**Categorical Data** (starting with 2 categories)

$p =$ True proportion of "*successes*" in the population (unknown)

$\hat{p}$ = Observed proportion of "*successes*" in the sample   $\hat{p} = x / n$  ( $x$= # of successes)

---

**NBC "Snap Poll":**   Have the debates caused you to change your vote?

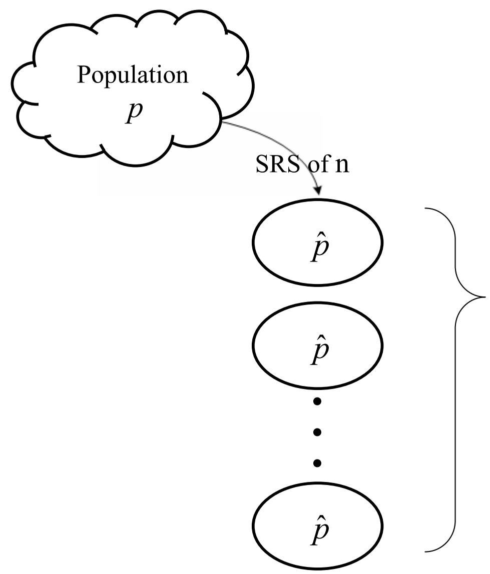
*n=400* questioned

*x=60* answered yes

The pollster's goal is to infer back to the population of voters.

e.g., Is the true proportion of all voters changing their mind less than 20 percent?

---

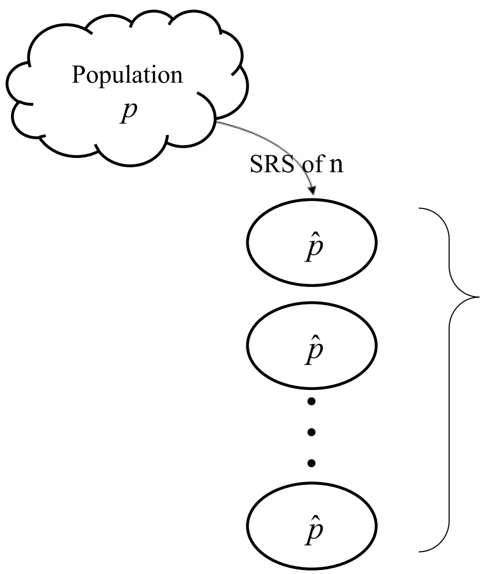



Sampling Distribution for $\hat{p}$:

approx. normal with

mean $= p$

$$\text{SD} = \sqrt{\frac{p(1-p)}{n}}$$




Sampling Distribution for $\hat{p}$:
approx. normal with

| | Mean | SD |
|---|---|---|
| under Ho | $p_o$ | $\sqrt{\frac{p_o(1-p_o)}{n}}$ |
| under Ha | $p_a$ | $\sqrt{\frac{p_a(1-p_a)}{n}}$ |
| CI | $\rightarrow$ | $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ |
| Study Design $\rightarrow$ | | $\sqrt{\frac{p^*(1-p^*)}{n}}$ |

Find the probability of a type II error if the true value for $p$ is really *0.13* (i.e., $\beta(0.13)$ )

1) Write the rejection region in terms of z-scores
2) Write the rejection region in terms of $\hat{p}$
3) $\beta(0.13) = P(Accept\ H_0 \mid H_a\ is\ true)$ → $p_a = 0.13$ is used to standardize $\hat{p}$ assuming Ha is true

