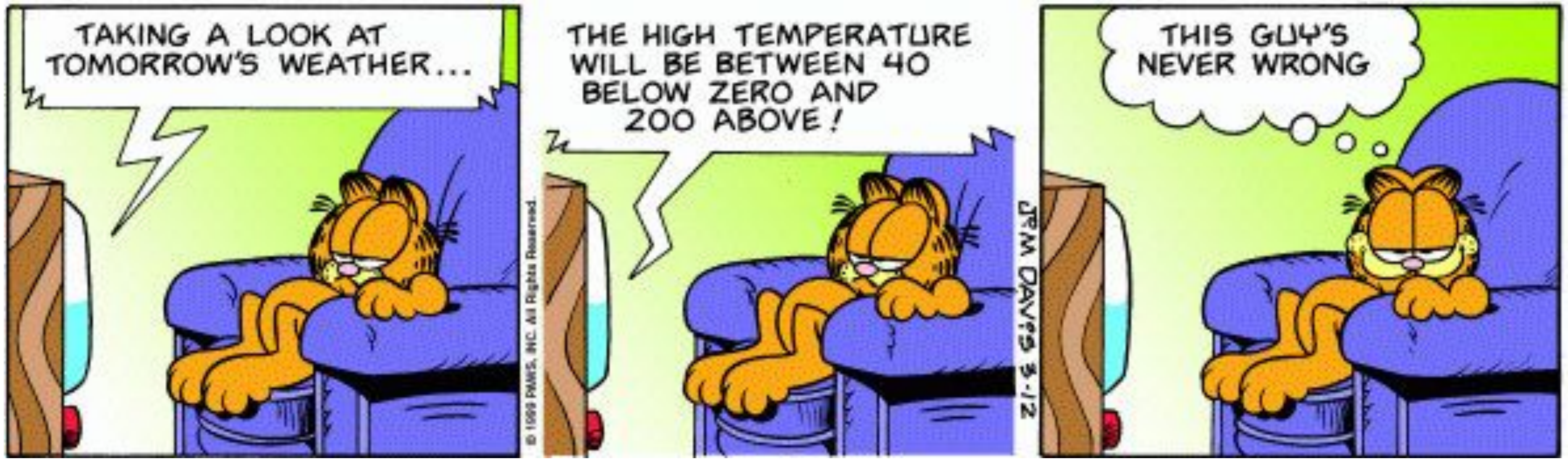
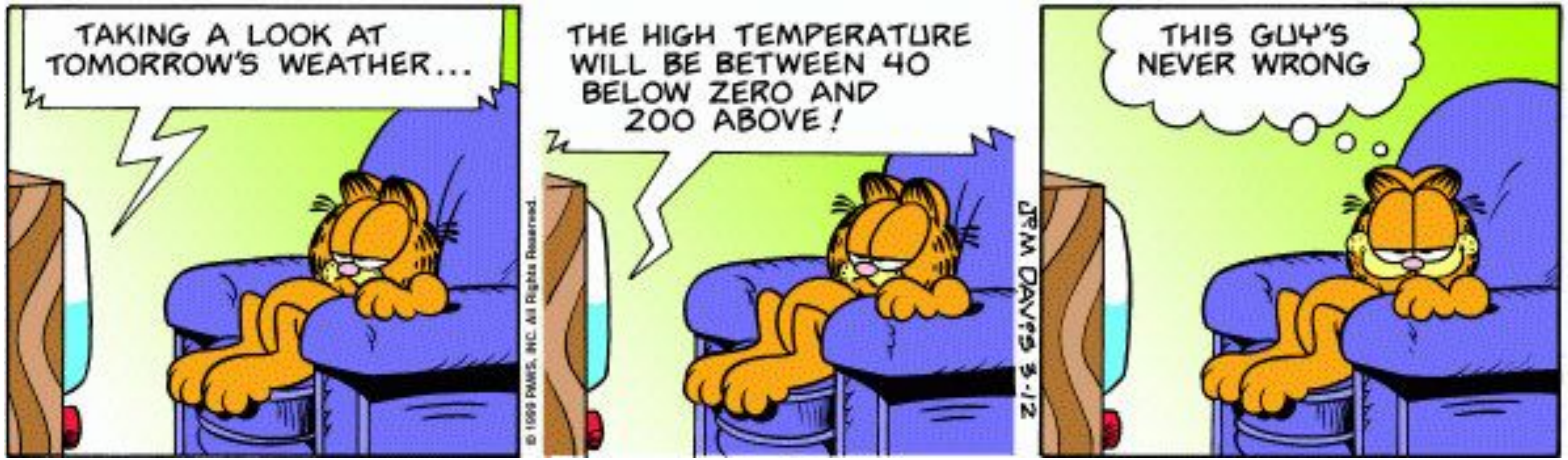
Lecture 15: Two-Sample Confidence Intervals



