Inferences for Proportion

Sravani Vadlamani

QMB

09/29/2020

Review

- Comparing a sample mean to a population and the population mean and standard deviation are **known**
  - z-test
- Comparing a sample mean to a population and the population mean is **known**, but the standard deviation is **unknown**
  - *t*-test
- Confidence intervals are used to give a sense of how uncertain we are about the population mean. Confidence interval will shrink as *n* increases (holding everything else constant)

Review

- Dependent Samples *t*-test
  - There is a logical connection between the data
    - e.g., repeated measures data
  - Calculate difference scores, mean of difference scores, and standard deviation of difference scores
  - Conduct a One-sample t-test with the information obtained from the difference scores with 0 as the population mean

Review

- Independent Samples *t*-test
  - Most common version of the *t*-test
  - Calculate Mean, Standard Deviation for each independent sample
  - Calculate the pooled variance, use pooled variance to calculate the *appropriate* standard error
  - Divide mean difference by the standard error and determine if value is greater than the critical value given the degrees of freedom

Proportions

Introduction

- So far we looked at only continuous variables
- How about inference for categorical data?
- Examples
  - Difference in response to survey questions when they are ordered different
  - What proportion of respondents support a newly introduced health insurance plan?
  - According to a survey 50% believe Candidate A can beat Candidate B
  - Percent survival for a Cancer trial on new drug vs. typical regimen

Introduction

- Sample Proportions are well characterized by a nearly normal distribution when certain conditions are met.
- Therefore, the tools already used on normal distributions can be used with proportions.
  - Z test
  - T test

Inferences for a single proportion

Example

- In 2014, a doctor who had treated Ebola patients in Guinea was admitted to a NY hospital with Ebola. A reputable polling firm found that 82% of NY ers favored a mandatory 21 day quarantine for anyone who came in contact with Ebola. The poll included 1,042 NY adults.
- Parameter of Interest: Proportion of all NY ers who approve of mandatory quarantine: P = population proportion.
- Point Estimate: Proportion of sampled NY ers who approve of a mandatory quarantine: $\hat{p}$ = a sample proportion.

What is proportion?

- Sample proportion can be described as a sample mean
- If success is denoted as 1 and failure as 0
- Sample proportion ($\hat{p}$) is the sample mean of these numerical outcomes.
  - $\hat{p} = \frac{0+1+1+\ldots 0}{1042} = 0.82$
- The term success & failure need not mean something positive or negative – just words to frequently describe proportions.

Sampling Distribution of Proportion

- Sample distribution of proportion is nearly normal when
  - Sample observations are independent
  - There are at least 10 successes and 10 failures in our sample
    - $np \geq 10$
    - $n(1-p) \geq 10$
  - If above conditions are met
    - SE = $\sqrt{\frac{p\,(1-p)}{n}}$

Confidence Interval for a Proportion

- Point estimate ± z * SE
- p ± z * SE
- SE = $\sqrt{\frac{p\,(1-p)}{n}}$

Confidence Interval for a Proportion

- Because we don't know the true population proportion (P) we use the sample proportion ($\hat{p}$) with the confidence interval and hypothesis tests to check the sample against the population.
  - Ebola Example: ($n = 1042$, $\hat{p} = 0.82$)
- Verify the sample distribution is nearly normal.
  - Independent: random samples and < 10% of the NY population.
  - Success-Failure Condition:
  - $n(\hat{p}) = 1042 \times \hat{p}$ or $1042 \times 0.82 = 854$ successes
  - $n(1-\hat{p}) = 1042 \times (1-\hat{p})$ or $1042 \times (1-.82)$; $1042 \times .18 = 188$ failures
  - Both > 10

Confidence Interval for a Proportion

- CI = Point estimate ± critical value(SE)
- SE = $\sqrt{\frac{p\,(1-p)}{n}}$ = $\sqrt{\frac{0.82\,(1-0.82)}{1042}}$ = 0.012

- CI = proportion ± Z*(SE)
- CI = 0.82 ± 1.96 x 0.012
- (0.796, 0.844)

Sample Size Calculation

- A recent estimate of congress approval rating was 19%. What sample size does this estimate suggest we should use for a margin of error of 0.04 with 95% confidence?

Hypothesis Testing for a Proportion

- Determine the null and alternate hypothesis
- Compute the test statistic (similar to the normal distribution model)
- $Z = \frac{point\ estimate - null\ value}{SE}$
- SE = $\sqrt{\frac{p_0\,(1-p_0)}{n}}$
- Use the normal probability tables to determine the p-value
- Decide to reject/retain the null hypothesis

Hypothesis Testing for a Proportion

- Determine the null and alternate hypothesis
- Compute the test statistic (similar to the normal distribution model)
- $Z = \frac{point\ estimate - null\ value}{SE}$ ⟵ You do not know the null value
- SE = $\sqrt{\frac{p_0\,(1-p_0)}{n}}$ ⟵ You do not know the null value
- Use the normal probability tables to determine the p-value
- Decide to reject/retain the null hypothesis

When Expected Proportion is Unknown

- No previous study or population data.
- Then 50 – 50 is a good guess.
- Gives the most conservative estimate – highest possible proportion.

