STA 2023

Module 10
Comparing Two Proportions

Learning Objectives

Upon completing this module, you should be able to:

1. Perform large-sample inferences (hypothesis test and confidence intervals) to compare two population proportions.
2. Describe the relationship between the sample sizes, confidence level, and margin of error for a confidence interval for the difference between two population proportions.
3. Determine the sample size required for a specified confidence level and margin of error for the estimate of the difference between two population proportions.

Population Proportions

In this module, we are going to learn how to compare two population proportions.

Remember, a population proportion, p is simply the percentage of a population that has a specified attribute.

Quick Review on Population Proportion and Sample Proportion

Population Proportion and Sample Proportion

Consider a population in which each member either has or does not have a specified attribute. Then we use the following notation and terminology.

Population proportion, p: The proportion (percentage) of the entire population that has the specified attribute.

Sample proportion, $\hat{p}$: The proportion (percentage) of a sample from the population that has the specified attribute.

In short, a sample proportion is obtained by dividing the number of members sampled that have the specified attribute (x) by the total number of members sampled (n).

Sometimes, we refer to "x" as the number of successes and "n-x" as the number o failures.

Quick Review on One-Proportion *z*-Interval

- When the conditions are met, we are ready to find the confidence interval for the population proportion, *p*.
- The confidence interval is $\hat{p} \pm z^* \times SE(\hat{p})$

  where $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

- The critical value, z^*, depends on the particular confidence level, C, that you specify.

Comparing Two Proportions

- Comparisons between two percentages are much more common than questions about isolated percentages. And they are more interesting.
- We often want to know how two groups differ, whether a treatment is better than a placebo control, or whether this year's results are better than last year's.

Another Ruler

- In order to examine the difference between two proportions, we need another ruler—the standard deviation of the sampling distribution model for the difference between two proportions.
- Recall that standard deviations don't add, but variances do. In fact, the variance of the sum or difference of two independent random variables is the sum of their individual variances.

The Standard Deviation of the Difference Between Two Proportions

- Proportions observed in independent random samples *are* independent. Thus, we can add their variances. So…
- The standard deviation of the difference between two sample proportions is

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

- Thus, the standard error is

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

Assumptions and Conditions

- Independence Assumptions:
  - Randomization Condition: The data in each group should be drawn independently and at random from a homogeneous population or generated by a randomized comparative experiment.
  - The 10% Condition: If the data are sampled without replacement, the sample should not exceed 10% of the population.
  - Independent Groups Assumption: The two groups we're comparing must be independent *of each other*.

Assumptions and Conditions (cont.)

- Sample Size Condition:
  - *Each* of the groups must be big enough…
  - Success/Failure Condition: Both groups are big enough that both successes and failures are at least 5 have been observed in each.

The Sampling Distribution

- We already know that for large enough samples, each of our proportions has an approximately Normal sampling distribution.
- The same is true of their difference.

The Sampling Distribution (cont.)

- Provided that the sampled values are independent, the samples are independent, and the samples sizes are large enough, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is modeled by a Normal model with
  - Mean: $\mu = p_1 - p_2$
  - Standard deviation:

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Two-Proportion *z*-Interval

- When the conditions are met, we are ready to find the confidence interval for the difference of two proportions:
- The confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

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

- The critical value z^* depends on the particular confidence level, C, that you specify.

Pool or Not Pool?

- The typical hypothesis test for the difference in two proportions is the one of no difference. In symbols, H_0: $p_1 - p_2 = 0$.
- Since we are hypothesizing that there is no difference between the two proportions, that means that the standard deviations for each proportion are the same.
- Since this is the case, we combine (pool) the counts to get one overall proportion.

What is the Pooled Proportion?

- The pooled proportion is

$$\hat{p}_{pooled} = \frac{Success_1 + Success_2}{n_1 + n_2}$$

where $Success_1 = n_1 \hat{p}_1$ and $Success_2 = n_2 \hat{p}_2$

  - If the numbers of successes are not whole numbers, round them first. (This is the *only* time you should round values in the middle of a calculation.)

