CSCI 3022 Intro to Data Science
Two-Sample CIs

The General Social Survey is a sociological survey used to collect data on demographic characteristics and attitudes of the residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours per day. Suppose further that the known standard deviation of the characteristic is 2 hours per day. Find a 95% confidence interval for the amount of relaxation hours per day.

Opening sol:

Opening sol:

We want the CI!
The CLT tells us where the sample mean comes from: $\bar{X} \sim N(\mu, \frac{2^2}{1000})$,
...but we know $\bar{X} = 3.6$ and are asking about μ!

This is a CI of

$$\bar{X} \pm z_{.025} \frac{2}{\sqrt{1000}}$$

Opening sol:

We want the CI!
The CLT tells us where the sample mean comes from: $\bar{X} \sim N(\mu, \frac{2^2}{1000})$,
...but we know $\bar{X} = 3.6$ and are asking about μ!

This is a CI of

$$\bar{X} \pm z_{.025} \frac{2}{\sqrt{1000}}$$

$$= [3.48, 3.72]$$

Announcements and Reminders

- Practicum delayed to Monday after CEAS spring pause. Also a HW due that Friday, since that should be more than enough time for the practicum!

Opening Followup:

Concept Check: In the previous example we found a 95% CI for relaxation time to be $[3.48, 3.72]$. Which of the following statements are true?

1. 95% of Americans spend 3.48 to 3.72 hours per day relaxing after work.
2. 95% of random samples of 1000 residents will yield CIs that contain the true average number of hours that Americans spend relaxing after work each day.
3. 95% of the time the true average number of hours an American spends relaxing after work is between 3.48 and 3.72 hours per day.
4. We are 95% sure that Americans in this sample spend 3.48 to 3.72 hours per day relaxing after work.

Where we at?

Last time we used the Central Limit Theorem (TL; DR: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$) to write probability statements regarding *random intervals* covering the desired parameter: the population mean μ. These boiled down to the same form:

1. The confidence interval for the population mean μ was:

Where we at?

Last time we used the Central Limit Theorem (TL; DR: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$) to write probability statements regarding *random intervals* covering the desired parameter: the population mean μ. These boiled down to the same form:

1. The confidence interval for the population mean μ was: $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$

Where we at?

Last time we used the Central Limit Theorem (TL; DR: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$) to write probability statements regarding *random intervals* covering the desired parameter: the population mean μ. These boiled down to the same form:

1. The confidence interval for the population mean μ was:

$$\underbrace{\bar{X}}_{\textbf{Point}\text{ estimate for }\mu} \quad \pm \quad \overbrace{z_{\frac{\alpha}{2}}}^{\text{error/precision term}} \quad \cdot \quad \underbrace{\frac{\sigma}{\sqrt{n}}}_{\textbf{Standard Error}\text{ of the sample mean}}$$

Where we at?

Last time we used the Central Limit Theorem (TL; DR: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$) to write probability statements regarding *random intervals* covering the desired parameter: the population mean μ. These boiled down to the same form:

1. The confidence interval for the population mean μ was:

2. When we don't know σ, we use s instead:

$$\underbrace{\bar{X}}_{\textbf{Point}\text{ estimate for }\mu} \pm \overbrace{z_{\frac{\alpha}{2}}}^{\text{error/precision term}} \cdot \underbrace{\frac{s}{\sqrt{n}}}_{\text{Estimated }\textbf{Standard Error}\text{ of the sample mean}}$$

Large Sample Confidence Intervals

A difficulty in using our previous equation for confidence intervals is that it uses the value σ of which will rarely be known.

Also, we may want a CI for a mean from some other non-normal distribution.

Large Sample Confidence Intervals

In this instance, we need to work with the sample standard deviation s. Remember from our first lesson that the standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$$

With this, we instead work with the standardized random variable:

Large Sample Confidence Intervals

In this instance, we need to work with the sample standard deviation s. Remember from our first lesson that the standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

With this, we instead work with the standardized random variable:

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Large Sample Confidence Intervals

Previously, there was randomness only in the numerator of Z by virtue of the estimator __.

