CSCI 3022-002 Intro to Data Science
Two-Sample CIs

$\mu = 12 \quad \sigma = .2 \quad n = 100$

A hardware store receives a shipment of bolts that are supposed to be 12cm long. When manufactured, the mean is indeed 12cm, and the standard deviation is 0.2cm. For quality control, the hardware store chooses 100 bolts at random to measure. They will declare the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 or greater than 12.04 cm. Fine the probability that the shipment is found defective.

Opening sol:

we want $P(\bar{x} < 11.97 \text{ OR } \bar{x} > 12.04)$

$\bar{X} \sim N(\mu, \sigma^2/n)$

We want the probability $1 - \boxed{P(11.97 < \bar{X} < 12.04)}$ for $X \sim N(12, .2^2)$; $n = 100$.

$\bar{X} \overset{apx}{\sim} N\left(12, \frac{(.2)^2}{100}\right)$

this is normal!!

$\sqrt{\frac{(.2)^2}{100}} = .2/10$

$Z = \frac{\bar{X} - 12}{.2/10}$

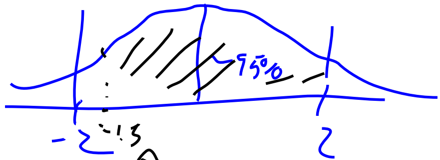
$= 1/5 \cdot \frac{1}{10} = 1/50$

$P(11.97 < \bar{X} < 12.04)$

$= P\left(\frac{11.97 - 12}{.2/10} < \frac{\bar{X} - 12}{.2/10} < \frac{12.04 - 12}{.2/10}\right)$

$Z \sim N(0,1)$

Opening sol:



We want the probability $P(11.97 < \bar{X} < 12.04)$ for $X \sim N(12, .2^2)$; $n = 100$.
We want to standardize: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, so

$$P\left(\frac{11.97-12}{0.2/10} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{12.04-12}{0.2/10}\right)$$

$$P\left(-1.5 < Z < 2.0\right)$$

cdf here

Opening sol:

We want the probability $P(11.97 < \bar{X} < 12.04)$ for $X \sim N(12, .2^2)$; $n = 100$.
We want to standardize: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, so

$$P\left(\frac{11.97-12}{0.2/10} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{12.04-12}{0.2/10}\right)$$

$$P\left(-1.5 < Z < 2.0\right)$$

$=$`stats.norm.cdf(2.0)-stats.norm.cdf(-1.5)`

Announcements and Reminders

- Exam posted tomorrow

95% CI on μ: $\boxed{[.1, .4]}$

CI Interpretation

1. The probability that the true mean is greater than 0 is at least 95 %. F

   μ is not random. It either >0 or not.

2. The probability that the true mean equals 0 is smaller than 5 %. F

   μ is not random

3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect.

   probability > .5

4. There is a 95 % probability that the true mean lies between 0.1 and 0.4. F

5. We can be 95 % "confident" that the true mean lies between 0.1 and 0.4.

   [.1, .4] are a random interval, .1 & .4 are volatile; vary from sample to sample.

6. If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between $\boxed{\text{0.1 and 0.4.}}$ F

   not this interval

→ if n large

Where we at?

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?

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?

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:

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

center

error from std. normal (CLT)

spread of $\bar{X}$

Where we at?

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 didn't know σ, we used s instead: if n large!

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

Comparing 2 Means

The natural estimator of $\mu_1 - \mu_2$ (for independent processes and samples X and Y) is ______.

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

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

Comparing 2 Means

The natural estimator of $\mu_1 - \mu_2$ (for independent processes and samples X and Y) is $\underline{\bar{X} - \bar{Y}}$.

μ_1 μ_2

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

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

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

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

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 $\bar{X}$ and $\bar{Y}$ have normal distributions.

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

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

Center: $\mu_1 - \mu_2$

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

Spread

Assumptions:
n, m large
σ_1^2, σ_2^2 known

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

$Z_{\alpha/2}$

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:

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

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

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

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:

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

replace σ_1^2, σ_2^2

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

$\boxed{\text{Center}} \pm z_{\alpha/2} \cdot \boxed{\text{spread of } \bar{X}-\bar{Y}}$

$\sim$ std normal

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:

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

CLT!

Comparing 2 Means: Large Sample

X $\qquad$ Y

$n = 50$ $\qquad$ $m = 40$ days

Views: $\bar{X} = 2$ m $\qquad$ $\bar{Y} = 2.25$ m

$\sigma_X = 1$ $\qquad$ $\sigma_Y = .5$

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).

difference!

Comparing 2 Means: Large Sample

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

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:

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

$\sigma/\sqrt{n}$

means $z_{.025}$

CI for μ_2:

$\sigma/\sqrt{n}$

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

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:

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

~~overlap!~~

not relevant.

