Two Main Ways to Conduct Frequentist Inference Review

Goal of Inference: Use a random sample drawn from a population to estimate some property of the population.

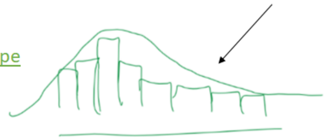
Most Common: Use a sample statistic (calculate from a random sample of size n drawn from the population) to estimate a population parameter.

	Confidence Intervals	Hypothesis Testing
What is it?	• A plausible range of values for the population parameter. • Centered around the ____________. • It's width is __________________.	Set up two competing claims about a population parameter θ (where θ_0, **the null value,** is the value that is involved in the claim). **Null Hypothesis** (status quo claim) $H_0 : \theta = \theta_0$ **Alternative Hypotheses** (claim you're trying to test) $H_A : \theta \neq \theta_0$
When you can use it?	When the sampling distribution (ie. the distribution of sample statistics drawn from samples of size n) is approximately normal.	When the sampling distribution (ie. the distribution of sample statistics drawn from samples of size n) is approximately normal.
General Format	$(sample\ statistic) \pm z_{1-\frac{\alpha}{2}}(standard\ error)$	See notes below.
Why does this work?	See notes below.	See notes below.
How to interpret.	"We are (1-α)*100% confident that the **population parameter** is between the lower bound and upper bound of the confidence interval."	See notes below.
What does "(1-α)*100% confident" mean?	If we took many many random samples all of the same size as the one we actually collected and calculated a confidence interval then we would expect that (1-α)*100% of these confidence intervals would contain the **population parameter.**	N/A

Recap of the relationship between a.) the population distribution, b.) the sample distribution, and c.) the sampling distribution of sample means.

Shapes mirror each other

Population Distribution Shape



X = enrollment

Sample Distribution Shape

X = enrollment

Population Distribution

Population Mean
μ

Population Standard Deviation
σ

	course	section	enrolled
0	adv307	A	37
1	badm210	A	215
2	badm210	B	178
3	badm210	C	197
4	cs105	A	345
5	cs105	B	201
6	stat107	A	197
7	stat207	A	53

Randomly sampled from population

Sample Distribution of Size n
(this is the sample we ACTUALLY collect)

Random Sample of n Course Enrollments *(drawn with replacement from population)*
197
37
345
201
178
37
53
201
201
197

Sample Mean
$\bar{x}$

Sample Standard Deviation
s

Randomly sampled from population

Randomly sampled from population

Randomly sampled from population

Randomly sampled from population

Random Sample of n Course Enrollments *(drawn with replacement from population)*	**Random Sample of n Course Enrollments** *(drawn with replacement from population)*	**Random Sample of n Course Enrollments** *(drawn with replacement from population)*	...	**Random Sample of n Course Enrollments** *(drawn with replacement from population)*
197	215	53	...	215
37	215	53	...	197
345	53	197	...	37
201	201	53	...	215
178	53	197	...	197
37	345	197	...	197
53	197	178	...	345
201	37	215	...	345
201	201	197	...	197
197	37	197	...	201
$\bar{x}$	$\bar{x}$	$\bar{x}$		$\bar{x}$

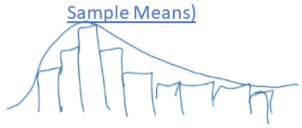
Sampling Distribution of Sample Means
(from random samples of size n)

Sample Means	164.7	155.4	153.7	...	214.6

Sampling Distribution Mean
(aka: Mean of Many, Many Sample Means)
μ

Sampling Standard Deviation (also called **Standard Error)**
(aka: Standard Deviation of Many, Many Sample Means)
$\frac{\sigma}{\sqrt{n}}$

Sampling Distribution Shape (ie. Shape of the Distribution of Sample Means)



$\bar{X}$= **Average** enrollment
(from a random sample of size n from the population)

Vs.

$\bar{X}$= **Average** enrollment
(from a random sample of size n from the population)

Shape is normal when:

- n>30 OR
- Population distribution is normal (ie. X~N(mean=μ, standard deviation=σ)

General Confidence Interval Framework and Example Creating a Confidence Interval for a Population Mean

The Theory Behind a Confidence Interval for Certain Population Parameters

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
   θ

   b) What sample statistic should be used as an estimate for this population parameter?
   $\widehat{\theta_0}$

   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
   $\widehat{\boldsymbol{\theta}}$

2. **What you have**
   $\widehat{\theta_0}$
   And other information from the random sample

3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) **Mean** of these sample statistics (ie. What is E[$\widehat{\boldsymbol{\theta}}$]?).