What is the Pooled Proportion? (Cont.)

- We then put this pooled value into the formula, substituting it for *both* sample proportions in the standard error formula:

$$SE_{pooled}\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}}$$

Compared to What?

- We'll reject our null hypothesis if we see a large enough difference in the two proportions.
- How can we decide whether the difference we see is large?
  - Just compare it with its standard deviation.
- Unlike previous hypothesis testing situations, the null hypothesis doesn't provide a standard deviation, so we'll use a standard error (here, pooled).

Two-Proportion *z*-Test

- The conditions for the two-proportion *z*-test are the same as for the two-proportion *z*-interval.
- We are testing the hypothesis H_0: $p_1 = p_2$.
- Because we hypothesize that the proportions are equal, we pool them to find

$$\hat{p}_{pooled} = \frac{Success_1 + Success_2}{n_1 + n_2}$$

Two-Proportion *z*-Test (cont.)

- We use the pooled value to estimate the standard error:

$$SE_{pooled}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}}$$

- Now we find the test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{pooled}(\hat{p}_1 - \hat{p}_2)}$$

- When the conditions are met and the null hypothesis is true, this statistic follows the standard Normal model, so we can use that model to obtain a P-value.

Quick Review

Let's look at the following one more time:

- How to find one proportion z-interval?
- How to perform a one proportion z-test?
- What is the sampling distribution of the difference between two sample proportions?
- How to perform a two proportion z-test?
- How to find a two proportion z-interval?

Let's review How to Construct a One Proportion *z*-interval?

One-Proportion *z*-Interval Procedure

Purpose To find a confidence interval for a population proportion, p

Assumptions

1. Simple random sample
2. The number of successes, x, and the number of failures, $n - x$, are both 5 or greater.

STEP 1 **For a confidence level of $1 - \alpha$, use Table II to find $z_{\alpha/2}$.**

STEP 2 **The confidence interval for p is from**

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n} \quad \text{to} \quad \hat{p} + z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$$

where $z_{\alpha/2}$ is found in Step 1, n is the sample size, and $\hat{p} = x/n$ is the sample proportion.

STEP 3 **Interpret the confidence interval.**

A Quick Review on How to Perform a One Proportion *z*-Test?

One-Proportion *z*-Test

Purpose To perform a hypothesis test for a population proportion, p

Assumptions

1. Simple random sample
2. Both np_0 and $n(1 - p_0)$ are 5 or greater

STEP 1 **The null hypothesis is H_0: $p = p_0$, and the alternative hypothesis is**

$$H_a: p \neq p_0 \text{ (Two tailed)} \quad \text{or} \quad H_a: p < p_0 \text{ (Left tailed)} \quad \text{or} \quad H_a: p > p_0 \text{ (Right tailed)}$$

STEP 2 **Decide on the significance level, α.**

STEP 3 **Compute the value of the test statistic**

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

and denote that value z_0.

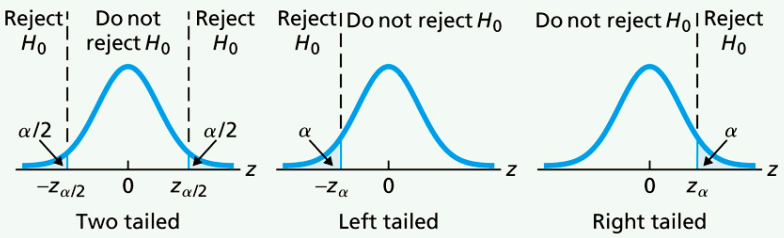
How to Perform a One Proportion *z*-Test ? (Cont.)

