MATH 1051H S61 - Lecture 12 - Proportion Testing

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?

1. All 1000 get the drug
2. 500 get the drug, 500 don't

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?

1. All 1000 get the drug
2. *500 get the drug, 500 don't*

Results from the GSS

The General Social Survey (GSS) collects information and keep a historical record of the concerns, experiences, attitudes, and practices of residents of the United States. Since 1972, the GSS has been monitoring societal change and studying the growing complexity of American society. Canada has been running a similar survey since 1985.

The GSS asks the question from the previous slide. 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

http://www.statcan.gc.ca/pub/89f0115x/89f0115x2013001-eng.htm

Parameter and Point Estimation

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

Parameter and Point Estimation

We would like to estimate the proportion of all Americans who have good intuition about experimental design, 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

p: a population proportion

Parameter and Point Estimation

We would like to estimate the proportion of all Americans who have good intuition about experimental design, 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

p: a population proportion

Point Estimate: proportion of **sampled** Americans who have good intuition about experimental design.

$\hat{p}$: a sample proportion

Inference on 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 question using a confidence interval, which we know is always of the form

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

where

$$\text{ME} = z^\star \times \text{SE}. \text{ (or } t^\star \times \text{SE)}$$

So what is the SE of our point estimate, $\text{SE}_{\hat{p}}$?

New Formula: SE of a Point Estimate $\hat{p}$

When we have a **sample proportion**, the standard error has a known formula:

$$\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

What are p and n?

1. n is the number of samples (it's a **sample proportion**)
2. p is the true underlying population proportion …

But we don't know p!

We "cheat" here, and replace p with $\hat{p}$. It mostly works.

Sample Proportions are Almost Normally Distributed

Remember the Central Limit Theorem (CLT).

Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\text{SE}_{\hat{p}}$ from the last slide. We can write this formally.

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

But, of course, this is only true under certain conditions … *any guesses?*

Sample Proportions are Almost Normally Distributed

Remember the Central Limit Theorem (CLT) from earlier.

Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\text{SE}_{\hat{p}}$ from the last slide. We can write this formally.

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

But, of course, this is only true under certain conditions … *any guesses?*

The requirements of the CLT! Independent observations, and "enough" samples

Where have you seen this … ?

R Assignment 2, and Workshop 8! We did this: flipped coins, and kept track of the **proportion** of heads. We then converted them into averages, but it's the same idea.

Rule of Thumb for Proportions

There is a rule of thumb for what “enough samples” means for a sample proportion inference:

1. At least 10 success cases
2. At least 10 failure cases

If you do not have the above, the CLT may not be a good approximation.

Back to the GSS Question

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?

Back to the GSS Question

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:

- $n = 670$
- $\hat{p} = 0.852$

Check the conditions!

Back to the GSS Question

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:

- $n = 670$
- $\hat{p} = 0.852$

Check the conditions!

1. **Independence**: The GSS is sampled randomly, and the population is much larger than the sample, so we can assume the responses are random.

Back to the GSS Question

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:

- $n = 670$
- $\hat{p} = 0.852$

Check the conditions!

1. **Independence**: The GSS is sampled randomly, and the population is much larger than the sample, so we can assume the responses are random.
2. **Enough Samples**: 571 people answered correctly (success) and 99 answered incorrectly (failure). Both numbers are greater than 10.

Practice

We are given $n = 670$, $\hat{p} = 0.852$, and we know that

$$\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

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

1. $0.852 \pm 1.96 \times \sqrt{\frac{0.85(0.15)}{670}}$
2. $0.852 \pm 1.65 \times \sqrt{\frac{0.85(0.15)}{670}}$
3. $0.852 \pm 1.96 \times \frac{0.85(0.15)}{\sqrt{670}}$
4. $571 \pm 1.96 \times \frac{571(99)}{670}$

Practice

We are given $n = 670$, $\hat{p} = 0.852$, and we know that

$$\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

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

1. $0.852 \pm 1.96 \times \sqrt{\frac{0.85(0.15)}{670}} = (0.825, 0.879)$
2. $0.852 \pm 1.65 \times \sqrt{\frac{0.85(0.15)}{670}}$
3. $0.852 \pm 1.96 \times \frac{0.85(0.15)}{\sqrt{670}}$
4. $571 \pm 1.96 \times \frac{571(99)}{670}$

New Idea: Choosing a Sample Size

We now have the formulas which allow us to compute confidence intervals using the normal approximation for sample proportions.