Announcements and reminders

- HW 3 posted! And due Monday 18 March (2 weeks)
- Quizlet 7 posted! And due Monday at 10 AM



Previously, on CSCI 3022…

Proposition: If X is a normally distributed random variable with mean μ and standard deviation σ, then Z follows a standard normal distribution if we define:

$$Z = \frac{X - \mu}{\sigma} \qquad \text{and} \qquad X = \sigma Z + \mu$$

The Central Limit Theorem: Let X_1, X_2, … , X_n be iid draws from some distribution. Then, as n becomes large…

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

prop : $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

A 100·(1-α)% confidence interval for the mean μ when the value of σ is known is given by

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad \text{or} \qquad \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

// $\hat{p} \pm z_\alpha$

Statistical Inference

Goal: Want to extract properties of an underlying population by analyzing sampled data

Last time:

- How to calc. a CI for pop. mean μ
- How to calc. a CI pop. proportion *p*

Today:

- How to get a CI for the **difference** between between the mean of two populations?
- … proportions …

Difference between population means

Question: How do two sub-populations compare? Are their means the same?

Classic motivating examples:

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

  → A/B testing

Difference between population means

Question: How do two sub-populations compare? Are their means the same?

Solution process: Collect samples from both sub-pops, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$

Basic Assumptions:

- $X_1, X_2, \dots, X_m$ is a random sample from distribution 1 with mean μ_1 and SD σ_1
- $Y_1, Y_2, \dots, Y_n$ is a random sample from distribution $\boxed{2}$ with mean $\boxed{\mu_2}$ and $\boxed{\text{SD } \sigma_2}$
- The X and Y samples are independent of one another

Difference between population means

$Var(x+y) = Var(x) + Var(y)$
$Var(cY) = c^2 Var(y)$

The natural estimator of μ$_1$ - μ$_2$ is the difference in sample means: $\boxed{\bar{x} - \bar{y}}$

… so is $\bar{x} - \bar{y}$ a good estimator for μ$_1$ - μ$_2$?

$(\mu_x - \mu_y)$

The expected value of $\bar{X} - \bar{Y}$ is given by: $E[\bar{X} - \bar{Y}] = \mu_1 - \mu_2$

The SD of $\bar{X} - \bar{Y}$ is given by:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y})$$
$$= \frac{\sigma_1^2}{m} + \frac{\sigma^2}{n}$$

$$\boxed{SD(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma^2}{n}}}$$

Difference between population means

The natural estimator of $\mu_1 - \mu_2$ is the difference in sample means: $\bar{x} - \bar{y}$

… so is $\bar{x} - \bar{y}$ a good estimator for $\mu_1 - \mu_2$?

The expected value of $\bar{X} - \bar{Y}$ is given by:

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2 \quad \leftarrow \text{unbiased estimator}$$

The SD of $\bar{X} - \bar{Y}$ is given by:

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

Normal populations with known SDs

If both populations are normal, then both $\bar{X}$ and $\bar{Y}$ are normally distributed

→ Indep. of the two samples implies that the samples' **means** are also indep.

→ The ***difference*** in sample means is normally distributed, for any sample sizes, with:

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

$100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$:

$$\bar{x} - \bar{y} \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma^2}{n}}$$

in the wild, if we don't know

Confidence intervals for the difference

stad: $\frac{\text{RV} - \text{central estimate}}{\text{stad error}}$

Standardized $\bar{X} - \bar{Y}$ gives a standard normal random variable!

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

if we don't know σ_1 & σ_2 & have "enough" samples ($n \geq 30$), then estimate using s_1 & s_2

So we can compute a 100·(1-α)% confidence interval for $\mu_1 - \mu_2$

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

Large-sample CIs for the difference

$\checkmark$ $n \geq 30$

If both m and n are large, then the CLT kicks in and our confidence interval for the difference of means is valid, even if the populations are not normally distributed

Furthermore, if m and n are large, and we don't know the SDs, we can replace them with the sample standard deviations:

Confidence intervals for the difference

$\left(1 - \frac{.05}{2}\right)$

$z_{\alpha/2}$ = stats.norm.ppf$\left(1 - \frac{.05}{2}\right)$

Example: S'pose 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).

