CSCI 3022 Intro to Data Science
CI Wrapup; Testing Intro

$\bar{X} \pm 5$ units

$\bar{x}-5 \qquad \bar{x} \qquad \bar{x}+5$

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 units?

Sample Size Calculations

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of ~~(at most)~~ 10 units?

z

CI: $\bar{X} \pm z_{\alpha/2} \left(\sigma/\sqrt{n} \right)$ or $\left(s/\sqrt{n} \right)$

≈ 1.96 $\sigma = 25$ choose n.

GOAL: $z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq 5$. units

1.96 25

Sample Size Calculations

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 units?
The width is $W = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. We want:

Sample Size Calculations

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 units?

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \cancel{10}\ 5$$

$$\implies z_{\alpha/2}\frac{\sigma}{\cancel{10}\ 5} < \sqrt{n}$$

$$\implies \left(z_{\alpha/2}\frac{\sigma}{\cancel{10}\ 5}\right)^2 < n$$

$$\approx \left(1.96 \cdot \frac{25}{5}\right)^2 \approx (10)^2 \approx 100$$

Note:
n is a sample size, round UP to next integer.

Announcements and Reminders

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

Where we at?

$$\sqrt{Var[\bar{X}]} = \frac{\sqrt{Var[X]}}{\sqrt{n}}$$

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

   normality   Stats.norm.ppf

2. When we don't know σ, we use s instead: $\underbrace{\bar{X}}_{\textbf{Point} \text{ estimate for } \mu} \pm \overbrace{z_{\frac{\alpha}{2}}}^{\text{error/precision term}} \cdot \underbrace{\frac{s}{\sqrt{n}}}_{\text{Estimated } \textbf{Standard Error} \text{ of the sample mean}}$

3. The same principle applies to a *proportion*, only now the $E[\hat{p}]$ and $SD[\hat{p}]$ come from the properties of a **binomial**

$$\underbrace{\hat{p}}_{\textbf{Point} \text{ estimate for } p} \pm \overbrace{z_{\alpha/2}}^{\text{error/precision term}} \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{Estimated } \textbf{Standard Error}}$$

look up!

$$\sqrt{\frac{\text{Var of a binomial}}{n^2}}$$

2 Sample CIs

For comparing two samples, we could ask "which mean is larger" by computing a $100(1-\alpha)\%$ CI on the difference in the means $\mu_1 - \mu_2$.

$\mu_1 - \mu_2 > 0 \Rightarrow \mu_1 > \mu_2$

$\mu_1 - \mu_2 < 0 \Rightarrow \mu_2 > \mu_1$

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

Standardizing our estimator gives:

add variances not std. dev.

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

This suggested the possibility of a **decision** rule. More on that, shortly...

2 Sample CIs

For comparing two samples, we could ask "which mean is larger" by computing a $100(1-\alpha)\%$ CI on the difference in the means $\mu_1 - \mu_2$.

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

Standardizing our estimator gives:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

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

This suggested the possibility of a **decision** rule. More on that, shortly...

2 Sample CIs

For comparing two samples, we could ask "which mean is larger" by computing a $100(1-\alpha)\%$ CI on the difference in the means $\mu_1 - \mu_2$.

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

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:

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

GOAL:
Conclude
$\mu_1 > \mu_2$
$\mu_2 > \mu_1$, or neither.

This suggested the possibility of a **decision** rule. More on that, shortly...

Comparing 2 Means: Large Sample

σ bad, S good.

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* $(1-\alpha)\cdot 100\%$.

S in for σ

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:

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

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:

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

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

$$(\bar{X}-\bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

added

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:

μ: $\bar{X} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

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:

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

2

CI for μ_2:

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

2.25

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

if we relax z : 80%

confidence, 1.96 would decrease.

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 $\mu_1 - \mu_2$:

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

middle: -.25

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.

If 0 is in the CI, the mean $(\mu_1 - \mu_2)$ might be zero $\Rightarrow$ Ads could be equivalent.