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

New Idea: Choosing a Sample Size

We now have the formulas which allow us to compute confidence intervals using the normal approximation for sample proportions.

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

$$\text{ME} = z^* \times \text{SE}$$

New Idea: Choosing a Sample Size

We now have the formulas which allow us to compute confidence intervals using the normal approximation for sample proportions.

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

$$\text{ME} = z^* \times \text{SE}$$

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

(since 1% = 0.01, 95% CI means use $z^* = 1.96$, and we're still interested in the GSS question)

New Idea: Choosing a Sample Size

We now have the formulas which allow us to compute confidence intervals using the normal approximation for sample proportions.

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

$$\text{ME} = z^* \times \text{SE}$$

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

$$(0.01)^2 \geq 1.96^2 \times \frac{0.85(0.15)}{n} \qquad \text{square both sides}$$

New Idea: Choosing a Sample Size

We now have the formulas which allow us to compute confidence intervals using the normal approximation for sample proportions.

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

$$\text{ME} = z^* \times \text{SE}$$

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

$$(0.01)^2 \geq 1.96^2 \times \frac{0.85(0.15)}{n} \qquad \text{square both sides}$$

$$n(0.01)^2 \geq 1.96^2 \times 0.85(0.15) \qquad \text{cross-multiply the denominator}$$

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

New Idea: Choosing a Sample Size

We now have the formulas which allow us to compute confidence intervals using the normal approximation for sample proportions.

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

$$\begin{aligned} \text{ME} &= z^* \times \text{SE} \\ 0.01 &\geq 1.96 \times \sqrt{\frac{0.85(0.15)}{n}} \\ n &\geq 1.96^2 \times \frac{0.85(0.15)}{0.01^2} \\ n &\geq 4898.04 \end{aligned}$$

Therefore we need at least 4,899 samples to cut the ME of a 95% confidence interval (for this problem) down to 1%.

A Tricky Bit

You may have noticed that we assumed that the ME for the new sample would be the same as the old sample: we used 0.85 and 0.15 in our formula!

This is common if we are repeating a study. The first study is sometimes called a **pilot study**, and it gives us a rough idea what our $\hat{p}$ might be.

What if we didn't have a pilot study to rely on?

A Tricky Bit

You may have noticed that we assumed that the ME for the new sample would be the same as the old sample: we used 0.85 and 0.15 in our formula!

This is common if we are repeating a study. The first study is sometimes called a **pilot study**, and it gives us a rough idea what our $\hat{p}$ might be.

What if we didn't have a pilot study to rely on?

Use the default $\hat{p} = 0.5$.

A Tricky Bit

Why would $\hat{p} = 0.50$ be a default?

- If you don't know any better, 50/50 guessing seems reasonable
- Using $\hat{p} = 0.50$ is the most conservative estimate, and gives the highest possible sample size number

(for those who have some calculus and are curious) 0.5 is a maximum because the formula has $\hat{p}(1 - \hat{p})$, which is a quadratic function of $\hat{p}$, which has a local maximum at $\hat{p} = 0.50$.

Confidence Intervals versus Hypothesis Testing

We have slightly different **Success-Failure Conditions** (for number of samples required):

- **CI**: at least 10 *observed* successes and failures
- **HT**: at least 10 *expected* successes and failures, under the null assumption

Confidence Intervals versus Hypothesis Testing

Standard Error

- **CI**: calculated using the sample proportion, $\hat{p}$

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

- **HT**: calculated using the null hypothesis value, p_0

$$\text{SE} = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

Practice

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

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

(we use a one-tailed hypothesis test because that is our question: “more than 80%”)

Practice

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

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

$$\text{SE} = \sqrt{\frac{0.80(0.20)}{670}} = 0.0154$$

Practice

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

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

$$\text{SE} = \sqrt{\frac{0.80(0.20)}{670}} = 0.0154$$

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

Practice (*p*-value)

Compute the *p*-value for this value of z.

Our Method: Use R!

```
1 - pnorm(q = 3.25)
```

```
## [1] 0.000577025
```

```
pnorm(q = 3.25, lower.tail = FALSE)
```

```
## [1] 0.000577025
```

So the *p*-value is 0.0006.

Practice: Conclusion

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?