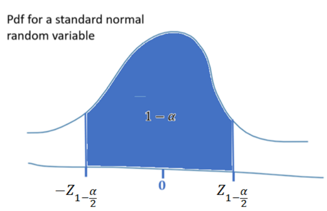
   b) **Standard deviation** of these sample statistics (ie What is SD[$\widehat{\boldsymbol{\theta}}$]?).).

   c) When is this **sampling distribution of these sample statistics normal**? (ie. when is $\widehat{\boldsymbol{\theta}}$ a normal random variable)?

4. ***When* $\widehat{\boldsymbol{\theta}}$ is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $\boldsymbol{Z} = \frac{\widehat{\theta} - E[\widehat{\theta}]}{\mathrm{SD}[\widehat{\theta}]}$ (ie. the **z-score of** $\widehat{\boldsymbol{\theta}}$ is a **standard normal random variable.**
<u>Aka</u>: $Z \sim \mathrm{N}(\mathrm{mean} = 0, \mathrm{standard\ deviation} = 1)$

b.) We know how to find $\boldsymbol{Z_{1-\frac{\alpha}{2}}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$\boldsymbol{Z} = \frac{\widehat{\boldsymbol{\theta}} - \boldsymbol{E}[\widehat{\boldsymbol{\theta}}]}{\mathbf{SD}[\widehat{\boldsymbol{\theta}}]}$$

c.) The area shaded under this pdf curve is the following probability.

$$P\left(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\widehat{\theta} - E[\widehat{\theta}]}{\mathrm{SD}[\widehat{\theta}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(\widehat{\boldsymbol{\theta}} - Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\widehat{\boldsymbol{\theta}}] < \boldsymbol{E}[\widehat{\boldsymbol{\theta}}] < \widehat{\boldsymbol{\theta}} + Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\widehat{\boldsymbol{\theta}}]\right) = 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α confidence interval equation:

$$\widehat{\theta_0} \pm Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\widehat{\boldsymbol{\theta}}]$$

<u>Ex</u>: Create a 90% confidence interval for the average number of hours UIUC students slept last night. Suppose we collect a random sample of size n=40 from the UIUC population that has a mean number of sleep hours of 7 and a standard deviation of 3.

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
   μ=mean number of hours all UIUC students slept last night
   b) What sample statistic should be used as an estimate for this population parameter?
   $\bar{x} = 7$

   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
   $\bar{X}$

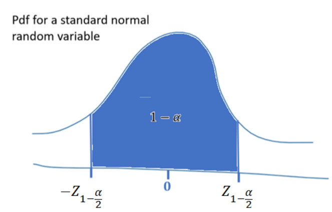
2. **What you have**
   - $\bar{x} = 7$, which is an instance of $\bar{X}$
   - n = 40
   - s = 3
3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) **Mean** of these sample statistics (ie. What is E[$\bar{X}$]?).
   $\mathrm{E}[\bar{X}] = \mu$

   b) **Standard deviation** of these sample statistics (ie What is SD[$\bar{X}$]?).).
   $\mathrm{SD}[\bar{X}] = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

   c) When is this **sampling distribution of these sample statistics normal**? (ie. when is $\bar{X}$ a normal random variable)?
   <u>Either when:</u>
   - When n>30 OR
   - When the population distribution (or equivalently the sample distribution) is normal.
4. ***When* $\bar{X}$ is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $\boldsymbol{Z} = \frac{\bar{X} - E[\bar{X}]}{\mathrm{SD}[\bar{X}]} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$(ie. the **z-score of** $\bar{X}$ a **standard normal random variable.**)
<u>Aka</u>: $Z \sim \mathrm{N}(\mathrm{mean} = 0, \mathrm{standard\ deviation} = 1)$

b.) We know how to find $\boldsymbol{Z_{1-\frac{\alpha}{2}}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$\boldsymbol{Z} = \frac{\bar{X} - \boldsymbol{E}[\bar{X}]}{\mathbf{SD}[\bar{X}]} = \frac{\bar{X} - \boldsymbol{\mu}}{\frac{\sigma}{\sqrt{n}}}$$

c.) The area shaded under this pdf curve is the following probability.

