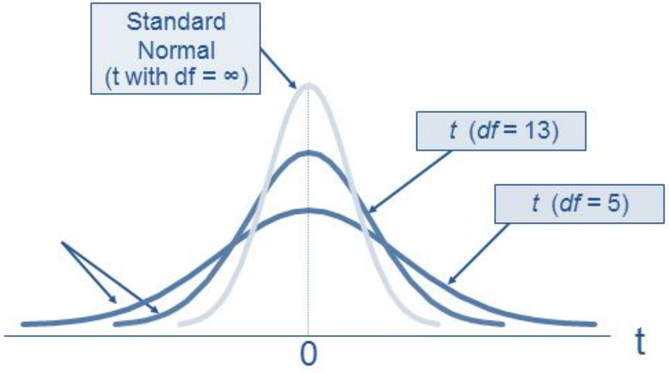
[Interactive T-Distribution]()



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



T-Distribution

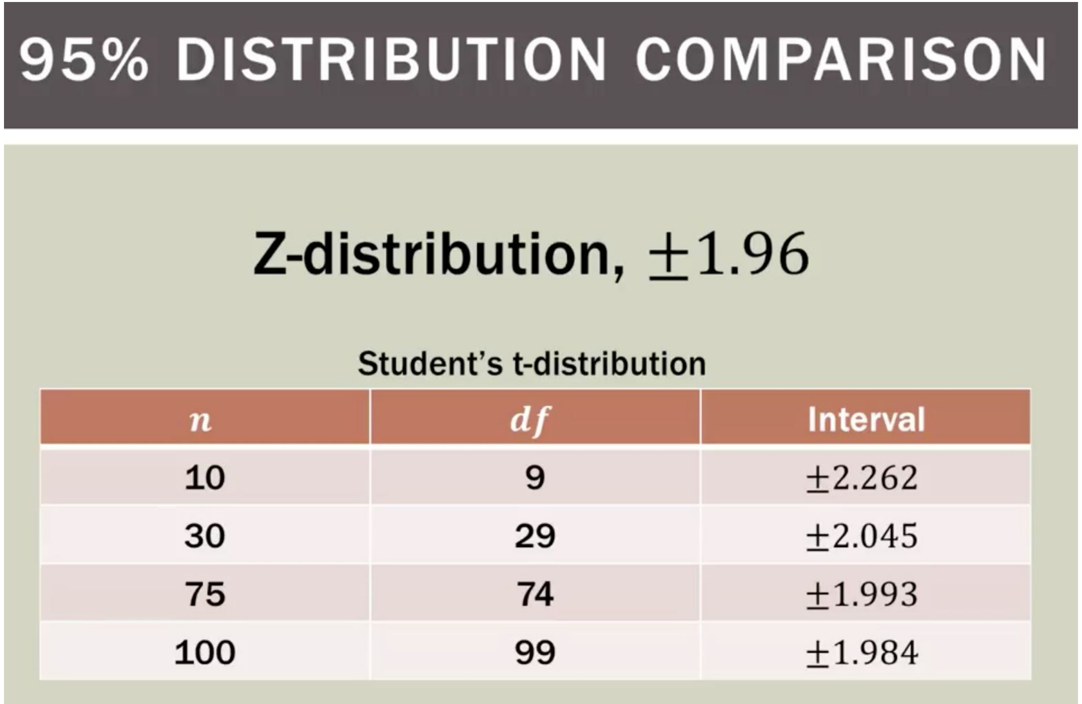


PDF:

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

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

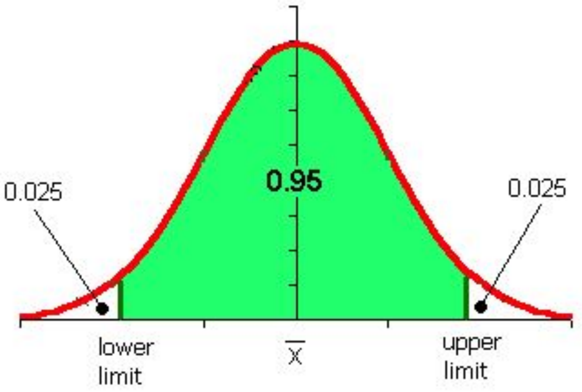
n	df	Interval
10	9	±2.262
30	29	±2.045
75	74	±1.993
100	99	±1.984

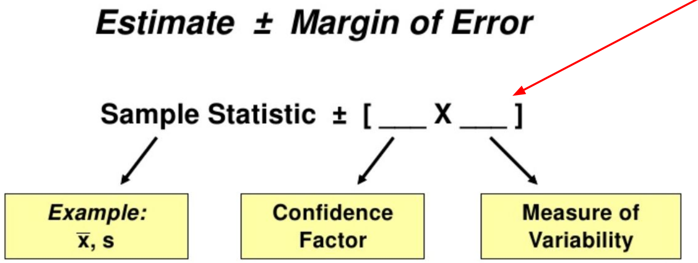
Confidence Intervals

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!)

