CSCI 3022 Intro to Data Science
Inference and Confidence

Review: What does the Central Limit Theorem even say?

Announcements and Reminders

return: people_infected : #/2/ days_last

- We will have class every day next week. (Wellness day is Thursday 25 Mar). Friday 26 Mar will be a relaxing notebook day, so if you wanted to take a 4-day weekend, feel free to pick that day and watch lecture at your own leisure later.
- 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!

#1) 2 things are random
- # of trees in a region (count)
- random locations (same HW #4/nb04)

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with known mean value and standard deviation . Then:

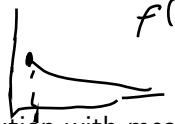
1) $E[\bar{X}] = \mu$

2) $Var[\bar{X}] = \frac{\sigma^2}{n}$

The standard deviation of the sample mean is:

$$s.e.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

This is also called the standard error of the mean.



Central Limit Theorem

Theorem: *Central Limit Theorem:* et $X_1, X_2, \ldots X_n$ be iid from a distribution with mean μ and variance σ^2. Then, for n *large enough*:

$$\bar{X} \overset{\sim}{\leftarrow} N\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger the value of n, the better the approximation! Typical rule of thumb: $n > 30$. **Idea:** The CLT provides insight into why many random variables have probability distributions that are approximately normal. As the sample size n increases, the sample mean $\bar{X}$ is close to normally distributed with expected value of the true population mean μ and with a *smaller* standard deviation $\sigma/\sqrt{n}$.

A result: Standardizing the sample mean by first subtracting the expected value and then dividing by the standard deviation yields a standard normal random variable.

and to data!

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

then $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$.

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

and to data!

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

1. Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
2. Is the sample proportion $\hat{p}$ a good approximation of the population proportion p?
3. Are two samples coming from populations with different means?

probability that it's close

group A

group B

and to data!

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

1. Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
2. Is the sample proportion $\hat{p}$ a good approximation of the population proportion p?
3. Are two samples coming from populations with different means?
4. **If Yes,** how sure or confident are we?

and to data!

What was the point of all this? We want the extract or infer properties of populations (like μ!) by analyzing samples. To do this, we ask:

1. Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
2. Is the sample proportion $\hat{p}$ a good approximation of the population proportion p?
3. Are two samples coming from populations with different means?
4. **If Yes,** how sure or confident are we?
5. How much data would we need to be sure or confident?

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

$$P(-1.96 < \quad < 1.96) = .95$$

(normals)

This is equivalent to:

know: $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is $N(0,1)$ if n large

random!

$$P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

$$.95 = P(-1.96 < Z < 1.96)$$

This is equivalent to:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

$$.95 = P(-1.96 < Z < 1.96)$$

This is equivalent to:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

algebra!

We want to know things about μ, however!

Conclude about μ given data $\bar{X}$.

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

$$.95 = P(-1.96 < Z < 1.96)$$

This is equivalent to:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

We want to know things about μ, however!
The 95% confidence interval for μ is the values of X that satisfy this inequality.

Solving for μ:

The interval:

Goal: $.95 = P(\#_a < \mu < \#_b)$

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$.95 = P\left(-1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

Solving for μ:

The interval:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$.95 = P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Solving for μ:

The interval:

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$.95 = P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu - \bar{X} > -1.96\frac{\sigma}{\sqrt{n}}\right)$$

Solving for μ:

The interval:

C.L.T. & standardizing

$$.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$.95 = P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu - \bar{X} > -1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Prob. interval for μ.

Confidence Interval for the Mean (SD known)

The interval

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Is called a 95% confidence interval for the mean.

This interval varies from sample to sample, as the sample mean varies. So, the interval itself is a random interval.

Which parts of the interval are random? $\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 \cdots + X_n}{n}$

Confidence Interval for the Mean (SD known)

The interval

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Is called a 95% confidence interval for the mean.

This interval varies from sample to sample, as the sample mean varies. So, the interval itself is a random interval.

Which parts of the interval are random? The two copies of $\bar{X}$

from $a < \mu < b$

Confidence Interval for the Mean (SD known)

$a = \bar{X} - 1.96\,\sigma/\sqrt{n}$ $\quad b = \bar{X} + 1.96\,\sigma/\sqrt{n}$

The CI is centered at $\bar{X}$ and extends $1.96\,\sigma/\sqrt{n}$ to each side in the x direction.

That width of $1.96 \cdot \sigma/\sqrt{n}$ is not random; only the location of the interval (its midpoint $\bar{X}$) is random.

