Section 10.1A "Comparing Two Proportions(CI's)"

- **What is the Difference Between Pair Differences and Difference between 2 Samples**

1) Paired Differences (i.e. Paired t-Test) LAST CHAPTER

- THE SAME SUBJECTS RECEIVE BOTH TREATMENTS AND THE TREATMENTS ARE RANDOMLY ASSIGNED

2) Difference between 2 Samples (i.e. Difference between 2 Proportions)

THERE ARE 2 APPLICATIONS

① COMPARE 2 DIFFERENT POPULATION TO COMPARE PROPORTIONS FOR THAT CHARACTERISTIC (p_1 and p_2)

② COMPARE THE EFFECTIVENESS OF 2 TREATMENTS IN A COMPLETELY RANDOMIZED EXPERIMENT. WE COMPARE PROPORTIONS FOR EACH TREATMENT (p_1 and p_2)

Population or treatment	Parameter	Statistic	Sample size
1	p_1	$\hat{p}_1$	n_1
2	p_2	$\hat{p}_2$	n_2

Population Parameter

Sample statistic

Remember PP and SS

Sampling Distribution for Difference of 2 Proportions

HW EXAMPLE #1: To explore the sampling distribution of the difference between two proportions, let's start with two populations having a known proportion of successes.

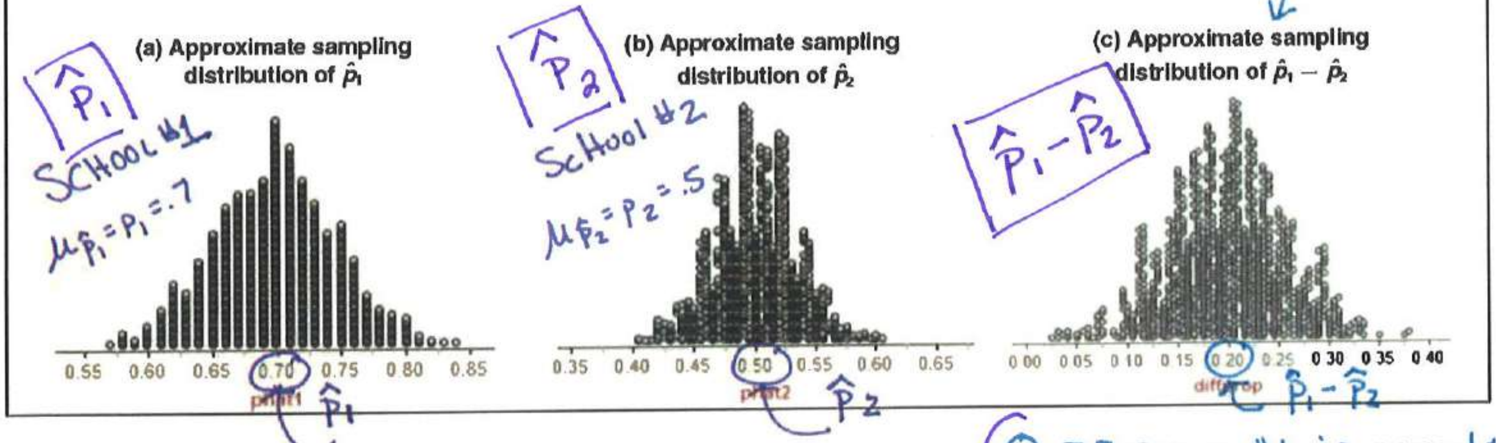
- ✓ At School 1, 70% of students did their homework last night
- ✓ At School 2, 50% of students did their homework last night.

$p_1 = .7$ $n_1 = 100$
$p_2 = .5$ $n_2 = 200$

Suppose the counselor at School 1 takes an SRS of 100 students and records the sample proportion that did their homework.

School 2's counselor takes an SRS of 200 students and records the sample proportion that did their homework.

Here are graphs of sampling distribution(each repeated 1000 times) for $\widehat{p1}$ and $\widehat{p2}$. What do you notice about the sampling distribution for for $\widehat{p1} - \widehat{p2}$?


