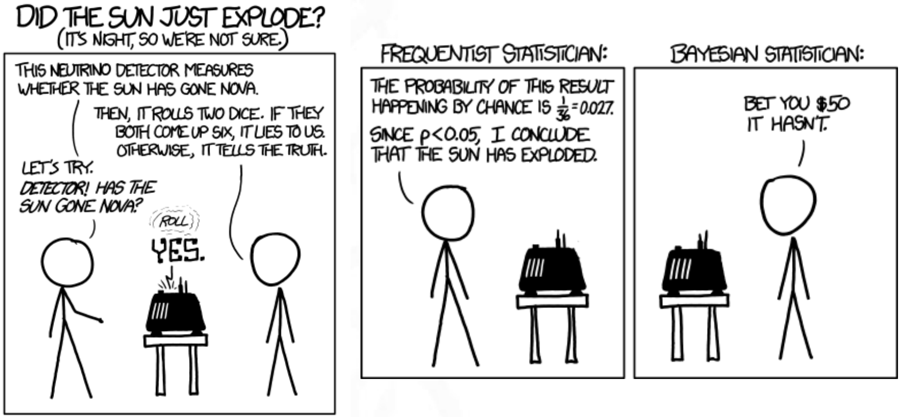
Inferential Statistics

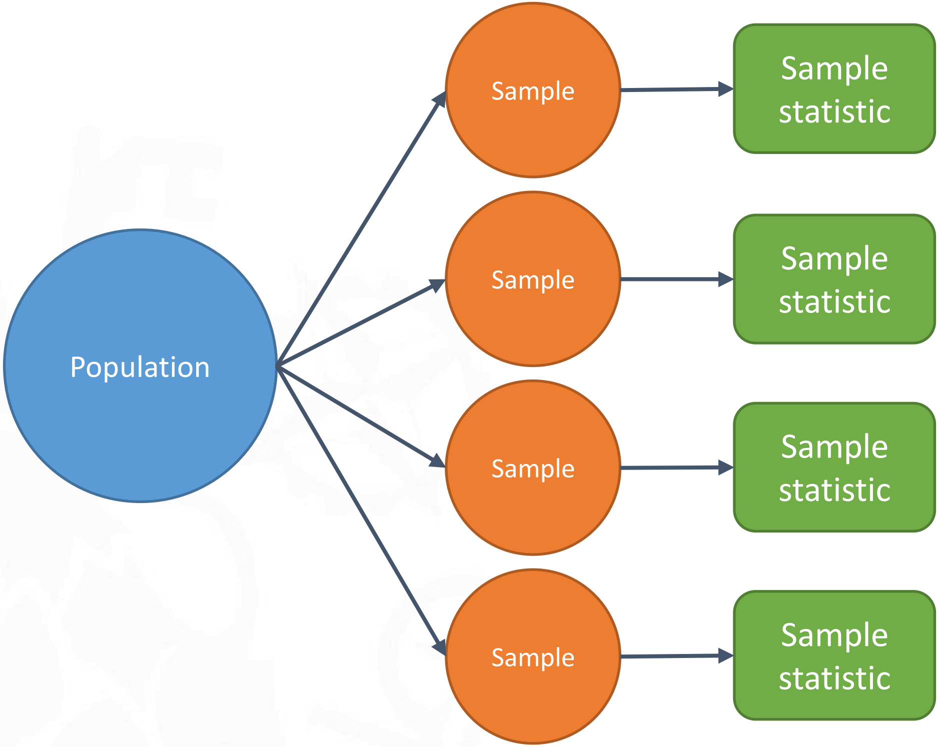

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Contents

1. Central Limit Theorem (CLT)
2. Confidence Intervals
3. Hypothesis testing
4. Comparing two means
5. Inference for proportions

Central Limit Theorem (CLT)

Sample distributions

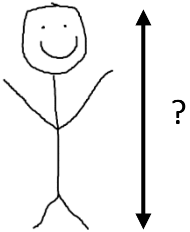


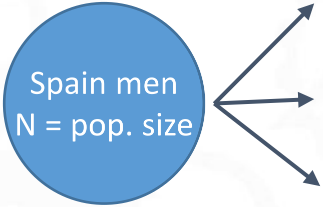
Sample distributions

Sampling distribution

Central Limit Theorem (CLT)
Standard Error

We want to obtain information about the height of men in Spain.





AR: $x_{AR,1}, x_{AR,2} + \cdots + x_{AR,1000}$ ⟶ $\bar{x}_{AR}$

…

MD: $x_{MD,1}, x_{MD,2} + \cdots + x_{MD,1000}$ ⟶ $\bar{x}_{MD}$

…

ZM: $x_{ZM,1}, x_{ZM,2} + \cdots + x_{ZM,1000}$ ⟶ $\bar{x}_{ZM}$

Sampling distribution

$$mean(\bar{x}) \approx \mu$$

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$\uparrow n$ ↓Standard error $SD(\bar{x}) < \sigma$

https://gallery.shinyapps.io/CLT_mean/

Central Limit Theorem (CLT)
Definition

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N(mean = \mu, SE = \frac{\sigma}{\sqrt{n}})$$

Usually, standard deviation of the population is not known, so standard deviation of sample is used:

$$\bar{x} \sim N(mean = \mu, SE = \frac{s}{\sqrt{n}})$$

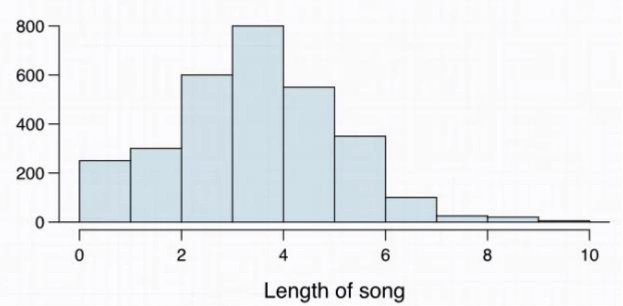
Conditions for the CLT:

1. **Independence**: Sampled observations must be independent.
   - Random sample/assignment
   - If sampling with replacement, n < 10% of population
2. **Sample size/skew**: Either the population distribution is normal, or the sample size is large (rule of thumb: n > 30)

Central Limit Theorem (CLT)

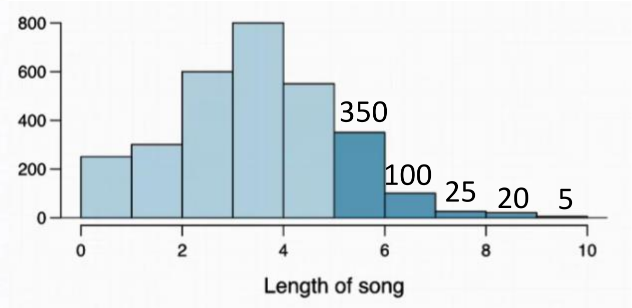
Example

Suppose my iPod has 3,000 songs. The histogram below shows a distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than five minutes.



Central Limit Theorem (CLT)
Example

Suppose my iPod has 3,000 songs. The histogram below shows a distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than five minutes.



x = length of one song

$$P(x > 5) = \frac{350 + 100 + 25 + 20 + 5}{3000} = \frac{500}{3000} \approx 0.17$$

Central Limit Theorem (CLT)
Example

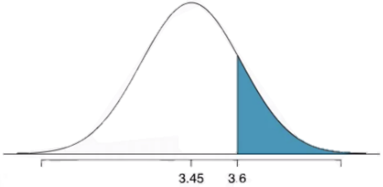
Another question. I'm about to take a trip to visit my parents and the drive is six hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive?

$$6\ hours = 360\ minutes$$

$$P(x_1 + x_2 + \ \ldots + x_{100} > 360\ min)?\ = P(\bar{x} > 3.6\ min)?$$

By CLT: $\bar{x} \sim N(mean = \mu = 3.45, SE = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{100}} = 0.163)$

$$\mathrm{Z} = \frac{3.6\ - 3.45}{0.163} = 0.92$$

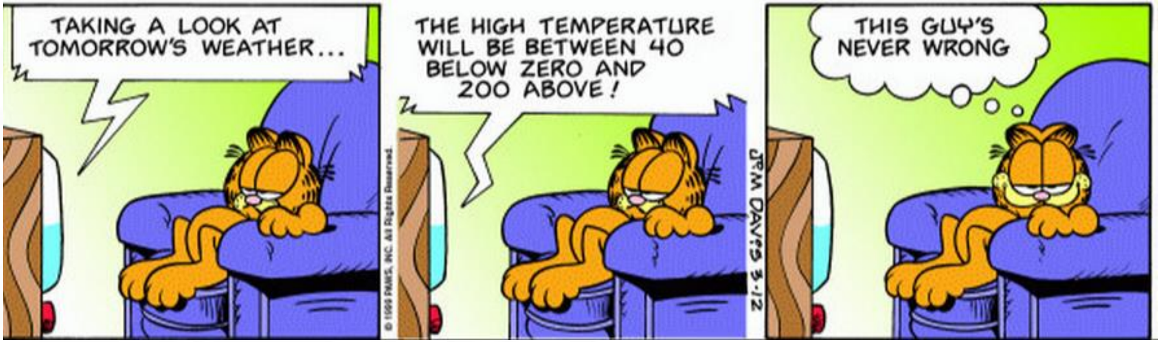


So we only need to calculate $P(Z > 0.92)$

Confidence Interval
Definition

A plausible range of values for the population parameter is called a **confidence interval**. Why are them important?

