BIOS:4120 – Introduction to Biostatistics
Unit 11: Inference on Proportions

Knute D. Carter

Department of Biostatistics
The University of Iowa

October 28, 2025

Overview

- Sampling Distribution of a Proportion
- Confidence Intervals
- Hypothesis Testing
- Sample Size Estimation
- Comparison of Two Proportions

Sampling Distribution of a Proportion

- Consider a population in which each object can be classified as a ‘success’ or a ‘failure’ based on a binary variable of interest.
- Suppose that a sample of size n is drawn from this population.
- Let X count the number of successes.
- Then X is a binomial random variable: $X \sim Bin(n, p)$.

Sampling Distribution of a Proportion

- Let $\hat{p} = X/n$ denote the sample proportion of successes.
- The sample proportion $\hat{p}$ serves as an estimator of the population proportion p.
- If one views the population data and the sample data as consisting of zeros (for the failures) and ones (for the successes), then one can interpret p as a population mean and $\hat{p}$ as a sample mean.

Sampling Distribution of a Proportion

Properties of the sampling distribution of $\hat{p}$:

- The mean (or expected value) of the statistic $\hat{p}$ is the same as the mean of the sampled population, p.
- The standard deviation of the statistic $\hat{p}$ is given by $\sqrt{p(1-p)/n}$, where n is the sample size.

These are general properties of the sampling distribution of $\hat{p}$ that hold for any sample size n.

Sampling Distribution of a Proportion

- The quantity $\sqrt{p(1-p)/n}$ is often called the *Standard Error of the Proportion*.
- The magnitude of this quantity reflects the accuracy of the sample proportion as an estimator of the population proportion.
- Central Limit Theorem for Proportions: If the sample size is 'large,' the sampling distribution of $\hat{p}$ is approximately normal.
- A 'large' sample is generally considered to be one where $np \geq 5$ and $n(1-p) \geq 5$.

Confidence Intervals

- If n is 'large,' we have

$$\begin{aligned}
P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq +z_{\alpha/2}\right) &= (1-\alpha) \\
&\vdots \\
P\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p}+z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) &= (1-\alpha)
\end{aligned}$$

Confidence Intervals

- The preceding suggests a confidence formula based on

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

- However, since p is unknown, we must replace p by $\hat{p}$ in the standard error. We then obtain

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

- The formula is valid provided that n is 'large.'

- We consider n to be 'large' if $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$.

Confidence Intervals

- The quantity

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

  is called the *Margin of Error*.

- Again, the general form of a two-sided confidence interval is as follows:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

Confidence Intervals

- A level $100(1-\alpha)\%$ *One-Sided Confidence Interval for p* is given by:

$$\left(0,\ \ \hat{p}+z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \qquad \text{Upper Bound}$$

$$\left(\hat{p}-z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \ 1\right) \qquad \text{Lower Bound}$$

Hypothesis Testing

Steps for a z-test on a Population Proportion

1. Label and describe the parameter of interest.
2. State the null hypothesis H_0 symbolically: $p = p_0$.
   State the alternative hypothesis H_A symbolically: $p \neq p_0$, $p < p_0$, $p > p_0$.
3. Select a value for α.

Hypothesis Testing

4. Specify the *Test Statistic* to be used. For the z-test, the test statistic is:

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

   This test statistic can be used provided that $np_0 \geq 5$ and $n(1 - p_0) \geq 5$.

5. Compute the numerical value of the test statistic.

Hypothesis Testing

6. Compute the *p*-value for the test using Table A.3.

H_A	*p*-value
$p \neq p_0$	$2P(Z > \|z\|)$
$p > p_0$	$P(Z > z)$
$p < p_0$	$P(Z < z)$

Hypothesis Testing

7. Arrive at a conclusion by either:
   (1) comparing the *p*-value to α; or
   (2) determining whether the test statistic falls into the rejection region.

H_A	Rejection Region
$p \neq p_0$	$z < -z_{\alpha/2}$ or $z > +z_{\alpha/2}$
$p > p_0$	$z > +z_\alpha$
$p < p_0$	$z < -z_\alpha$

8. State the conclusion: whether or not H_0 should be rejected.

Sample Size for Confidence Intervals

- Problem: Suppose we want to find the sample size n needed so that the margin of error does not exceed some bound m.
- Accordingly, we want to find the sample size n needed so that the width of a two-sided confidence interval does not exceed $2m$.
- We need to find the n that will ensure

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

Sample Size for Confidence Intervals

- Solution: Take

$$n \geq \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{m} \right)^2$$

- Note, however, that we would not know $\hat{p}$ at the time when we would be attempting to determine the desired sample size.

- The value of $\hat{p}$ could be based on a previous study or a 'pilot' study, or on an 'educated guess.'

Sample Size for Confidence Intervals

- However, a common approach is to use $\hat{p} = 0.5$.
- Rationale: the function $\sqrt{\hat{p}(1-\hat{p})}$ is maximized at $\hat{p} = 0.5$.
- Thus, using $\hat{p} = 0.5$ in the preceding formula results in a sample size that is at least as large (and probably larger) than the sample size required to meet our objective.
- Solution (based on $\hat{p} = 0.5$): Take

$$n \geq \left(\frac{z_{\alpha/2} \times 0.5}{m} \right)^2$$

Sample Size for Hypothesis Testing

- Problem: Suppose we are planning a hypothesis test for a population proportion p using a level of significance α.
- Let p_A denote a value of p represented under the alternative hypothesis, such that the effect size $|p_A - p_0|$ represents a clinically important difference.
- What sample size would be required in order for the power of the test to be at least $(1 - \beta)$?
- Note: The book uses p_1 in place of p_A.