(a) Approximate sampling distribution of $\hat{p}_1$

(b) Approximate sampling distribution of $\hat{p}_2$

(c) Approximate sampling distribution of $\hat{p}_1 - \hat{p}_2$

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

① IF SCHOOL #1 is normal + School #2 is normal then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is Normal

② to find the mean + STD DEV REVIEW Random Variables

③ Green Sheet

$\mu_{\hat{p}} = p$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

NEXT PAGE SHOWS THESE CALCULATIONS →

Write the General Formulas to Describe Sampling Distribution of a Difference Between Two Proportions:

Sampling distribution of $\hat{p}_1 - \hat{p}_2$:

- **Define Parameters**

  P_1 = population 1 or treatment 1

  P_2 = population 2 or treatment 2

- **Shape:** The distribution is approximately normal if all 4 are at least 10 (Show all 4 calculations)

  $n_1 p_1 \geq 10$ $\quad$ $n_2 p_2 \geq 10$

  $n_1 q_1 \geq 10$ $\quad$ $n_2 q_2 \geq 10$

- **Center:**

$$\mu_{\hat{p}_1 - \hat{p}_2} = P_1 - P_2$$

- **Spread:**

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

DO NOT memorize! IT IS ON GREEN SHEET

REQUIRED CONDITION!!

① To Calculate any σ, independence must be checked.

② You must check the 10% condition for both population 1 and population 2.

*When sampling without replacement.

Example #1 (continued): Who Does More Homework?

Back to our homework problem. We have 2 large high schools, each with more than 2,000 students, in a certain town. The counselors from School 1 and School 2 meet to discuss the results of their homework surveys.

a) Describe the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

We have 2 populations WITH $N = 2,000$ plus students

School 1: $p_1 = .7$ $n_1 = 100$ students

School 2: $p_2 = .5$ $n_2 = 200$ students

SHAPE:

MUST SHOW GREEN HIGHLIGHTS FOR FULL CREDIT

EITHER OR BOTH

$n_1 p_1 = 100(.7) = 70 \geq 10$ ✓

$n_1 q_1 = 100(.3) = 30 \geq 10$ ✓

$n_2 p_2 = 200(.5) = 100 \geq 10$ ✓

$n_2 q_2 = 200(.5) = 100 \geq 10$ ✓

- MUST SHOW ALL 4 CALCULATIONS
- MUST SAY: THE SAMPLING DISTRIBUTION $\hat{p}_1 - \hat{p}_2$ IS approximately normal because all successes + failures were at least 10.

CENTER: $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = .70 - .50 = \underline{\underline{.20}}$

SPREAD: Both schools meet the independence condition

School #1 = 100(10) = 1,000 students ✓

School #2 = 200(10) = 2,000 students ✓

Both schools had at least 2,000 students

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$ ← ON GREEN SHEET

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{(.7)(.3)}{100} + \frac{(.5)(.5)}{200}} = \underline{\underline{.058}}$$

Review Content

$\mu_{\hat{p}_1 - \hat{p}_2}$ IN Repeated samples, the mean difference in HW completion between the 2 schools would be about 20% in the long run

$\sigma_{\hat{p}_1 - \hat{p}_2}$ on average, there would be about a 5.8% difference from the mean (20%) for HW completion between the 2 schools.

Example #1 (continued): Who Does More Homework?

After the meeting, the 2 school counselor report to their principals that $\hat{p}_1 - \hat{p}_2 = 0.10$

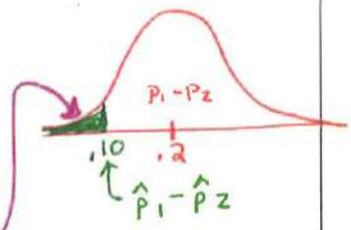
① ALWAYS SKETCH GRAPH

b) Find the probability of getting a difference in sample proportions $\hat{p}_1 - \hat{p}_2$ of 0.10 or less from the two surveys.

② FIND PROBABILITY:

$$P(\hat{p}_1 - \hat{p}_2 \le .10) = P\left(Z \le \frac{.10 - .20}{.058}\right) = P(Z \le -1.72)$$

$$= \boxed{.0427}$$



③ Calculcte Z score

normalcdf (-E99, -1.72, 0, 1)

c) Does the result in part (b) give us reason to doubt the counselors' reported value?