$\bar{x}_1 = 2$ $\qquad$ $\bar{x}_2 = 2.25$

$s_1 = 1$ $\qquad$ $s_2 = 0.5$

$n_1 = 50$ $\qquad$ $n_2 = 40$

$z_{\alpha/2} = 1.96$

Ad 1: $\bar{x}_1 \pm z_{\alpha/2} \cdot \frac{s_1}{\sqrt{n_1}} = 2 \pm 1.96 \cdot \frac{1}{\sqrt{50}}$

Ad 2: $\bar{x}_2 \pm z_{\alpha/2} \cdot \frac{s_2}{\sqrt{n_2}} = 2.25 \pm 1.96 \cdot \frac{0.5}{\sqrt{40}}$

Confidence intervals for the difference

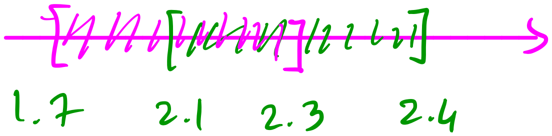
Example: S'pose you run two different email ad campaigns over many days and record the amount of traffic driven to your website on days that each ad was sent. Ad 1 was sent on 50 different days and generates an average of 2 million page views per day, with a SD of 1 million page views. Ad 2 was sent 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).

$$\bar{x} = 2, \sigma_1 = 1, \quad m = 50$$
$$\bar{y} = 2.25, \sigma_2 = 0.5, \quad n = 40$$
$$\alpha = 0.05 \to z_{0.025} = 1.96$$

$$\begin{aligned} \mathrm{CI}_1 &= \bar{x} \pm z_{\alpha/2}\frac{\sigma_1}{\sqrt{m}} \\ &= 2 \pm 1.96 \cdot \frac{1}{\sqrt{50}} \\ &= [1.723, \quad 2.277] \end{aligned}$$

$$\begin{aligned} \mathrm{CI}_2 &= \bar{y} \pm z_{\alpha/2}\frac{\sigma_2}{\sqrt{n}} \\ &= 2.25 \pm 1.96 \cdot \frac{0.5}{\sqrt{40}} \\ &= [2.095, \quad 2.405] \end{aligned}$$

Question: What does this tell us?



Confidence intervals for the difference

Example: S'pose 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 a 95% confidence interval for **the difference** in average page views per day (in units of millions of views).

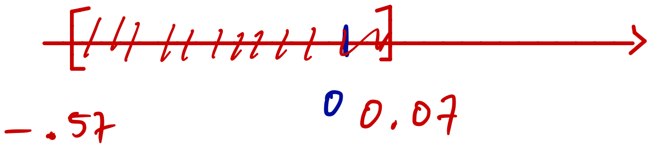
$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

CI $\mu_1 - \mu_2$

$$= 2 - 2.25 \pm 1.96\sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}}$$

$$= -0.25 \pm 1.96 \cdot 0.162$$

$$= [-0.568, 0.068]$$



Confidence intervals for the difference

Example: S'pose 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 a 95% confidence interval for **the difference** in average page views per day (in units of millions of views).

$$\bar{x} = 2, \sigma_1 = 1, \quad m = 50$$

$$\bar{y} = 2.25, \sigma_2 = 0.5, \quad n = 40$$

$$\alpha = 0.05 \rightarrow z_{0.025} = 1.96$$

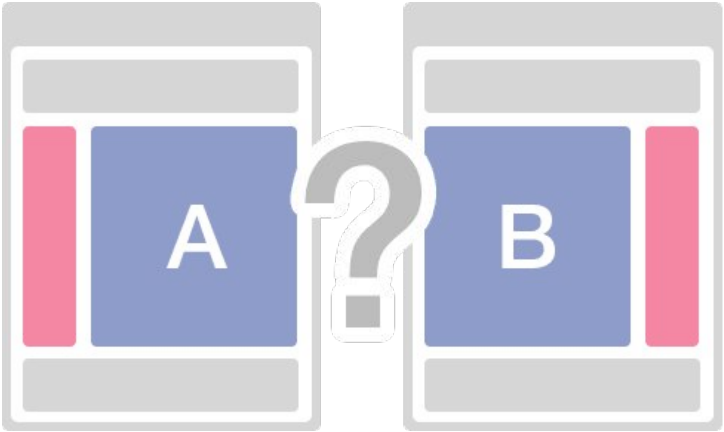
low CI, $z_{\alpha/2}$ will decrease

$$\begin{aligned}
\text{CI} &= \bar{x} - \bar{y} \pm z_{\alpha/2}\text{SD} \\
&= 2 - 2.25 \pm 1.96 \cdot \sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}} \\
&= -0.25 \pm 1.96 \cdot 0.162 \\
&= [-0.568, \quad 0.068]
\end{aligned}$$