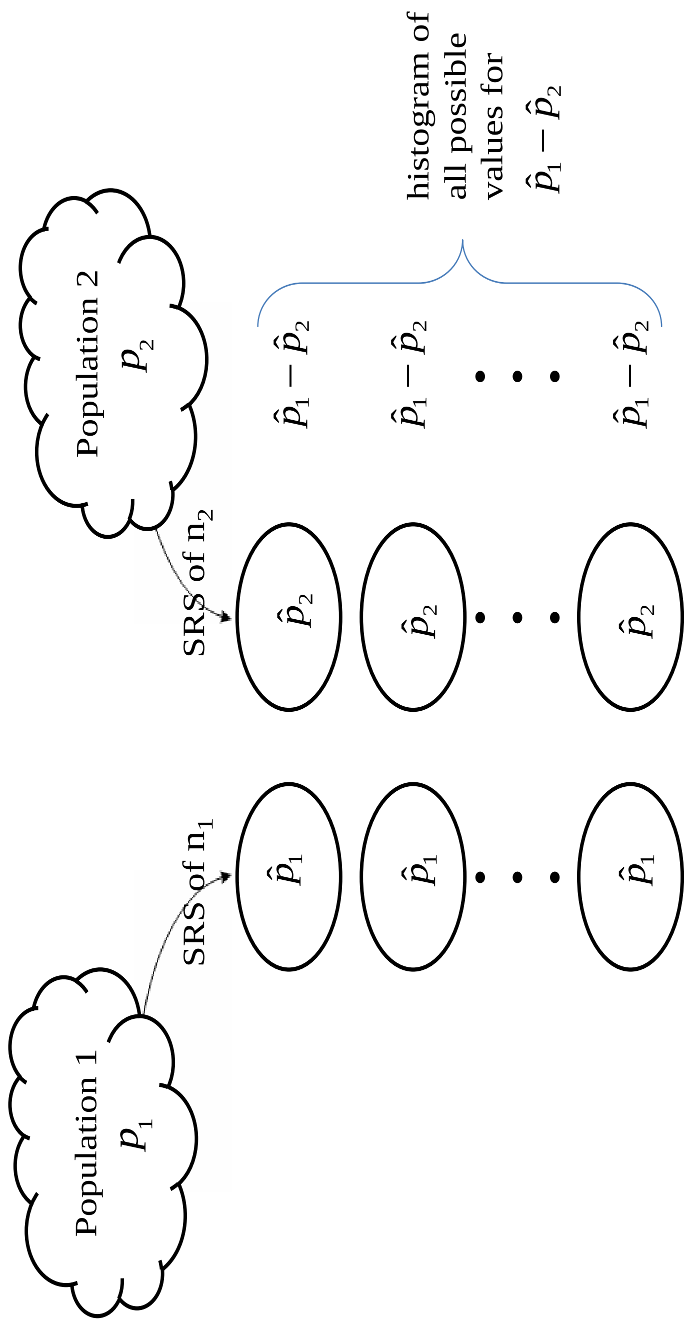
---

**Conditions for a Valid CI and Hypothesis test**

1) The data must be a SRS from the population
2) The population must be large enough ( >10 times the sample size)
3) The sample size must be large enough

$n\hat{p} \geq 5$ *and* $n(1-\hat{p}) \geq 5$ for CIs

$np_0 \geq 5$ *and* $n(1-p_0) \geq 5$ for Hypothesis tests




**Comparing Two Proportions**

$\hat{p}_1 = x_1 / n_1$ proportion of "*successes*" in sample 1

$\hat{p}_2 = x_2 / n_2$ proportion of "*successes*" in sample 2

The *sampling distribution* for …

$\hat{p}_1$ is approximately Normal with *Mean*= $p_1$, $SD = \sqrt{\frac{p_1(1-p_1)}{n_1}}$, & *Variance*= $\frac{p_1(1-p_1)}{n_1}$

$\hat{p}_2$ is approximately Normal with *Mean*= $p_2$, $SD = \sqrt{\frac{p_2(1-p_2)}{n_2}}$, & *Variance*= $\frac{p_2(1-p_2)}{n_2}$

$\hat{p}_1 - \hat{p}_2$ is approx Normal with *Mean*= $p_1 - p_2$, $SD = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

---

For testing $H_0 : p_1 - p_2 = 0$ we need the *sampling distribution* for $\hat{p}_1 - \hat{p}_2$ <u>when $H_0$ is true.</u>

$$H_0 \rightarrow p_1 = p_2 \ (= p) \rightarrow \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$z_s = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Testing to see if a new flu vaccine is more effective

$n_1$=75 get the new vaccine $x_1$=12 develop the flu over the next 6 months

$n_2$=80 get the old vaccine $x_2$=24 develop the flu over the next 6 months

| | Number with Event | |
|---|---|---|
| Event | Prevastatin ($n_1 = 900$) | Placebo ($n_2 = 411$) |
| ~~Cardiac Chest Pain~~ | ~~36~~ | ~~14~~ |
| Dermatologic Rash | 36 | 5 |
| Headache | 56 | 16 |

**<u>Conditions for a Valid CI and Hypothesis test</u>**

1) The samples are independent Simple Random Samples from the population

2) The populations must be large enough ( >10 times the sample sizes)

3) The sample sizes must be large enough:

$$n_1\hat{p}_1,\ n_1(1-\hat{p}_1),\ n_2\hat{p}_2,\ n_2(1-\hat{p}_2) \quad all\ \geq 5$$