THE COUNSELOR'S RESULTS ARE SUSPICIOUS.

THERE IS ONLY ABOUT A 4% CHANCE OF GETTING A DIFFERENCE IN SAMPLE PROPORTIONS FOR HW COMPLETION AS SMALL OR SMALLER THAN THE 10% REPORTED BY THE COUNSELORS.

■ Example 2 "You Try This One ☺": Who Does More Homework #2?

Suppose that the counselors at School 1, Mike and Lynn, independently take a random sample 100 of students from there school and record the proportion of students that did their homework last night. When they were finished they find their difference in proportions was .08. They are surprised to get such a big difference, considering they were sampling from the same population.

$\hat{P}_m$ = MIKE %

$\hat{P}_L$ = LYNN %

a) Describethe shape,center,and spreadof the samplingdistributonof $\hat{p}_M - \hat{p}_L$.

BOTH COUNSELORS ARE FROM THE SAME SCHOOL.

SCHOOL 1 (FROM EXAMPLE #1) : $P_1 = .3 \quad 1 - P_1 = .7 \quad n_1 = 100$

$\boxed{\text{CENTER}}$ $\mu_{\hat{P}_m - \hat{P}_L} = P_1 - P_1 = .7 - .7 = \boxed{0}$

$\boxed{\text{SPREAD}}$ $p_1 = .3 \quad q_1 = .7 \quad n_1 = 100$

$$\sigma_{\hat{P}_m - \hat{P}_L} = \sqrt{\frac{.7(.3)}{100} + \frac{.7(.3)}{100}} = \boxed{.065}$$

$\boxed{\text{NORMAL}}$

$n_m p_m = 100(.7) = 70 \geq 10$ ✓

$n_m q_m = 100(.3) = 30 \geq 10$ ✓

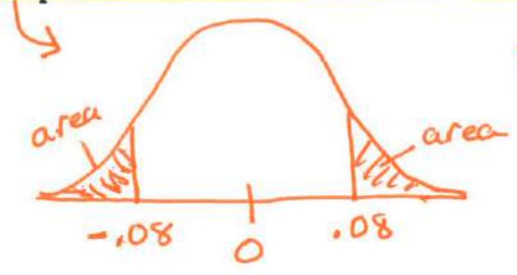
$n_L p_L = 100(.7) = 70 \geq 10$ ✓

$n_L q_L = 100(.3) = 30 \geq 10$ ✓

THERE FORE THE DISTRIBUTION IS APPROXIMATLY NORMAL

b) Find the probability of getting a two proportions that are at least .08 apart

Keep in mind their differences could be lower or higher.



$P(\hat{P}_m - \hat{P}_L \leq -.08)$ OR $P(\hat{P}_m - \hat{P}_L \geq .08) =$

$.1092 + .1092 = \boxed{.2184}$

MAKE SURE YOU KNOW HOW TO FIND THE ZSCORE! $Z = \dfrac{0 - .08}{.065} = -1.23$ *

* NOTE: ARE AREA MAY BE SLIGHTLY DIFFERENT DUE TO ROUNDING

CALC: normalcdf (-E99, -.08, 0, .065) = .1092

OR normalcdf (-E99, -1.23, 0, 1) = .1093

c) Should the counselors have been so surprised to get a difference this big? Explain.

SINCE THE PROBABILITY IS NOT VERY SMALL (approx. 22%), WE SHOULD NOT BE SURPRISED TO GET A DIFFERENCE BETWEEN THE 2 COUNSELORS HW PROPORTION OF 0.08 OR LARGER BY CHANCE, EVEN WHEN THE SAMPLING IS FROM THE SAME POPULATION.