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 $\mu_1 - \mu_2$:

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

still includes zero

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

Var of bin: $np(1-p)$.

Comparing 2 Means: Proportions

Recall: CI for p on one sample:

Sample

Pop

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

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

sample proportion

$(100-\alpha)\%$ CI: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

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

Variances add (if independent)

$$\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}$$

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

CI for $p_1 - p_2$:

So, a $(1-\alpha)\cdot 100\%$ confidence interval for $\hat{p}_1 - \hat{p}_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \underbrace{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}_{\text{Spread}}$$

This interval can safely be used as long as

$$n_1\hat{p}_1;\ n_1(1-\hat{p}_1);\ n_2\hat{p}_2;\ n_2(1-\hat{p}_2);$$

Same as before

are all at least 10.

10 successes AND 10 failures for each group

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.

$\hat{p}_1 = 76/154$

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?

$98/164 = \hat{p}_2$

m

$z_{.005}$

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

$\frac{p(1-p)}{n}$

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

≈ 0.0555

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

wider than $z_{.025} = 1.96$

CI for $p_1 - p_2$:

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

$<50\%$ $\approx 60\%$

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	σ unknown
Underlying Non-Normal Distribution	σ known	σ known
	σ unknown	σ unknown

C.L.T.: n large

OR

n small AND everything starts normal AND σ known

Center $\pm$ Zvalue $\cdot$ Spread

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

if $s \to \sigma$, this is Z

normalized:

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

np.var(ddof!)

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

1) X is normal
2) n is small
3) s is estimating σ

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

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 $\underline{\bar{X}}$)

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 $\underline{s}$).

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

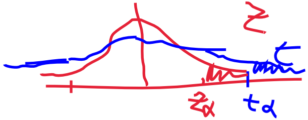
μ (with $\underline{\bar{X}}$)

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

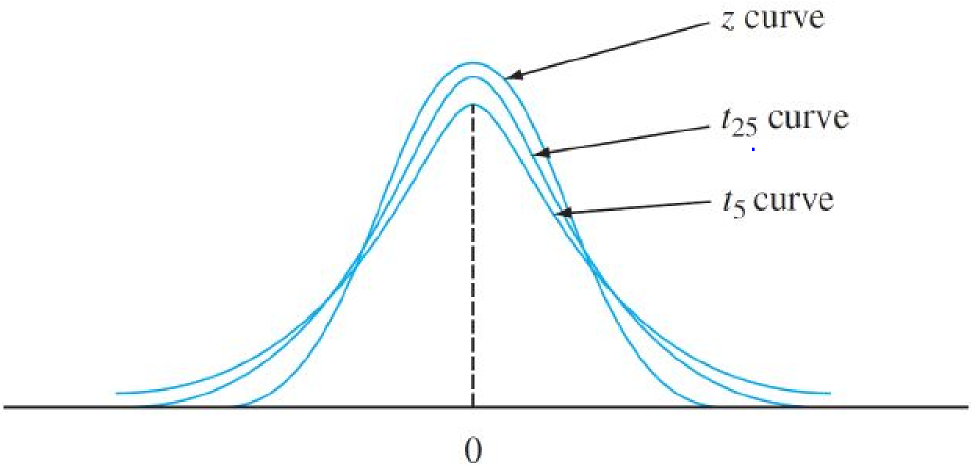
When our sample size is small, this is often a costly approximation, and as a result we have to *widen* our confidence intervals.

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 = ____)

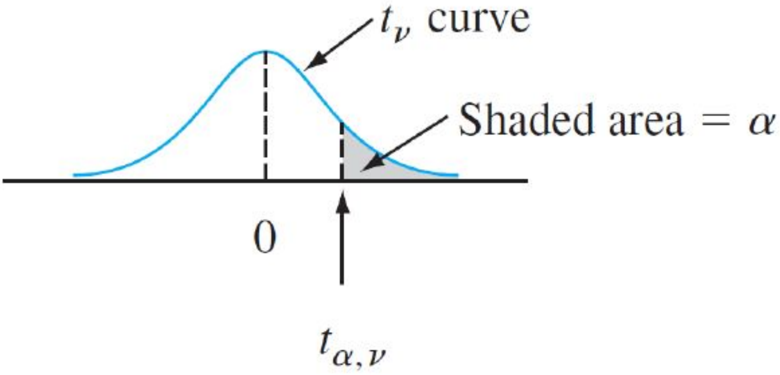
Properties of the t

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. Comparing multiple large or normal samples for equivalence of the mean! Also, how to handle single-samples that are underlying normal but $n < 30$ and unknown variance.

Moving forward:

- nb day Friday

Next time in lecture:

- Hypotheses!