Why isn't σ random?

$\sigma :=$ St. dev of population

$S :=$ St. dev of a sample

For now:

to find a Conf. interval for μ, we assume/need to know σ.

Confidence Interval for the Mean (SD known)

The CI is centered at $\bar{X}$ and extends $1.96 \cdot \sigma/\sqrt{n}$ to each side in the x direction.

That width of $1.96 \cdot \sigma/\sqrt{n}$ is not random; only the location of the interval (its midpoint $\bar{X}$) is random.

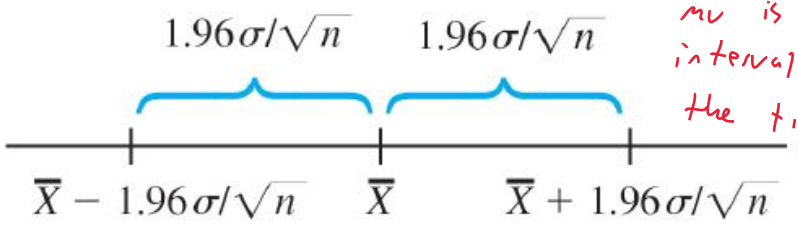
Confidence Interval for the Mean (SD known)

.95: P(a CI C b)

The CI is centered at $\bar{X}$ and extends _____ to each side in the x direction.

That width of $1.96 \cdot \sigma/\sqrt{n}$ is not random; only the location of the interval (its midpoint $\bar{X}$) is random.

Prob statement:
"μ is inside this interval 95% of the time"



Confidence Interval for the Mean (SD known)

As we showed, for a given sample, the CI can be expressed as

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

A couple of concise expressions for the interval are

where the left endpoint is the lower limit and the right endpoint is the upper limit.

Confidence Interval for the Mean (SD known)

As we showed, for a given sample, the CI can be expressed as

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

A couple of concise expressions for the interval are

$$[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$$

where the left endpoint is the lower limit and the right endpoint is the upper limit.

Confidence Interval for the Mean (SD known)

As we showed, for a given sample, the CI can be expressed as

$$.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

A couple of concise expressions for the interval are

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

CI for μ.

where the left endpoint is the lower limit and the right endpoint is the upper limit.

Interpreting CIs

(μ)

We are "95% confident" that the true parameter is in this interval.

What does that mean??

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability.

Interpreting CIs

We are "95% confident" that the true parameter is in this interval.

What does that mean??

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability.

In **repeated** sampling, 95% of the confidence intervals obtained from all samples will actually contain μ. The other 5% of the intervals will not.

... In practice, we have 1 interval & 1 $\bar{x}$.

Interpreting CIs

$P(2 < \mu < 8) = \begin{cases} 0, & \text{it is or it isn't.} \\ 1, & \end{cases}$

We are "95% confident" that the true parameter is in this interval.

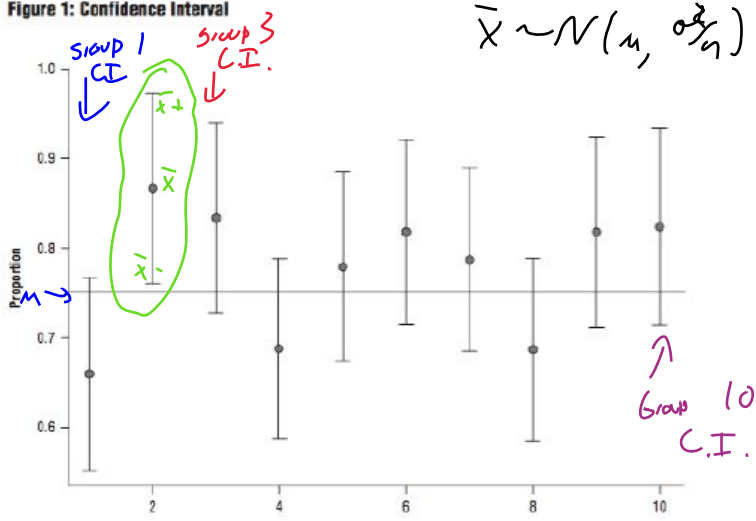
What does that mean??

$P(\bar{x} - m < \mu < \bar{x} + \dots)$

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability.

The confidence level is not a statement about any particular interval instead it pertains to what would happen if a very large number of like intervals were to be constructed using the same CI formula.