In the new standardized variable, both __ and _ vary in value from one sample to another.

When n is large, the substitution of s for σ adds little extra variability, so nothing needs to change.

When n is smaller, the distribution of this new variable should be wider than the normal to reflect the extra uncertainty. (We talk more about this soon!)

Large Sample Confidence Intervals

Previously, there was randomness only in the numerator of Z by virtue of the estimator $\underline{\bar{X}}$.

In the new standardized variable, both $\underline{\bar{X}}$ and $\underline{s}$ vary in value from one sample to another.

When n is large, the substitution of s for σ adds little extra variability, so nothing needs to change.

When n is smaller, the distribution of this new variable should be wider than the normal to reflect the extra uncertainty. (We talk more about this soon!)

Large Sample Confidence Intervals

Large Sample CI: If n is sufficiently large $(n \geq 30)$, the standardized random variable

has approximately a standard normal distribution. This implies that

is a large-sample confidence interval for μ with confidence level approximately $100(1-\alpha)\%$. This formula is valid regardless of the population distribution for sufficiently large n.

Large Sample Confidence Intervals

Large Sample CI: If n is sufficiently large $(n \geq 30)$, the standardized random variable

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has approximately a standard normal distribution. This implies that

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

is a large-sample confidence interval for μ with confidence level approximately $100(1-\alpha)\%$. This formula is valid regardless of the population distribution for sufficiently large n.

CI overview

1. The confidence interval with σ applied when we knew σ, and *either* the sample was large or we knew it was coming from a normal distribution.

2. The interval using s to approximate σ applies **only when the sample was large**.

	$n \geq 30$	$n < 30$
Underlying Normal Distribution	σ known	σ known
	σ unknown	σ unknown
Underlying Non-Normal Distribution	σ known	σ known
	σ unknown	σ unknown

Method:
Z or approximately Z by Central Limit Theorem

Special Cases: Populations

Let p denote the proportion of "successes" in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of n individuals is selected, and X is the number of successes in the sample.

Then, X can be modeled as a ________ rv with mean of __ and

variance of ________

Special Cases: Populations

Let p denote the proportion of "successes" in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of n individuals is selected, and X is the number of successes in the sample.

Then, X can be modeled as a Binomial rv with mean of $\underline{np}$ and

variance of $\underline{np(1-p)}$

Special Cases: Populations

Let p denote the proportion of "successes" in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of n individuals is selected, and X is the number of successes in the sample.

Then, X can be modeled as a Binomial rv with mean of $\underline{np}$ and

variance of $\underline{np(1-p)}$

If both $np > 10$ and $n(1-p) > 10$, X has approximately a normal distribution.

Special Cases: Populations

The estimator of p is: $\hat{p}=$ ____

Standardizing the estimator yields:

and a resulting CI is:

Special Cases: Populations

The estimator of p is: $\hat{p}= \underline{X/n}$

Standardizing the estimator yields:

and a resulting CI is:

Special Cases: Populations

The estimator of p is: $\hat{p}= \underline{X/n}$

This estimator is approximately normally distributed and:

$$E[\hat{p}] = p \qquad Var[\hat{p}] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Standardizing the estimator yields:

and a resulting CI is:

Special Cases: Populations

The estimator of p is: $\hat{p}= \underline{X/n}$

This estimator is approximately normally distributed and:

$$E[\hat{p}] = p \qquad Var[\hat{p}] = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

Standardizing the estimator yields:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

and a resulting CI is:

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

Special Cases: Populations

Example:
The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 (63.5%) of these sampled households to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportional of homes with indoor radon levels above 4 pCi/L.

Special Cases: Populations

Example:
The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 (63.5%) of these sampled households to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportional of homes with indoor radon levels above 4 pCi/L.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}};$$ `stats.norm.ppf(0.995) = 2.57;`

use $\hat{p}$ where we must; $$= 0.635 \pm 2.57\sqrt{\frac{0.635(1-0.635)}{200}}$$

$$= [0.548, 0.722]$$

How about a pair?