- If we report a point estimate, hitting the exact value for the population parameter is highly unlikely.
- If we report a range of plausible values, having the exact value for the population inside the range is much more probable.


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

Confidence Interval

For a mean

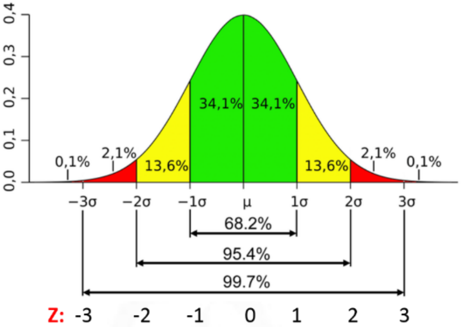
By CLT:

$$\bar{x} \sim N(mean = \mu, SE = \frac{s}{\sqrt{n}})$$

So an approximate 95% confidence interval for μ is:

$$CI_{(95\%)} \approx \bar{x} \pm 2 \cdot SE$$

Where $2 \cdot SE$ is called the margin of error (ME)



Confidence Interval

For a mean

For different confidence levels, the confidence interval for a population mean is defined as the sample mean plus/minus a margin of error calculated as the critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution:

$$CI_{(XX\%)} = \bar{x} \pm z^* \cdot \frac{s}{\sqrt{n}}$$

A confidence value of XX% determines that for XX out of 100 random samples, the true value of the population parameter would fall inside the confidence interval of each random sample.

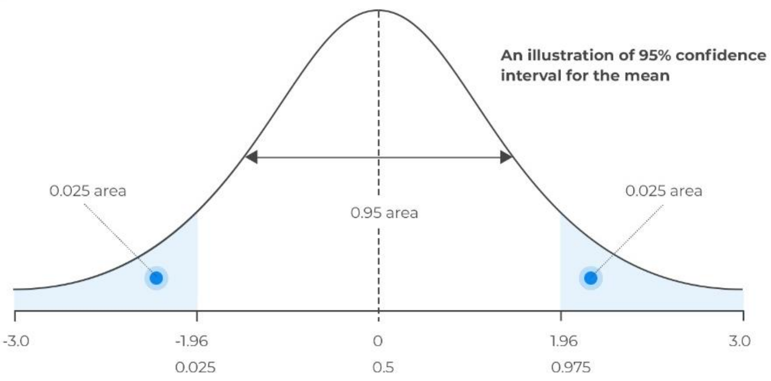
Conditions for this confidence interval:

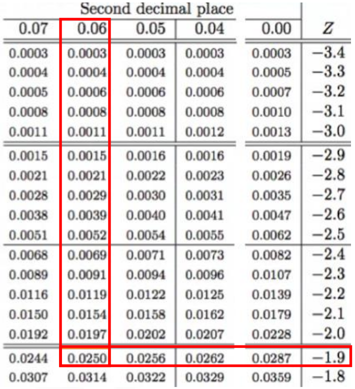
1. **Independence**: Sampled observations must be independent.
   - Random sample/assignment
   - If sampling with replacement, n < 10% of population
2. **Sample size/skew**: n > 30, larger if the population is severely skewed

Confidence Interval

Finding the critical value

For the 95% confidence interval:





Second decimal place					
0.07	0.06	0.05	0.04	0.00	Z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

$$z^* = 1.96$$

Accuracy vs precision

Commonly used critical values are 90%, 95%, 98% and 99%.

What happens to the confidence Interval when the confidence level increases?

1. Confidence Interval is wider, as it captures more values
2. Confidence Interval is narrower, as it is more confident in the calculation

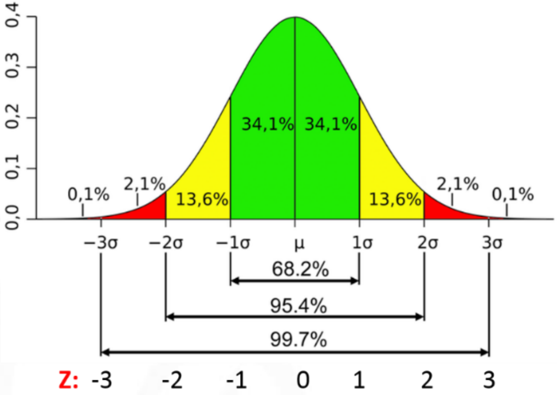
Confidence Interval

Accuracy vs precision

Commonly used critical values are 90%, 95%, 98% and 99%.

What happens to the confidence Interval when the confidence level increases?

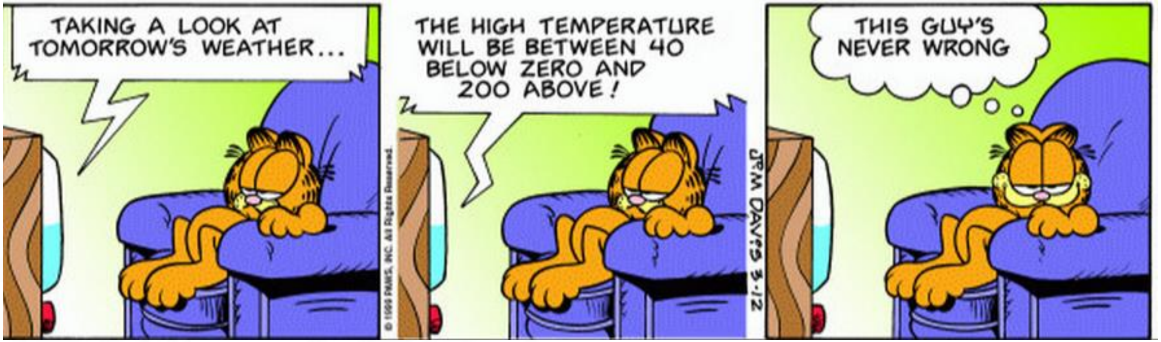
1. Confidence Interval is wider, as it captures more values
2. Confidence Interval is narrower, as it is more confident in the calculation



A **wider interval is more accurate**, as it is more probable to capture the output statistic true value.

Confidence Interval
Accuracy vs precision

However, **a wider interval is less informative**, i.e., **it loses precision**. Remember:


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

To get the best of both worlds we need to **increase the sample size**, as the margin error would decrease:

$$CI_{(XX\%)} = \bar{x} \pm z^{*} \cdot \frac{s}{\sqrt{n}}$$

Confidence Interval

Accuracy vs precision

The General Social Survey, the GSS, is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 U.S. residents. Based on the survey results, a 95% confidence interval for the average number of hours Americans have to relax or pursue activities that they enjoy after an average workday, was found to be 3.53 to 3.83 hours.

Determine if each of the following statements are true or false:

1. 95% of Americans spend between 3.53 to 3.83 hours relaxing after a work day.
2. 95% of random samples of 1,154 Americans will yield confidence intervals that contain the true average number of hours Americans spend relaxing after a work day.
3. 95% of the time the true average number of hours Americans spend relaxing after a work day is between 3.53 and 3.83 hours.

