

Lecture 30

Inference for Proportions I

STAT 8010 Statistical Methods I November 6, 2019

> Whitney Huang Clemson University

Inference for Categorical Data



In the next few lectures we will focus on categorical data analysis:

- Inference for a single proportion p
- Comparison of two proportions p_1 and p_2
- χ^2 tests: Inference for Multi-category data and contingency tables

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 adminis-

Dichotomous (two-category) outcomes: "success" & "failure"

tered this treatment.

 Similar to the inferential problem for μ, the population mean, we would like to infer p, the population proportion of success ⇒ point estimate, interval estimate, hypothesis testing



Point/Interval Estimation



Point estimate:

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

Recall the Binomial random variable, we have $\mathbb{E}[X] = np$ where $X \sim \text{Bin}(n,p) \Rightarrow \mathbb{E}[\hat{p}] = \mathbb{E}[\frac{X}{n}] = \frac{1}{n}\mathbb{E}[X] = p$

• $100(1-\alpha)\%$ CI:

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

Why?

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

Motivated Example Revisited



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 who would survive at least 5 years after being administered this treatment.
- Construct a 95% CI for p

Another Example



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.

- What is the point estimate for p (Proportion of U.S. voters who are concerned about the spread of bird flu?
- Construct a 99% CI for p
- Is it reasonable to conclude that *p* is .600? in the United States)

Margin of error & Sample Size Calculation



Margin of error:

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

$$\Rightarrow$$
 CI for $p = \hat{p} \pm$ margin of error

Sample size determination:

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

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.

- How large a sample does she need if she thinks the true proportion is about .9?
- Output
 <p
- How large a sample does she need if she wants to use the most conservative estimate?

Hypothesis Testing for p



State the null and alternative hypotheses:

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

Ompute the test statistic:

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

Make the decision of the test:

Rejection Region/ P-Value Methods

Oraw the conclusion of the test:

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

Bird Flu Example Revisited



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 if p > .667.

Another CI for p: Wilson Score Confidence Interval



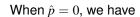


Idea: Solving
$$p=\hat{p}\pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Rightarrow (p-\hat{p})^2=z_{\alpha/2}^2\frac{p(1-p)}{n}$$

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

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





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

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

$$(0,3/n),$$
 if $\hat{p} = 0$
 $(1-3/n,1),$ if $\hat{p} = 1$

Summary



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 two proportions p_1 and p_2