Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

September 14, 2019

Notes

Topics

Notes

Objectives

Objectives:

- Carry out a two-sample z test for two population proportions.
- Compute and interpret a two-sample z CI for the difference between two population proportions.

Notes

Two-Sample Z Test for Two Population Proportions p_1 and p_2 (9.4)

- Suppose we have random samples of sizes m and n from **two populations** of **successes** and **failures**.
- We'll see how to use the samples to decide if the **population proportions** of **successes** p_1 and p_2 are different.

  The appropriate test is called the ***two-sample z test for*** $p_1 - p_2$.

Notes

- The **null hypothesis** is that no difference between the population proportions p_1 and p_2:

Null Hypothesis:

$$H_0 : p_1 - p_2 \; = \; 0$$

Notes

- The **alternative hypothesis** will depend on what we're trying to "prove":

Alternative Hypothesis: The alternative hypothesis will be one of

1. $H_a : p_1 - p_2 > 0$ (**one-sided, upper-tailed**)
2. $H_a : p_1 - p_2 < 0$ (**one-sided, lower-tailed**)
3. $H_a : p_1 - p_2 \neq 0$ (**two-sided, two-tailed**)

depending on what we're trying to verify using the data.

Notes

The Sampling Distribution of $\hat{P}_1 - \hat{P}_2$

- Suppose we have random samples of sizes m and n from **two populations** whose **proportions** of **successes** are p_1 and p_2.
- The difference $\hat{P}_1 - \hat{P}_2$ between the two **sample proportions** is an **estimator** of $p_1 - p_2$.

Notes

- Because $\hat{P}_1$ and $\hat{P}_2$ are (approximately) **normal** random variables when m and n are both **large** (Class Notes 4), and linear combinations of normal random variables are themselves normal (Class Notes 1), we have the following fact.

Notes

Proposition

If we have random samples of sizes m and n (drawn *independently* of each other) from **two populations** whose **proportions** of **successes** are p_1 and p_2, then if m and n are both **large**,

$$\hat{P}_1 - \hat{P}_2 \sim \mathsf{N}\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}\right)$$

(approximately). In this case,

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}} \sim \mathsf{N}(0,1) \quad (1)$$

(approximately).

Notes

- This follows because

  $$\hat{P}_1 \sim N\left(p_1,\ \sqrt{\frac{p_1(1-p_1)}{m}}\right) \quad \text{and} \quad \hat{P}_2 \sim N\left(p_2,\ \sqrt{\frac{p_2(1-p_2)}{n}}\right)$$

  (approximately), and so $\hat{P}_1 - \hat{P}_2$ is a linear combination of two independent normal random variables.

Notes

- It follows (from the proposition) that **when** $\boldsymbol{H_0 : p_1 - p_2 = 0}$ is **true**, the random variable

  $$Z = \frac{\hat{P}_1 - \hat{P}_2 - 0}{\sqrt{p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim \mathsf{N}(0,1)$$

  (approximately), where p is the **common value** of p_1 and p_2.

Notes

Two-Sample z Test Statistic for $p_1 - p_2$:

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - 0}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

where $\hat{P}$ is a ***pooled estimator of*** p, defined as

$$\hat{P} = \frac{\text{Total number of } \textit{successes} \text{ in both samples combined}}{\text{Total sample size when both samples are combined}}$$

$$= \frac{m}{m+n} \cdot \hat{P}_1 + \frac{n}{m+n} \cdot \hat{P}_2$$

Notes

- Z measures how many standard errors $\hat{P}_1 - \hat{P}_2$ is away from $\mathbf{0}$.
- $\hat{P}_1 - \hat{P}_2$ is an estimator of the unknown difference $\boldsymbol{p_1 - p_2}$, so ...
  1. Z will be approximately **zero** (most likely) if $p_1 - p_2 = 0$.
  2. It will be **positive** (most likely) if $p_1 - p_2 > 0$.
  3. It will be **negative** (most likely) if $p_1 - p_2 < 0$.

Notes

1. **Large positive** values of Z provide **evidence against $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_a : p_1 - p_2 > 0}$.**
2. **Large negative** values of Z provide **evidence against $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_a : p_1 - p_2 < 0}$.**
3. **Large positive *and* large negative** values of Z provide **evidence against $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_a : p_1 - p_2 \neq 0}$.**

Notes

Sampling Distribution of the Test Statistic Under $\boldsymbol{H_0}$: If Z is the two-sample z test statistic, then when m and n are both large and

$$H_0 : p_1 - p_2 = 0$$

is true,

$$Z \sim N(0,\, 1)$$

(approximately).

Notes

- The $\boldsymbol{N(0,\, 1)}$ curve gives us:
  - The ***rejection region*** as the **extreme 100α% of z values** (in the direction(s) specified by H_a).
  - The ***p-value*** as the **tail area(s) beyond the observed z value** (in the direction(s) specified by H_a).

Notes

Exercise

A study appearing in the journal *Science* investigated whether there's a link between television violence and aggressive behavior by those who watch a lot of TV.

The researchers randomly sampled **707** families in New York state and made follow-up observations over 17 years.

The table on the next slide shows results about whether a sampled teenager later conducted any aggressive act against another person.

Notes

Time Watching TV	Sample Size	Aggressive Act Yes	Aggressive Act No
Less than 1 hr per day	$m = 88$	5	83
More than 1 hr per day	$n = 619$	154	465

The **sample proportions** that conducted **aggressive acts** are

$$\hat{P}_1 = \frac{5}{88} = \mathbf{0.057} \quad \text{and} \quad \hat{P}_2 = \frac{154}{619} = \mathbf{0.249}$$

Notes

Carry out the **two-sample z test** of the hypotheses:

$$\begin{aligned} H_0 : p_1 - p_2 &= 0 \\ H_a : p_1 - p_2 &< 0 \end{aligned}$$

where p_1 is the **proportion** that conduct aggressive acts in the **population** that watches **less than 1 hour** of TV per day, and p_2 is the **proportion** in the **population** that watches **more than 1 hour** per day.

Hints: You should get the pooled estimate $\hat{P} = 0.225$, a test statistic $Z = -4.04$, and a p-value **0.0000**.

Notes

Two-Sample Z Confidence Interval for $p_1 - p_2$

Two-Sample Z CI: For independent samples of sizes m and n from two populations whose proportions of *successes* are p_1 and p_2, a $100(1-\alpha)\%$ ***two-sample z confidence interval for $p_1 - p_2$*** is

$$\hat{P}_1 - \hat{P}_2 \;\pm\; z_{\alpha/2} \cdot \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{m} + \frac{\hat{P}_2(1-\hat{P}_2)}{n}}\,.$$

Notes

- The CI is valid as long as the sample sizes m and n **are both large**

Notes

Exercise

Consider again the study of TV violence and aggressive behavior in people who watch TV.

a) Give a (point) **estimate** of the (unknown) difference $p_1 - p_2$.

b) Compute and interpret a **95% CI** for $p_1 - p_2$.

Hints: The z **critical value** is $z_{0.025} = 1.96$ and you should get $-0.192 \pm 0.059 = (-0.251, -0.133)$.

Notes

Notes

Notes