Confidence intervals for the difference

Looking forward: What does our CI tell us about the effectiveness of the two advertisements?

from our two sample 95% CI, we dont have enough evidence to rule to 0 as possible



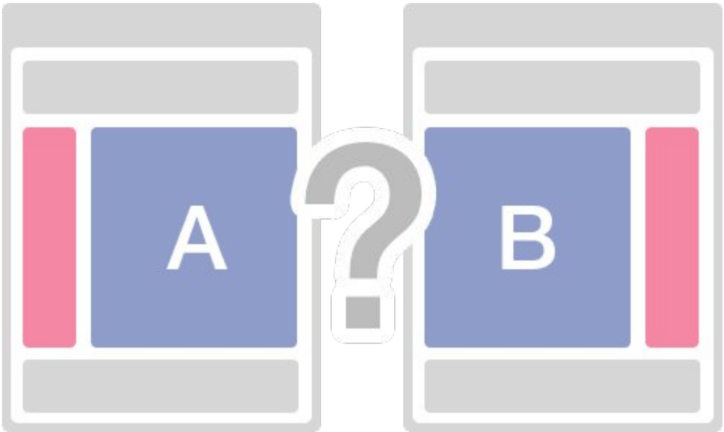
Confidence intervals for the difference

Looking forward: What does our CI tell us about the effectiveness of the two advertisements?

→ Suggests Ad 2 might be better

→ But CI covers 0, so there's a reasonable chance there is no significant difference

* want to do a two sample CI -> direct & easier



Difference between population proportions

CI : central est. ± $z_{\alpha/2}$ · STD error

$\sqrt{\text{sample mean var 1} + \text{sample mean var}_2}$

What if we want to compare population **proportions** instead of means?

S'pose a sample of size m is selected from pop 1, and sample of size n from pop 2

Let X denote the number of units with the characteristic of interest in pop 1 (# "successes"), and let Y denote … in pop 2

Reasonable estimators for the population proportions are: $\hat{p}_1 \approx p_1$ and $\hat{p}_2 \approx p_2$

Natural estimator for the difference in population proportions p_1 - p_2 is: $\hat{p}_1 - \hat{p}_2$

(central)

Difference between population proportions

Now, let $\hat{p}_1 = \frac{X}{m}$ and $\hat{p}_2 = \frac{Y}{n}$, where X ~ Bin(m, p_1) and Y ~ Bin(n, p_2)

Assuming that X and Y are independent, we can show that the **expected value** and **SD** are estimated by:

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

and

$$Var(\hat{p}_1 - \hat{p}_2) = \dots = \frac{\overbrace{p(1-p_1)}^{var(\hat{p}_1)}}{n_1} + \frac{\overbrace{\hat{p}_2(1-p_2)}^{var(\hat{p}_2)}}{n_2}$$

square root of this

Difference between population proportions

Now, let $\hat{p}_1 = \frac{X}{m}$ and $\hat{p}_2 = \frac{Y}{n}$, where X ~ Bin(m, p_1) and Y ~ Bin(n, p_2)

Assuming that X and Y are independent, we can show that the **expected value** and **SD** are estimated by:

$$E[\hat{p_1} - \hat{p_2}] = E[\hat{p_1}] - E[\hat{p_2}] = \frac{1}{m}E[X] - \frac{1}{n}E[Y] = \frac{1}{m}mp_1 - \frac{1}{n}np_2 = p_1 - p_2$$

and

$$Var(\hat{p}_1 - \hat{p}_2) =$$

Difference between population proportions

Difference between population proportions

$$
\begin{aligned}
Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(-\hat{p}_2) \\
&= Var(\hat{p}_1) + Var(\hat{p}_2) \\
&= Var\left(\frac{X}{m}\right) + Var\left(\frac{Y}{n}\right) \\
&= \frac{1}{m^2}Var(X) + \frac{1}{n^2}Var(Y) \\
&= \frac{1}{m^2}mp(1-p) + \frac{1}{n^2}np(1-p) \\
&= \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}
\end{aligned}
$$

$$
\left.
\begin{aligned}
SD(\hat{p}_1 - \hat{p}_2) &= \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}} \\
&\approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}
\end{aligned}
\right]
$$