Univariate data is pretty boring. We often want to be able to compare options and reach a decision:

1. Is a drug's effectiveness the same in children and adults?
2. Does cigarette brand X contain more nicotine than brand Y?
3. Does a class perform better when taught using method One or method Two?
4. Does organizing a website give better user exp. using format A or format B?... or more clicks/customers?

How about a pair?

Univariate data is pretty boring. We often want to be able to compare options and reach a decision:

1. Is a drug's effectiveness the same in children and adults?
2. Does cigarette brand X contain more nicotine than brand Y?
3. Does a class perform better when taught using method One or method Two?
4. Does organizing a website give better user exp. using format A or format B?... or more clicks/customers?

$\Longrightarrow$ **"A/B testing"**

Comparing 2 Means

How do two populations compare, in terms of their means?

To try to answer this question, we collect samples from both populations and perform inference on both samples to draw conclusions about $\mu_1 - \mu_2$.

Comparing 2 Means

Basic Assumptions:

Note: We haven't made any distributional assumptions, for now.

Comparing 2 Means

Basic Assumptions:

1. $X_1, X_2, \ldots X_n$ are a random sample from distribution 1 with mean μ_1 (or μ_X) and SD σ_1.
2. $Y_1, Y_2, \ldots Y_m$ are a random sample from distribution 2 with mean μ_2 and SD σ_2.
3. The X and Y sample are independent of one another.

Note: We haven't made any distributional assumptions, for now.

Comparing 2 Means

The natural estimator of $\mu_1 - \mu_2$ is ______.

Inferential procedures are based on standardizing estimators, so we'll need the mean and standard deviation of ______.

Comparing 2 Means

The natural estimator of $\mu_1 - \mu_2$ is $\underline{\bar{X} - \bar{Y}}$.

Inferential procedures are based on standardizing estimators, so we'll need the mean and standard deviation of $\underline{\bar{X} - \bar{Y}}$.

Comparing 2 Means

Mean of $\bar{X} - \bar{Y}$:

Variance/Standard Deviation of $\bar{X} - \bar{Y}$:

Comparing 2 Means

Mean of $\bar{X} - \bar{Y}$:

$$E[\bar{X} - \bar{Y}] = E\left[\frac{\sum_i X_i}{n} - \frac{\sum_j Y_j}{m}\right] = \cdots = \mu_1 - \mu_2$$

Variance/Standard Deviation of $\bar{X} - \bar{Y}$:

$$Var\left[\bar{X} - \bar{Y}\right] = Var\left[\frac{\sum_i X_i}{n} - \frac{\sum_j Y_j}{m}\right] = Var[\bar{X}] + Var[\bar{Y}] = \ldots$$

$$= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$$

Comparing 2 Means

Normal Populations with known variances:

If both populations are normal, both __ and __ have normal distributions.

Further if the samples are independent, then the sample means are independent of one another.

Thus, ______ is normally distributed with expected value ______ and standard deviation:

Comparing 2 Means

Normal Populations with known variances:
If both populations are normal, both $\underline{\bar{X}}$ and $\underline{\bar{Y}}$ have normal distributions.

Further if the samples are independent, then the sample means are independent of one another.

Thus, $\underline{\bar{X} - \bar{Y}}$ is normally distributed with expected value $\underline{\mu_1 - \mu_2}$ and standard deviation:

$$\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Comparing 2 Means

$$So : \left(\bar{X} - \bar{Y}\right) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Standardizing our estimator gives:

Therefore, the $(1-\alpha) \cdot 100\%$ confidence interval is:

Comparing 2 Means

$$So : (\bar{X} - \bar{Y}) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Standardizing our estimator gives:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

Therefore, the $(1 - \alpha) \cdot 100\%$ confidence interval is:

Comparing 2 Means

$$So : (\bar{X} - \bar{Y}) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Standardizing our estimator gives:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

Therefore, the $(1 - \alpha) \cdot 100\%$ confidence interval is:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Comparing 2 Means: Large Sample

If both n_1 and n_2 are large then the CLT implies that our confidence interval is valid even without the assumption of normal populations. In this case, the confidence level is *approximately* $(1-\alpha)\cdot 100\%$.
Further, we can replace sample standard deviations for population standard deviations:

So the $(1-\alpha)\cdot 100\%$ confidence interval is:

Comparing 2 Means: Large Sample

If both n_1 and n_2 are large then the CLT implies that our confidence interval is valid even without the assumption of normal populations. In this case, the confidence level is *approximately* $(1-\alpha)\cdot 100\%$.
Further, we can replace sample standard deviations for population standard deviations:

$$Z = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

So the $(1-\alpha)\cdot 100\%$ confidence interval is:

Comparing 2 Means: Large Sample

If both n_1 and n_2 are large then the CLT implies that our confidence interval is valid even without the assumption of normal populations. In this case, the confidence level is *approximately* $(1-\alpha)\cdot 100\%$.
Further, we can replace sample standard deviations for population standard deviations:

$$Z = \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}}$$

So the $(1-\alpha)\cdot 100\%$ confidence interval is:

$$\left(\bar{X}-\bar{Y}\right) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}$$

Comparing 2 Means: Large Sample

Example:
Suppose you run two different email ad campaigns over many days and record the amount of traffic driven to your website on days that each ad was sent. Ad 1 was sent on 50 different days and generates an average of 2 million page views per day, with a SD of 1 million page views. Ad 2 was sent on 40 different days and generates an average of 2.25 million page views per day, with SD of half a million views. Find 95% confidence intervals for the average page views for each ad (in units of millions of views).

Comparing 2 Means: Large Sample

Example: $\bar{X} = 2,\ s_1 = 1,\ n = 50; \bar{Y} = 2.25,\ s_2 = 0.5,\ m = 40;$
CI for μ_1:

CI for μ_2:

Comparing 2 Means: Large Sample

Example: $\bar{X} = 2,\ s_1 = 1,\ n = 50; \bar{Y} = 2.25,\ s_2 = 0.5,\ m = 40;$
CI for μ_1:

$$\bar{X} \pm 1.96 \frac{s_X}{\sqrt{n}} = 2 \pm 1.96 \frac{1}{\sqrt{50}} = [1.723, 2.277]$$

CI for μ_2:

Comparing 2 Means: Large Sample

Example: $\bar{X} = 2,\ s_1 = 1,\ n = 50; \bar{Y} = 2.25,\ s_2 = 0.5,\ m = 40;$
CI for μ_1:

$$\bar{X} \pm 1.96\frac{s_X}{\sqrt{n}} = 2 \pm 1.96\frac{1}{\sqrt{50}} = [1.723, 2.277]$$

CI for μ_2:

$$\bar{Y} \pm 1.96\frac{s_Y}{\sqrt{m}} = 2.25 \pm 1.96\frac{0.5}{\sqrt{40}} = [2.095, 2.405]$$

Comparing 2 Means: Large Sample

Example: $\bar{X} = 2,\ s_1 = 1,\ n = 50; \bar{Y} = 2.25,\ s_2 = 0.5,\ m = 40;$
CI for μ_1:

$$\bar{X} \pm 1.96 \frac{s_X}{\sqrt{n}} = 2 \pm 1.96 \frac{1}{\sqrt{50}} = [1.723, 2.277]$$

CI for μ_2:

$$\bar{Y} \pm 1.96 \frac{s_Y}{\sqrt{m}} = 2.25 \pm 1.96 \frac{0.5}{\sqrt{40}} = [2.095, 2.405]$$

What does this tell us?

Comparing 2 Means: Large Sample

A: **Not much!** These things overlap, which makes it hard to tell if that .25 million difference matters. So we should instead be asking about $\mu_1 - \mu_2$! CI for $\mu_1 - \mu_2$:

A: While ad 2 looks a little better than ad 1, at our chosen tolerance for errors (at most 5%!), there's a reasonable chance that the difference we're observing was simple random volatility, and there is no **significant** difference.

Comparing 2 Means: Large Sample

A: **Not much!** These things overlap, which makes it hard to tell if that .25 million difference matters. So we should instead be asking about $\mu_1 - \mu_2$! CI for $\mu_1 - \mu_2$:

$$\bar{X} - \bar{Y} \pm 1.96\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} = -.25 \pm 1.96\sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}} = [-0.568, 0.068]$$

What does this tell us?

