

Handout 09-Inference for Proportions

Jin, Ick Hoon

July 29th 2009

- The same principles used for the confidence interval for the mean are used for the confidence interval of the population proportion.
- Here we want to obtain a plausible range of values for the population proportion, π .
 - Keep in mind, π , should have a value between 0 and 1
- Previously, we used p as an estimate of π , so that initially we might consider p when trying to construct confidence intervals for π .
- However, using p can lead to confidence intervals which contain values outside of 0 and 1
 - Why would this be a problem?

- Recall that **sample proportion** is:

$$p = \frac{X}{n},$$

where X is the counts and n is the sample size. The **standard error** of p is:

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

and the **margin of error** of confidence level C is: $m = z^* SE_p$ and an **approximate level C confidence interval** for p is: $p \pm m$.

- Thus, we can rewrite the confidence interval for a population proportion explicitly in this form:

$$p \pm z^* \sqrt{\frac{p(1-p)}{n}}$$

This formula should be used when both $np \geq 15$ and $n(1-p) \geq 15$.

Radio Example

- A recent fire in a warehouse that contained 100,000 radios damaged an unknown number of the radios. A freight broker who purchases damaged goods offers to purchase the entire contents from the insurance company that provides coverage for the warehouse. The freight broker will eventually sort through all the radios and sell those that are not damaged. Before the broker makes an offer to the insurance company, he would like to know what proportion of the 100,000 radios are damaged and cannot be sold (from Graybill, Iyer and Burdick, *Applied Statistics*, 1998).
- Find a 99% confidence interval for the proportion of the 100,000 radios which are damaged.

Radio Example Suppose a random sample of 200 radios is taken from the warehouse and 34 of them were damaged.

- We first obtain p : $p = \frac{X}{n} = \frac{34}{200} = 0.17$.
- Next we use the formula

$$\begin{aligned} p \pm z^* \sqrt{\frac{p(1-p)}{n}} &= 0.17 \pm 2.576 \sqrt{\frac{0.17(1-0.17)}{200}} \\ &= (0.101, 0.239) \end{aligned}$$

This is a 99% Confidence Interval for the true proportion.

- For one single proportion, we want to test $H_0 : \pi = \pi_0$, v.s. $H_a : \pi \neq \pi_0$ or $\pi > \pi_0$ or $\pi < \pi_0$.
- The first thing recommended is to check the following **rules of thumb**:
 - $n\pi_0 \geq 10$
 - $n(1 - \pi_0) \geq 10$
 - where n is the sample size and π_0 is the hypothesized proportion.

- 1 State the null hypothesis: $H_0: \pi = \pi_0$
- 2 State the alternative hypothesis: $H_a: \pi \neq \pi_0$ or $\pi > \pi_0$ or $\pi < \pi_0$
- 3 State the level of significance.
 - RECALL: we assume $\alpha = 0.05$ unless otherwise stated.
- 4 Calculate the test statistic (z statistics):

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

5 Find the P-Value

- For a Two-sided Test: $H_a: \pi \neq \pi_0$

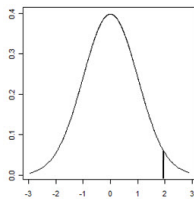
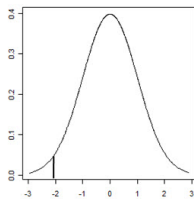
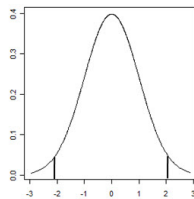
$$p - \text{Value} = P(Z \geq |z| \text{ or } Z \leq -|z|) = 2P(Z \geq |z|)$$

- For a Two-sided Test: $H_a: \pi > \pi_0$

$$p - \text{Value} = P(Z \geq z)$$

- For a Two-sided Test: $H_a: \pi < \pi_0$

$$p - \text{Value} = P(Z \leq z)$$

 $P(Z > z)$  $P(Z < z)$  $2P(Z > |z|)$

- 6 Reject or fail to reject H_0 based on the P-value.
 - If the P-value is less than or equal to α , reject H_0 .
 - If the P-value is greater than α , fail to reject H_0 .

- 7 State your conclusion.
 - Your conclusion should reflect your original statement of the hypotheses.
 - Furthermore, your conclusion should be stated in terms of the alternative hypotheses, e.g.
 - If H_0 is rejected, “there is significant statistical evidence that the population proportion is different than π_0 .”
 - If H_0 is not rejected, “there is not significant statistical evidence that the population mean is different than π_0 .”

