Elementary Statistics – Inference for Categorical Data Pt. 1

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Inference for comparing two proportions

Warm-Up: Hypothesis Testing

Pieces of a Hypothesis test:

1. ***Two competing and complementary claims about the world***
2. ***Test Statistic***
3. ***Null Distribution***
4. ***P-value***

The ***z-score*** of an observation characterizes the number of standard deviations it falls above or below the population average if the null hypothesis is true.

for a sample mean, $\bar{x}$, $Z = \frac{\bar{x}-\mu}{\sigma}$

for a sample proportion, $\hat{p}$, $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

Warm-Up: Hypothesis Testing

Pieces of a Hypothesis test:

1. ***Two competing and complementary claims about the world***
2. ***Test Statistic***
3. ***Null Distribution***
4. ***P-value***

The ***z-score*** of an observation characterizes the number of standard deviations it falls above or below the population average if the null hypothesis is true.

for a sample mean, $\bar{x}$, $Z = \frac{\bar{x}-\mu}{\sigma}$

for a sample proportion, $\hat{p}$, $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

You suspect more the 50% of WSU students are commuters. You perform an experiment to statistically test this suspicion, and sample 100 students. You find 33 of them are commuters.

Your hypotheses are: $H_0: p = 0.5$, $H_A: p > 0.5$

What is Z? What p-value does Z imply? Should you reject your null hypothesis?

Inference for One Proportion

Our sample statistic ($\hat{p}$) represents our best guess for the true population parameter, (p). We know this best guess is not perfect; we expect error (variability) due to the sampling process.

Because we can't know the truth directly, we infer the truth via:

1. A confidence interval
2. A hypothesis test

Inference for One Proportion

In either case, we need the sampling distribution for $\hat{p}$.

We can approximate it with the normal distribution as long as:

1. The sample's observations are **independent**
2. There are at least 10 successes and 10 failures in the sample, i.e. the **success-failure condition** is met. To check, verify that $(best\ guess\ for\ p)n \geq 10, (1 - (best\ guess\ for\ p))n \geq 10$

When the conditions are met so that the distribution of $\hat{p}$ is nearly normal, variability of $\hat{p}$ is well described by:

$$SE(\hat{p}) = \sqrt{\frac{(best\ guess\ of\ p)(1-best\ guess\ of\ p)}{n}}$$

Inference for One Proportion

Confidence Interval for One Proportion

For confidence intervals, we use $\hat{p}$ as the best guess of p, so

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

Inference for One Proportion

Confidence Interval for One Proportion

For confidence intervals, we use $\hat{p}$ as the best guess of p, so

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

We use SE to compute margin of error for our confidence interval: $(\hat{p} - z^* SE, \hat{p} + z^* \mathrm{SE})$

Inference for One Proportion

Confidence Interval for One Proportion

For confidence intervals, we use $\hat{p}$ as the best guess of p, so

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

We use SE to compute margin of error for our confidence interval:
$(\hat{p} - z^* SE, \hat{p} + z^* \mathrm{SE})$

z^* is calculated from a specified percentile on the normal distribution.

Ex. 5^{th} percentile for a 95% confidence

Inference for One Proportion

Confidence Interval for One Proportion

For confidence intervals, we use $\hat{p}$ as the best guess of p, so

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

We use SE to compute margin of error for our confidence interval:
$(\hat{p} - z^* SE, \hat{p} + z^* \mathrm{SE})$

z^* is calculated from a specified percentile on the normal distribution.

Ex. 5th percentile for a 95% confidence

Find z^* for a 95% CI, a 90% CI, and a 99% CI

Confidence Interval for One Proportion

Inference for One Proportion

For confidence intervals, we use $\hat{p}$ as the best guess of p, so

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

We use SE to compute margin of error for our confidence interval:
$(\hat{p} - z^* SE, \hat{p} + z^* \mathrm{SE})$

z^* is calculated from a specified percentile on the normal distribution.

Ex. 5^{th} percentile for a 95% confidence

You suspect more the 50% of WSU students are commuters. You perform an experiment to statistically test this suspicion, and sample 100 students. You find 33 of them are commuters. Calculate a 95% CI for p from your $\hat{p}$

Inference for One Proportion

Hypothesis Test for One Proportion

When the conditions are met so that the distribution of $\hat{p}$ is nearly normal, variability of $\hat{p}$ is well described by:

$$SE(\hat{p}) = \sqrt{\frac{(best\ guess\ of\ p)(1-best\ guess\ of\ p)}{n}}$$

For hypothesis tests, we use p form H_0 as the best guess of p, so

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

Inference for One Proportion

Hypothesis Test for One Proportion

When the conditions are met so that the distribution of $\hat{p}$ is nearly normal, variability of $\hat{p}$ is well described by:

$$SE(\hat{p}) = \sqrt{\frac{(best\ guess\ of\ p)(1-best\ guess\ of\ p)}{n}}$$

For hypothesis tests, we use p form H_0 as the best guess of p, so

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

So, another way to express the z-score is:

$$Z = \frac{\hat{p} - p_0}{SE}$$

Comparing Two Proportions

So far, we've done inference to see if our population proportion differs from some hypothesized value.