A: While ad 2 looks a little better than ad 1, at our chosen tolerance for errors (at most 5%!), there's a reasonable chance that the difference we're observing was simple random volatility, and there is no **significant** difference.

Comparing 2 Means: Proportions

Now consider the comparison of two population proportions. Just as before, an individual or object is a success if some characteristic of interest is present ("graduated from college", a refrigerator "with an icemaker", etc.).

Let:
$p_1 =$ the true proportion of successes in population 1
$p_2 =$ the true proportion of successes in population 2

Goal: Determine whether one proportion is bigger than the other. In other words: we make an interval for $p_1 - p_2$.

Comparing 2 Means: Proportions

Mean of $\hat{p}_1 - \hat{p}_2$:

Variance/Standard Deviation of $\hat{p}_1 - \hat{p}_2$:

Comparing 2 Means: Proportions

Mean of $\hat{p_1} - \hat{p_2}$:

$$E[\hat{p_1} - \hat{p_2}] = p_1 - p_2$$

Variance/Standard Deviation of $\hat{p_1} - \hat{p_2}$:

$$Var[\hat{p_1} - \hat{p_2}] = Var[\hat{p_1}] + Var[\hat{p_2}] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Comparing 2 Means: Proportions

Mean of $\hat{p_1} - \hat{p_2}$:

$$E[\hat{p_1} - \hat{p_2}] = p_1 - p_2$$

Variance/Standard Deviation of $\hat{p_1} - \hat{p_2}$:

$$Var[\hat{p_1} - \hat{p_2}] = Var[\hat{p_1}] + Var[\hat{p_2}] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$SD : \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

Comparing 2 Means: Proportions

So, a $(1-\alpha)\cdot 100\%$ confidence interval for $\hat{p_1}-\hat{p_2}$ is:

This interval can safely be used as long as

$$n_1\hat{p_1};\ n_1(1-\hat{p_1});\ n_2\hat{p_2};\ n_2(1-\hat{p_2});$$

are all at least 10.

Comparing 2 Means: Proportions

So, a $(1-\alpha) \cdot 100\%$ confidence interval for $\hat{p_1} - \hat{p_2}$ is:

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

This interval can safely be used as long as

$$n_1\hat{p_1};\ n_1(1-\hat{p_1});\ n_2\hat{p_2};\ n_2(1-\hat{p_2});$$

are all at least 10.

Comparing 2 Means: Proportions

Example:
The authors of the article "Adjuvant Radiotherapy and Chemotherapy in Node- Positive Premenopausal Women with Breast Cancer" (New Engl. J. of Med., 1997: 956–962) reported on the results of an experiment designed to compare treating cancer patients with chemotherapy only to treatment with a combination of chemotherapy and radiation.

Of the 154 individuals who received the chemotherapy-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least that long. What is the 99% confidence interval for this difference in proportions?

Comparing 2 Means: Large Sample

Example: $\hat{p_1} = 76/154$, $\hat{p_2} = 98/165$, $z_{0.005} = 2.576$

CI for $p_1 - p_2$:

Comparing 2 Means: Large Sample

Example: $\hat{p_1} = 76/154$, $\hat{p_2} = 98/165$, $z_{0.005} = 2.576$
The pooled standard deviation estimator is

$$\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = \sqrt{\frac{0.\hat{4}94(1-0.\hat{4}94)}{154} + \frac{0.\hat{5}98(1-0.\hat{5}98)}{165}}$$

≈ 0.0555

CI for $p_1 - p_2$:

Comparing 2 Means: Large Sample

Example: $\hat{p_1} = 76/154$, $\hat{p_2} = 98/165$, $z_{0.005} = 2.576$
The pooled standard deviation estimator is

$$\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = \sqrt{\frac{0.\hat{4}94(1-0.\hat{4}94)}{154} + \frac{0.\hat{5}98(1-0.\hat{5}98)}{165}}$$

≈ 0.0555

CI for $p_1 - p_2$:

$$\frac{76}{154} - \frac{98}{165} \pm 2.576 \cdot 0.0555 = [-0.247, 0.039]$$

What does this tell us?