Orange Trees Example

- The owner of an orange grove wants to determine if the proportion of diseased trees in the grove is more than 10%. He will use this information to determine if it will be cost effective to spray the entire grove. The owner would like to know the exact value of π , but he realizes that he cannot know the exact value unless he examines every one of the 6,010 trees, which would be too expensive. He decides to take a simple random sample of 150 trees and examine them for the disease. He finds that 12 of the 150 trees are diseased. (adapted from Graybill, Iyer and Burdick, *Applied Statistics*, 1998).

Orange Trees Example Information we have:

- $n = 150, X = 12, \pi_0 = 0.10$
- $p = \frac{X}{n} = \frac{12}{150} = 0.08.$
- We first check the rules of thumb.
 - $n\pi_0 = 150(0.10) = 15 > 10$
 - $n(1 - \pi_0) = 150(1 - 0.10) = 135 > 10$
- The assumptions for the test are approximately met.

Orange Trees Example

- State the null hypothesis: $H_0: \pi = 0.10$
- State the alternative hypothesis: $H_a: \pi > 0.10$
- State the level of significance: assume $\alpha = 0.05$.
- Calculate the test statistic

$$\begin{aligned} z &= \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.08 - 0.10}{\sqrt{\frac{0.10(1-0.10)}{150}}} \\ &= -0.82 \end{aligned}$$

Orange Trees Example

- Find the p-Value

$$P\text{-Value} = P(Z \geq z) = P(Z \geq -0.82) = 1 - P(Z < -0.82) = 0.7939$$

- Do we reject or fail to reject H_0 based on the P-value?

P-value = 0.7939 is greater than $\alpha = 0.05$.

- State the conclusion.

We fail to reject H_0 and conclude that: "There is not significant statistical evidence that the true proportion of diseased orange trees is greater than 10%."

Immunization Shots Example

- The superintendent of a large school district wants to know if the proportion of first graders in her district that have received their immunization shots is different from last year. Last year, 74% of the first grade children had received their immunization shots. The superintendent randomly selects 100 first grade students and 77 of them have received their immunization shots.

Immunization Shots Example Information we have:

- $n = 100, X = 77, \pi_0 = 0.74$
- $p = \frac{X}{n} = \frac{77}{100} = 0.77.$
- We first check the rule of thumb.
 - $n\pi_0 = 100(0.74) = 74 > 10$
 - $n(1 - \pi_0) = 100(1 - 0.74) = 26 > 10$
- The assumptions for the test are approximately met.

Immunization Shots Example

- State the null hypothesis: $H_0: \pi = 0.74$
- State the alternative hypothesis: $H_a: \pi \neq 0.74$
- State the level of significance: assume $\alpha = 0.05$.
- Calculate the test statistic

$$\begin{aligned} z &= \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.77 - 0.74}{\sqrt{\frac{0.74(1-0.74)}{100}}} \\ &= 0.68 \end{aligned}$$

Immunization Shots Example

- Find the p-Value

$$\begin{aligned} P - \text{Value} &= 2P(Z \geq |z|) = 2P(Z \geq |0.68|) = 2(1 - P(Z < 0.68)) \\ &= 0.4966 \end{aligned}$$

- Do we reject or fail to reject H_0 based on the P-value?

P-value = 0.4966 is greater than $\alpha = 0.05$.

- State the conclusion.

We fail to reject H_0 and conclude that: “There is not significant statistical evidence that the true proportion of first grade students with immunization shots is different than last year’s proportion of 0.74”

- To get a desired margin of error (m) by adjusting the sample size n , we need:
 - Determine the desired margin of error (m).
 - Use the following formula:

$$n = \left[\left(\frac{z^*}{m} \right)^2 p^*(1 - p^*) \right]$$

where p^* is a guessed value for the proportion of successes in the future sample. z^* is the critical value corresponding to the confidence level C . (e.g. $z^* = 1.96$ for $\alpha = 0.05$.)

- If we have no idea what p^* could be, we choose $p^* = 0.5$ to get a conservative sample size.

Radio Example (Revisited)

- A confidence interval of (0.101, 0.239) is not very narrow.
- How large of a sample size would be needed to have a margin of error (m) of 0.01?

$$\begin{aligned}n &= \left[\left(\frac{z^*}{m} \right)^2 p^*(1 - p^*) \right] = \left[\left(\frac{2.576}{0.01} \right)^2 0.17(1 - 0.17) \right] \\ &= 9363.08\end{aligned}$$

- We can use 0.17 as a good guess since it is estimated from the data.
- The freight broker should use a sample of size 9,364 (always round UP in sample size calculation) to achieve a margin of error of 0.01.