Sample Size for Hypothesis Testing

To complete the sample size calculation, the following quantities are required:

- α and β;
- for a two-sided test, the critical value $z_{\alpha/2}$;
  for a one-sided test, the critical value z_{α};
- the critical value z_{β}; and
- the hypothesized proportions p_0 and p_A.

Sample Size for Hypothesis Testing

- For a two-sided test, take

$$n \geq \left(\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_A(1-p_A)}}{|p_A - p_0|} \right)^2$$

- For a one-sided test, take

$$n \geq \left(\frac{z_{\alpha}\sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_A(1-p_A)}}{|p_A - p_0|} \right)^2$$

Sample Size: Example

- A paper published in *Cancer* in 1991 considered the five-year survival for individuals under the age of 40 who have been diagnosed with lung cancer.
- In a random sample of 52 individuals, only 6 survived five years (Jubelirer and Wilson, 1991).

a) Using the preceding data, find a 95% two-sided confidence interval for the true proportion of individuals diagnosed with lung cancer who survive 5 years.

Sample Size: Example

Sample Size: Example

b) Is the sample size large enough to justify the use of the z formula?

c) How large of a sample would be required to obtain a 95% confidence interval having a width no greater than 0.10? Compute the required sample size in two ways. First, consider the reported results as being based on a pilot study, and use the sample proportion as $\hat{p}$. Second, use 0.5 as $\hat{p}$ to obtain a conservative sample size estimate.

Sample Size: Example

Comparison of Two Proportions

- The inferential procedures we will discuss in this section require independent samples.

- Let p_1 and p_2 denote the proportions for the two populations.

- Suppose that a sample of size n_1 is drawn from the first population, and a sample of size n_2 is drawn from the second population.

- Let $\hat{p}_1$ and $\hat{p}_2$ denote the sample proportions for the two samples.

Comparison of Two Proportions

Steps for a z-test on a Population Proportions (Independent Samples)

1. Label and describe the parameters of interest.
2. State the null hypothesis H_0 symbolically: $p_1 = p_2$.
   State the alternative hypothesis H_A symbolically: $p_1 \neq p_2$, $p_1 < p_2$, $p_1 > p_2$.
3. Select a value for α.

Comparison of Two Proportions

4. Specify the *Test Statistic* to be used. For the z-test, the test statistic is:

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

In the preceding, $\hat{p}$ denotes the proportion of 'successes' in both samples combined:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Note: This z-test statistic can be used when both sample sizes are 'large.' The sample sizes are considered 'large' if all of the following inequalities hold:

$$n_1\hat{p}_1 \geq 5, \;\; n_1(1-\hat{p}_1) \geq 5, \;\; n_2\hat{p}_2 \geq 5, \;\; n_2(1-\hat{p}_2) \geq 5$$

Comparison of Two Proportions

5. Compute the numerical value of the test statistic.

6. Compute the *p*-value for the test using Table A.3.

H_A	*p*-value
$p_1 \neq p_2$	$2P(Z > \|z\|)$
$p_1 > p_2$	$P(Z > z)$
$p_1 < p_2$	$P(Z < z)$

Comparison of Two Proportions

7. Arrive at a conclusion by either:
   (1) comparing the *p*-value to α; or
   (2) determining whether the test statistic falls into the rejection region.

H_A	Rejection Region
$p_1 \neq p_2$	$z < -z_{\alpha/2}$ or $z > +z_{\alpha/2}$
$p_1 > p_2$	$z > +z_{\alpha}$
$p_1 < p_2$	$z < -z_{\alpha}$

8. State the conclusion: whether or not H_0 should be rejected.

z Confidence Interval Procedures (Independent Samples)

- A level $100(1-\alpha)\%$ *Two-Sided Confidence Interval for* $(p_1 - p_2)$ based on the Z distribution is given by

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

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

Comparison of Two Proportions

- A level $100(1-\alpha)\%$ *One-Sided Confidence Interval for* $(p_1 - p_2)$ based on the Z distribution is given by

  Lower bound:

$$\left((\hat{p}_1 - \hat{p}_2) - z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ \ 1\right)$$

  Upper bound:

$$\left(-1,\ \ (\hat{p}_1 - \hat{p}_2) + z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

Example: Salk Polio Vaccine Study

- The polio epidemic first hit the U.S. in 1916. During the next 40 years, the disease claimed tens of thousands of lives. The majority of those who died of polio were children.

- By the 1950's, several vaccines against the disease had been proposed. The one which showed the most promise was the one developed by Jonas Salk.

- In 1954, the Public Health Service (PHS) decided to organize a large-scale clinical trial to test the vaccine. The PHS wanted to gear the study towards children, so it selected a random sample of nearly one million first, second, and third graders, and contacted their parents for permission to include the children in the trial.

Example: Salk Polio Vaccine Study

- About 600,000 children had parents who refused the PHS. The remaining 400,000 children were randomly divided into a *treatment group* and a *control group*.
- The treatment group was given the vaccine and the control group was given a placebo (which consisted of a harmless salt-water injection).
- The study was run in a double-blind manner: neither the children nor the doctors who monitored the children knew who received the vaccine or who received the placebo.

Example: Salk Polio Vaccine Study

- A year later, the prevalence of polio was measured in each group.
- The results are featured in the table which follows.

Group	*n*	Cases
Treatment	200,000	57
Control	200,000	142
Total	400,000	199

- Do the results indicate that the vaccine is effective in reducing the risk of contracting polio? Test using $\alpha = 0.001$.

Example: Salk Polio Vaccine Study