Ex. Is the proportion of WSU students who commute greater than 50%?

Sometimes our research question instead focuses on **comparing proportions from two groups**.

Ex. Is the proportion of WSU students who commute different than the proportion of Springfield College students who commute?

Comparing Two Proportions

So far, we've done inference to see if our population proportion differs from some hypothesized value.

Ex. Is the proportion of WSU students who commute greater than 50%?

Sometimes our research question instead focuses on **comparing proportions from two groups**.

Ex. Is the proportion of WSU students who commute different than the proportion of Springfield College students who commute?

Just like with one proportion, we can use sample statistics to infer the population level answer to this question with
(a) a confidence interval and/or
(b) a hypothesis test

Comparing Two Proportions

To compare two proportions, we look at their difference.

Ex. Is the proportion of WSU students who commute different than the proportion of Springfield College students who commute?

To answer this, we need to look at $\hat{p}_{WSU} - \hat{p}_{SC}$

Inference for Two Proportions

Confidence Interval for Difference Between Two Proportions

Conditions for $\hat{p}_1 - \hat{p}_2$ to be approximated with the normal distribution:

1. Data are **independent within and between** the two **groups**
2. The **success-failure condition** is met in each group. To check, verify that $(best\ guess\ for\ p_1)n_1 \geq 10, \big(1 - (best\ guess\ for\ p_1)\big)n_1 \geq 10, (best\ guess\ for\ p_2)n_2 \geq 10, \big(1 - (best\ guess\ for\ p_2)\big)n_2 \geq 10$

When the conditions are met so that the distribution of $\hat{p}_1 - \hat{p}_2$ is nearly normal, variability of $\hat{p}_1 - \hat{p}_2$ is well described by:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(best\ guess\ of\ p_1)(1-best\ guess\ of\ p_1)}{n_1} + \frac{(best\ guess\ of\ p_2)(1-best\ guess\ of\ p_2)}{n_2}}$$

p_1, p_2 are population proportions for each group and n_1, n_2 are the sample sizes for each group

Inference for Two Proportions

Confidence Interval for Difference Between Two Proportions

For confidence intervals, we use $\hat{p}_1$ and $\hat{p}_2$ as the best guesses of p_1 and p_2 so

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Inference for Two Proportions

Confidence Interval for Difference Between Two Proportions

For confidence intervals, we use $\hat{p}_1$ and $\hat{p}_2$ as the best guesses of p_1 and p_2 so

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

We use SE to compute margin of error for our confidence interval:
$\left((\hat{p}_1 - \hat{p}_2) - z^* SE, (\hat{p}_1 - \hat{p}_2) + z^* \mathrm{SE}\right)$

You suspect different proportions of WSU and SC students commute. You perform an experiment to statistically test this suspicion. You sample 120 WSU students and find 40 of them are commuters. You sample 130 SC students and find 35 of them are commuters. Calculate a 95% CI for $p_{SC} - P_{WSU}$ from your $\hat{p}_{SC}$ and $\hat{p}_{WSU}$.

Inference for Two Proportions

Hypothesis Test for Difference Between Two Proportions

For a hypothesis test, our null hypothesis will be that there is no difference between proportions.

Inference for Two Proportions

Hypothesis Test for Difference Between Two Proportions

For a hypothesis test, our null hypothesis will be that there is no difference between proportions.

You suspect different proportions of WSU and SC students commute. You perform an experiment to statistically test this suspicion. You sample 120 WSU students and find 40 of them are commuters. You sample 130 SC students and find 35 of them are commuters.

You want to perform a hypothesis test to see if there is a difference in these proportions. What is H_0 in terms of $\hat{p}_{SC} - \hat{p}_{WSU}$? What is H_A?

Inference for Two Proportions

Hypothesis Test for Difference Between Two Proportions

For a hypothesis test, our null hypothesis will be that there is no difference between proportions.

When this is our null hypothesis, we will use a pooled proportion to check the success-failure condition for approximating the distribution of $\hat{p}_1 - \hat{p}_2$ with the normal distribution, and for calculating Z.

$$\hat{p}_{pool} = \frac{number\ of\ successes}{number\ of\ cases} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Inference for Two Proportions

Hypothesis Test for Difference Between Two Proportions

For a hypothesis test, our null hypothesis will be that there is no difference between proportions.

When this is our null hypothesis, we will use a pooled proportion to check the success-failure condition for approximating the distribution of $\hat{p}_1 - \hat{p}_2$ with the normal distribution, and for calculating Z.

$$\hat{p}_{pool} = \frac{number\ of\ successes}{number\ of\ cases} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Hypothesis Test for Difference Between Two Proportions

Inferenc
Two
Proportions

You suspect different proportions of WSU and SC students commute. You perform an experiment to statistically test this suspicion. You sample 120 WSU students and find 40 of them are commuters. You sample 130 SC students and find 35 of them are commuters.

Finish the hypothesis test. Calculate Z, find the p-value, and compare to an α of 0.05.

$$\hat{p}_{pool} = \frac{number\ of\ successes}{number\ of\ cases} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Inference Practice

Work with a small group on the proportion-inference practice problems under Demos on the course website.

Be prepared to share your answers with the class.