CRITICAL-VALUE APPROACH or **P-VALUE APPROACH**

STEP 4 The critical value(s) are

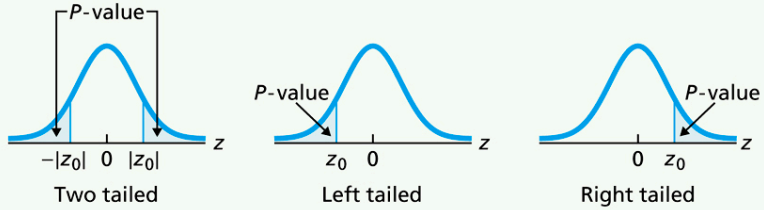
$\pm z_{\alpha/2}$ (Two tailed) or $-z_{\alpha}$ (Left tailed) or z_{α} (Right tailed)

Use Table II to find the critical value(s).



STEP 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 4 Use Table II to obtain the *P*-value.



STEP 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

STEP 6 Interpret the results of the hypothesis test.

The Sampling Distribution of the Difference Between Two Sample Proportions

The Sampling Distribution of the Difference Between Two Sample Proportions for Independent Samples

For independent samples of sizes n_1 and n_2 from the two populations,

- $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$,
- $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$, and
- $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed for large n_1 and n_2.

What does it mean?

For large independent samples, the possible differences between two sample proportions have approximately a normal distribution with mean $p_1 - p_2$ and standard deviation as above.

How to Perform a Two Proportion *z*-Test?

Two-Proportions *z*-Test

Purpose To perform a hypothesis test to compare two population proportions, p_1 and p_2

Assumptions

1. Simple random samples
2. Independent samples
3. x_1, $n_1 - x_1$, x_2, and $n_2 - x_2$ are all 5 or greater

STEP 1 **The null hypothesis is H_0: $p_1 = p_2$, and the alternative hypothesis is**

$$H_a: p_1 \neq p_2 \text{ (Two tailed)} \quad \text{or} \quad H_a: p_1 < p_2 \text{ (Left tailed)} \quad \text{or} \quad H_a: p_1 > p_2 \text{ (Right tailed)}$$

STEP 2 **Decide on the significance level, α.**

STEP 3 **Compute the value of the test statistic**

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

where $\hat{p}_p = (x_1 + x_2)/(n_1 + n_2)$. Denote the value of the test statistic z_0.

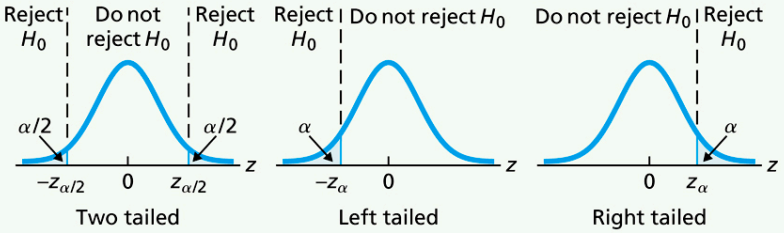
How to Perform a Two Proportion *z*-Test? (Cont.)

CRITICAL-VALUE APPROACH or ***P*-VALUE APPROACH**

STEP 4 The critical value(s) are

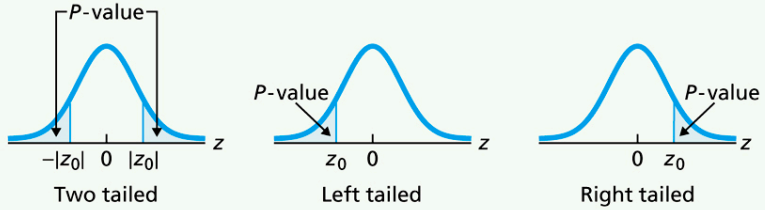
$\pm z_{\alpha/2}$ (Two tailed) or $-z_{\alpha}$ (Left tailed) or z_{α} (Right tailed)

Use Table II to find the critical value(s).



STEP 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 4 Use Table II to obtain the *P*-value.



STEP 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

STEP 6 Interpret the results of the hypothesis test.

