

# ECON 3818: Introduction to Statistics with Computer Applications

---

Kyle Butts

November 4, 2020

## **Chapter 22: Inference about a Population Proportion**

---

# Inference about a Population Proportion

Previously we have discussed making inference in population *means*

This chapter talks about questions where we're interested in the proportion of an outcome

- Percent of adults that have multiple partners
  - single population proportion
- Is there a difference between the proportion of male students and proportion of female students that smoke cigarettes
  - comparing two population proportions

# The Sample Proportion, $\hat{p}$

The statistic that estimates the population proportion,  $p$ , is the **sample proportion**:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

For example:

Say we want to estimate the proportion of heterosexual adults who have had more than one sexual partners in the past year. To estimate this proportion, a researcher collected survey data and contacted 2673 people, and 170 said they had multiple sex partners

$$\hat{p} = \frac{170}{2673} = 0.0636$$

# Binomial Distribution Review

# Sampling Distribution of a Sample Proportion

We can think of random variables as binary (take only two values) and use Bernoulli distribution

- Assign one outcome 0 and the other outcome 1
- $X \sim B(1, p)$
- This means  $p$  is the **unobserved** probability of outcome 1

We use the sample statistic  $\hat{p} = \frac{\text{number of successes}}{\text{total observations}}$

- The **mean** of the sampling distribution is  $p$
- The **standard deviation** of the sampling distribution is  $\sqrt{\frac{p(1-p)}{n}}$

# Sampling Distribution of a Sample Proportion

Say we draw a simple random sample of size  $n$  from a large population that contains  $p$  proportion of successes. Let  $\hat{p}$  be the **sample proportion** of successes,

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

The **Central Limit Theorem** tells us that with a large enough sample size, the standardized value of  $\hat{p}$  will be approximately normal:

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

## Clicker Question

A study investigated ways to prevent staph infections in surgery patients. In a first step, the researchers examined the nasal secretions of a random sample of 6771 patients admitted to various hospitals for surgery. They found that 1251 of these patients tested positive for *Staphylococcus aureus*, a bacterium responsible for most staph infections.

What is the population and what is the parameter  $p$ ?

Calculate the statistic  $\hat{p}$  that estimates  $p$

- (a) 5.41
- (b) 0.185
- (c) 0.341





# Election Tomorrow

A poll by YouGov asked 1360 voters in Pennsylvania if they were going to vote for Biden or Trump. We will code a vote for Biden = 1, so the proportion  $\hat{p}$  is the proportion of people who will vote for Biden. Biden will win Pennsylvania if the population portion is  $p > .5$ .

They find that  $\hat{p} = .53$ . What is the sampling distribution of  $\hat{p}$ ?

Source: <https://projects.fivethirtyeight.com/polls/>



## What's the probability Biden Wins PA?

Using the sampling distribution, what's the probability that  $p > .50$ ?



# Confidence Intervals for a Population Proportion

We follow the same path from sampling distribution to confidence interval as we did for  $\bar{X}$

Note, the standard deviation of  $\hat{p}$  depends on the parameter  $p$  – a value that we don't know. We therefore estimate the standard deviation with the standard error of  $\hat{p}$ :

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

# Confidence Intervals for a Population Proportion

Say we draw a simple random sample of size  $n$  from a large population that contains an unknown proportion  $p$  of successes. An approximate  $C\%$  confidence interval for  $p$  is:

$$\hat{p} \pm Z_{\frac{1-C}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

What do we mean by large? Can only use this confidence interval when number of successes and failures in the sample are both at least 15 (to remember, half of 30 each)

## Example

A poll by YouGov asked 1360 voters in Pennsylvania if they were going to vote for Biden or Trump. We will code a vote for Biden = 1, so the proportion  $\hat{p}$  is the proportion of people who will vote for Biden. Biden will win Pennsylvania if the population portion is  $p > .5$ .

They find that  $\hat{p} = .53$ . What is the sampling distribution of  $\hat{p}$ ?

Check the conditions:

- SRS ✓
- number of success ( $1360 * 0.53$ ) and failures ( $1360 * 0.47$ ) are both larger than 15 ✓

So we can go ahead and calculate a 95% confidence interval for the population parameter  $p$ ...