- “A national survey is conducted to compare the percentage of clergy favoring the ordination of women to the priesthood to the percentage of nonclergy in favor of such a move.” (from Milton, McTeer, and Corbet, *Introduction to Statistics*, 1997)
- Comparing high school drop-out rates for two races or for gender.
- Comparing students 'to non-students' views on a particular policy concerning parking

- The confidence interval is for the difference of population proportions ($\pi_1 - \pi_2$) for two groups.
- An approximate confidence interval for $\pi_1 - \pi_2$ is

$$p_1 - p_2 \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where $p_1 = \frac{X_1}{n_1}$ and $p_2 = \frac{X_2}{n_2}$

Flu Shots Example In an experiment designed to study the effect of fear, psychologists performed the following experiment: 400 randomly selected students were randomly divided into two groups of 200 each. Each group was urged to get flu shots. Group I was shown slides and given a gory verbal description of the effects of flu. The presentation was made in such a way that it was extremely frightening. Group II was simply given a brochure describing the disease, and no attempt was made to induce fear in the subjects. Of the 200 subjects in group I, 44 elected to receive the vaccine, whereas 39 of those in group II did so. (from Milton, McTeer, and Corbet, *Introduction to Statistics*, 1997)

- Find a 95% confidence interval for $\pi_1 - \pi_2$

Flu Shots Example

- **Sample Proportions:**

$$p_1 = \frac{X_1}{n_1} = \frac{44}{200} = 0.22 \quad p_2 = \frac{X_2}{n_2} = \frac{38}{200} = 0.19$$

- **Confidence Intervals**

$$\begin{aligned} p_1 - p_2 \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\ = .22 - .19 \pm 1.96 \sqrt{\frac{0.22(1-0.22)}{200} + \frac{0.19(1-0.19)}{200}} \\ = (-.049, .109) \end{aligned}$$

- **Significance Test:** The test statistic for comparing two proportions is

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

where $p_1 = \frac{X_1}{n_1}$, $p_2 = \frac{X_2}{n_2}$, and $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ is called the **pooled estimate** because it pools the information from both samples.

- The null hypothesis can be any of the following:

$$H_0 : \pi_1 = \pi_2, \quad H_0 : \pi_1 \leq \pi_2, \quad \text{or} \quad H_0 : \pi_1 \geq \pi_2$$

- The alternative hypothesis can be any of the following (depending on the question being asked):

$$H_a : \pi_1 \neq \pi_2, \quad H_a : \pi_1 > \pi_2, \quad \text{or} \quad H_a : \pi_1 < \pi_2$$

The other steps are the same as those used for the tests we have looked at previously.

Discipline Example

- “A child protection agency conducted a poll of adults in the state and asked them the question, ‘It is a hard spanking sometimes necessary to discipline a child?’ The agency wants to determine whether or not there is a difference between the proportion of men who believe that a hard spanking is sometimes necessary and the proportion of women who have this belief.” Out of 571 men asked the question, 421 responded ‘Yes’. Of 611 women asked, 367 answered ‘Yes’ (from Graybill, Iyer and Burdick, *Applied Statistics*, 1998)
- Perform a test to see if there is any difference between the true proportions of men and women at a 0.01 significance level.

Discipline Example Information we have:

- Sample Size: $n_1 = 571$, $n_2 = 611$; $X_1 = 421$, $X_2 = 367$.
- $p_1 = \frac{X_1}{n_1} = \frac{421}{571} = 0.737$, $p_2 = \frac{X_2}{n_2} = \frac{367}{611} = 0.601$.
- $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{421 + 367}{571 + 611} = 0.667$

- State the null hypothesis: $H_0: \pi_1 = \pi_2$
- State the alternative hypothesis: $H_a: \pi \neq \pi_2$
- State the level of significance: assume $\alpha = 0.01$.
- Calculate the test statistic

$$\begin{aligned} z &= \frac{p_1 - p_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{0.737 - 0.601}{\sqrt{.667(1 - .667) \left(\frac{1}{571} + \frac{1}{611} \right)}} \\ &= 4.96 \end{aligned}$$

- Find the p-Value

$$\begin{aligned} P\text{-Value} &= 2P(Z \geq |z|) = 2P(Z \geq 4.96) = 2(1 - P(Z < 4.96)) \\ &\approx 0.0000 \end{aligned}$$

- Do we reject or fail to reject H_0 based on the P-value?

P-value ≈ 0.0000 is smaller than $\alpha = 0.01$.

- State the conclusion.

We reject H_0 and conclude: "There is significant statistical evidence (at the 0.01 level of significance) that there is a difference between the proportion of men and that of women who believe that a hard spanking is sometimes necessary."