Inference for categorical data

- Inference for a single proportion
- Comparing two proportions

Inference for a Single Proportion

- Confidence interval for a single proportion
- Hypothesis testing for a single proportion

Confidence interval

for a proportion

from the 2010 GSS

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

experimental design	
bad intuition	99
good intuition	571
total	670

What percent of Americans have good intuition about experimental design?

Parameter of interest

percent of *all* Americans who have good intuition about experimental design.

p

a population proportion

Point estimate:

percent of *sampled* Americans who have good intuition about experimental design.

$\hat{p}$

a sample proportion

estimating a proportion

What percent of all Americans have good intuition about experimental design, 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

point estimate ± ME

And we also know that *ME = critical value times the SE* of the point estimate.

$SE_{\hat{p}} = ?$

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 mean, *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.

The GSS found that 571 out of 670 (85%) of 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.

1. *Independence*: The sample is random, and 670 < 10% of all Americans, therefore we can assume that one respondent's response is independent of another.
2. *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

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

95% confidence interval of *p?*

We are 95% confident that 82.3% to 87.7% of all Americans have good intuition about experimental design.

choosing a sample size

The margin of error for the previous confidence interval was 2.7%.

If, for a new confidence interval based on a new sample, we wanted to reduce the margin of error to 1% while keeping the confidence level the same, at least how many people should you sample?

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%?

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

$$0.01 \geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}}$$ → *Use estimate for $\hat{p}$ from previous study*

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

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

$$n \geq 4898.04$$ → *n should be at least 4,899*

calculating the required sample size for desired ME

remember $$ME = z^* \times SE = z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- If there is a previous study that we can rely on the value of $\hat{p}$ use that in the clculation of the required sample size

- If not, 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