$$P\left(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - E[\bar{X}]}{\mathrm{SD}[\bar{X}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\bar{X}] < \boldsymbol{E}[\bar{X}] < \bar{X} + Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\bar{X}]\right) = 1-\alpha$$
$$P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \boldsymbol{\mu} < \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α=90% confidence interval equation:

$$\bar{x} \pm Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$
$$\approx \bar{x} \pm Z_{1-\frac{0.10}{2}}\frac{s}{\sqrt{n}}$$
$$7 \pm Z_{0.95}\frac{3}{\sqrt{40}}$$
norm.ppf(0.95)
$$7 \pm 1.645\frac{3}{\sqrt{40}}$$
(6.2, 7.78)

We are 90% confident that the average time ALL UIUC students spent sleeping last night (ie. μ) is between 6.2 and 7.78 hours.

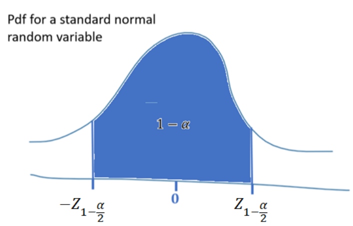
General Confidence Interval Framework and Example Creating a Confidence Interval for a Population Proportion

The Theory Behind a Confidence Interval for Certain Population Parameters

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
   θ
   b) What sample statistic should be used as an estimate for this population parameter?
   $\widehat{\theta_0}$
   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
   $\widehat{\boldsymbol{\theta}}$
2. **What you have**
   $\widehat{\theta_0}$
   And other information from the random sample
3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) Mean of these sample statistics (ie. What is E[$\widehat{\boldsymbol{\theta}}$]?).
   b) Standard deviation of these sample statistics (ie What is SD[$\widehat{\boldsymbol{\theta}}$]?).).
   c) When is this **sampling distribution of these sample statistics normal**? (ie. when is $\widehat{\boldsymbol{\theta}}$ a normal random variable)?
4. ***When $\widehat{\boldsymbol{\theta}}$* is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $\boldsymbol{Z} = \frac{\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}]}{\mathbf{SD}[\widehat{\boldsymbol{\theta}}]}$ (ie. the **z-score of $\widehat{\boldsymbol{\theta}}$** is a **standard normal random variable.**
Aka: $Z \sim N(\text{mean} = 0, \text{standard deviation} = 1)$

b.) We know how to find $\boldsymbol{Z_{1-\frac{\alpha}{2}}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$\boldsymbol{Z} = \frac{\widehat{\boldsymbol{\theta}} - \boldsymbol{E}[\widehat{\boldsymbol{\theta}}]}{\mathbf{SD}[\widehat{\boldsymbol{\theta}}]}$$

c.) The area shaded under this pdf curve is the following probability.

$$P\left(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}]}{\mathbf{SD}[\widehat{\boldsymbol{\theta}}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\widehat{\boldsymbol{\theta}} - Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\widehat{\boldsymbol{\theta}}] < \boldsymbol{E}[\widehat{\boldsymbol{\theta}}] < \widehat{\boldsymbol{\theta}} + Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\widehat{\boldsymbol{\theta}}]\right) = 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α confidence interval equation:

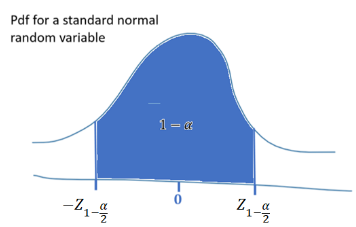
$$\widehat{\theta_0} \pm Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\widehat{\boldsymbol{\theta}}]$$

Ex: Create a 95% confidence interval for the proportion of adults living in the US that approve of Donald Trump. We have a random sample of 1503 adults living in the US, in which 38.1% approve.