Confidence Interval

Accuracy vs precision

The General Social Survey, the GSS, is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 U.S. residents. Based on the survey results, a 95% confidence interval for the average number of hours Americans have to relax or pursue activities that they enjoy after an average workday, was found to be 3.53 to 3.83 hours.

Determine if each of the following statements are true or false:

1. 95% of Americans spend between 3.53 to 3.83 hours relaxing after a work day.
2. 95% of random samples of 1,154 Americans will yield confidence intervals that contain the true average number of hours Americans spend relaxing after a work day.
3. 95% of the time the true average number of hours Americans spend relaxing after a work day is between 3.53 and 3.83 hours.

Confidence Interval

Backtracking to n from ME

Given a target margin of error, confidence level, and information on the variability of the sample or the population, we want to determine the required sample size to achieve the desired margin of error.

Let’s check an example

Confidence Interval

Backtracking to n from ME

Suppose a group of researchers want to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQs of three-year-old children born to mothers who were on this medication during their pregnancy.

Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points.

How many such children should the researches sample in order to obtain a 90% confidence interval with a margin of error less than or equal to four points?

$ME = 4$

$CL = 90\%$

$z^* = 1.96$

$\sigma = 18$

Confidence Interval

Backtracking to n from ME

Suppose a group of researchers want to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQs of three-year-old children born to mothers who were on this medication during their pregnancy.

Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points.

How many such children should the researches sample in order to obtain a 90% confidence interval with a margin of error less than or equal to four points?

$ME = 4$
$CL = 90\%$
$z^* = 1.96$
$\sigma = 18$

$$4 = 1.65 \cdot \frac{18}{\sqrt{n}} \rightarrow n = \left(\frac{1.65 \cdot 18}{4}\right)^2 = 55.13 \rightarrow 56\ children$$

Backtracking to n from ME

Earlier we found that we needed at least 56 children in the sample to achieve a maximum margin of error of four points. How would the required sample size change if we want to further decrease the margin of error by half?

$$\frac{1}{2} \cdot ME = z^* \cdot \frac{s}{\sqrt{n}} \cdot \frac{1}{2}$$

Backtracking to n from ME

Earlier we found that we needed at least 56 children in the sample to achieve a maximum margin of error of four points. How would the required sample size change if we want to further decrease the margin of error by half?

$$\frac{1}{2} \cdot ME = z^* \cdot \frac{s}{\sqrt{n}} \cdot \frac{1}{2}$$

$$\frac{1}{2} \cdot ME = z^* \cdot \frac{s}{\sqrt{4 \cdot n}}$$

So, to reduce the ME by half, we need to duplicate the sample size!

Hypothesis

Null – H0	Often either a skeptical perspective or a claim to be tested	=
Alternative – Ha	Represents an alternative claim under consideration and is often represented by a range of possible parameter values	<, >, ≠

The skeptic will not abandon H0 unless the evidence in favor of Ha is so strong that H0 must be rejected in favor of Ha

Hypothesis testing

Hypothesis via CI

A previous study found that a 95% confidence interval for the average number of exclusive relationships college students have been in to be 2.7 to 3.7. Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than three exclusive relationships?

Hypothesis via CI

A previous study found that a 95% confidence interval for the average number of exclusive relationships college students have been in to be 2.7 to 3.7. Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than three exclusive relationships?

Being X the number of exclusive relationships college students have been in, what are the hypothesis?

Hypothesis testing

Hypothesis via CI

A previous study found that a 95% confidence interval for the average number of exclusive relationships college students have been in to be 2.7 to 3.7. Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than three exclusive relationships?

Being X the number of exclusive relationships college students have been in, what are the hypothesis?

H0: $\mu = 3$ College students have been in 3 exclusive relationships, on average
Ha: $\mu > 3$ College students have been in more than 3 exclusive relationships, on average.

Hypothesis via CI

A previous study found that a 95% confidence interval for the average number of exclusive relationships college students have been in to be 2.7 to 3.7. Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than three exclusive relationships?

Being X the number of exclusive relationships college students have been in, what are the hypothesis?

H0: $\mu = 3$ College students have been in 3 exclusive relationships, on average
Ha: $\mu > 3$ College students have been in more than 3 exclusive relationships, on average.

Note that hypothesis are always about population parameters, never about sample statistics. Also… could you reject H0 based on your knowledge?

Hypothesis testing

p-value

The p-value is the probability of observed or more extreme outcome, given that the null hypothesis is true.

$$p - value = P(observed\ or\ more\ extreme\ outcome \mid H_0)$$

In the previous example, the CI interval comes from a sample of 50 students with a mean of 3.2 and a standard deviation of 1.74:

$$n = 50$$
$$\bar{x} = 3.2$$
$$s = 1.74$$

With this data and using CLT, we know that:

$$\bar{x} \sim N(\mu = 3, SE = \frac{1.74}{50} = 0.246)$$

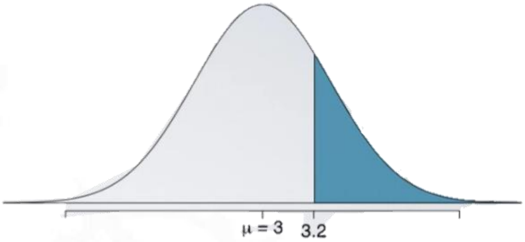
So, what is the probabilty of observing the sample mean or a more extreme outcome, given the null hypothesis?

$$P(\bar{x} > 3.2 \mid \mu = 3)$$

Hypothesis testing

p-value

So let's represent the probability we want to calculate:



The z-score of the point of interest is:

$$z = \frac{3.2 - 3}{0.246} = 0.81$$

This z-score would be the test statistic, because is the statistic used to calculate the p-score:

$$P(z > 0.81) = 0.209$$

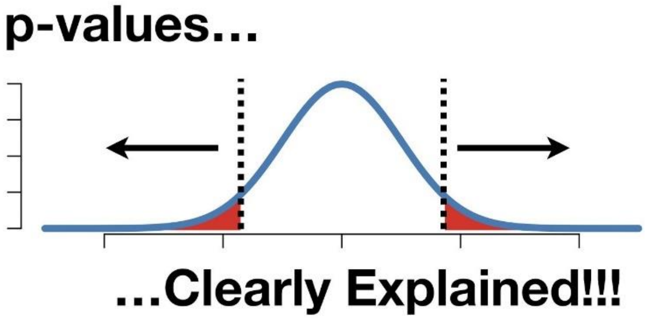
Since p-value is high, H0 can not be rejected.

Hypothesis testing

Interpreting the p-value

For this example, the p-value means that if in fact college students have been in 3 exclusive relationships on average, that's the equivalent of saying if in fact the null hypothesis is true, there is a 21% chance that a random sample of 50 college students would still yield a sample mean of 3.2 or higher.

Since this is a pretty high probability, a sample mean of 3.2 or more exclusive relationships is likely to happen simply by chance or sampling variability.



Two-sided tests

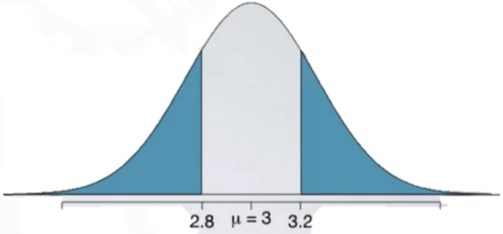
Instead of looking for a divergence from the null hypothesis in a specific direction, it might be interesting the divergence in any direction.

This type of hypothesis tests are called two-sided (or two-tailed).

The p-value “at least as extreme as the observed outcome” probability is extended in both directions.

For the example, the p-value would be:

$$P(\bar{x} > 3.2 \; OR \; \bar{x} < 2.8 \mid H_0\text{:}\, \mu = 3)$$

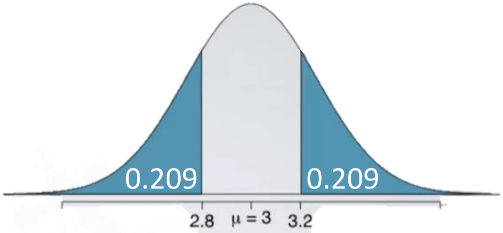


Hypothesis testing

Two-sided tests

The p-value would be:

$$P(\bar{x} > 3.2\ OR\ \bar{x} < 2.8 \mid H_0\colon \mu = 3)$$



$$p - value = P(z > 0.81) + P(z < -0.81) = 2 \cdot 0.209 = 0{,}418$$

Other estimators

Nearly normal sampling distributions

The previous method, doing hypothesis tests and confidence intervals, can be easily adapted for any point estimator that has a nearly normal sampling distribution:

- Sample mean: $\bar{x}$
- Difference between sample means: $\bar{x}_1 - \bar{x}_2$
- Sample proportion: $\hat{p}$
- Difference between sample proportions: $\hat{p}_1 - \hat{p}_2$

Point estimators are unbiased, i.e., they do not generally over or underestimate the population parameter true value, only gives a good estimate.