Any guesses what this measure will be???





Estimate ± Margin of Error

Sample Statistic ± [___ X ___]

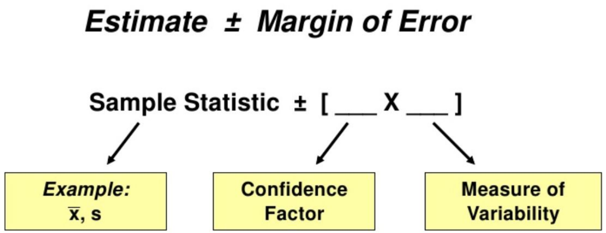
Example: **$\overline{x}$, s**

Confidence Factor

Measure of Variability

Confidence Intervals

Assuming a 95% confidence interval….



Estimate ± Margin of Error

Sample Statistic ± [___ X ___]

Example: **$\bar{x}$, s**

Confidence Factor

Measure of Variability

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)

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

What would $t_{alpha/2}$ be if we had a sample size of 25?

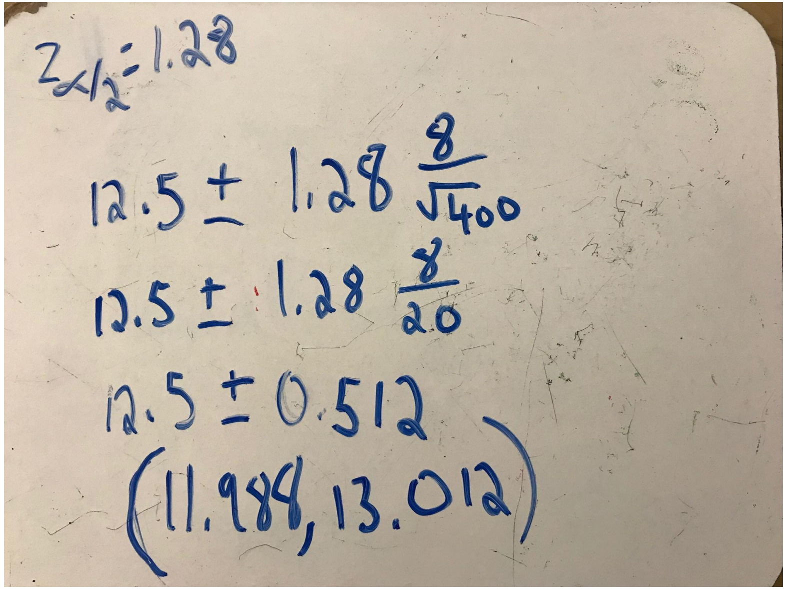
Confidence Interval Citi Bike Example

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?

Hint: Look at using scipy.stats.norm

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?