1. Yes
2. No
3. Can't 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?

1. Yes
2. *No*
3. Can't tell

Recap: Inference for One Proportion

Population parameter p, Point Estimate $\hat{p}$

Recap: Inference for One Proportion

Population parameter p, Point Estimate $\hat{p}$

Conditions:

- independence
  - random sample, and less than 10% of population
- at least 10 successes and 10 failures
  - if not, we can't use the normal approximation $\rightarrow$ use randomization/permutation instead

Recap: Inference for One Proportion

Population parameter p, Point Estimate $\hat{p}$

Conditions:

- independence
  - random sample, and less than 10% of population
- at least 10 successes and 10 failures
  - if not, we can't use the normal approximation $\rightarrow$ use randomization/permutation instead
- Standard Error (SE) = $\sqrt{\frac{p(1-p)}{n}}$
  - for CI, use $\hat{p}$
  - for HT, use p_0

Example: Libraries

Do the majority of voters in a large city favour increased funding for public libraries? Suppose a poll of 250 randomly selected voters in this city found that 140 of them favoured increased funding for public libraries.

Hypotheses:

$$H_0 : p = 0.50 \qquad \text{versus} \qquad H_A : p > 0.50$$

Check assumptions

- The sample is random and independent.
- Large sample size: $np_0 = n(1 - p_0) = 250(0.5) = 125 \geq 10$
- Large population: the city is large, so the number of voters is at least $10 \times 250 = 2500$

Computing the test statistic

Start with:

$$\hat{p} = \frac{140}{250} = 0.56$$

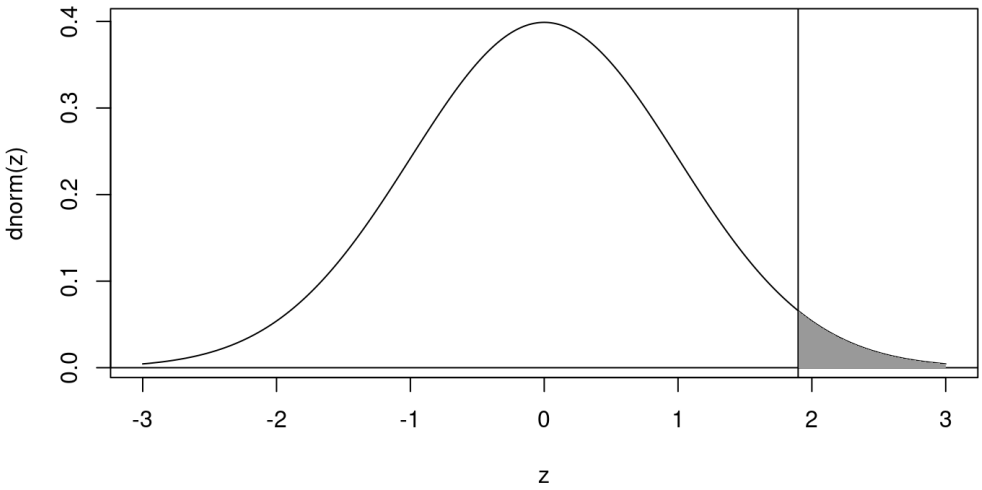
and then compute the SE:

$$\text{SE} = \sqrt{\frac{0.5(0.5)}{250}} = 0.031623.$$

Combine these to make the test statistic:

$$Z_{\text{test}} = \frac{0.56 - 0.50}{0.031623} = 1.8974$$

Visualizing the *p*-value



Computing the *p*-value

```
1 - pnorm(q = 1.8974)
```

```
## [1] 0.02888758
```

```
pnorm(q = 1.8974, lower.tail = FALSE)
```

```
## [1] 0.02888758
```

Computing the confidence interval

Recall the formula:

$$\hat{p} \pm z^* \times \text{SE}$$

so we have

$$0.56 \pm z* \times 0.031623$$

For a 95% confidence interval, we use

```
qnorm(0.95)
```

```
## [1] 1.644854
```

so

$$0.56 \pm 1.644854 \times 0.031623 = (0.508, 0.620)$$

Conclusion

So, since $p < \alpha$ (0.05), we reject the null. The same conclusion comes from the confidence interval not including the null hypothesis.

Thus, we reject the null hypothesis, and conclude that there is evidence to support a majority (more than half) of voters in a large city favouring increased funding for libraries.

Banking Example

A large bank polled 2,000 customers who have access to the Internet. Of these, 1,210 did their banking online.

1. What is the point estimate for the true proportion of the bank's customers who do their banking online?
2. Construct a 95% confidence interval for the proportion of the bank's customers who do their banking online.
3. If you were a polling agency, how would you report your results?