To construct CI of unbiased estimators:

$$point\ estimate\ \pm z^* \cdot SE$$

SE calculation might be different for different point estimators. More about this in following sections!

Other estimators
Example

A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch the Daily Show. An American late-night TV Show. The standard error of this estimate is 0.014. We are asked to estimate the 95% confidence interval for the proportion of college graduates who watch The Daily Show.

Other estimators
Example

A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch the Daily Show. An American late-night TV Show. The standard error of this estimate is 0.014. We are asked to estimate the 95% confidence interval for the proportion of college graduates who watch The Daily Show.

$$\hat{p} = 0.33$$
$$SE = 0.014$$

Other estimators

Example

A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch the Daily Show. An American late-night TV Show. The standard error of this estimate is 0.014. We are asked to estimate the 95% confidence interval for the proportion of college graduates who watch The Daily Show.

$$\hat{p} = 0.33$$
$$SE = 0.014$$
$$CI_{95\%}: 0.33 \pm 1.96 \cdot 0.014 = (0.303, 0.357)$$

Second example

The third national health and nutrition examination survey NHANES, collected body fat percentage and gender data from over 13,000 subjects in ages between 20 to 80. The average body fat percentage for the 6,580 men in the sample was 23.9%. And this value was 35% for the, for the 7,021 women. The standard error for the difference between the average male and female body fat percentages was 0.114. Do these data provide convincing evidence that men and women have different average body fat percentages? You may assume that the distribution of the point estimate is nearly normal.

Other estimators
Second example

The third national health and nutrition examination survey NHANES, collected body fat percentage and gender data from over 13,000 subjects in ages between 20 to 80. The average body fat percentage for the 6,580 men in the sample was 23.9%. And this value was 35% for the, for the 7,021 women. The standard error for the difference between the average male and female body fat percentages was 0.114. Do these data provide convincing evidence that men and women have different average body fat percentages? You may assume that the distribution of the point estimate is nearly normal.

$H_0: \mu_{men} = \mu_{women}; H_A: \mu_{men} \neq \mu_{women}$ → Two-tailed test

Other estimators

Second example

The third national health and nutrition examination survey NHANES, collected body fat percentage and gender data from over 13,000 subjects in ages between 20 to 80. The average body fat percentage for the 6,580 men in the sample was 23.9%. And this value was 35% for the, for the 7,021 women. The standard error for the difference between the average male and female body fat percentages was 0.114. Do these data provide convincing evidence that men and women have different average body fat percentages? You may assume that the distribution of the point estimate is nearly normal.

$H_0: \mu_{men} = \mu_{women}; H_A: \mu_{men} \neq \mu_{women}$ → Two-tailed test

Point estimate: $\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

Other estimators

Second example

The third national health and nutrition examination survey NHANES, collected body fat percentage and gender data from over 13,000 subjects in ages between 20 to 80. The average body fat percentage for the 6,580 men in the sample was 23.9%. And this value was 35% for the, for the 7,021 women. The standard error for the difference between the average male and female body fat percentages was 0.114. Do these data provide convincing evidence that men and women have different average body fat percentages? You may assume that the distribution of the point estimate is nearly normal.

$H_0: \mu_{men} = \mu_{women}; H_A: \mu_{men} \neq \mu_{women}$ → Two-tailed test

Point estimate: $\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

This point estimate is distributed:

$$\bar{x}_{men} - \bar{x}_{women} \sim N(\mu_{men} - \mu_{women}, SE)$$

Second example

The third national health and nutrition examination survey NHANES, collected body fat percentage and gender data from over 13,000 subjects in ages between 20 to 80. The average body fat percentage for the 6,580 men in the sample was 23.9%. And this value was 35% for the, for the 7,021 women. The standard error for the difference between the average male and female body fat percentages was 0.114. Do these data provide convincing evidence that men and women have different average body fat percentages? You may assume that the distribution of the point estimate is nearly normal.

$H_0: \mu_{men} = \mu_{women}; H_A: \mu_{men} \neq \mu_{women}$ → Two-tailed test

Point estimate: $\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

This point estimate is distributed:

$$\bar{x}_{men} - \bar{x}_{women} \sim N(\mu_{men} - \mu_{women}, SE) = N(0, 0.114)$$

Other estimators

Second example

The third national health and nutrition examination survey NHANES, collected body fat percentage and gender data from over 13,000 subjects in ages between 20 to 80. The average body fat percentage for the 6,580 men in the sample was 23.9%. And this value was 35% for the, for the 7,021 women. The standard error for the difference between the average male and female body fat percentages was 0.114. Do these data provide convincing evidence that men and women have different average body fat percentages? You may assume that the distribution of the point estimate is nearly normal.

$H_0 : \mu_{men} = \mu_{women}; H_A : \mu_{men} \neq \mu_{women}$ → Two-tailed test

Point estimate: $\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

This point estimate is distributed:

$$\bar{x}_{men} - \bar{x}_{women} \sim N(\mu_{men} - \mu_{women}, SE) = N(0, 0.114)$$

So, the p-value is basically 0 and H0 can be rejected.

Decision errors

Types of errors

		Decision	
		Fail to reject H0	Reject H0
Truth	H0 true	✓	**Type I error**
	Ha true	**Type II error**	✓

- Type I error is rejecting H0 when H0 is true
- Type II error is failing to reject H0 when HA is true

Decision errors

Which is worse?

In general, reducing both types of errors is not feasible, so we must try to determine which type of error we want to minimize. So, which is worse?

- Type II: Declaring the defendant innocent when they are actually guilty.
- Type I: Declaring the defendant guilty when they are actually innocent.

The answer is… depends on the problem!

Decision errors

Significance rate

Usually, the significance rate is chosen at $\alpha = 0.05$, meaning that we reject H0 when the p-value is less than 0.05, i.e., given that H0 is true there is a probability of less than 5% of seeing an observed value or a more extreme value.

This significance level means that, for those cases where H0 is actually true, no more than 5% of the times H0 is incorrectly rejected.

In other words:

$$P(Type\ I\ error\,|\,H_0\ true) = \alpha$$

If type I errors are dangerous or especially costly, choose a smaller significance level. In the other hands, if type II errors are dangerous, choose a higher significance level.

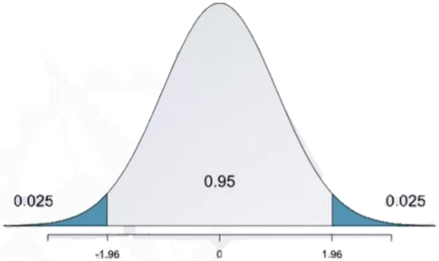
Decision errors

Significance vs confidence

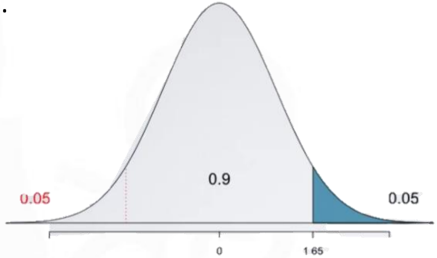
In general, significance and confidence interval are complementary:

$$1 - \alpha = \beta$$

This is related to the area below the PDF for a two-tailed test with a 95% confidence interval.



However, this is not the case for a one-sided test. For the same covered area, the confidence level is 90%:



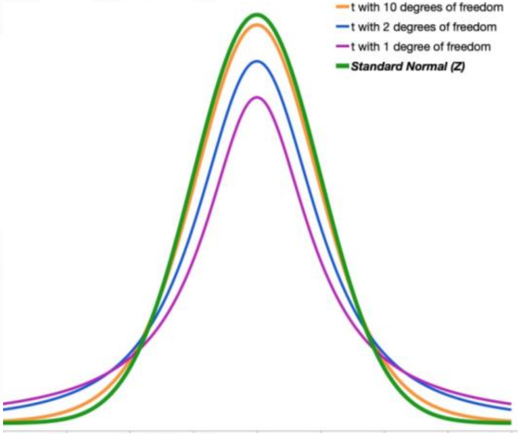
t-distribution

Introduction

When **population standard deviation σ is unknown**, t-distribution can be used to address the **uncertainty of the standard error** estimate. t-distribution is usually called student’s t.