Confidence Intervals for $p_1 - p_2$ – Understanding the Formulas

1) When the Independent condition is met, the standard deviation of the statistic $\hat{p}1 - \hat{p}2$ is :

SD WHEN WE KNOW P_1 and P_2

$$\underbrace{\sigma_{\hat{p}_1 - \hat{p}_2}}_{} = \sqrt{\frac{p1(1-p1)}{n1} + \frac{p2(1-p2)}{n2}}$$

THIS IS THE POPULATION PARAMETER

statistic ± (critical value) · (standard deviation of statistic) ← ON GREEN SHEET

2) Because IF we don't know the values of the parameters $p1$ and $p2$, we replace them with $\hat{P}_1$ + $\hat{P}_2$ in the standard deviation formula with the sample proportions. The result is the **standard error** of the statistic $(\underbrace{\hat{p}1 - \hat{p}2}_{})$ $\sqrt{\frac{\hat{p}1(1-\hat{p}1)}{n1} + \frac{\hat{p}2(1-\hat{p}2)}{n2}}$

$SE(\hat{p}_1 - \hat{p}_2)$

This is the statistic

FINDING THE CONFIDENCE INTERVAL:

3) If the Normal condition is met, we find the critical value z* for the given confidence level from the standard Normal curve. Our confidence interval for $p_1 - p_2$ is:

statistic ±(critical value) ×standard deviation of statistic)

FORMULA

Name: 2 Sample CI for p → $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

4) Summarize the conditions to check for a confidence interval for $p_1 - p_2$:

CONDITIONS

① Random: EITHER... EACH SAMPLE IS RANDOMLY SELECTED

OR TWO GROUPS ARE IN A RANDOMIZED EXPERIMENT

② NORMAL: The distribution is approximately normal if the counts of "successes" and "failures" in each Sample OR TREATMENT GROUP are at least 10.

$n_1\hat{p}_1 \geq 10$ $\quad n_2\hat{p}_2 \geq 10$

$n_1\hat{q}_1 \geq 10$ $\quad n_2\hat{q}_2 \geq 10$

③ Independent:

- When sampling without replacement, check the 10% condition for BOTH samples.
- Both samples or groups are independent as well as the individual observations within are independent.

Example #3: *Teens & Adults on Social Networks -- Construct a Confidence Interval*

As part of the Pew Internet and American Life Project, researchers conducted two surveys in late 2009. The first survey asked a random sample of 800 U.S. teens about their use of social media and the Internet. A second survey posed similar questions to a random sample of 2253 U.S. adults. In these two studies, 73% of teens and 47% of adults said that they use social-networking sites. Use these results to construct and interpret a 95% confidence interval for the difference between the proportion of all U.S. teens and adults who use social-networking sites.

Teens
$\hat{p} = .73$
$n = 800$

Adults
$\hat{p} = .47$
$n = 2253$

1) Define Parameters:

p_1 = true proportion of U.S. Teens who use social network sites
p_2 = true proportion of U.S Adults who use social network Sites.

2) Check Conditions:

Random: 2 different random samples
- random sample of teens ($n_1 = 800$)
- random sample of adults ($n_2 = 2,253$)

Independent
- We clearly have 2 independent samples - teens and adults
- It is reasonable, there are
$800(10) = 8,000$ U.S. teens and
$2,253(10) = 22,530$ U.S. adults.

Normal We check the counts of "succeses" and "failures" for Both sample and found the normal conditions were met

$\hat{p}_1 = .73$
$n_1 = 800$

$n_1\hat{p}_1 = (800)(.73)$
$= 584 \geq 10$ ✓

$n_1\hat{q}_1 = (800)(.37) =$
$216 \geq 10$ ✓

$n_2\hat{p}_2 = (2253)(.47)$
$= 1059 \geq 10$ ✓

$n_2\hat{q}_2 = (2253)(.53)$
$1194 \geq 10$ ✓

$\hat{p}_2 = .47$
$n_1 = 225$

Example 3 (continued)

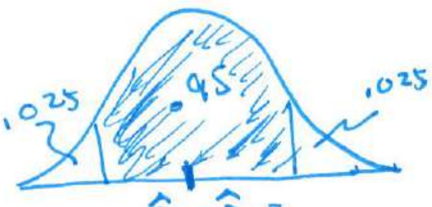
3) Calculations:

$\hat{P}_1 = .73$ (TEENS) $N_1 = 800$

$\hat{P}_2 = .47$ (ADULTS) $n_2 = 2253$

STATE "2 Sample Z interval for a difference between 2 proportions"