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.).

$p_1 - p_2$

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

$\hat{p}_1 - \hat{p}_2$

Samples

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

$\hat{p}_1 - \hat{p}_2$ $p_1 - p_2$

$$E[\hat{p}_1 - \hat{p}_2] = p_1 - p_2$$

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

indep.

$$Var[\hat{p}_1 - \hat{p}_2] = Var[\hat{p}_1] + Var[\hat{p}_2] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Sample 1 Sample 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 $\hat{p_1} - \hat{p_2}$:

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

Pop.

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

Sample

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:

point/guess

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

sample variances

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.

10 success

10 failures

$\Rightarrow$ use normal

(otherwise use binomial).

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

$$\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = \sqrt{\frac{0.\hat{494}(1-0.\hat{494})}{154} + \frac{0.\hat{598}(1-0.\hat{598})}{165}}$$

≈ 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

$$\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = \sqrt{\frac{0.\hat{494}(1-0.\hat{494})}{154} + \frac{0.\hat{598}(1-0.\hat{598})}{165}}$$

≈ 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.

	$n \geq 30$	$n < 30$
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

← underlying normal
using s, not σ.

small
sample

Method:
Z or approximately Z by Central Limit Theorem

The t Distribution

$\sigma/\sqrt{n}$ or $s/\sqrt{n}$

We've danced around the idea that we can't just replace $\boxed{\sigma}$ with $\boxed{s}$ 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 $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the random variable

$$t_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Not $N(0,1)$ ... but it's kinda close to it....

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 $\bar{X}$)

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

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

normal: μ: $\bar{X} \pm z_{\alpha/2}\, s/\sqrt{n}$

t: μ: $\bar{X} \pm$ more $s/\sqrt{n}$

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.

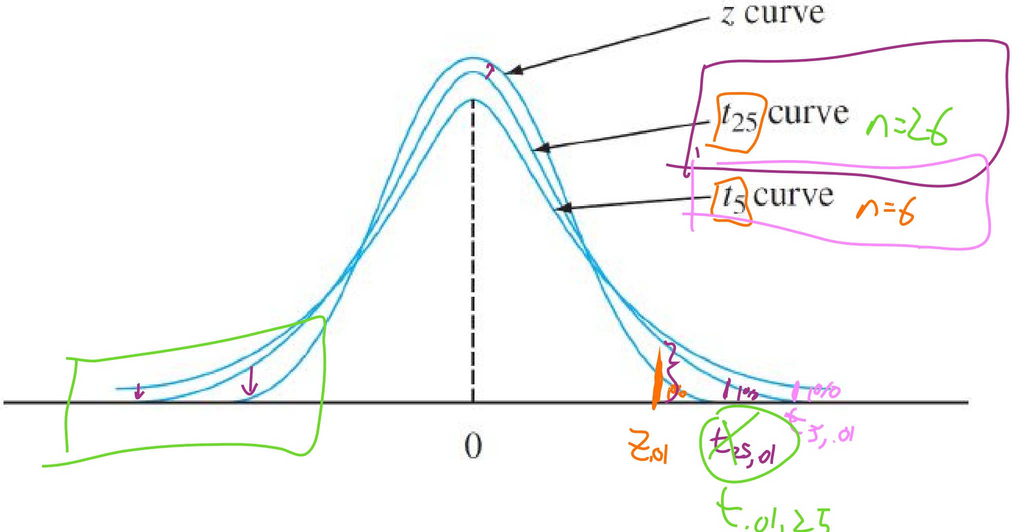
The cost of this approximation scales with n, so as n is smaller, we need to widen our intervals even more.

Intuition: Should t_α be greater or less than z_α?