Observations are more likely to fall beyond 2 SDs from the mean than in an standard normal distribution, helping mitigating the effect of a less reliable estimate for the SE.

While normal distribution has two parameters (µ and σ), student’s t has only one: **degrees of freedom (df)**.



Inference for a mean

In a study, researchers evaluated the **relationship between being distracted and recall of food consumed and snacking**, with the idea that if you are distracted while you are eating, you may **not remember what you eat**. They also hypothesized that failure to recall food consumed might lead to **increased snacking later on**. The sample for this study consisted of **44 volunteer** patients, **half men, half women**. These 44 patients were **randomized into two groups**. **One group** was asked to play solitaire on the computer while eating and **was asked to win as many games as possible**, and the **other group** was asked to **eat lunch without any distractions**. Both groups were provided the **same amount of lunch** and then **after lunch they were offered biscuits** to snack on.

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Inference for a mean

The goal in this example is to estimate the **average snacking level for distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Inference for a mean

The goal in this example is to estimate the **average snacking level for distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Confidence interval using the t-distribution:

t-distribution

Inference for a mean

The goal in this example is to estimate the **average snacking level for distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Confidence interval using the t-distribution:

$$point\ estimate\ \pm margin\ of\ error$$

t-distribution

Inference for a mean

The goal in this example is to estimate the **average snacking level for distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Confidence interval using the t-distribution:

$$point\ estimate\ \pm margin\ of\ error$$

$$\bar{x} \pm t_{df}^{*} \cdot SE_{\bar{x}}$$

Inference for a mean

The goal in this example is to estimate the **average snacking level for distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Confidence interval using the t-distribution:

$$point\ estimate\ \pm margin\ of\ error$$

$$\bar{x} \pm t_{df}^{*} \cdot SE_{\bar{x}}$$

$$\bar{x} \pm t_{df}^{*} \cdot \frac{s}{\sqrt{n}}$$

The degrees of freedom for inference on one sample mean is the number of samples minus 1: $df = n - 1$ (t_{df}^{*}: t-score)

Inference for a mean

The goal in this example is to estimate the **average snacking level for distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

In the example:

$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$
$$t^*_{21} = 2.08$$

$$CI: \bar{x} \pm t^*_{21} \cdot \frac{s}{\sqrt{n}} = 52.1 \pm 2.08 \cdot \frac{45.1}{\sqrt{22}} \approx 52.1 \pm 20$$

$$CI: (32.1\ g, 72.1\ g)$$

t-distribution

Hypothesis testing for a mean

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$
$$df = 21$$
$$\text{SE} = \frac{45.1}{\sqrt{22}} = 9.62$$

t-distribution

Hypothesis testing for a mean

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$
$$df = 21$$
$$\mathrm{SE} = \frac{45.1}{\sqrt{22}} = 9.62$$

Hypothesis:

$$H_0: \mu = 30$$
$$H_A: \mu \neq 30$$

Hypothesis testing for a mean

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

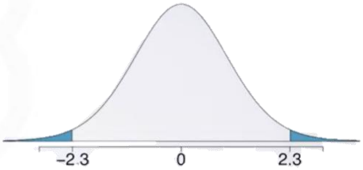
$\bar{x} = 52.1\ g$
$s = 45.1\ g$
$n = 22$
$df = 21$
$\text{SE} = \frac{45.1}{\sqrt{22}} = 9.62$

Hypothesis:

$H_0: \mu = 30$
$H_A: \mu \neq 30$

t-statistic (not the same as t-score):

$$T = \frac{\bar{x} - \mu}{SE} = \frac{52.1 - 30}{9.62} = 2.3$$



t-distribution

Hypothesis testing for a mean

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$
$$df = 21$$
$$\text{SE} = \frac{45.1}{\sqrt{22}} = 9.62$$

Hypothesis:

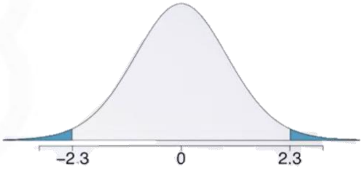
$$H_0: \mu = 30$$
$$H_A: \mu \neq 30$$

t-statistic (not the same as t-score):

$$T = \frac{\bar{x} - \mu}{SE} = \frac{52.1 - 30}{9.62} = 2.3$$

And calculate:

$$p - value: P(T > 2.3\ OR\ T < -2.3 \mid t_{21}) \approx 0.0318$$



t-distribution

Inference for difference of two independent means

Estimate the **difference of the average snacking level for distracted and non-distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Confidence interval using the t-distribution for two independent means:

$$point\ estimate\ \pm margin\ of\ error$$

Inference for difference of two independent means

Estimate the **difference of the average snacking level for distracted and non-distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Confidence interval using the t-distribution for two independent means:

$$point\ estimate\ \pm margin\ of\ error$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

t-distribution

Inference for difference of two independent means

Estimate the **difference of the average snacking level for distracted and non-distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

Confidence interval using the t-distribution for two independent means:

$$point\ estimate\ \pm margin\ of\ error$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

Where:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad df = \min(n_1 - 1, n_2 - 1)$$

t-distribution

Inference for difference of two independent means

Estimate the **difference of the average snacking level for distracted and non-distracted eaters**:

Biscuit intake	$\bar{x}$	s	n
Solitaire	52.1 g	45.1 g	22
No distraction	27.1 g	26.4 g	22

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{21} \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

$$= (52.1 \ - 27.1) \pm 2.08 \cdot \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}}$$

$$= 25 \pm 2.08 \cdot 11.14 = (1.83\ g, 48.17\ g)$$

t-distribution

Hypothesis testing for two independent means

Is there a **difference between the average post-meal snack consumption between groups**?

$\bar{x}_1 = 52.1\ g$

$s_1 = 45.1\ g$

$n_1 = 22$

$\bar{x}_2 = 27.1\ g$

$s_2 = 26.4\ g$

$n_2 = 22$

$df = 21$

$\mathrm{SE} = 11.14$

t-distribution

Hypothesis testing for two independent means

Is there a **difference between the average post-meal snack consumption between groups**?

$\bar{x}_1 = 52.1\ g$
$s_1 = 45.1\ g$
$n_1 = 22$
$\bar{x}_2 = 27.1\ g$
$s_2 = 26.4\ g$
$n_2 = 22$

$df = 21$
$\mathrm{SE} = 11.14$

Hypothesis:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

t-distribution

Hypothesis testing for two independent means

Is there a **difference between the average post-meal snack consumption between groups**?

$\bar{x}_1 = 52.1\ g$
$s_1 = 45.1\ g$
$n_1 = 22$
$\bar{x}_2 = 27.1\ g$
$s_2 = 26.4\ g$
$n_2 = 22$

$df = 21$
$\mathrm{SE} = 11.14$

Hypothesis:

$H_0: \mu_1 - \mu_2 = 0$
$H_A: \mu_1 - \mu_2 \neq 0$

t-statistic:

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE} = \frac{25 - 0}{11.14} = 2.24$$

t-distribution

Hypothesis testing for two independent means

Is there a **difference between the average post-meal snack consumption between groups**?

$$\bar{x}_1 = 52.1\ g$$
$$s_1 = 45.1\ g$$
$$n_1 = 22$$
$$\bar{x}_2 = 27.1\ g$$
$$s_2 = 26.4\ g$$
$$n_2 = 22$$

$$df = 21$$
$$\text{SE} = 11.14$$

Hypothesis:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

t-statistic:

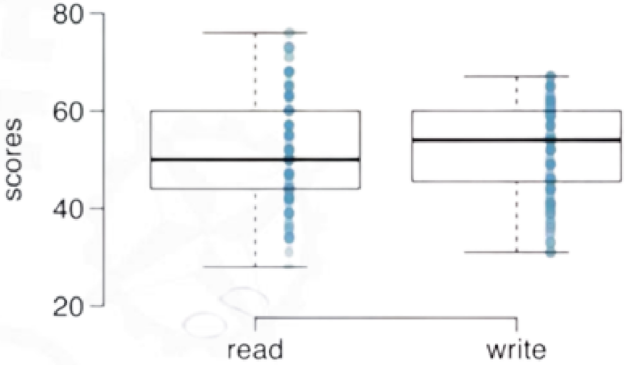
$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE} = \frac{25 - 0}{11.14} = 2.24$$

And calculate:

$$p - value: P(T > 2.24\ OR\ T < -2.24 \mid t_{21}) \approx 0.036$$

Hypothesis testing for two paired means

200 observations were randomly sampled from the High School and Beyond survey. The same students took a reading and a writing test.