GRAPH



$\hat{P}_1 - \hat{P}_2 =$
$.73 - .44 =$
$\underline{.26}$

$\boxed{Z^* = \pm 1.96}$

GREEN SHEET

$$(\hat{P}_1 - \hat{P}_2) \pm Z^* \sqrt{\frac{\hat{P}_1(1-\hat{P}_2)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

$$(.73 - .47) \pm 1.96 \sqrt{\frac{(.73)(.27)}{800} + \frac{(.47)(.53)}{2253}}$$

$$\overset{\hat{P}_1 - \hat{P}_2}{.26} \pm 1.96\ (\underset{SE(\hat{P}_1 - \hat{P}_2)}{.01889})$$

$$.26 \pm \underset{\text{Margin of Error (ME)}}{.037}$$

$$\boxed{\text{INTERVAL } (.223, .297)}$$

CALCULATOR

(STAT) (TESTS)

B) 2-PROPZINT

$X_1 = 584$ ← $(.73)(800)$

$n_1 = 800$

$X_2 = 1059$ ← $(.47)(2253)$

$n_2 = 2253$

C-LEVEL: .95

⇓

(.22399, .29802)

4) Conclusion:

WE ARE 95% CONFIDENT THAT THE INTERVAL .223 TO .297 CAPTURES THE TRUE DIFFERENCE IN THE PROPORTION OF US TEENS AND US ADULTS WHO USE SOCIAL NETWORKING SITES.

** THIS INTERVAL SUGGEST THAT MORE TEENS THAN ADULTS IN THE U.S. ENGAGE IN SOCIAL NETWORKING BY 22.3 TO 29.7 PERCENTAGE POINTS.

Review Finding Sample Size:

Example #4: Suppose that researchers want to estimate the difference in proportions of people who are against the death penalty in Texas & in California. If the two sample sizes are the same, what size sample is needed to be within 2% of the true difference at 90% confidence?

KEY INFO

- Need "n" for TX and CA
- ME = 2% = .02
- CL = .90

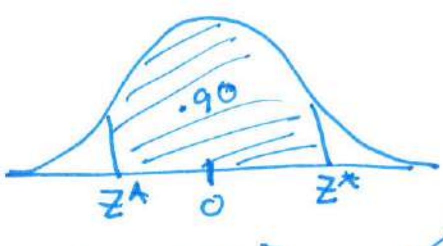
* PROPORTION PARAMETER NOT GIVEN

What proportion to use?

* Since "p" was NOT Given we use a conservative Proportion → $\boxed{p = 1/2}$

GREEN SHEET

$$\underbrace{\hat{p}_1 - \hat{p}_2}_{\text{STATISTIC}} \pm \underbrace{z^* \underbrace{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}_{SE(\hat{p}_1 - \hat{p}_2)}}_{ME}$$



