Chapter 6

Inference for categorical data[1]

Department of Mathematics & Statistics
North Carolina A&T State University

[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

Inference for a single proportion

Practice

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

A) All 1000 get the drug.

B) 500 get the drug, 500 don't

Practice

Two scientists want to know if a certain druh is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

A) All 1000 get the drug.

B) 500 get the drug, 500 don't

Results from the GSS

The GSS asks the same question, below is the distribution of responses from the 2010 survey:

All 1000 get the drug	99
500 get the drug, 500 don't	571
Total	670

Parameter and point estimate

We would like to estimate the proportion of all Americans who have good intuition about experimental desgin, i.e. would answer "500 get the drug, 500 don't"? What are the parameter of interest and the point estimate?

Parameter and point estimate

We would like to estimate the proportion of all Americans who have good intuition about experimental desgin, i.e. would answer "500 get the drug, 500 don't"? What are the parameter of interest and the point estimate?

- **Parameter of interest:** Proportion of **all** Americans who have good intuition about experimental design.

$\mathbf{p}$ (a population proportion)

Parameter and point estimate

We would like to estimate the proportion of all Americans who have good intuition about experimental desgin, i.e. would answer "500 get the drug, 500 don't"? What are the parameter of interest and the point estimate?

- **Parameter of interest:** Proportion of **all** Americans who have good intuition about experimental design.

$$\mathbf{p} \text{ (a population proportion)}$$

- **Point estimate:** Proportion of **sampled** Americans who have good intuition about experimental design.

$$\hat{\mathbf{p}} \text{ (a sample proportion)}$$

Inference on a proportion

What percent of all Americans have good intuition about experimental desgin, i.e. would answer "500 get the drug, 500 don't"?

Inference on a proportion

What percent of all Americans have good intuition about experimental desgin, i.e. would answer "500 get the drug, 500 don't"?

- We can answer this research question using a confidence interval, which we know is always of the form

$$\textbf{point estimate} \pm \textbf{ME}$$

- And we also know that ME = critical value $\times$ standard error of the point estimate

Standard error of a sample proportion

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

Sample proportions are also nearly normally distributed

Central limit theorem for proportions

Sample proportions will be nearly normally distributed with mean equal to the population proportion, p, and standard error equal to $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

But of course this is true only under certain conditions...

Any guesses?

Sample proportions are also nearly normally distributed

Central limit theorem for proportions

Sample proportions will be nearly normally distributed with mean equal to the population proportion, p, and standard error equal to $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

But of course this is true only under certain conditions...

Any guesses?

Independent observations, at least 10 successes and 10 failures

---

Note: If p is unknown (most cases), we use $\hat{p}$ in the calculation of the standard error.

Back to experimental design...

The GSS found that 571 out of 670 (87%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Back to experimental design...

The GSS found that 571 out of 670 (87%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Given: $n = 670, \hat{p} = 0.85$. First, check conditions.

- **Independence**: The sample is random, and 670 $<$ 10% of all Americans, therefore we can assume that one respondent's response is independent of another.
- **Success-Failure**: 571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.

Practice

We are given that $n = 670, \hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Which of the below is the correct calculation of the 95% confidence interval?

A) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

B) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

C) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$

D) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

Practice

We are given that $n = 670, \hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Which of the below is the correct calculation of the 95% confidence interval?

A) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}} \rightarrow (0.82, 0.88)$

B) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

C) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$

D) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

Choosing a sample size

How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%?

Choosing a sample size

How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%?

$$ME = z^* \times SE$$

$$0.01 \geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \text{Use } \hat{p} \text{ from previous study}$$

$$0.01^2 \geq 1.96^2 \times \frac{0.85 \times 0.15}{n}$$

$$n \geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2}$$

$$n \geq 4898.04 \rightarrow \text{ n should be at least 4,899}$$

What if there isn't a previous study?

... use $\hat{p} = 0.5$

Why?

What if there isn't a previous study?

... use $\hat{p} = 0.5$

Why?

- If you don't know any better, 50-50 is a good guess.
- $\hat{p} = 0.5$ gives the most conservative estimate $-$ highest possible sample size.

CI vs. HT for proportions

- Success-Failure conditions:
  - CI: At least 10 **observed** successes and failures.
  - HT: At least 10 **expected** successes and failures, calculated using the null value.
- Standard error:
  - CI: Calculate using observed sample proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
  - HT: Calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

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

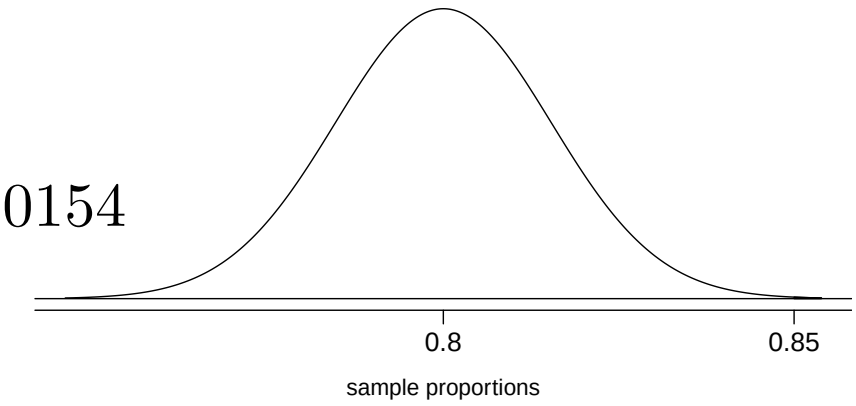
The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

$$H_0 : p = 0.80 \quad H_A : p > 0.80$$

$$SE = \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154$$

$$Z = \frac{0.85 - 0.80}{0.0154} = 3.25$$

$$p - value = 1 - 0.9994 = 0.0006$$



Since the p-value is low, we reject H_0. The data provide convincing evidence that more than 80% of Americans have a good intuition on experimental design.

Practice

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

A) Yes

B) No

C) Cannot tell

Practice

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

A) Yes

B) No

C) Cannot tell

Recap - Inference for one proportion

- Popular parameter: p. point estimate: $\hat{p}$
- Conditions:
  - Independence
    - Random sample and 10% condition
  - At least 10 successes and failures
    - If not $\rightarrow$ randomization
- Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - for CI: use $\hat{p}$
  - for HT: use p_0