CSCI 3022-002 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: $\bar{X} = 3.6$ $\quad \cancel{\sigma}\, S = 2$ $\quad n = 1000$

$36 \pm 1.96 \cdot 2/1000$

$2/1000$

0.002

CI: $.95 = P(\underbrace{-1.96 < Z < 1.96}_{})$

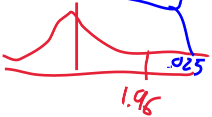
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$\bar{X} - 1.96 \cdot \sigma/\sqrt{n}, \quad \bar{X} + 1.96\, \sigma/\sqrt{n}$

μ is in that interval for a random $\bar{X}$ 95% of the time.

Opening sol:

recall: $1.96 = z_{.025}$



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

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

$$= [3.48, 3.72]$$

Opening Followup:

mean of the population,
not shape

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?

→ Prob on the Population

1. 95% of Americans spend 3.48 to 3.72 hours per day relaxing after work.

→ we only have 1 random sample!

[✓ 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.

μ is a number. any interval either contains it or not.

4. We are 95% su~~r~~e that Americans in this sample spend 3.48 to 3.72 hours per day relaxing after work.

(shhhh. - -

Average $\bar{X}$ estimates μ
average

Announcements and Reminders

- Homework 5 tonight!
- Exam posted this week: likely Wednesday or Thursday

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:

$$\underset{\substack{\uparrow\\ \text{estimates}\\ \mu}}{\bar{X}} \pm \underbrace{z_{\alpha/2}}_{\substack{\text{error tolerance}\\ 1.96 \text{ if } \alpha=.05}} \cdot \underbrace{\frac{\sigma}{\sqrt{n}}}_{\substack{\text{C.L.T. stuff:}\\ \bar{X} \text{ is } \underline{\text{less}} \text{ variable}\\ \text{than 1 individual}\\ X\text{-value.}}}$$

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 didn't know σ, we used s instead:

np.sqrt(np.var(ddof=1))

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

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	$\cancel{\sigma}$ known
	σ unknown	σ unknown
Underlying Non-Normal Distribution	σ known	σ known
	σ unknown	σ unknown

→ new dist!

Special cases!

→ simulate!

large n.

Method:

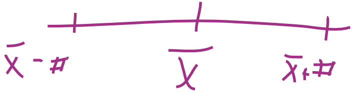
Z or approximately Z by Central Limit Theorem

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?

$$\bar{X} \pm \underbrace{1.96 \cdot \sigma/\sqrt{n}}_{\text{this is half the interval.}}$$



$\sigma = 25$

$$\text{width} = 2(1.96 \cdot \sigma/\sqrt{n})$$

GOAL: choose n so that

$$2(1.96 \cdot \sigma/\sqrt{n}) \leq 10$$

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 $W \overset{2\cdot}{=} z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. We want:

Sample Size Calculations

$> 25 \cdot (\text{stats.norm.ppf}(.975))^2$

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

$> 25 \cdot (z_{.025})^2$

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?

"smallest integer n so that"

$n > \left(\frac{2}{10}\right)^2 \cdot \sigma^2$

$\boxed{1.96^2}$

$n > \left(\frac{2 \cdot 25}{10}\right)^2 (1.96)^2$

$> 25 \cdot (1.96)^2$

$$2\left(\overset{1.96}{z_{\alpha/2}}\frac{\sigma}{\sqrt{n}}\right) < 10$$

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

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

(mult both sides $\cdot \frac{\sqrt{n}}{10}$)

$\left(2 \cdot 1.96 \cdot \frac{\sigma}{10}\right)^2 < \left(\sqrt{n}\right)^2$

$\sim \left(\frac{4 \cdot \sigma}{10}\right)^2 < n \Rightarrow \left(\frac{100}{10}\right)^2 < n$

$100 < n$

Special Cases: Populations

$\bar{x} \pm \underbrace{z}_{\text{error}} \cdot \underbrace{\sigma/\sqrt{n}}_{\text{std. dev of } \bar{x}}$