$z^* = 1.645$
$\hat{p} = 1/2$
ME = .02
$n = n_1 = n_2$

$$.02 = 1. \sqrt{\frac{(.5)(.5)}{n} + \frac{(.5)(.5)}{n}}$$

$$.02 = 1.645 \sqrt{\frac{2(.5)(.5)}{n}}$$

$$.02 = 1.645 \cdot \frac{\sqrt{.5}}{\sqrt{n}}$$

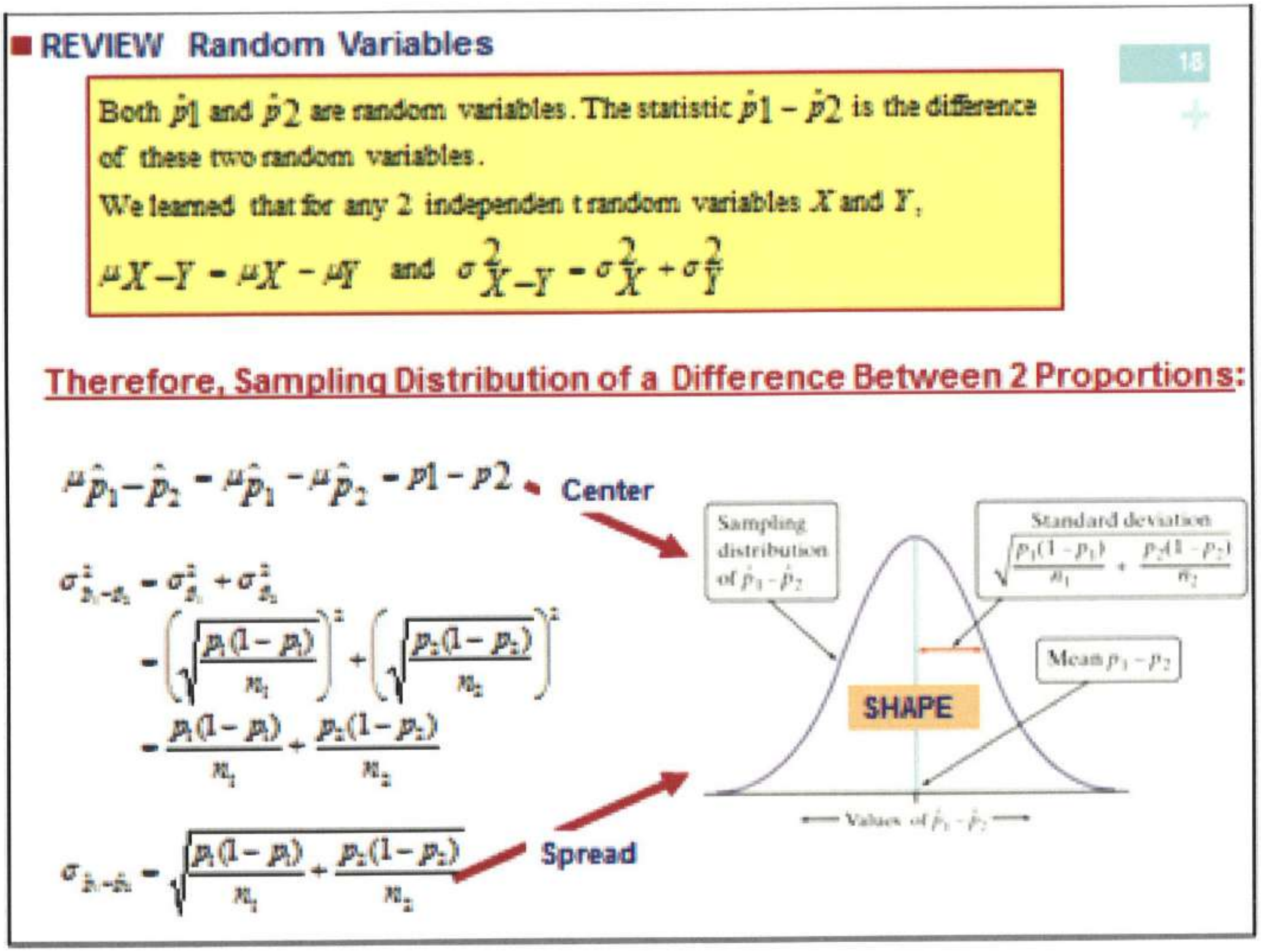
$$\left(\sqrt{n}\right)^2 = \left(\frac{(1.645)(\sqrt{.5})}{.02}\right)^2$$

$$\boxed{n \simeq 3{,}382.5}$$

IMPORTANT! You must always round up to ENSURE ME is met

CONCLUSION: We must have sample sizes of 3,383 from BOTH Texas and California.

APPENDIX:



REVIEW Random Variables

Both $\hat{p}1$ and $\hat{p}2$ are random variables. The statistic $\hat{p}1 - \hat{p}2$ is the difference of these two random variables.

We learned that for any 2 independen t random variables X and Y,

$\mu_{X-Y} = \mu_X - \mu_Y$ and $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$

Therefore, Sampling Distribution of a Difference Between 2 Proportions:

$$\mu_{\hat{p}_1-\hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p1 - p2$$

Center

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \left(\sqrt{\frac{p_1(1-p_1)}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2(1-p_2)}{n_2}}\right)^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Spread

Sampling distribution of $\hat{p}_1 - \hat{p}_2$
Standard deviation $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Mean $p_1 - p_2$
SHAPE
Values of $\hat{p}_1 - \hat{p}_2$

Free Response Example

FRAPPY 2009B

Question 3