Inference for Proportions

IS381 - Statistics and Probability with R

Jason Bryer, Ph.D.

March 30, 2026

One Minute Paper Results

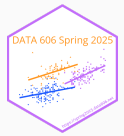
What was the most important thing you learned during this class?

What important question remains unanswered for you?

Example

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1,000 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?

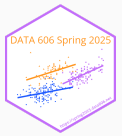
> - **500 get the drug, 500 don't**



Survey of Americans

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

Response	n
All 1000 get the drug	99
500 get the drug 500 don't	571
Total	**670**



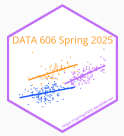
Parameter of Interest

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

$$p(population\ proportion)$$

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

$$\hat{p}(sample\ proportion)$$



Inference for a proportion

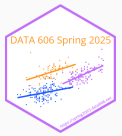
What percent of all Americans have good intuition about experimental design (i.e. would answer "500 get the drug 500 don't?"

- Using a confidence interval

$$point\ estimate \pm ME$$

- We know that ME = critical value x standard error of the point estimate.

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



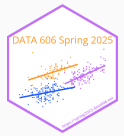
Central limit theoreom for proportions

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

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

This is true given the following conditions:

- independent observations
- at least 10 successes and 10 failures



Simulating the CLT

Let's consider a population of 1,000,000 where the true proportion is 0.85.

```
pop_prop <- 0.85
N <- 1000000
pop <- c(rep(0, N * (1 - pop_prop)),
         rep(1, N * pop_prop))
```

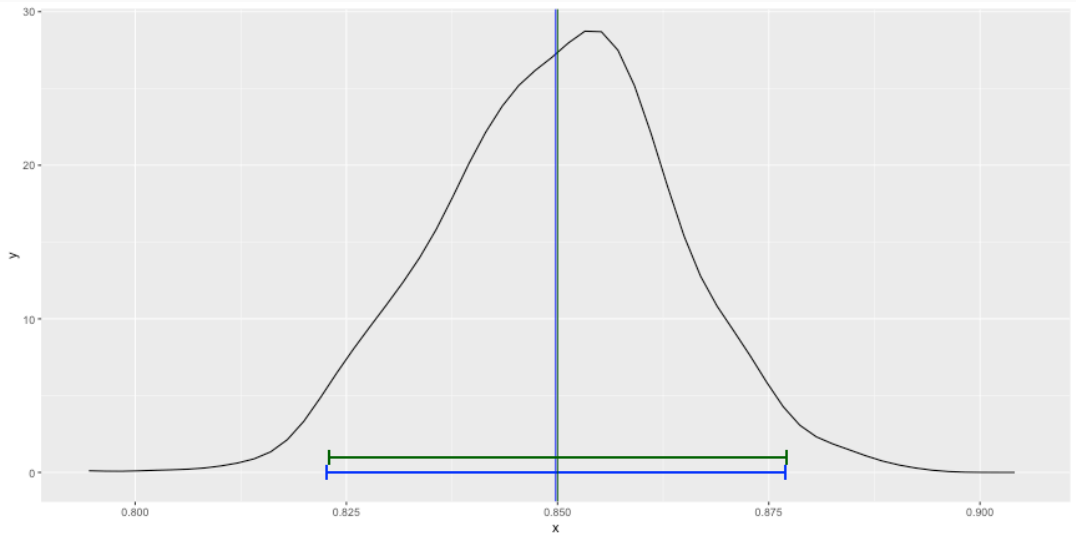
We can estimate the sampling distribution by taking 1,000 random samples of size 30.

```
n <- 670
samp_dist <- numeric(1000)
for(i in 1:length(samp_dist)) {
    samp_dist[i] <- sample(pop, size = n) |> mean()
}
```

Calculate the standard error using one sample.

```
samp_se <- sqrt((0.85 * (1 - 0.85)) / 670)
```

The figure represents the sampling distribution. The blue line is from the estimated sampling distribution. The green line is from the one sample (i.e using the SE formula).



Back to the Survey

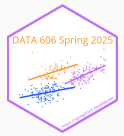
- 571 out of 670 (85%) of Americans answered the question on experimental design correctly.
- Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

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

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.

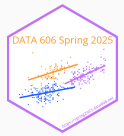


Calculating Confidence Interval

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

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

We are 95% confidence the true proportion of Americans that have a good intuition about experimental designs is betwee 82% and 88%.



How many should we sample?

Suppose you want a 3% margin of error, how many people would you have to survey?

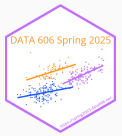
Use $\hat{p} = 0.5$

- 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

$$0.03 = 1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}}$$

$$0.03^2 = 1.96^2 \times \frac{0.5 \times 0.5}{n}$$

$$n \approx 1,068$$



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%?

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

Using $\hat{p}$ from previous slides.

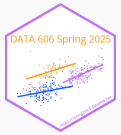
$$0.01 \geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}}$$

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

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

$$n \geq 4,898.04$$

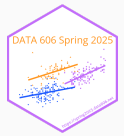
n needs to be at least 4,899 to have a 1% margin of error.



Example: Two Proportions

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

Response	GSS	Duke
A great deal	454	69
Some	124	40
A little	52	4
Not at all	50	2
Total	680	105



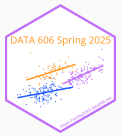
Parameter and Point Estimate

Parameter of interest: Difference between the proportions of *all* Duke students and *all* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

Point estimate: Difference between the proportions of *sampled* Duke students and *sampled* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{Duke} - \hat{p}_{US}$$



Everything else is the same...

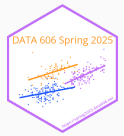
- CI: $point\ estimate \pm margin\ of\ error$
- HT: $Z = \frac{point\ estimate - null\ value}{SE}$

Standard error of the difference between two sample proportions

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

Conditions:

1. Independence within groups: The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well. $n_{Duke} < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.
2. Independence between groups: The sampled Duke students and the US residents are independent of each other.
3. Success-failure: At least 10 observed successes and 10 observed failures in the two groups.



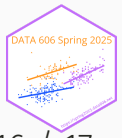
95% Confidence Interval

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap ($p_{Duke} - p_{US}$).

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z* \times \sqrt{\frac{p_{Duke}\left(1 - p_{Duke}\right)}{n_{Duke}} + \frac{p_{US}\left(1 - p_{US}\right)}{n_{US}}}$$

$$(0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} = (-0.108,\ 0.086)$$



One Minute Paper

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?

https://forms.gle/bz8GvYWfKdMggRv38