CIs

we generally don't know p_1 & p_2, so est use $\hat{p}_1$ & $\hat{p}_2$

CIs for the difference of proportions

The $100 \cdot (1-\alpha)\%$ confidence interval for $p_1 - p_2$ is then given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

CIs for the difference of proportions

The 100·(1-α)% confidence interval for $p_1 - p_2$ is then given by

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

CIs for the difference of proportions

we look at proportion 76 of 154

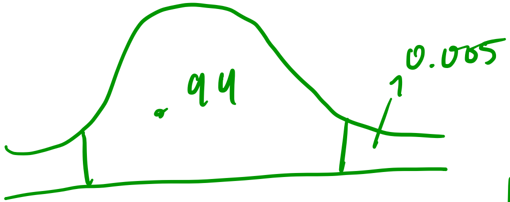
Example: A study was published in the New England Journal of Medicine in 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemo and radiation. Of the 154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 165 patients who received the hybrid treatment survived at least 15 years. What is the 99% CI for this difference in proportions?

$\hat{p}_1 = \frac{76}{154} \quad \hat{p}_2 = \frac{98}{165}$

$n_1 = 154 \quad n_2 = 165$

99% CI : $\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$= \frac{76}{154} - \frac{98}{165} \pm 2.576\sqrt{\quad}$



$z_{\alpha/2} = ppf(.099 + 0.005) = 2.576$

CIs for the difference of proportions

Example: A study was published in the New England Journal of Medicine in 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemo and radiation. Of the 154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 165 patients who received the hybrid treatment survived at least 15 years. What is the 99% CI for this difference in proportions?

$$\hat{p}_1 = \frac{76}{154} \approx 0.494$$

$$\hat{p}_2 = \frac{98}{165} \approx 0.598$$

$$\alpha = 0.01 \rightarrow z_{\alpha/2} = z_{0.005} = 2.576$$

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

$$= \sqrt{\frac{0.494(1-0.494)}{154} + \frac{0.598(1-0.598)}{165}}$$

$$\approx 0.0555$$

not enough evident to conclue that they are different CI

$$CI = \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} SD = 0.494 - 0.598 \pm 2.576 \cdot 0.0555$$

$$= -0.104 \pm 0.143 = [-0.247, \quad 0.039] \rightarrow \text{contain } 0$$

Writing autograders

Example: S'pose you're a TA for Intro to Data Science, and your prof-boss has tasked you with writing an autograder for a HW assignment that asks students to write a simulation to estimate the expected winnings in the game of Chuck-a-Luck.

Writing autograders

Example: S'pose you're a TA for Intro to Data Science, and your prof-boss has tasked you with writing an autograder for a HW assignment that asks students to write a simulation to estimate the expected winnings in the game of Chuck-a-Luck.

Answer:

- We know the true mean of Chuck-a-Luck -- we calculated it!
- So run student code many times
- … and compute a CI for student code's mean
- … is the true mean in the CI?

Writing autograders

Example: Now s'pose that your prof-boss asks you to write an autograder for a simulation of Miniopoly. Specifically, she asks you to check solutions to the function that estimates the probability that a player goes Bankrupt within the first 20 turns of the game.

How is this problem different from the Chuck-a-Luck problem? What should you do?

* p is unknown
* have a solution code -> generation solution

Writing autograders

Example: Now s'pose that your prof-boss asks you to write an autograder for a simulation of Miniopoly. Specifically, she asks you to check solutions to the function that estimates the probability that a player goes Bankrupt within the first 20 turns of the game.

How is this problem different from the Chuck-a-Luck problem? What should you do?

Answer:

- This is about **proportions** instead of means.
- We don't have a true proportion, but we do have a correct "solution" simulation
- Compute $\hat{p}_1$ from student code via m simulations, and $\hat{p}_2$ from correct code via n sims
- Compute CI for diff in proportions -- does it contain 0?
- If not, run the codes again (a bunch of times)

What just happened?

- ***Two-sample*** **confidence intervals** happened!
    - Sometimes we want to compare two groups
    - See if there is some significant difference between them
    - Two overlapping one-sample CIs is tough/impossible to interpret
      One two-sample CI works better!