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
   p = proportion of ALL adults living in the US that approve of Trump.
   b) What sample statistic should be used as an estimate for this population parameter?
   $\hat{p} = 0.381$
   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
   $\hat{p}$
2. **What you have**
   - $\hat{p} = .381$, which is an instance of the random variable $\hat{p}$
   - n = 1503
3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) Mean of these sample statistics (ie. What is E[$\hat{p}$]?).
   E[$\hat{p}$] =
   b) Standard deviation of these sample statistics (ie What is SD[$\bar{X}$]?).).
   SD[$\hat{p}$] =
   c) When is this **sampling distribution of these sample statistics normal**? (ie. when is $\hat{p}$ a normal random variable)?
   When:
   - When $np \geq 10$ and $n(1-p) \geq 10$
     (If you don't know p, then check:
     $n\hat{p} \geq 10$ *and* $n(1-\hat{p}) \geq 10$)
4. ***When $\hat{p}$* is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $\boldsymbol{Z} = \frac{\hat{p} - E[\hat{p}]}{\mathbf{SD}[\hat{p}]} =$ (ie. the **z-score of $\hat{p}$** is a **standard normal random variable.**
Aka: $Z \sim N(\text{mean} = 0, \text{standard deviation} = 1)$

b.) We know how to find $\boldsymbol{Z_{1-\frac{\alpha}{2}}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$\boldsymbol{Z} = \frac{\hat{p} - \boldsymbol{E}[\hat{p}]}{\mathbf{SD}[\hat{p}]} =$$

c.) The area shaded under this pdf curve is the following probability.

$$P\left(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\hat{p} - E[\hat{p}]}{\mathbf{SD}[\hat{p}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\hat{p} - Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\hat{p}] < \boldsymbol{E}[\hat{p}] < \hat{p} + Z_{1-\frac{\alpha}{2}}\mathbf{SD}[\hat{p}]\right) = 1-\alpha$$

$$= 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α=95% confidence interval equation:

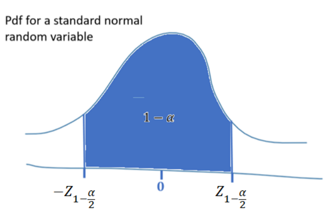
General Confidence Interval Framework and Example Creating a Confidence Interval for the Difference Between Two Population Means

The Theory Behind a Confidence Interval for Certain Population Parameters

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
      θ
   b) What sample statistic should be used as an estimate for this population parameter?
      $\widehat{\theta_0}$
   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
      $\widehat{\boldsymbol{\theta}}$
2. **What you have**
   $\widehat{\theta_0}$
   And other information from the random sample
3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) **Mean** of these sample statistics (ie. What is E[$\widehat{\boldsymbol{\theta}}$]?).
   b) **Standard deviation** of these sample statistics (ie What is SD[$\widehat{\boldsymbol{\theta}}$]?).).
   c) When is this **sampling distribution of these sample statistics normal**? (ie. when is $\widehat{\boldsymbol{\theta}}$ a normal random variable)?
4. ***When* $\widehat{\boldsymbol{\theta}}$ is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $Z = \frac{\widehat{\theta}-E[\widehat{\theta}]}{\mathrm{SD}[\widehat{\theta}]}$ (ie. the **z-score of** $\widehat{\boldsymbol{\theta}}$ is a **standard normal random variable.**
Aka: $Z \sim N(\text{mean} = 0, \text{standard deviation} = 1)$

b.) We know how to find $Z_{1-\frac{\alpha}{2}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$Z = \frac{\widehat{\theta} - E[\widehat{\theta}]}{\mathrm{SD}[\widehat{\theta}]}$$

c.) The area shaded under this pdf curve is the following probability.

$$P(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}) = 1-\alpha$$
$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\widehat{\theta}-E[\widehat{\theta}]}{\mathrm{SD}[\widehat{\theta}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(\widehat{\theta} - Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\theta}] < E[\widehat{\theta}] < \widehat{\theta} + Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\theta}]\right) = 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α confidence interval equation:

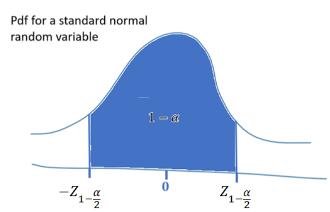
$$\widehat{\theta_0} \pm Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\theta}]$$

Ex: Create a 95% confidence interval for the difference in the average AGE of ALL children exposed to low levels of lead and the AGE of ALL children exposed to high levels of lead.