Interpreting CIs



Interpreting CIs

Some reading on the common misinterpretations of CIs:

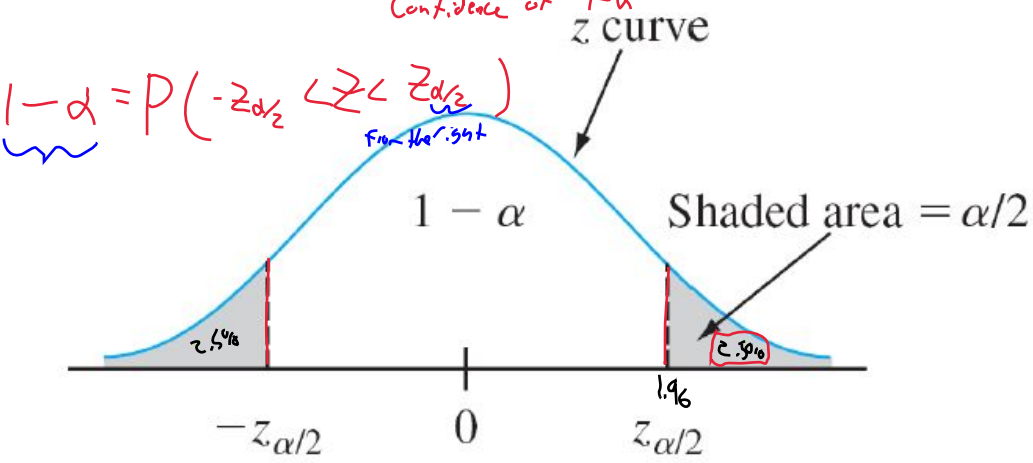
`http://www.ejwagenmakers.com/inpress/HoekstraEtAlPBR.pdf`

Other Levels of Confidence

.95 = P(−1.96 < Z < 1.96)

A confidence level of $1-\alpha$ can be achieved by using another $z_{\alpha/2}$ in place of $z_{0.025} = 1.96$:

error of α (5%)
Confidence of $1-\alpha$



Other Levels of Confidence

Common: 90, 95, 99,

$z_{\alpha/2}$: $\alpha/2 = .05$, $= .025$, $= .005$

A $100(1-\alpha)\%$ confidence interval for the mean when the value of α is known is given by:

Or, equivalently, by:

Other Levels of Confidence

A $100(1-\alpha)\%$ confidence interval for the mean when the value of α is known is given by:

$$1-\alpha = P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Or, equivalently, by:

Other Levels of Confidence

A $100(1-\alpha)\%$ confidence interval for the mean when the value of α is known is given by:

Or, equivalently, by:

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

not always 1.96.

Confidence Interval for the Mean (SD known)

$z_{.05}$ 

Example:

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

1. Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

$\sigma = 0.1$ $\alpha = .1$

$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5.426 \pm z_{.05} \frac{.1}{\sqrt{40}}$

stats.norm.ppf(.95) $1-.05$

2. What about the 99% confidence interval?

3. What are the advantages and disadvantages to a wider confidence interval?

Confidence Interval for the Mean (SD known)

Example:

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

1. Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

$$5.426 \pm z_{.05}\frac{0.1}{\sqrt{40}}$$

90%

2. What about the 99% confidence interval?

.005 or .5% error above

.5% error below $\leq .005$

3. What are the advantages and disadvantages to a wider confidence interval?

Confidence Interval for the Mean (SD known)

Example:

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

1. Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

$$5.426 \pm z_{.05} \frac{0.1}{\sqrt{40}}$$

2. What about the 99% confidence interval?

norm.

$$5.426 \pm \texttt{scipy.stats.ppf(.995)} \frac{0.1}{\sqrt{40}}$$

3. What are the advantages and disadvantages to a wider confidence interval?

Confidence Interval for the Mean (SD known)

Example:

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

1. Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

2. What about the 99% confidence interval?

3. What are the advantages and disadvantages to a wider confidence interval?
   Idea: This is a tradeoff of *accuracy* versus *precision*.

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?

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 $W = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. 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?

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

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

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

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 np and

variance of $np(1-p)$

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 np and

variance of $np(1-p)$

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.

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

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 $\bar{X}$ 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

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

Standardizing our estimator gives:

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

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:

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:

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

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.

Daily Recap

Today we learned

1. Making *inference* on the mean or means.

Moving forward:

- **Lecture** This Friday

Next time in lecture:

- CIs for other models and relaxing assumptions.