Example

- Do a majority of Americans support nuclear arms reduction? Set up a one-sided hypothesis test to evaluate this question.
- A simple random sample of 1028 US adults in March 2013 found that 56% support nuclear arms reduction. Does this provide convincing evidence that a majority of Americans supported nuclear arms reduction at the 5% significance level?

Inferences for difference of two proportions

Difference of Two Proportions

- To make conclusions about the difference in two population proportions
- Difference in response to survey questions when they are ordered different

Example

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

(a) A great deal
(b) Some
(c) A little
(d) Not at all

Example

The GSS asks the same question, below are the distributions of responses from the 2010 GSS as well as from a group of introductory statistics students at Duke University:

	GSS	Duke
A great deal	454	69
Some	124	30
A little	52	4
Not at all	50	2
Total	680	105

Parameter and Point Estimate

Parameter of interest: Difference between the proportions of all Duke students and all Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

Point estimate: Difference between the proportions of sampled Duke students and sampled Americans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{Duke} - \hat{p}_{US}$$

Sampling Distribution of Difference of Two Proportions

- Difference of two proportions $(\hat{p}_1 - \hat{p}_2)$is normal when
  - Each proportion separately follows a normal distribution
  - Two samples are independent of each other

Confidence Interval for Difference of Proportions

- CI = Point estimate ± z * SE
- $(\hat{p}_1 - \hat{p}_2) \pm z * SE$
- SE = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ⟸ Can this be used? NO
- SE = $\sqrt{SE_{p_1}{}^2 + SE_{p_2}{}^2}$
- SE = $\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$

Conditions for Confidence Interval for Difference of Proportions

Independence within groups

The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well.

- n_{Duke} < 10% of all Duke students and 680 < 10% of all Americans.
- We can assume that the attitudes of Duke students in the sample are independent of each other, and attitudes of US residents in the sample are independent of each other as well.

Independence between groups

The sampled Duke students and the US residents are independent of each other.

Success-failure

At least 10 observed successes and 10 observed failures in the two groups.

For the Melting Ice Example

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap ($p_{Duke} - p_{US}$).

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

For the Melting Ice Example

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap ($p_{Duke} - p_{US}$).

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$= -0.011 \pm 1.96 \times 0.0497$$

$$= -0.011 \pm 0.097$$

$$= (-0.108, 0.086)$$

Hypothesis Testing for Difference of Proportions

Which of the following is the correct set of hypotheses for testing if the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do?

(a) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} \neq p_{US}$

(b) $H_0 : \hat{p}_{Duke} = \hat{p}_{US}$
$H_A : \hat{p}_{Duke} \neq \hat{p}_{US}$

(c) $H_0 : p_{Duke} - p_{US} = 0$
$H_A : p_{Duke} - p_{US} \neq 0$

(d) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} < p_{US}$

Hypothesis Testing for Difference of Proportions

Which of the following is the correct set of hypotheses for testing if the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do?

(a) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} \neq p_{US}$

(b) $H_0 : \hat{p}_{Duke} = \hat{p}_{US}$
$H_A : \hat{p}_{Duke} \neq \hat{p}_{US}$

(c) $H_0 : p_{Duke} - p_{US} = 0$
$H_A : p_{Duke} - p_{US} \neq 0$

(d) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} < p_{US}$

Both A & C are correct

Hypothesis Testing for Difference of Proportions

- Determine the null and alternate hypothesis
- Compute the test statistic (similar to the normal distribution model)
- $Z = \frac{point\ estimate - null\ value}{SE}$
- SE = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$
- Use the normal probability tables to determine the p-value
- Decide to reject/retain the null hypothesis

Pooled Proportion

Use the pooled proportion estimate when H_0 is $p_1 - p_2 = 0$

When the null hypothesis is that the proportions are equal, use the pooled proportion ($\hat{p}$) to verify the success-failure condition and estimate the standard error:

$$\hat{p} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

For the Melting Ice Example

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{Duke}$ or $\hat{p}_{US}$) the pooled estimate is closer to? Why?

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

For the Melting Ice Example

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{Duke}$ or $\hat{p}_{US}$) the pooled estimate is closer to? Why?

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

$$= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666$$

For the Melting Ice Example

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

For the Melting Ice Example

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$Z = \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}}$$

$$= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22$$

$$p - value = 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82$$