$t_\alpha > z_\alpha$

The t

stats.t.ppf(prob, $\boxed{\text{dof}}$) .25, .5, etc. $\boxed{n-1}$



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. (like $N(0,1)$).
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.
   ($\nu = n-1$)
4. As ν $\underline{\to \infty}$ the sequence of t_ν curves approaches the standard normal curve (so the z curve is the t curve with df = $\underline{\infty}$)

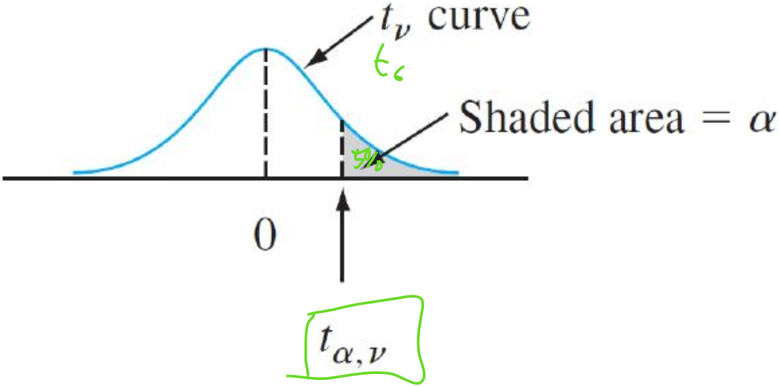
Properties of the t

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)
```

(prob to the left, d.o.f. = n-1)

The t Confidence Interval

Let ___$\bar{X}$___ and ___s___ 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

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

$z_{\alpha/2}$: $t_{\alpha/2, n-1}$

or, more compactly:

$$\bar{X} \pm \boxed{t_{\alpha/2}}_{n-1}\frac{s}{\sqrt{n}}$$

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.

n

$\bar{x}$

s

$\alpha = .1$

$t_{\alpha/2, n-1}$

$t_{.05, 22}$

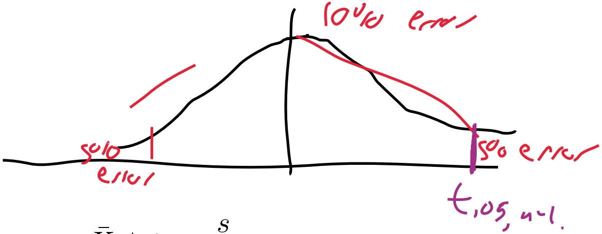
The t Confidence Interval

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.



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

.1 error α=.05

.975

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

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

Now what?

Decomposing an interval like the interval from our two-sample proportion test

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

into a yes or no decision is how we transition into statistical hypothesis testing. Based on our confidence interval on $p_1 - p_2$, we can try to answer whether $p_1 = p_2$, $p_1 < p_2$, etc.

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

Examples:

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

Examples:

1. Company A makes parts that last longer than company B.

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

Examples:

1. Company A makes parts that last longer than company B.
2. In Boulder, it's usually a colder maximum daily temperature in February than June.

Statistical Hypotheses

Definition: *Statistical Hypothesis*
A *Statistical Hypothesis* is a claim about the value of a parameter or population characteristic.

Examples:

1. Company A makes parts that last longer than company B.
2. In Boulder, it's usually a colder maximum daily temperature in February than June.
3. Students in Zach's sections are generally much more dashing, resourceful, and socially meritorious than students in other sections.

Statistical Hypotheses

One example statisticians often revisit: is a coin fair?
This is a real world question!

`https://www.newscientist.com/article/dn1748-euro-coin-accused-of-unfair-flipping/`

As the Euro was introduced, Polish Mathematicians claimed that the Belgian 1 Euro coin was weighted so that it was more likely ot return a heads!
Suppose I handed you such a coin. How would you decide whether it was fair?

Logic of Hypothesis Testing

Analogy: Jury in a criminal trial.

When a defendant is accused of a crime, the jury (is supposed to) presumes that she is not guilty (not guilty; that's the "null hypothesis").

Then, we gather evidence. If the evidence is seems implausible under the assumption of non-guilt, we might reject non-guilt and claim that the defendant is (likely) guilty.

Logic of Hypothesis Testing

Important Question: Is there strong evidence for the alternative?

The burden of proof is placed on those who believe in the alternative claim.

The initially favored claim, the null hypothesis H_0, will not be rejected in favor of the alternative hypothesis, H_a or H_1, unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions:

Logic of Hypothesis Testing

Important Question: Is there strong evidence for the alternative?

The burden of proof is placed on those who believe in the alternative claim.

The initially favored claim, the null hypothesis H_0, will not be rejected in favor of the alternative hypothesis, H_a or H_1, unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions:
Fail to Reject the null hypothesis if there is insufficient statistical evidence to do so.

Logic of Hypothesis Testing

Important Question: Is there strong evidence for the alternative?

The burden of proof is placed on those who believe in the alternative claim.