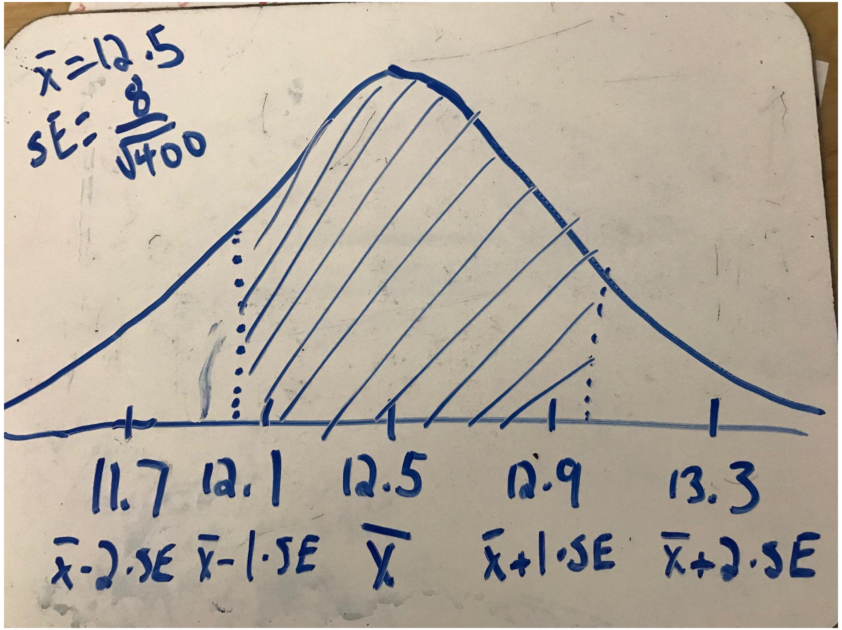
$Z_{\alpha/2} = 1.28$

$12.5 \pm 1.28 \frac{8}{\sqrt{400}}$

$12.5 \pm 1.28 \frac{8}{20}$

12.5 ± 0.512

$(11.988, 13.012)$



$\bar{x} = 12.5$

$SE = \frac{8}{\sqrt{400}}$

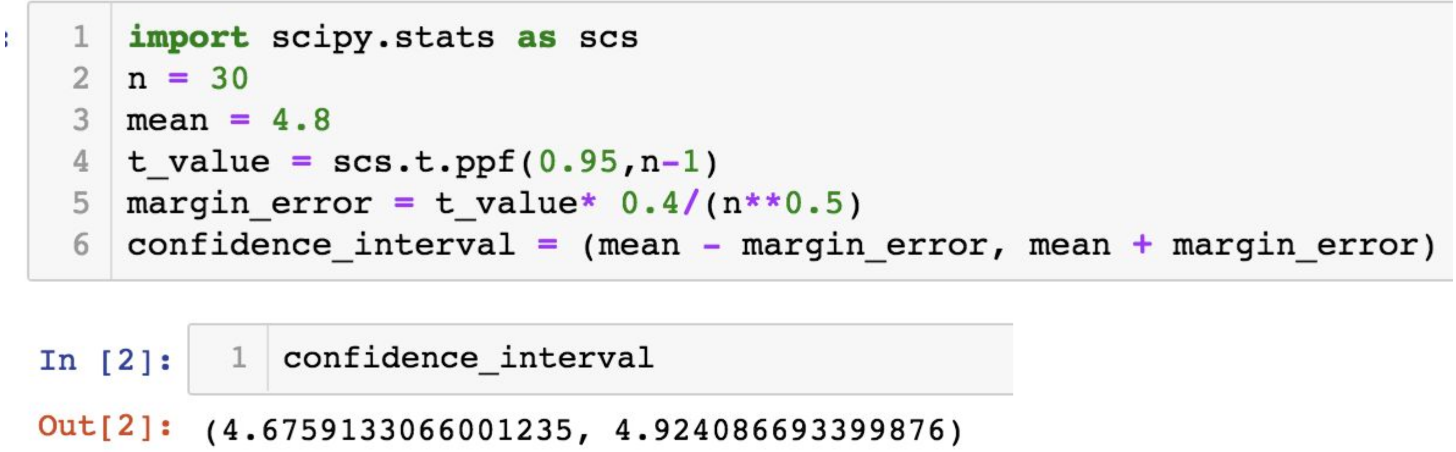
Confidence Interval Factory Example

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. What are the bounds of your confidence interval? Draw results on a sampling distribution

Hint: Look at using scipy.stats.t

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. What are the bounds of your confidence interval? Draw results on a sampling distribution



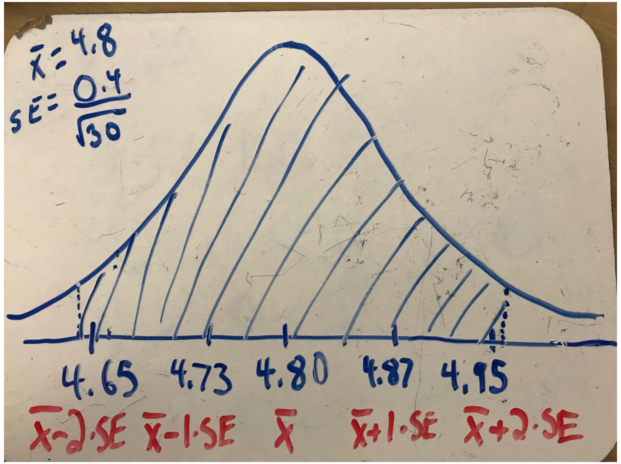
```
import scipy.stats as scs
n = 30
mean = 4.8
t_value = scs.t.ppf(0.95,n-1)
margin_error = t_value* 0.4/(n**0.5)
confidence_interval = (mean - margin_error, mean + margin_error)
```

```
In [2]:  confidence_interval

Out[2]: (4.675913306601235, 4.924086693399876)
```

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. What are the bounds of your confidence interval? Draw results on a sampling distribution.



Any Questions?