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
      $\mu_{lo} - \mu_{hi}$
   b) What sample statistic should be used as an estimate for this population parameter?
      $\bar{x}_{lo} - \bar{x}_{hi}$
   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
      $$\bar{X}_{lo} - \bar{X}_{hi}$$
2. **What you have**
   - $\bar{x}_{lo} - \bar{x}_{hi} = 9.33 - 8.27 = 1.06$., which is an instance of $\bar{X}_{lo} - \bar{X}_{hi}$
   - $n_{lo} = 78, n_{hi} = 46$
   - $s_{lo} = 0.404, s_{hi} = 0.503$
3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) **Mean** of these sample statistics (ie. What is E[$\bar{X}_{lo} - \bar{X}_{hi}$]?).
      E[$\bar{X}_{lo} - \bar{X}_{hi}$] =
   b) **Standard deviation** of these sample statistics (ie What is SD[$\bar{X}_{lo} - \bar{X}_{hi}$])?
      SD[$\bar{X}_{lo} - \bar{X}_{hi}$] =
   c) When is this **sampling distribution of these sample statistics normal**? (ie. when is $\bar{X}_{lo} - \bar{X}_{hi}$ a normal random variable)?
      Either when:
      - When n>30 OR
      - When the population distribution (or equivalently the sample distribution) is normal.
4. ***When* $\bar{X}_{lo} - \bar{X}_{hi}$ is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $Z = \frac{(\bar{X}_{lo}-\bar{X}_{hi})-E[\bar{X}_{lo}-\bar{X}_{hi}]}{\mathrm{SD}[\bar{X}_{lo}-\bar{X}_{hi}]} =$
(ie. the **z-score of** $\bar{X}_{lo} - \bar{X}_{hi}$ a **standard normal random variable.**
Aka: $Z \sim N(\text{mean} = 0, \text{standard deviation} = 1)$

b.) We know how to find $Z_{1-\frac{\alpha}{2}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$Z = \frac{(\bar{X}_{lo}-\bar{X}_{hi}) - E[\bar{X}_{lo} - \bar{X}_{hi}]}{\mathrm{SD}[\bar{X}_{lo} - \bar{X}_{hi}]}$$

c.) The area shaded under this pdf curve is the following probability.

$$P(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}) = 1-\alpha$$
$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{(\bar{X}_{lo}-\bar{X}_{hi})-E[\bar{X}_{lo}-\bar{X}_{hi}]}{\mathrm{SD}[\bar{X}_{lo}-\bar{X}_{hi}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\bar{X}_{lo} - \bar{X}_{hi}]E[\bar{X}_{lo} - \bar{X}_{hi}] < \bar{X} + Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\bar{X}_{lo} - \bar{X}_{hi}]\right) = 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α=95% confidence interval equation:

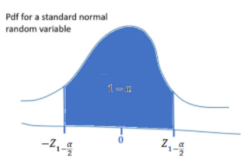
General Confidence Interval Framework and Example Creating a Confidence Interval for the Difference Between Two Population Proportions

The Theory Behind a Confidence Interval for Certain Population Parameters

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
   θ
   b) What sample statistic should be used as an estimate for this population parameter?
   $\widehat{\theta_0}$
   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
   $\widehat{\boldsymbol{\theta}}$
2. **What you have**
   $\widehat{\theta_0}$
   And other information from the random sample
3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) **Mean** of these sample statistics (ie. What is E[$\widehat{\boldsymbol{\theta}}$]?).
   b) **Standard deviation** of these sample statistics (ie What is SD[$\widehat{\boldsymbol{\theta}}$]?).).
   c) When is this **sampling distribution of these sample statistics normal**? (ie. when is $\widehat{\boldsymbol{\theta}}$ a normal random variable)?
4. ***When*** $\widehat{\boldsymbol{\theta}}$ **is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $Z = \frac{\widehat{\theta} - E[\widehat{\theta}]}{SD[\widehat{\theta}]}$ (ie. the **z-score of** $\widehat{\theta}$ is a **standard normal random variable.**
Aka: $Z \sim N(\text{mean} = 0, \text{standard deviation} = 1)$

b.) We know how to find $Z_{1-\frac{\alpha}{2}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$Z = \frac{\widehat{\theta} - E[\widehat{\theta}]}{SD[\widehat{\theta}]}$$

c.) The area shaded under this pdf curve is the following probability.