Comparing 2 Means: Proportions

On occasion an inference concerning $p_1 - p_2$ may have to be based on samples for which at least one sample size is small.

Appropriate methods for such situations are not as straightforward as those for large samples, and there is more controversy among statisticians as to recommended procedures.

One frequently used test, called the Fisher–Irwin test, is based on the hypergeometric distribution.

Your friendly neighborhood statistician can be consulted for more information.

CI overview

1. The first interval with σ applied when we knew σ, and *either* the sample was large or we knew it was coming from a normal distribution.

2. The second interval with s applied only when the sample was large.

3. **What do we do if the sample size is small?**

	$n \geq 30$	$n < 30$
Underlying Normal Distribution	σ known	σ known
	σ unknown	$\boxed{\sigma \text{ unknown}}$
Underlying Non-Normal Distribution	σ known	σ known
	σ unknown	σ unknown

Method:

Z or approximately Z by Central Limit Theorem

The t Distribution

We've danced around the idea that we can't just replace σ with s when the sample size is small, even if we know the underlying population is normal. Let's formalize!

The results on which large sample inferences are based introduces a new family of probability distributions called **t distributions**.

When __ is the mean of a random sample of size n from a normal distribution with mean _, the random variable

has a probability distribution called a t Distribution with n–1 degrees of freedom (df).

The t Distribution

We've danced around the idea that we can't just replace σ with s when the sample size is small, even if we know the underlying population is normal. Let's formalize!

The results on which large sample inferences are based introduces a new family of probability distributions called **t distributions**.

When $\underline{\bar{X}}$ is the mean of a random sample of size n from a normal distribution with mean $\underline{\mu}$, the random variable

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a probability distribution called a t Distribution with n–1 degrees of freedom (df).

The t Distribution

Main idea:

With the t-distribution, we're accounting for a second approximation. Not only do we have to approximate

μ (with __)

We also now have to approximate σ (with _).

The t Distribution

Main idea:
With the t-distribution, we're accounting for a second approximation. Not only do we have to approximate

μ (with $\underline{\bar{X}}$)

We also now have to approximate σ (with $\underline{s}$).

The t Distribution

Main idea:
With the t-distribution, we're accounting for a second approximation. Not only do we have to approximate

μ (with $\underline{\bar{X}}$)

We also now have to approximate σ (with $\underline{s}$).

When our sample size is small, this is often a costly approximation, and as a result we have to *widen* our confidence intervals.

The cost of this approximation scales with n, so as n is smaller, we need to widen our intervals even more.

The t Distribution

Main idea:
With the t-distribution, we're accounting for a second approximation. Not only do we have to approximate

μ (with $\underline{\bar{X}}$)

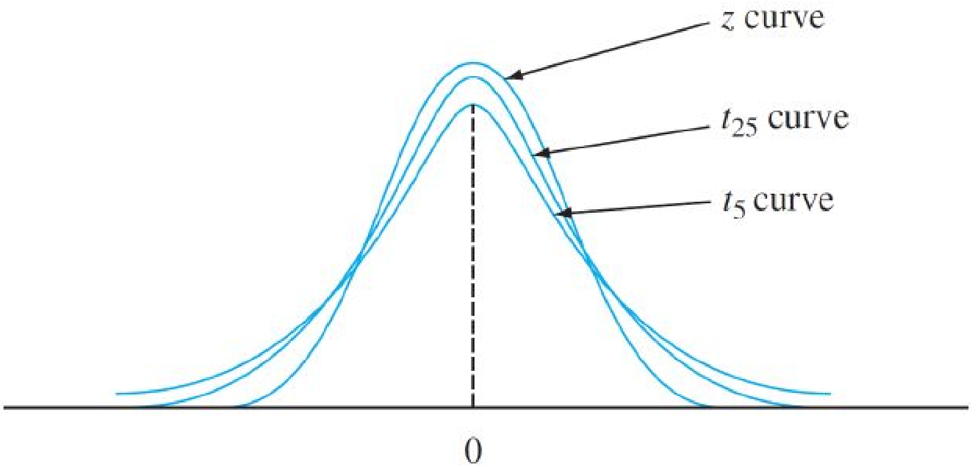
We also now have to approximate σ (with $\underline{s}$).