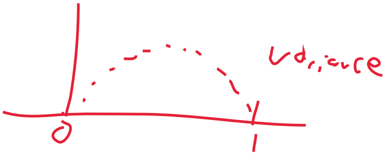
Let p denote the proportion of "successes" in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of $\underline{n}$ individuals is selected, and X is the number of successes in the sample.

fixed

Then, X can be modeled as a binomial rv with mean of np and

(on avg. we get prop of success. # of tries successes)

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 $\underline{np}$ 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 <u>Binomial</u> rv with mean of $\underline{np}$ and

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

E[failures]

E[successes]

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

Special Cases: Populations

recall: $\bar{X} \pm z_{\alpha/2} \cdot \underbrace{\sigma/\sqrt{n}}_{\text{std. dev. of } \bar{X}}$

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

pop proportion

X := 'success'

n := all trials.

$E[X] = np$

$Var[X] = np(1-p)$

$SD(X) = \sqrt{np(1-p)}$

Standardizing the estimator yields:

X is approx normal... so

$$\frac{X - E[X]}{SD[X]} \sim N(0,1).$$

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

$\hat{p}$: estimate for sample

p: true proportion for binom

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

This estimator is approximately normally distributed and:

$E[X] = np$

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

Standardizing the estimator yields:

$$\frac{\hat{p} - E[\hat{p}]}{s.D[\hat{p}]} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

and a resulting CI is:

estimate p:

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

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.

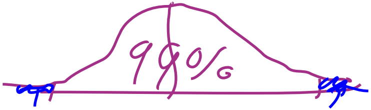
$n = 200$

$\hat{p} = .635$

$Var[\hat{p}] = \frac{p(1-p)}{n} = \frac{.635 \cdot (1-.635)}{200}$

$CI: \quad \hat{p} \pm Z_{.005} \cdot \sqrt{\frac{.635(1-.635)}{200}}$

Special Cases: Populations



combine to be 1%

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;

$$= 0.635 \pm 2.57\sqrt{\frac{0.635(1-0.635)}{200}}$$

$$= [0.548, 0.722]$$

$2.57 > 1.96$

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?

X vs. Y

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

μ_X vs. μ_Y

Comparing 2 Means

Basic Assumptions:

old assumption: $X_1, X_2, \dots X_n \overset{iid}{\sim}$ from X

pdf f_X

or some dist.

mean μ_X var σ_X^2

X & Y

are also independent

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

$\bar{X}$ estimates μ_1

$\bar{Y}$ '' μ_2

pop 1 pop2

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

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 $\bar{X} - \bar{Y}$: $E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}]$

$= \mu_X - \mu_Y$.

$\underbrace{Var[2Y]}_{2^2 \cdot Var[Y]}$

X & Y indep, so

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

$$Var[\bar{X} - \bar{Y}] = Var[\bar{X} \boxed{+} (-\bar{Y})]$$

$$= Var[\boxed{\bar{X}}] + \underbrace{Var[(-)\bar{Y}]}$$

$$= \frac{\sigma_X^2}{n} \qquad + (-1)^2 Var[\bar{Y}]$$

$$= \sigma_X^2/n + \sigma_Y^2/n$$

Comparing 2 Means

n: # of X

m: # of Y's

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

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

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

$$SD[\bar{X} - \bar{Y}] = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

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 $\mu_X-\mu_Y$ and standard deviation:

$$\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$$

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

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

Standardizing our estimator gives:

$$\frac{\boxed{(\bar{X}-\bar{Y}) - (\mu_1-\mu_2)}}{\boxed{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}} \sim \boxed{N(0,1)}$$

Center

std. error

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

CI for $\mu_1-\mu_2$: $(\bar{X}-\bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$

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:

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

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

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.

Daily Recap

Today we learned

1. Comparing multiple large or normal samples for equivalent of the mean!

Moving forward:

- nb day Friday

Next time in lecture:

- More: how we can use that it's all normal!