STAT:2010/4200
Statistical Methods and Cmputing

Differences between Population Proportions
Introduction to Contingency Tables

Lecture 20
April 15, 2020

Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

Comparing two proportions

Recall: In a *two-independent sample* problem, we want to compare two populations or the responses to two different treatments using data from two independent samples.

When we are interested in comparing the *proportions of successes* in two groups, the notation is:

Population	Population proportion	Sample size	Sample proportion
1	p_1	n_1	$\hat{p}_1$
2	p_2	n_2	$\hat{p}_2$

We compare the populations by doing inference about the difference

$$p_1 - p_2$$

between the population proportions.

The *statistic* that *estimates* this difference is

$$\hat{p}_1 - \hat{p}_2$$

the difference between the two sample proportions.

Example: Do seatbelts affect the survival of children during car accidents?

- study of deaths among children involved in car accidents during an 18-month period
- two simple random samples
  - one sample from population of children who were wearing seatbelts at the time of car accident
  - one sample from population of children who were not wearing seatbelts at the time of car accident
- parameters of interest: proportions of children who die in car accidents from each of these populations

Population	Population proportion	Sample size	Sample proportion
seatbelts	p_1	123	$\frac{3}{123} = 0.024$
no seatbelts	p_2	290	$\frac{13}{290} = 0.045$

To determine whether the study provides significant evidence that seatbelts affect the proportion of kids who die if they are involved in a car accident, we test the hypotheses:

$$H_0 : p_1 - p_2 = 0 \quad or \quad H_0 : p_1 = p_2$$
$$H_a : p_1 - p_2 \neq 0 \quad or \quad H_a : p_1 \neq p_2$$

To estimate how large the difference is, we compute a confidence interval for the difference $p_1 - p_2$.

The sampling distribution of $\hat{p}_1 - \hat{p}_2$

- When both samples are large, the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.
- The mean of this normal distribution is $p_1 - p_2$.
- The standard deviation of the difference is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Because we don't know p_1 and p_2, we must replace them with estimates. These estimates will be different for confidence intervals versus hypothesis tests.

Confidence intervals for comparing two proportions

To compute a c.i., we estimate the population proportions p_1 and p_2 by their corresponding sample proportions $\hat{p}_1$ and $\hat{p}_2$.

The resulting *standard error* of $\hat{p}_1 - \hat{p}_2$ is

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

The approximate level-C two-sided confidence interval is

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

where z^* is the upper $\frac{1-C}{2}$ standard normal cutoff.

Rules of thumb for using this confidence interval:

1. Both populations are at least 10 times as large as the samples.
2. The counts of successes and failures are 10 or more in each sample.

Car accident example

Population	Population proportion	Sample size	Sample proportion
Seatbelts	p_1	123	$\frac{3}{123} = 0.024$
No seatbelts	p_2	290	$\frac{13}{290} = 0.045$

$$\hat{p}_1 - \hat{p}_2 = -0.021$$

$$\begin{aligned} SE &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= \sqrt{\frac{(0.024)(0.976)}{123} + \frac{(0.045)(0.955)}{290}} \\ &= 0.0184 \end{aligned}$$

The 95% two-sided confidence interval is

$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) \pm z^* SE &= \\ (0.024 - 0.045) \pm (1.96)(0.0184) &= \\ -0.021 \pm 0.0.036 &= \\ (-0.057, 0.015) \end{aligned}$$

We are 95% confident that this interval covers the true difference between the proportions of kids who die from car accidents in the population who were wearing seatbelts at the time of the accident vs. the population who were not.

The interval includes the value 0, so it is plausible based on this data that there is no difference!

The hypothesis test

For the formal hypothesis test, the hypotheses are:

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

Suppose we had set $\alpha = .05$ when we were designing the study.

- We must standardize $\hat{p}_1 - \hat{p}_2$ to get a z statistic.
- We do this under the assumption that H_0 is true, that is that p_1 and p_2 have the same value p.
  - Instead of estimating p_1 and p_2 separately in the standard deviation of the difference, we *pool* the two samaples and use the overall sample proportion to estimate the single population parameter p.
  - The **pooled sample proportion** is

$$\hat{p} = \frac{total\ count\ of\ successes\ in\ both\ samples}{n_1 + n_2}$$

The test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) \ - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Car accident example

The pooled sample proportion is:

$$\hat{p} = \frac{3+13}{123+290} = \frac{16}{413} = 0.039$$

The z statistic is:

$$z = \frac{(0.024 - 0.045) \ - 0}{\sqrt{(0.039)(0.961)\left(\frac{1}{123} + \frac{1}{290}\right)}}$$

$$= -1.01$$

To get the p-value for the two sided test, we look for the area under a standard normal curve that is farther away from 0 than -1.01 in either direction.

Table A gives .156 as the area to the left of -1.01.

$$p - value = 2 * (0.156) = 0.312$$

We cannot reject the null hypothesis. This particular set of sample data does not provide evidence that the proportion of children dying in car accidents differs between the population of those wearing seatbelts at the time of the accident and the population of those not.

We do not trust these results because the rules of thumb are not met.

Contingency Tables and the Chi-square test

An equivalent way of comparing two population proportions, that generalizes to more than two populations.

Begin by presenting the data as a two-way table, with rows representing levels of one variable and columns representing levels of the other.