$$P\left(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\widehat{\theta} - E[\widehat{\theta}]}{SD[\widehat{\theta}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(\widehat{\theta} - Z_{1-\frac{\alpha}{2}}SD[\widehat{\theta}] < E[\widehat{\theta}] < \widehat{\theta} + Z_{1-\frac{\alpha}{2}}SD[\widehat{\theta}]\right) = 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α confidence interval equation:

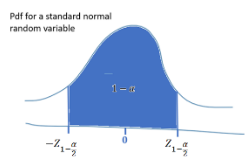
$$\widehat{\theta_0} \pm Z_{1-\frac{\alpha}{2}}SD[\widehat{\theta}]$$

Ex: Create a 90% confidence interval for DIFFERENCE in a.) the proportion of ALL adults living in the US registered as "independent" that approve of Trump and b.) the proportion of ALL adults living in the US that "have no political preference" that approve of Trump. We have a random sample of 183 who are independent and random sample of 15 with "no political preference." 34.5% in the "independent" sample approve of Trump and 36.6% in the "no political preference" sample approve of Trump.

1. **Definitions**
   a) What is the population parameter you're trying to make an inference about?
   $p_{ind} - p_{no\,pref}$
   b) What sample statistic should be used as an estimate for this population parameter?
   $\widehat{p}_{ind} - \widehat{p}_{no\,pref}$
   c) What random variable represents the experiment of randomly calculating one of these sample statistics?
   $\widehat{p}_{ind} - \widehat{p}_{no\,pref}$
2. **What you have**
   - $\widehat{p}_{ind} - \widehat{p}_{no\,pref} = .345 - .366$
   - $n_{ind} = 183, n_{ind} = 15$
3. **Information about the Sampling Distribution of These Sample Statistics** If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?
   a) **Mean** of these sample statistics (ie. What is E[$\widehat{p}_{ind} - \widehat{p}_{no\,pref}$]?).
   E[$\widehat{p}_{ind} - \widehat{p}_{no\,pref}$] =
   b) **Standard deviation** of these sample statistics (ie What is SD[$\widehat{p}_{ind} - \widehat{p}_{no\,pref}$]?).).
   SD[$\widehat{p}_{ind} - \widehat{p}_{no\,pref}$] =
   c) When is this **sampling distribution of these sample statistics normal**? (ie. when IS $\widehat{p}_{ind} - \widehat{p}_{no\,pref}$ a normal random variable)?
   When:
   - $n_{ind}p_{ind} \geq 10$ and $n_{ind}(1-p_{ind}) \geq 10$
   - AND $n_{no\,pref}p_{no\,pref} \geq 10$ and $n_{no\,pref}(1-p_{no\,pref}) \geq 10$
4. ***When*** $\widehat{p}_{ind} - \widehat{p}_{no\,pref}$ **is a normal random variable (ie. when the sampling distribution of the sample statistics is normal), this is what we know.**

a.) $Z = \frac{(\widehat{p}_{ind} - \widehat{p}_{no\,pref}) - E[\widehat{p}_{ind} - \widehat{p}_{no\,pref}]}{SD[\widehat{p}_{ind} - \widehat{p}_{no\,pref}]} =$
(ie. the **z-score of** $\widehat{p}_{ind} - \widehat{p}_{no\,pref}$ is a **standard normal random variable.**
Aka: $Z \sim N(\text{mean} = 0, \text{standard deviation} = 1)$

b.) We know how to find $Z_{1-\frac{\alpha}{2}}$, which is the x-axis value that produces a **left area of 1-α/2** under the standard normal pdf curve with in Python with:
norm.ppf(1-α/2)



$$Z = \frac{(\widehat{p}_{ind} - \widehat{p}_{no\,pref}) - E[\widehat{p}_{ind} - \widehat{p}_{no\,pref}]}{SD[\widehat{p}_{ind} - \widehat{p}_{no\,pref}]} =$$

c.) The area shaded under this pdf curve is the following probability.