Can the reading and writing scores for a giving student be independent from each other?

Hypothesis testing for two paired means

- When two sets of observations have this special correspondence, or in other words they are not independent, they are said **to be paired**.
- To analyze paired data, it is often **useful to look at the difference in outcomes** of each pair of observations.
    - So, in this example, for each student, we subtract their writing score from their reading score

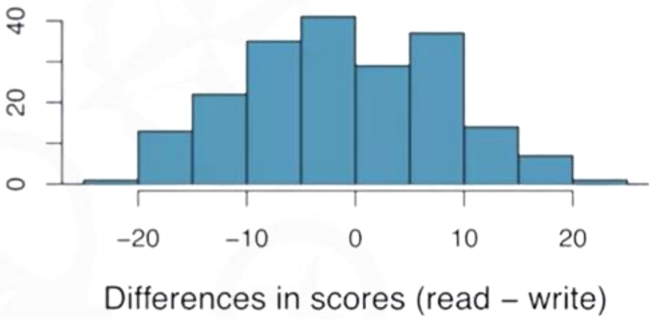
ID	Read	Write	Diff
0	57	52	5
1	44	33	11
…	…	…	…

t-distribution

Hypothesis testing for two paired means

If in fact there was no difference between the reading and writing scores, what would we expect the differences to be?

$$\bar{x}_{diff} = -0.545$$
$$s_{diff} = 8.887$$
$$n_{diff} = 200$$



t-distribution

Hypothesis testing for two paired means

If in fact there was no difference between the reading and writing scores, what would we expect the differences to be?

$$\bar{x}_{diff} = -0.545$$

$$s_{diff} = 8.887$$

$$n_{diff} = 200$$

$$df = 199$$

Hypothesis:

$$H_0: \mu_{diff} = 0$$
$$H_A: \mu_{diff} \neq 0$$

t-statistic:

$$T = \frac{\bar{x} - \mu}{SE} = \frac{-0.545 - 0}{8.887/\sqrt{200}} = -0.87$$

And calculate:

$$p - value: P(T > 0.87\ OR\ T < -0.87 \mid t_{199}) \approx 0.385$$

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

1. p-value is the probability that the average scores on the reading and writing exams are equal.
2. p-value is the probability that the average scores on the reading and writing exams are different.
3. p-value is the probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.87 in either direction, if in fact the true average difference between the scores is zero.
4. p-value is the probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

1. p-value is the probability that the average scores on the reading and writing exams are equal. → P(H0)
2. p-value is the probability that the average scores on the reading and writing exams are different.
3. p-value is the probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.87 in either direction, if in fact the true average difference between the scores is zero.
4. p-value is the probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

1. p-value is the probability that the average scores on the reading and writing exams are equal. → P(H0)
2. p-value is the probability that the average scores on the reading and writing exams are different. → P(HA)
3. p-value is the probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.87 in either direction, if in fact the true average difference between the scores is zero.
4. p-value is the probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

t-distribution

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

1. p-value is the probability that the average scores on the reading and writing exams are equal. → P(H0)
2. p-value is the probability that the average scores on the reading and writing exams are different. → P(HA)
3. p-value is the probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.87 in either direction, if in fact the true average difference between the scores is zero.
4. p-value is the probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

t-distribution

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

1. p-value is the probability that the average scores on the reading and writing exams are equal. → P(H0)
2. p-value is the probability that the average scores on the reading and writing exams are different. → P(HA)
3. p-value is the probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.87 in either direction, if in fact the true average difference between the scores is zero.
4. p-value is the probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true. → P(Type I error)

Compare n means

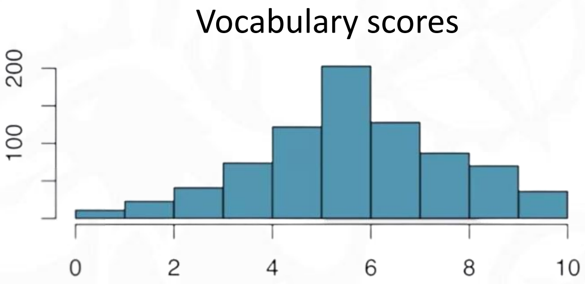
The example for this section comes from the general social survey. The variable of interest are vocabulary scores and self-identified social class. Vocabulary score is calculated based on a set of 10 question vocabulary test, where a higher score means better vocabulary, and self identified social class has four levels, lower, working, middle, and upper class.

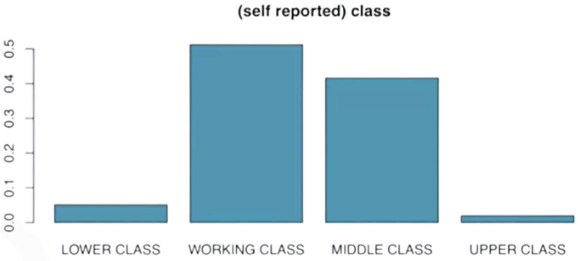
Voc. scores	Class
6	Middle class
9	Working class
…	…
7	Upper class

ANOVA

Compare n means

The example for this section comes from the general social survey. The variable of interest are vocabulary scores and self-identified social class. Vocabulary score is calculated based on a set of 10 question vocabulary test, where a higher score means better vocabulary, and self identified social class has four levels, lower, working, middle, and upper class.





ANOVA

Compare n means

The example for this section comes from the general social survey. The variable of interest are vocabulary scores and self-identified social class. Vocabulary score is calculated based on a set of 10 question vocabulary test, where a higher score means better vocabulary, and self identified social class has four levels, lower, working, middle, and upper class.

	n	mean	sd
Lower class	41	5.07	2.24
Working class	407	5.75	1.87
Middle class	331	6.76	1.89
Upper class	16	6.19	2.34
Overall	795	6.14	1.98

ANOVA
Hypothesis

In an ANOVA test, hypothesis are:

$$H_0\text{: } mean\ outcome\ of\ a\ variable\ is\ the\ same\ across\ all\ k\ groups\ (categories)$$

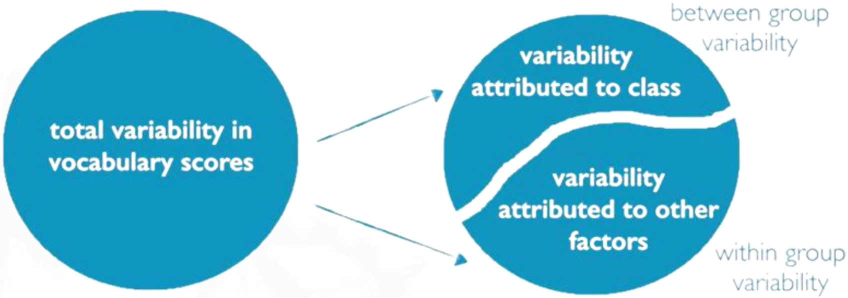
$$\mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_A\text{: } at\ least\ one\ pair\ of\ means\ are\ different$$

t-test	ANOVA
Compare means from two groups to see whether they're so far apart that the observed difference cannot reasonably be attributed to sampling variability.	Compare means from more than two groups to see whether they're so far apart that the observed differences cannot all reasonably be attributed to sampling variability.

