

Lecture 21

Categorical Data Analysis I: Inference for Proportions

Text: Chapter 10

STAT 8010 Statistical Methods I

November 10, 2020

Whitney Huang
Clemson University

In the next few lectures we will focus on **categorical data analysis**, i.e, statistical inference for categorical data

- Inference for a single proportion p
- Comparison of two proportions p_1 and p_2
- Inference for multi-category data and multivariate category data

Inference for a single proportion: Motivated Example

Researchers in the development of new treatments for cancer patients often evaluate the effectiveness of new therapies by reporting the **proportion** of patients who survive for a specified period of time after completion of the treatment. A new genetic treatment of 870 patients with a particular type of cancer resulted in 330 patients surviving at least 5 years after treatment. **Estimate** the proportion of all patients with the specified type of cancer who would survive at least 5 years after being administered this treatment.

- Binary (two-category) outcomes: “success” & “failure”
- Similar to the inferential problem for μ , we would like to infer p , the population proportion of success \Rightarrow **point estimate, interval estimate, hypothesis testing**

- Point estimate:

$$\hat{p} = \frac{X(\text{\# of "successes"})}{n}$$

Recall: $X \sim \text{Bin}(n, p) \Rightarrow E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{np}{n} = p$

- $100(1 - \alpha)\%$ CI for p :

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{(\hat{p})(1 - \hat{p})}{n}}$$

Why?

- CLT approximation: $\hat{p} \approx N(p, \sigma_{\hat{p}}^2)$ where n “sufficiently large”
 $\Rightarrow \min(np, n(1 - p)) \geq 5$
- $\sigma_{\hat{p}}^2 = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{1}{n^2} n(p)(1 - p) = \frac{p(1-p)}{n}$

A new genetic treatment of 870 patients with a particular type of cancer resulted in 330 patients surviving at least 5 years after treatment.

- 1 Estimate the proportion of all patients who would survive at least 5 years after being administered this treatment.
- 2 Construct a 95% CI for p

Among 900 randomly selected registered voters nationwide, 63% of them are somewhat or very concerned about the spread of bird flu in the United States.

- 1 What is the point estimate for p , the proportion of U.S. voters who are concerned about the spread of bird flu?
- 2 Construct a 95% CI for p

Margin of Error & Sample Size Calculation

- Margin of error (ME):

$$z_{\alpha/2} \sqrt{\frac{n\hat{p}(1-\hat{p})}{n}}$$

\Rightarrow CI for $p = \hat{p} \pm \text{ME}$

- Sample size determination:

$$n = \frac{\tilde{p}(1-\tilde{p}) \times z_{\alpha/2}^2}{\text{ME}^2},$$

What value of \tilde{p} to use?

- An educated guess
- A value from previous research
- Use a pilot study
- The “most conservative” choice is to use $\tilde{p} = 0.5$

Example

A researcher wants to estimate the proportion of voters who will vote for candidate A. She wants to estimate to within 0.05 with 90% confidence.

- 1 How large a sample does she need if she thinks the true proportion is about .9?
- 2 How large a sample does she need if she thinks the true proportion is about .6?
- 3 How large a sample does she need if she wants to use the most conservative estimate?

- 1 State the null and alternative hypotheses:

$$H_0 : p = p_0 \text{ vs. } H_a : p > \text{ or } \neq \text{ or } < p_0$$

- 2 Compute the test statistic:

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- 3 Make the decision of the test:

Rejection Region/ P-Value Methods

- 4 Draw the conclusion of the test:

We (do/do not) have enough statistical evidence to conclude that (H_a in words) at α significant level.

Among 900 randomly selected registered voters nationwide, 63% of them are somewhat or very concerned about the spread of bird flu in the United States. Conduct a hypothesis test at .01 level to assess the research hypothesis: $p > .6$.