The initially favored claim, the null hypothesis H_0, will not be rejected in favor of the alternative hypothesis, H_a or H_1, unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions:
Fail to Reject the null hypothesis if there is insufficient statistical evidence to do so.
Reject the null hypothesis in favor of the alternative if there is statistically *significant* cause to do so.

Logic of Hypothesis Testing

Notation and general process:

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

2. State the alternative to be tested: H_a :
   $\theta > \theta_0$ **OR** $\theta < \theta_0$ **OR** $\theta \neq \theta_0$

3. Draw a decision based on how improbable or probable the actual data looks if the null hypothesis is true. If the observed data is very unlikely, it might be because our hypothesis was wrong!

Why *assume* the null hypothesis?

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

Why *assume* the null hypothesis?

1. Burden of proof
2. We know how to calculate probabilities when we *know* θ!

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

1. $H_a : \quad \theta \neq \theta_0$

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

1. $H_a : \quad \theta \neq \theta_0$

2. $H_a : \quad \theta > \theta_0$

3. $H_a : \quad \theta < \theta_0$

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

Example: Suppose a company is considering putting a new type of coating on bearings that it produces.

The true average wear life with the current coating is known to be 1000 hours. With denoting the true average life for the new coating, the company would not want to make any (costly) changes unless evidence strongly suggested that exceeds 1000.

Logic of Hypothesis Testing

Example: An appropriate problem formulation would involve testing:

H_0:

H_a:

The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon."

Logic of Hypothesis Testing

Example: An appropriate problem formulation would involve testing:

H_0: New company lifetime average is 1000

H_a: New company lifetime exceeds 1000

The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon."

Test Statistics: The Evidence

Definition: *Test Statistic*
A test statistic is a a quantity derived based on sample data and calculated under the null hypothesis. It is used in a decision about whether to reject H_0.

We can think of a test statistic as our evidence. Next, we need to quantify whether we think our evidence is "rare" under the null hypothesis.

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data
3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data
3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?
4. Decide whether $\hat{p}$ is **too far** from $p = .5$, and make a decision accordingly.

Test Statistics: The Evidence

Which test statistic is "best"?

There are an infinite number of possible tests that could be devised, so we have to limit this in some way or total statistical madness will ensue!

In the previous example, we might use $\hat{p}$.

Rejection Regions

How would we know when the test statistic is "sufficiently rare" under the null hypothesis such that we might regard the null as false?

We could define a **rejection region**: a range of values of the test statistic that leads a researcher to reject the null hypothesis.

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 10 heads mean?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 10 tails mean?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 6 heads mean?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- Is there a difference between 60% heads if we flip 10 times and 60% heads if we flip 1000 times?

What is extreme: let's compute these!

Bring back α!

Definition: The **Significance level** α of a hypothesis test is the largest *probability* of a test statistic under the null hypothesis that would lead you to reject the null hypothesis.

Equivalently, it's the probability of the entire rejection region!

We thought of α last week during CIs as a term that widened or shrank as our tolerance for error grew, now it's very literally an *error rate*. Specifically, it's the probability of rejecting the null hypothesis when we were not supposed to do so.

Daily Recap

Today we learned

1. Comparing means.

Moving forward:

- **Lecture** This Friday

Next time in lecture:

- CIs for other models and relaxing assumptions.