ANOVA
Variability partitioning

For the example:


total variability in vocabulary scores
variability attributed to class
between group variability
variability attributed to other factors
within group variability

		Df	Sum Sq	Mean Sq	F value	P-value
Group	**Class**	3	236.56	78.855	21.735	<0.0001
Error	**Residuals**	791	2869.80	3.628		
	Total	794	3106.36			

ANOVA

Variability partitioning

For the example:

			Sum Sq			
Group	**Class**		236.56			
Error	**Residuals**		2869.80			
	Total		3106.36			

Sum of Squares Total or SST

- Measures the total variability of the output response
- Calculated similar to variance (not scaled by sample size)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

ANOVA
Variability partitioning

For the example:

			Sum Sq			
Group	**Class**		236.56			
Error	**Residuals**		2869.80			
	Total		3106.36			

Sum of Squares Group or SSG

- Measures the variability between groups
- Explained variability: squared deviation of group means from overall means, weighted by sample size

$$SSG = \sum_{j=1}^{n} n_j \left(\bar{y}_j - \bar{y}\right)^2$$

ANOVA

Variability partitioning

For the example:

			Sum Sq			
Group	**Class**		236.56			
Error	**Residuals**		2869.80			
	Total		3106.36			

Sum of Squares Errors or SSE

- Measures the variability within groups
- Unexplained variability by the group variable

$$SSE = SST \ - SSG$$

ANOVA

Variability partitioning

For the example:

		df	Sum Sq			
Group	**Class**	3	236.56			
Error	**Residuals**	791	2869.80			
	Total	794	3106.36			

Degrees of freedom associated with ANOVA

- Total: $df_T = n - 1$
- Group: $df_G = k - 1$
- Error: $df_E = df_T - df_G$

ANOVA
Variability partitioning

For the example:

		df	Sum Sq	Mean Sq		
Group	**Class**	3	236.56	78.855		
Error	**Residuals**	791	2869.80	3.628		
	Total	794	3106.36			

Mean Squares

- Average variability between and within groups.
  - Group: $MSG = SSG/df_G$
  - Error: $MSE = SSE/df_E$

ANOVA
Variability partitioning

For the example:

		df	Sum Sq	Mean Sq	F value	
Group	**Class**	3	236.56	78.855	21.735	
Error	**Residuals**	791	2869.80	3.628		
	Total	794	3106.36			

F statistic

- Ratio of the average between group and within group variabilities:

$$F = \frac{MSG}{MSE}$$

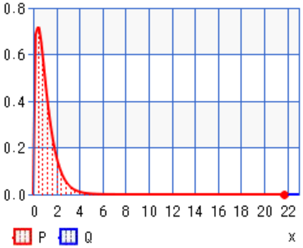
ANOVA
Variability partitioning

For the example:

		df	Sum Sq	Mean Sq	F value	P-value
Group	**Class**	3	236.56	78.855	21.735	<0.0001
Error	**Residuals**	791	2869.80	3.628		
	Total	794	3106.36			

P-value

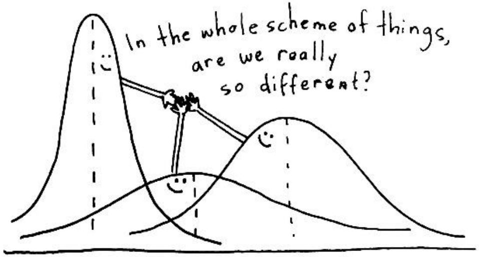
- Probability of at least as large a ratio between the “between” and “within” group variabilities if in fact the means of all groups are equal.
- Area under the F curve with degrees of freedom df_G and df_E



ANOVA

Conditions

1. Independence
   - Within groups: sampled observations must be independent
   - Between groups: groups must be independent from each other
2. Approximate normality: distributions should be nearly normal within each group
3. Equal variance: groups should have roughly equal variability: **homoscedastic groups**.



Multiple comparisons

Which means differ

ANOVA is great to check if any mean differs from the rest, but does not tell which pair of means differ.

Solution, use multiple t-tests:

- Two sample t tests for differences in each possible pair of groups.
- Multiple tests → inflated Type I error rate
- Need to use a modified significance level → Bonferroni correction.
    - Adjust α by the number of comparisons being considered

$$\alpha^* = \frac{\alpha}{K}; K: number\ of\ comparisons = \frac{k \cdot (k-1)}{2}$$

Multiple comparisons

Which means differ

Apart from using a corrected significance level, other conditions must be considered:

- Constant variance → corrected standard error and degrees of freedom for all tests

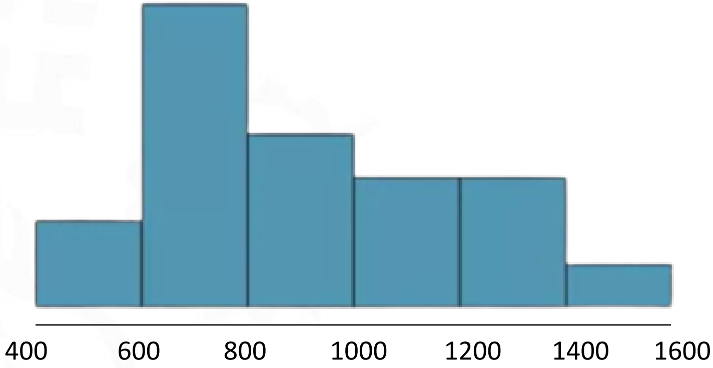
$$SE = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

$$df = df_E$$

Bootstrapping

Example

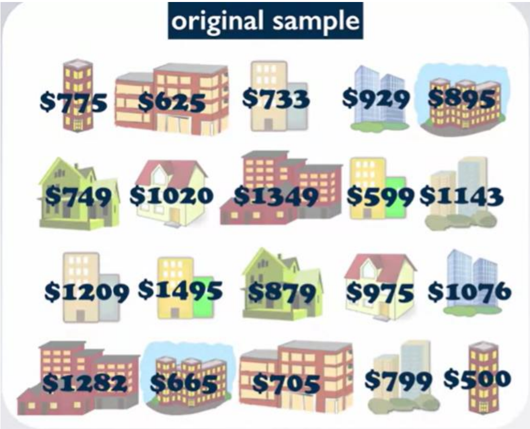
We have a dataset of 20 apartments randomly selected from Craigslist housing ads. These are apartments with at least one bedroom in Durham, North Carolina.



Which is the best central measure?

Example

The sample median in the original sample is $887.



In bootstrapping, we assume that, for each observation in the sample, there may be others like it in the population. The bootstrap population can be thought as a population where each observation from the sample appears many times.

Bootstrapping scheme

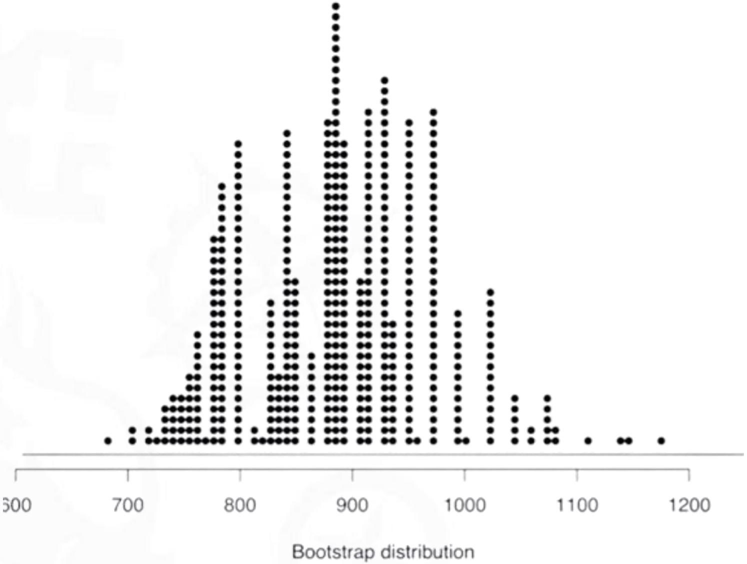
So, how is bootstrap performed:

1. Take a bootstrap sample: random sample with replacement from the original sample, of the same size as the original sample.
2. Calculate the bootstrap statistic: mean, median, proportion, … from the bootstrap samples.
3. Repeat steps 1 and 2 many times to create a bootstrap statistics distribution.