Seatbelt example:

Seatbelts	Died	Did not die	Total
Yes	3	120	123
No	13	277	290
Total	16	397	413

To test the hypotheses

$$H_0 : p_1 - p_2 = 0 \quad or \quad H_0 : p_1 = p_2$$
$$H_a : p_1 - p_2 \neq 0 \quad or \quad H_a : p_1 \neq p_2$$

using the two-way table, we must compute the **expected counts**. These are the counts we would expect (except for random variation) if H_0 were true.

$$expected\ count = \frac{row\ total \times column\ total}{table\ total}$$

If H_0 were true, there would be just one p shared by both populations.

Our best estimate is again the pooled sample proportion $\hat{p} = 0.039$

Seatbelt example

Observed counts (for reference)

Seatbelts	Died	Did not die	Total
Yes	3	120	123
No	13	277	290
Total	16	397	413

Expected counts

Seatbelts	Died	Did not die	Total
Yes	4.8	118.2	123
No	11.3	278.7	290
Total	16	397	413

The Chi-Square Test

- Recall that the expected counts were computed under the assumption that *the null hypothesis was true.*
- We can test the null hypothesis by determining whether the differences between the observed and expected counts are too large to be likely to be due to chance.
- Notation
  - O_i is the observed count in cell i
  - E_i is the expected count in cell i

The Chi-square statistic

The *statistic* that we use for this test is the sum over all the cells in the table of $\frac{(observed\ count - expected\ c}{expected\ count}$

The formula in mathematical notation is

$$X^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$

where rc is the total number of cells in the table.

- r is the number of rows
- c is the number of columns

- Think of X^2 as a measure of the *distance* of the observed counts from the expected counts.
- Like any distance, X^2
  - is always zero or positive
  - is zero only when the observed counts are exactly equal to the expected counts
- Large values of X^2 are evidence *against* H_0.
  - indicate that observed counts are *far away* from what we would expect if H_0 were true.
- The Chi-square statistic X^2 follows a *Chi-square distribution* (χ^2 distribution) with $(r-1)(c-1)$ degrees of freedom.

The Chi-Square test for the car-accidents example

$$\begin{aligned} X^2 &= \frac{(3-4.8)^2}{4.8} + \frac{(120-118.2)^2}{118.2} \\ &\quad + \frac{(13-11.3)^2}{11.3} + \frac{(277-278.7)^2}{278.7} \\ &= 0.969 \end{aligned}$$

Since we have 2 rows and 2 columns in our table, the degrees of freedom is

$$(r-1)(c-1) = 1(1) = 1$$

We will carry out our hypothesis test at $\alpha = .05$. According to Table E, the .05 cutoff under the Chi-square distribution with 1 degree of freedom is 3.84.

The Chi-square test is always 2-sided. For the Chi-square test, we always reject if the test statistic is *larger* than the cutoff value.

Our computed value, 0.969, is *smaller* than this cutoff. Therefore we cannot reject H_0. This result is consistent with what we got with both the confidence interval and the z test.

When does Chi-square test give accurate enough inference?

- rule of thumb: when *expected* counts in all cells are ≥ 5
- not quite satisfied in this example

SAS for the Chi-square test

```
data seatbelts ;
input seatbelts $ died $ count ;
datalines ;
Y Y 3
Y N 120
N Y 13
N N 277
;
run ;

proc freq data = seatbelts ;
tables seatbelts * died ;
weight count ;
run ;

proc freq data = seatbelts ;
tables seatbelts * died / expected ;
weight count ;
run ;


proc freq data = seatbelts ;
       tables seatbelts * died / chisq ;
weight count ;
run ;
```

The FREQ Procedure

Table of seatbelts by died

seatbelts	died		
Frequency | Percent | Row Pct | Col Pct	N	Y	Total
N	277	13	290
	67.07	3.15	70.22
	95.52	4.48	
	69.77	81.25	
Y	120	3	123
	29.06	0.73	29.78
	97.56	2.44	
	30.23	18.75	
Total	397	16	413
	96.13	3.87	100.00

The FREQ Procedure

Table of seatbelts by died

seatbelts	died		
Frequency | Expected | Percent | Row Pct | Col Pct	N	Y	Total
N	277	13	290
	278.77	11.235	
	67.07	3.15	70.22
	95.52	4.48	
	69.77	81.25	
Y	120	3	123
	118.23	4.7651	
	29.06	0.73	29.78
	97.56	2.44	
	30.23	18.75	
Total	397	16	413
	96.13	3.87	100.00

The FREQ Procedure

Table of seatbelts by died

seatbelts	died		
Frequency | Percent | Row Pct | Col Pct	N	Y	Total
N	277	13	290
	67.07	3.15	70.22
	95.52	4.48	
	69.77	81.25	
Y	120	3	123
	29.06	0.73	29.78
	97.56	2.44	
	30.23	18.75	
Total	397	16	413
	96.13	3.87	100.00

Statistics for Table of seatbelts by died

Statistic	DF	Value	Prob
Chi-Square	1	0.9687	0.3250
Likelihood Ratio Chi-Square	1	1.0553	0.3043
Continuity Adj. Chi-Square	1	0.4976	0.4805
Mantel-Haenszel Chi-Square	1	0.9664	0.3256
Phi Coefficient		-0.0484	
Contingency Coefficient		0.0484	
Cramer's V		-0.0484	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Cell (1,1) Frequency (F)	277
Left-sided Pr <= F	0.2468
Right-sided Pr >= F	0.9019
Table Probability (P)	0.1488
Two-sided Pr <= P	0.4126

Sample Size = 413