22S:30/105
Statistical Methods and Cmputing

Differences between Population Proportions
Introduction to Contingency Tables

Lecture 21
April 11, 2008

Kate Cowles
374 SH, 335-0727
kcowles@stat.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 protect 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 5 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

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