## Clicker Question

We are given that  $n=670$ ,  $\hat{p} = 0.85$ , we will use the standard error of the sample proportion as

$$SE_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/n}$$

Which of the following is the correct calculation for a 95% confidence interval?

- (a)  $0.85 \pm 1.96 \cdot \sqrt{\frac{0.85 \cdot 0.15}{670}}$
- (b)  $0.85 \pm 1.645 \cdot \sqrt{\frac{0.85 \cdot 0.15}{670}}$
- (c)  $0.85 \pm 1.96 \cdot \frac{0.85 \cdot 0.15}{\sqrt{670}}$
- (d)  $571 \pm 1.96 \cdot \sqrt{\frac{571 \cdot 99}{670}}$

# Hypothesis Testing

We design a hypothesis test such as:

$$H_0 : \hat{p} = p_0 \text{ vs. } H_1 : \hat{p} \neq p_0$$

Or one-sided alternatives, such as:  $\hat{p} < p_0$  or  $\hat{p} > p_0$ .

We reject  $H_0$  if our p-value is lower than our level of significance

- p-value: probability of calculating the sample proportion, or more extreme value, if the null hypothesis is true
- level of significance: predetermined by the size of the test.

# Test Statistic

Draw an SRS of size  $n$  from a large population that contains an unknown proportion  $p$  of successes. To test the hypothesis  $H_0 : \hat{p} = p_0$ , compute the following z-statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Use this test when the sample size  $n$  is so large that both  $n \cdot p_0$  and  $n \cdot (1 - p_0) = 15$  or more.

A survey found that 571 out of 670 (85%) of Americans answered a question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

$$H_0 : p = 0.8$$

$$H_1 : p > 0.8$$



## Breakout Group

A survey found that 571 out of 670 (85%) of Americans answered a question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

$$H_0 : p = 0.8$$

$$H_1 : p > 0.8$$

Calculate the the p-value:

$$P(\hat{p} > 0.85 \mid p = 0.8)$$
$$P(Z > \frac{0.85 - 0.8}{\sqrt{\frac{0.8 \cdot 0.2}{670}}}) = P(Z > 3.25) = 0.0006$$

Since p-value = 0.0006 <  $\alpha$  = 0.05, reject  $H_0$ .

## Yesterday's Election: Hypothesis Testing

On Nov. 1st, the New York Times and Siena College released a poll for Wisconsin with  $n = 1253$  and the sample proportion of people supporting Biden was  $\hat{p} = 0.52$ . On election day, we learned the population proportion supporting Biden was  $p = 0.496$ . Would we have rejected the following hypothesis at the  $\alpha = 0.05$  significance level.

$$H_0 : p = 0.496$$

$$H_1 : p > 0.496$$





# Election Polling and Simple Random Sample

Why did we reject the null hypothesis which was true? Which of the following problems do we think occurred (list from ch. 9)?

- **Undercoverage**: when some groups in the population are left out of the process of choosing the sample
- **Oversampling**: when some groups are sampled more often than others in a way that is not representative of the population
- **Nonresponse**: when an individual chosen for the sample can't be contacted or refuses to participate
- **Response Bias**: a systematic pattern of incorrect responses in a sample survey
- **Wording Effect**: a systematic pattern of responses due to poor (or manipulated) wording of survey questions

## Breakout Group

Suppose you are an epidemiologist studying cancer incidence in an old manufacturing town. It is believed the cancer incidence in this town is above average. You know that the proportion of the national population that has a certain cancer is 0.03. The manufacturing town has an observed cancer incidence of 0.045 among a sample of 400 residents. Test the following hypothesis at the  $\alpha = 0.05$  significance level.

$$H_0 : p = 0.03$$

$$H_1 : p > 0.03$$

- (a) Reject  $H_0$
- (b) Fail to reject  $H_0$

## Breakout Group

Suppose you are an epidemiologist studying cancer incidence in an old manufacturing town. It is believed the cancer incidence in this town is above average. You know that the proportion of the national population that has a certain cancer is 0.03. The manufacturing town has an observed cancer incidence of 0.045 among a sample of 400 residents. Test the following hypothesis at the  $\alpha = 0.05$  significance level.

For what region of sample proportions  $\hat{p}$  will you reject the following null hypothesis at the  $\alpha = 0.05$  significance level.

$$H_0 : p = 0.03$$

$$H_1 : p > 0.03$$