Recap: Inference for p

- Point estimate:

$$\hat{p} = \frac{x}{n}$$

where x is the number of “successes” in a sample with sample size n , and the probability of success, p , is the parameter of interest

- $100(1 - \alpha)\%$ confidence interval:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{(\hat{p})(1 - \hat{p})}{n}}$$

- Hypothesis Testing: $H_0 : p = p_0$ vs. $H_a : p >$ or \neq or $< p_0$

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Under $H_0 : p = p_0$, $z^* \sim N(0, 1)$

Another CI for p : Wilson Score Confidence Interval

- The actual coverage probability of $100(1 - \alpha)\%$ CI

$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}}$ is usually **falls below** $(1 - \alpha)$ 😞

- E.B. Wilson proposed one solution in 1927

Idea: Solving $\frac{p - \hat{p}}{\sqrt{\frac{p(1-p)}{n}}} = \pm z_{\alpha/2}$ for p

$$\Rightarrow (p - \hat{p})^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n}$$

Another CI for p : Wilson Score Confidence Interval

- The actual coverage probability of $100(1 - \alpha)\%$ CI

$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}}$ is usually **falls below** $(1 - \alpha)$ 😞

- E.B. Wilson proposed one solution in 1927

Idea: Solving $\frac{p - \hat{p}}{\sqrt{\frac{p(1-p)}{n}}} = \pm z_{\alpha/2}$ for p

$$\Rightarrow (p - \hat{p})^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n}$$

$100(1 - \alpha)\%$ Wilson Score Confidence Interval:

$$\frac{X + \frac{z_{\alpha/2}^2}{2}}{n + z_{\alpha/2}^2} \pm \frac{z_{\alpha/2}}{n + z_{\alpha/2}^2} \sqrt{\frac{X(n-X)}{n} + \frac{z_{\alpha/2}^2}{4}}$$

Example

Suppose we would like to estimate p , the probability of being vegetarian (for all the CU student). We take a sample with sample size $n = 25$ and none of them are vegetarian (i.e., $x = 0$). Construct a 95% CI for p .

Rule of Three: An Approximate 95% CI for p When $\hat{p} = 0$ or 1

When $\hat{p} = 0$, we have

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 0 \pm z_{\alpha/2} \times 0 = (0, 0)$$

Similarly, when $\hat{p} = 1$, we have

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 1 \pm z_{\alpha/2} \times 0 = (1, 1)$$

These Wald CIs degenerate to a point, which do not reflect the estimation uncertainty. Here we could apply the **rule of three** to approximate 95% CI:

$$(0, 3/n), \quad \text{if } \hat{p} = 0$$

$$(1 - 3/n, 1), \quad \text{if } \hat{p} = 1$$

Comparing Two Population Proportions p_1 and p_2

- We often interested in comparing two groups, e.g., does a particular treatment increase the survival probability for cancer patients ?
- We would like to infer $p_1 - p_2$, the difference between two population proportions \Rightarrow point estimate, interval estimate, hypothesis testing

- Parameters

- p_1, p_2 : population proportions
- $p_1 - p_2$: the difference between two population proportions

- Sample Statistics

- n_1, n_2 : sample sizes
- $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$: sample proportions

$$\Rightarrow \hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

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

- Point estimate:

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

- $100(1 - \alpha)\%$ CI based on CLT:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{(\hat{p}_1)(1 - \hat{p}_1)}{n_1} + \frac{(\hat{p}_2)(1 - \hat{p}_2)}{n_2}}$$

Hypothesis Testing for $p_1 - p_2$

- 1 State the null and alternative hypotheses:

$$H_0 : p_1 - p_2 = 0 \text{ vs. } H_a : p_1 - p_2 > \text{ or } \neq \text{ or } < 0$$

- 2 Compute the test statistic:

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}},$$

where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$

- 3 Make the decision of the test:

Rejection Region/ P-Value Methods

- 4 Draw the conclusion of the test:

We (do/do not) have enough statistical evidence to conclude that (H_a in words) at $\alpha\%$ significant level.

Example

A Simple Random Sample of 100 CU graduate students is taken and it is found that 79 “strongly agree” that they would recommend their current graduate program. A Simple Random Sample of 85 USC graduate students is taken and it is found that 52 “strongly agree” that they would recommend their current graduate program. At 5 % level, can we conclude that the proportion of “strongly agree” is higher at CU?

In this lecture, we learned statistical inference for population proportion p :

- Point estimate
- Interval estimate
- Hypothesis testing

In next lecture we will learn statistical inference for multi-category data and bivariate categorical data