Bootstrapping

Bootstrap distribution

For the example, the median bootstrap distribution for 500 simulations:



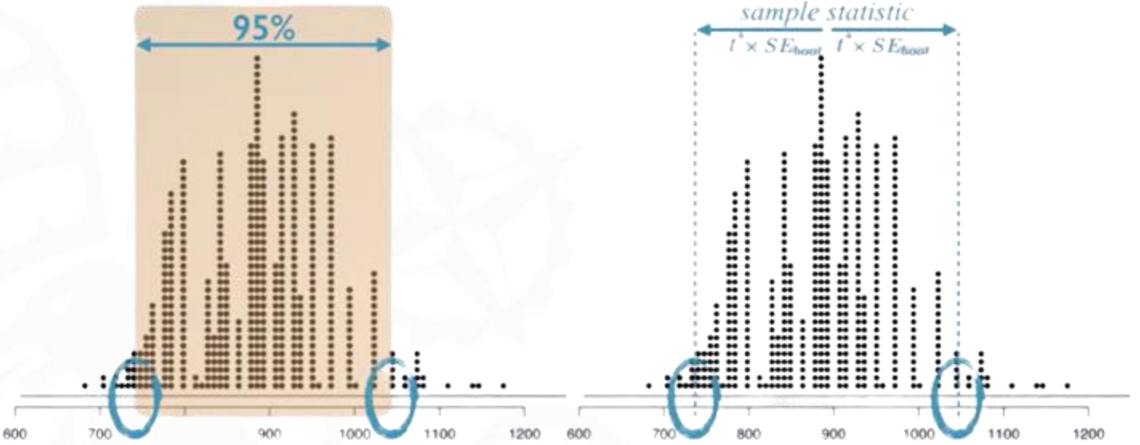
Bootstrapping

Confidence interval

Confidence interval can be calculated using two methods:

percentile method

standard error method



The DF in the standard error methods would be n-1 being n the original sample size.

Bootstrapping

Limitations

Bootstrap does not present as rigid conditions as CLT based methods. However:

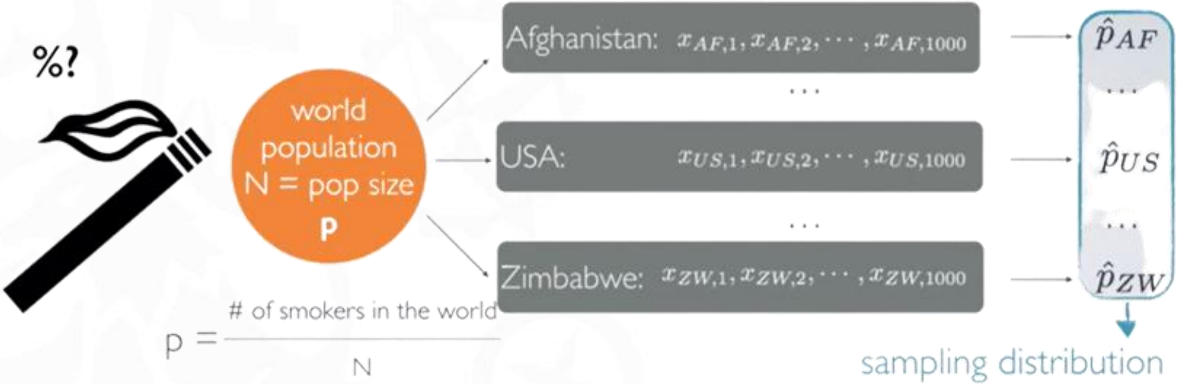
- If the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be unreliable.
- A representative sample is still required. If the sample is biased, the estimates would also be biased.

Categorical variables
Sampling distribution

In order to estimate the world smoker population proportion, we can ask individual people in each country and calculate the proportion of smokers for each country:

Smoker categorical

% Smoker numerical


%?
world population
N = pop size
p
Afghanistan: $x_{AF,1}, x_{AF,2}, \cdots, x_{AF,1000}$
$\cdots$
USA: $x_{US,1}, x_{US,2}, \cdots, x_{US,1000}$
$\cdots$
Zimbabwe: $x_{ZW,1}, x_{ZW,2}, \cdots, x_{ZW,1000}$
$\hat{p}_{AF}$
$\hat{p}_{US}$
$\hat{p}_{ZW}$
$p = \frac{\text{\# of smokers in the world}}{N}$
sampling distribution

$$mean(\hat{p}) \approx p$$

Categorical variables

CLT for proportions

The distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size:

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p \cdot (1-p)}{n}} \right)$$

Conditions for CLT:

1. **Independence**: Sampled observations must be independent
2. **Sample size/skew**: There should be at least 10 successes and 10 failures in the sample: $n \cdot p \geq 10$ and $n \cdot (1-p) \geq 10$
   1. If p is unknown, use $\hat{p}$

Confidence interval for a proportion

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to a thousand people with high blood pressure, and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

1. All 1000 get the drug
2. 500 get the drug, 500 don’t

Categorical variables

Confidence interval for a proportion

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to a thousand people with high blood pressure, and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

1. All 1000 get the drug
2. 500 get the drug, 500 don’t

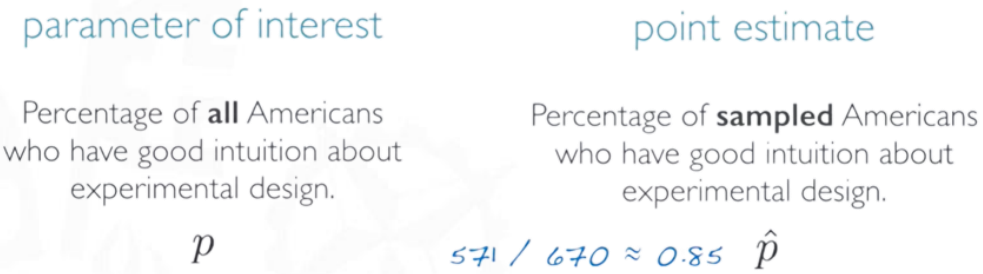
This same question was asked to 670 americans:

Option 1 (bad int.)	99
Option 2 (good int.)	571
Total	670

Categorical variables

Confidence interval for a proportion

Estimate the proportion of all americans who have good intuition about experimental design:



parameter of interest

Percentage of **all** Americans who have good intuition about experimental design.

p

point estimate

Percentage of **sampled** Americans who have good intuition about experimental design.

571 / 670 ≈ 0.85 $\hat{p}$

As always:

$$point\ estimate \pm margin\ of\ error$$
$$\hat{p} \pm z^* \cdot SE_{\hat{p}}$$

Where:

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

Categorical variables

Confidence interval for a proportion

$$CI_{95\%}\colon \hat{p} \pm z^* \cdot SE_{\hat{p}}$$

$$= 0.85 \pm 1.96 \cdot \sqrt{\frac{0.85 \cdot 0.15}{670}}$$

$$= 0.85 \pm 1.96 \cdot 0.0138$$

$$= (0.823, 0.877)$$

We are 95% confident that 82.3% to 87.7% of all americans have good intuition about experimental design.

Categorical variables

Required sample size for a given ME

$$ME = z^* \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Using this formula, we can obtain n for a given ME. If we do not have a sample yet:

- If there is a previous study, rely on the value calculated in that study for $\hat{p}$.
- If not, use $\hat{p} = 0.5$
  - If you do not know any better, 50-50 is a good guess
  - Gives the most conservative estimate for n

Remember to always **round up** n.

Categorical variables

Hypothesis testing for a proportion

1. Set the hypothesis
   - $H_0: p = null\ value$
   - $H_A: p < or > or \neq null\ value$
2. Calculte the point estimate
3. Check conditions
   1. Independence
   2. Sample size/skew
4. Calculate test statistic

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

5. Make a decision based on p-value and α

Categorical variables

Confidence interval for a proportion

A 2013 Pew Research poll found that 60% of 1,983 randomly sampled American adults believe in evolution. Does this provide convincing evidence that majority of Americans believe in evolution?

$H_0: p = 0.5$
$H_A: p > 0.5$
$\hat{p} = 0.6$
$n = 1983$

$$\hat{p} \sim N\left(0.5, \sqrt{\frac{0.5 \cdot (1 - 0.5)}{1983}} \approx 0.0112\right)$$

$$Z = \frac{0.6 - 0.5}{0.0112} \approx 8.92 \gg 3 \rightarrow p - value \approx 0$$

Alberto Aguilera 23, E-28015 Madrid - Tel: +34 91 542 2800 - http://www.iit.comillas.edu