$$P\left(-Z_{1-\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{(\widehat{p}_{ind} - \widehat{p}_{no\,pref}) - E[\widehat{p}_{ind} - \widehat{p}_{no\,pref}]}{SD[\widehat{p}_{ind} - \widehat{p}_{no\,pref}]} < Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left((\widehat{p}_{ind} - \widehat{p}_{no\,pref}) - Z_{1-\frac{\alpha}{2}}SD[(\widehat{p}_{ind} - \widehat{p}_{no\,pref})] < E[(\widehat{p}_{ind} - \widehat{p}_{no\,pref})] < (\widehat{p}_{ind} - \widehat{p}_{no\,pref}) + Z_{1-\frac{\alpha}{2}}SD[(\widehat{p}_{ind} - \widehat{p}_{no\,pref})]\right) = 1-\alpha$$

d.) Therefore, we are able to find (or approximate) each of these pieces in our 1-α=95% confidence interval equation:

The Theory Behind Hypothesis Testing (Two-Tailed Tests)

1. Definitions (for two tailed-hypothesis testing)

a) What is the population parameter you're trying to make an inference about?

θ

b) What sample statistic should be used as an estimate for this population parameter?

$\widehat{\theta_0}$

c) What random variable represents the experiment of randomly calculating one of these sample statistics?

$\hat{\theta}$

d) What is your null hypothesis about this population parameter? (*This hypothesis assumes the status quo, no effect, and/or nothing is happening).*

$H_0: \theta = \theta_0$

e) What is your alternative hypothesis about this population parameter? (*This hypothesis assumes the claim you are trying to test/some effect/something is happening).*

$H_A: \theta \neq \theta_0$

2. What you have:

- $\widehat{\theta_0}$ which is an instance of the random variable $\widehat{\boldsymbol{\theta}}$
- other information about the random sample

3. Information about the Sampling Distribution of These Sample Statistics

If we were to form a distribution of many, many sample statistics (each collected in the same way (ie. same sample size(s)) as the sample statistic we have, what do we know about the following?

a) **Mean** of these sample statistics (ie. What is E[$\widehat{\boldsymbol{\theta}}$]?).

b) **Standard deviation** of these sample statistics (ie What is SD[$\widehat{\boldsymbol{\theta}}$]?).).

c) When is this **sampling distribution of these sample statistics $\widehat{\boldsymbol{\theta}}$ normal**? (ie. when is $\widehat{\boldsymbol{\theta}}$ a normal random variable)?

4. Assumptions:

Let's assume that $\boldsymbol{H_0: \theta = \theta_0}$.

5. Things we Know

When** is $\widehat{\boldsymbol{\theta}}$ **is a normal random variable (ie. when the sampling distribution of the sample statistics** is $\widehat{\boldsymbol{\theta}}$ **is normal),** ***assume that $\boldsymbol{H_0: \theta = \theta_0}$**, the following is what we know.**

a.) $Z = \frac{\hat{\theta}-E[\hat{\theta}]}{\mathrm{SD}[\hat{\theta}]}$ (ie. the **z-score of** is $\hat{\theta}$ is a **standard normal random variable.**)

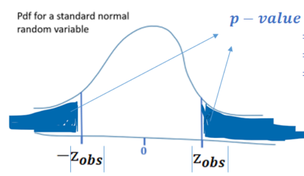
Aka: $Z \sim N(\text{mean} = 0, \text{standard deviation} = 1)$

b) We can form a 1-α confidence interval for $\boldsymbol{\theta}$ using $(\widehat{\theta_0} - Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\boldsymbol{\theta}}], \widehat{\theta_0} + Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\boldsymbol{\theta}}])$

c.) We can also calculate the z-score of the sample statistic, $\hat{\theta}$, that we observed in our sample. Because we're assuming $\boldsymbol{H_0: \theta = \theta_0}$, we can actually get a fixed number for this. We call this the **z-statistic.**

$$z_{obs} = \frac{\widehat{\theta_0} - E[\widehat{\boldsymbol{\theta}}]}{\mathrm{SD}[\widehat{\boldsymbol{\theta}}]} = \frac{\widehat{\theta_0} - \boldsymbol{\theta}}{\mathrm{SD}[\widehat{\boldsymbol{\theta}}]} = \frac{\widehat{\theta_0} - \boldsymbol{\theta_0}}{\mathrm{SD}[\widehat{\boldsymbol{\theta}}]}$$