When our sample size is small, this is often a costly approximation, and as a result we have to *widen* our confidence intervals.

The cost of this approximation scales with n, so as n is smaller, we need to widen our intervals even more.

Intuition: Should t_α be greater or less than z_α?

The t



Properties of the t

Let t_ν denote the t distribution with ν df.

1. Each t_ν curve is bell-shaped and centered at 0.
2. Each t_ν curve is more spread out than the standard normal (z) curve.
3. As ν increases, the spread of the corresponding t_ν curve decreases.
4. As ν ______ the sequence of t_ν curves approaches the standard normal curve (so the z curve is the t curve with df $=$ ____)

Properties of the t

Let t_ν denote the t distribution with ν df.

1. Each t_ν curve is bell-shaped and centered at 0.

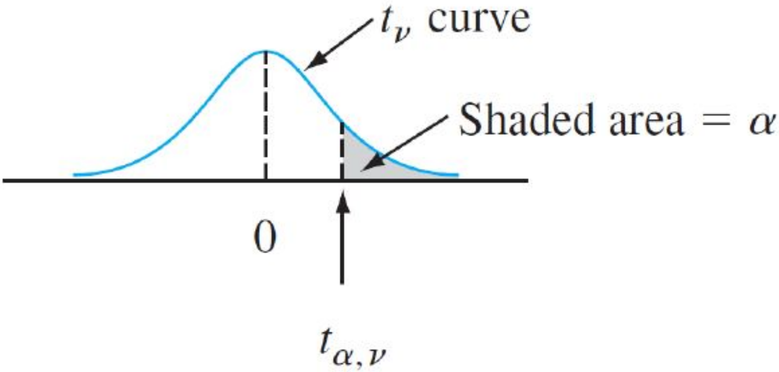
2. Each t_ν curve is more spread out than the standard normal (z) curve.

3. As ν increases, the spread of the corresponding t_ν curve decreases.

4. As ν $\underline{\to \infty\quad}$ the sequence of t_ν curves approaches the standard normal curve (so the z curve is the t curve with df = $\underline{\infty\quad}$)

The t

Let $t_{\alpha,\nu}$ = the number on the measurement axis for which the area under the t curve with ν df to the right of t_ν is α;
$t_{\alpha,\nu}$ is called a t critical value.



For example, $t_{.05,6}$ is the t critical value that captures an upper-tail area of .05 under the t

Finding t-values:

The probabilities of t curves are found in a similar way as the normal curve.

Example: obtain $t_{.05,15}$

Finding t-values:

The probabilities of t curves are found in a similar way as the normal curve.

Example: obtain $t_{.05,15}$

```
stats.t.ppf(.95,15)
```

The t Confidence Interval

Let ________ and ________ be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ. Then a $100(1-\alpha)\%$ t-confidence interval for the mean μ is

$$\left[\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

or, more compactly:

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

The t Confidence Interval

Example: Example: Suppose that the GPA measurements for 23 students follow a normal distribution. The sample mean is 3.146. The sample standard deviation is 0.308. Calculate a 90% CI for the mean GPA.

The t Confidence Interval

Example: Example: Suppose that the GPA measurements for 23 students follow a normal distribution. The sample mean is 3.146. The sample standard deviation is 0.308. Calculate a 90% CI for the mean GPA.

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

The t Confidence Interval

Example: Example: Suppose that the GPA measurements for 23 students follow a normal distribution. The sample mean is 3.146. The sample standard deviation is 0.308. Calculate a 90% CI for the mean GPA.

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$3.146 \pm 1.7171 \cdot \frac{.308}{\sqrt{23}}$$

since `stats.t.ppf(.95,22)`$= t_{.05} = 1.7171$ (compare to $z_{.05} = 1.644!$)

Daily Recap

Today we learned

1. Making *inference* via confidence intervals on the mean or means.

Moving forward:

- **Hypthesis Testing** next week

Next time in lecture:

- Finishing up CI and relaxing assumptions.