Banking Example: Q1

The point estimate is the proportion of customers polled:

$$\hat{p} = \frac{1210}{2000} = 0.605.$$

```
p_hat <- 1210 / 2000
p_hat
```

```
## [1] 0.605
```

Banking Example: Q2

To construct the confidence interval, we need the Margin of Error, ME.

```
z_star <- qnorm(p = 0.95 + (1 - 0.95) / 2, lower.tail = TRUE)
n <- 2000
ME <- z_star * sqrt( p_hat * (1 - p_hat) / n )
ME
```

```
## [1] 0.02142443
```

That is, the formula:

$$\text{ME} = z^\star \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0214.$$

Banking Example: Q2

Then the 95% confidence interval is:

```
LCL <- p_hat - ME
UCL <- p_hat + ME
```

```
## (0.5836,0.6264)
```

Banking Example: Q3

Let’s phrase this as a survey result:

60.5% of the bank’s customers use internet banking. These results are obtained from a survey of 2000 randomly selected clients, and the estimate is considered accurate to within plus-or-minus 2.1 percentage points, 19 times out of 20.

Final Example: Cyber Security

Based on information from the National Cyber Security Alliance, 93% of computer owners believe they have antivirus programs installed on their computers.

In a random sample of 400 scanned computers, it is found that 380 of them (or 95%) actually have antivirus software programs.

Use the sample data from the scanned computers to test the claim that 93% of computers have antivirus software.

Requirements

1. The 400 computers were randomly selected (check!)
2. There is a fixed number of independent trials, two possible outcomes (check!)
3. Is $np \geq 10$? Is $n(1-p) \geq 10$?

$$np = (400)(0.93) = 372$$
$$n(1-p) = (400)(1-0.93) = 28$$

Check!

Hypotheses

Write the hypotheses:

$$\mathbf{H_0}: p = 0.93 \qquad \text{versus} \qquad \mathbf{H_A}: p \neq 0.93$$

Significance Level, Test Statistic

Since we didn't have a specified level, choose $\alpha = 0.05$. We are testing a claim about a **population proportion**, so we will use a normal approximation:

$$z_{\text{test}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{380}{400} - 0.93}{\sqrt{\frac{0.93(0.07)}{400}}}$$

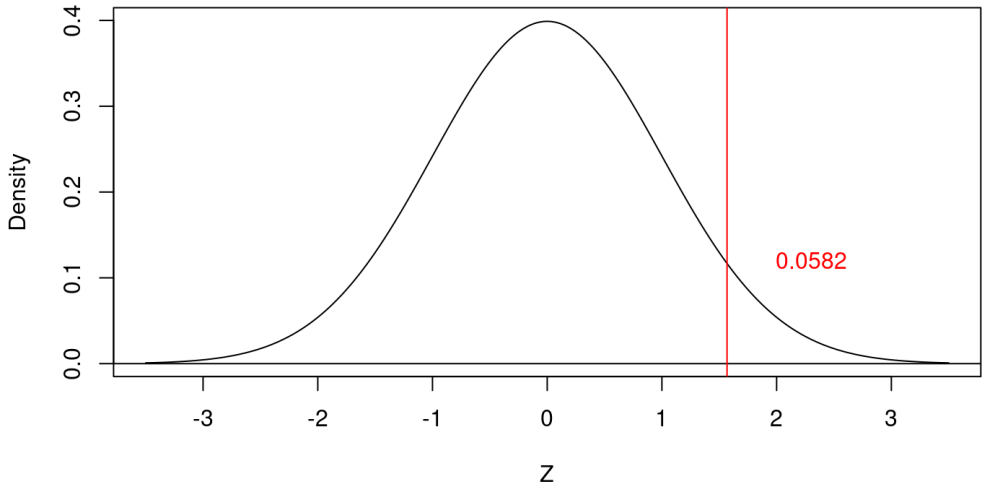
```
z_test <- ( 380/400 - 0.93 ) / sqrt( 0.93 * 0.07 / 400 )
z_test
```

```
## [1] 1.567724
```

The p-value

```
pnorm(z_test, lower.tail = FALSE) * 2
```

```
## [1] 0.1169457
```



Conclusion

Thus, since $p > \alpha$, we do not have evidence at the 95% level to conclude that the population proportion of computers having antivirus software is not 93%. In other words, there is not sufficient evidence to warrant rejection of this claim.