How to Perform a Two Proportion *z*-Interval?

Two-Proportions *z*-Interval Procedure

Purpose To find a confidence interval for the difference between two population proportions, p_1 and p_2

Assumptions

1. Simple random samples
2. Independent samples
3. x_1, $n_1 - x_1$, x_2, and $n_2 - x_2$ are all 5 or greater

STEP 1 **For a confidence level of $1 - \alpha$, use Table II to find $z_{\alpha/2}$.**

STEP 2 **The endpoints of the confidence interval for $p_1 - p_2$ are**

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

STEP 3 **Interpret the confidence interval.**

What is the Margin of Error for the Estimate of p_1 - p_2?

Margin of Error for the Estimate of $p_1 - p_2$

The margin of error for the estimate of $p_1 - p_2$ is

$$E = z_{\alpha/2} \cdot \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}.$$

What does it mean?

The margin of error equals half the length of the confidence interval. It represents the precision with which the difference between the sample proportions estimates the difference between the population proportions at the specified confidence level.

How to Find the Sample Size for Estimating $p_1 - p_2$?

Sample Size for Estimating $p_1 - p_2$

A $(1 - \alpha)$-level confidence interval for the difference between two population proportions that has a margin of error of at most E can be obtained by choosing

$$n_1 = n_2 = 0.5\left(\frac{z_{\alpha/2}}{E}\right)^2$$

rounded up to the nearest whole number. If you can make educated guesses, $\hat{p}_{1g}$ and $\hat{p}_{2g}$, for the observed values of $\hat{p}_1$ and $\hat{p}_2$, you should instead choose

$$n_1 = n_2 = \left(\hat{p}_{1g}(1 - \hat{p}_{1g}) + \hat{p}_{2g}(1 - \hat{p}_{2g})\right)\left(\frac{z_{\alpha/2}}{E}\right)^2$$

rounded up to the nearest whole number.

What Can Go Wrong?

Don't Misstate What the Interval Means:

- Don't suggest that the parameter varies.
- Don't claim that other samples will agree with yours.
- Don't be certain about the parameter.
- Don't forget: It's the parameter (not the statistic).
- Don't claim to know too much.
- Do take responsibility (for the uncertainty).

What Can Go Wrong? (cont.)

Margin of Error Too Large to Be Useful:

- We can't be exact, but how precise do we need to be?
- One way to make the margin of error smaller is to reduce your level of confidence. (That may not be a useful solution.)
- You need to think about your margin of error when you design your study.
  - To get a narrower interval without giving up confidence, you need to have less variability.
  - You can do this with a larger sample…

What Can Go Wrong? (cont.)

Violations of Assumptions:

- Watch out for biased sampling.
- Think about independence.

What Can Go Wrong? (cont.)

- Don't base your null hypothesis on what you see in the data.
  - *Think* about the situation you are investigating and develop your null hypothesis appropriately.
- Don't base your alternative hypothesis on the data, either.
  - Again, you need to *Think* about the situation.

What Can Go Wrong? (Cont.)

- Don't use two-sample proportion methods when the samples aren't independent.
  - These methods give wrong answers when the independence assumption is violated.
- Don't apply inference methods when there was no randomization.
  - Our data must come from representative random samples or from a properly randomized experiment.
- Don't interpret a significant difference in proportions causally.
  - Be careful not to jump to conclusions about causality.

What have we learned?

We have learned to:

1. Perform large-sample inferences (hypothesis test and confidence intervals) to compare two population proportions.
2. Describe the relationship between the sample sizes, confidence level, and margin of error for a confidence interval for the difference between two population proportions.
3. Determine the sample size required for a specified confidence level and margin of error for the estimate of the difference between two population proportions.

Credit

Some of these slides have been adapted/modified in part/whole from the slides of the following textbooks.

- Weiss, Neil A., Introductory Statistics, 8th Edition
- Bock, David E., Stats: Data and Models, 3rd Edition