d.) Our z_{obs} can be considered an observation randomly drawn from a standard normal curve. Therefore, we can find the probability that z_{obs} is "extreme" with respect to our alternative hypothesis. For the alternative hypothesis $\boldsymbol{H_A: \theta \neq \theta_0}$, this means we want to find the probability that our $-z_{obs}$ *and* z_{obs} are "sufficiently far away" from 0 (ie. the center) of the standard normal distribution on either side. This probability is called the **p-value.**



$$\begin{aligned} \boldsymbol{p-value} &= P(|Z| > |Z_{obs}|) \\ &= P(Z > |Z_{obs}| or\ Z < -|Z_{obs}|) \\ &= P(Z > |Z_{obs}|) + P(Z < -|Z_{obs}|) \\ &= 2P(Z < -|Z_{obs}|) \end{aligned}$$

6. Make a Conclusion About your Null and Alternative Hypotheses

$H_0: \theta = \theta_0$

$H_A: \theta \neq \theta_0$

Inference Method #1: Use a confidence interval.

If $\boldsymbol{\theta_0 \in (\widehat{\theta_0} - Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\theta}], \widehat{\theta_0} + Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\theta}])}$

- ***What this tells us:*** θ_0 is considered a "plausible value" for θ, when using a 1-α confidence interval.
- ***Therefore we say:***
  - *"We fail to reject the null hypothesis."*
  - *"There is NOT sufficient evidence to suggest the alternative hypothesis."*

Equivalent $\leftrightarrow$

If $\boldsymbol{\theta_0 \notin (\widehat{\theta_0} - Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\theta}], \widehat{\theta_0} + Z_{1-\frac{\alpha}{2}}\mathrm{SD}[\widehat{\theta}])}$

- ***What this tells us:*** θ_0 is NOT considered a "plausible value" for θ, when using a 1-α confidence interval.
- ***Therefore we say:***
  - *"We reject the null hypothesis."*
  - *"There is sufficient evidence to suggest the alternative hypothesis."*

Equivalent $\leftrightarrow$

Inference Method #2: Use the z-statistic

If $\left|\frac{\hat{\theta}-\theta_0}{SD(\hat{\theta})}\right| \leq \boldsymbol{Z_{1-\frac{\alpha}{2}}}$

- ***What this tells us:*** θ_0 is considered a "plausible value" for θ, when using a 1-α confidence interval.
- ***Therefore we say:***
  - *"We fail to reject the null hypothesis."*
  - *"There is NOT sufficient evidence to suggest the alternative hypothesis."*

Equivalent $\leftrightarrow$

If $\left|\frac{\hat{\theta}-\theta_0}{SD(\hat{\theta})}\right| > \boldsymbol{Z_{1-\frac{\alpha}{2}}}$

- ***What this tells us:*** θ_0 is NOT considered a "plausible value" for θ, when using a 1-α confidence interval.
- ***Therefore we say:***
  - *"We reject the null hypothesis."*
  - *"There is sufficient evidence to suggest the alternative hypothesis."*

Equivalent $\leftrightarrow$

Inference Method #3: Use the p-value

- ***If the p-value is not low (specifically greater than or equal to some threshold α),***
  - ***What this tells us:*** *IF we assume our null hypothesis* $\boldsymbol{H_0: \theta = \theta_0}$, *then the sample statistic that we observed* $\widehat{\theta_0}$ *would have been NOT unlikely enough to make us doubtful that our null hypothesis was actually true* $\boldsymbol{H_0: \theta = \theta_0}$.
  - ***Therefore we say:***
    - *"We fail to reject the null hypothesis."*
    - *"There is NOT sufficient evidence to suggest the alternative hypothesis."*
- ***If the p-value is low (specifically less than some threshold α),***
  - ***What this tells us:*** *IF we assume our null hypothesis* $\boldsymbol{H_0: \theta = \theta_0}$, *then the sample statistic that we observed* $\widehat{\theta_0}$ *would have been so unlikely, that this make us doubtful that our null hypothesis was actually true* $\boldsymbol{H_0: \theta = \theta_0}$.
  - ***Therefore we say:***
    - *"We reject the null hypothesis."*
    - *"There is sufficient evidence to suggest the alternative hypothesis."*