Two sample proportions

Announcements

- Lab due dates adjusted - full credit as long as they are turned in by the beginning of RRR week
- Statistics is everywhere extra credit will stay open until Friday April 16th
- Poll for window of time for final exam - please answer by next Monday - we will schedule based on the responses

Roadmap

Part III started with continuous outcomes and categorical predictors

How many groups?	Independent?	parametric?	test
1	yes	yes	Z or one sample T
2	yes	yes	Two sample T
2	yes	no	Wilcox rank sum
2	no	yes	Paired T
2	no	no	Wilcox sign rank
3 or more	yes	yes	ANOVA
3 or more	yes	no	Kruskal Wallis

Then we addressed continuous outcomes with continuous predictors (correlation and linear regression)

Last Lecture we introduced categorical outcomes (binary/proportions)

Today we will look at comparing two proportions.

Comparing two proportions (B & M text Chapter 20)

- Two SRS from independent populations

Notation:

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

CI for the difference in two proportions

Large-sample confidence interval for the difference of two proportions

- Use when the number of observed successes and failures are > 10 for both samples

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Just like for the difference between two means, the SE of the difference is the square root of the sum of the variances.
- This large-sample interval often has a lower confidence level than the one specified. That is, if you repeated the method several times < 95 of the 100 created intervals would contain the true value for the difference between the proportions for a 95% CI.

Example using the large sample method

Patients in a randomized controlled trial were severely immobilized and randomly assigned to either Fragamin (to prevent blood clots) or to placebo. The number of patients experiencing deep vein thrombosis (DVT) was recorded

	DVT	no DVT	Total	$\hat{p}$
Fragamin	42	1476	1518	0.0277
Placebo	73	1400	1473	0.0496

- We can apply the large study method because the sample sizes are large and the number of observed successes and failures are > 10 (i.e., 42, 73, 1476, and 1400 all > 10).

Example using the large sample method

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.0496 - 0.0277) \pm z^* \sqrt{\frac{0.0496_1(1-0.0496)}{1473} + \frac{0.0277(1-0.0277)}{1518}}$$

$$0.0219 \pm 1.96 \times 0.0071 = 0.008 \text{ to } 0.0358$$

Plus 4 method for the comparison of two proportions

- When the assumptions of the large sample method are not satisfied, we use the plus four method.
- When you have two samples this method says: add 4 observations, 1 success and 1 failure to each of the two samples.

$$\tilde{p}_1 = \frac{\text{no. of successes in pop1} + 1}{n_1+2} \quad \tilde{p}_2 = \frac{\text{no. of successes in pop2} + 1}{n_2+2}$$

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

- Use when the sample size is at least five, with any counts of success and failure (can even use when number of successes or failures $= 0$)
- Much more accurate when the sample sizes are small
- May be conservative (giving a higher level of confidence than the one specified)

Example using the plus four method

	Flu	no Flu	Total	$\hat{p}$
Vaccine	4	96	100	0.04
Placebo	11	89	100	0.11

Here, we don't have 10 "successes" (flu) in both groups, so we cannot use the Normal approximation method.

Example using the plus four method

$$\tilde{p}_1 = \frac{\text{no. of successes in pop1} + 1}{n_1+2} = \frac{5}{102}$$

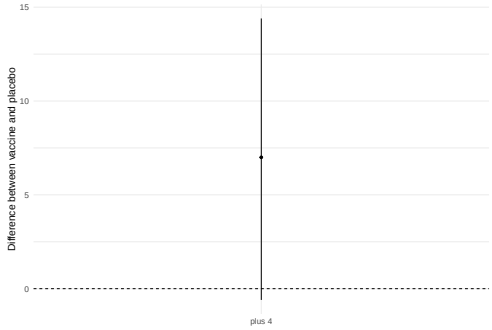
$$\tilde{p}_2 = \frac{\text{no. of successes in pop2} + 1}{n_2+2} = \frac{12}{102}$$

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

$$(\frac{12}{102} - \frac{5}{102}) \pm 1.96 \times 0.0384 = \text{-0.6\% to 14.4\%}$$

Example using the plus four method (continued)

The 95% CI of the difference ranged from -0.6 percentage points to 14.4% percentage points. While this CI contains 0 (the null hypothesized value for no difference) most of the values contained within it are positive, perhaps suggesting support for the alternative hypothesis. In this case, we might want to collect more data to create a more precise CI.



Hypothesis testing - two samples binary data

Hypothesis testing when you have two samples and binary data

$H_0 : p_1 = p_2$ or $p_1 - p_2 = 0$

H_a :

- $p_1 \neq p_2$ or $p_1 - p_2 \neq 0$ (two-sided)
- $p_1 > p_2$ or $p_1 - p_2 > 0$ (one sided upper tail)
- $p_1 < p_2$ or $p_1 - p_2 < 0$ (one sided lower tail)

What does it mean to assume the null is true?

- If the null hypothesis is true, then p_1 is truly equal to p_2. In this case, our best estimate of the underlying proportion that they are both equal to is

$$\hat{p} = \frac{\text{no. successes in both samples}}{\text{no. individuals in both samples}}$$

- Also, our best guess at the SE for $\hat{p}$ is:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

$$\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

This is the formula for the SE for the difference between two proportions but we have substituted $\hat{p}$ for p_1 and p_2.

Hypothesis testing when you have two samples and binary data

Using the information from the previous slide, we can create the z-test for the difference between two proportions as:

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

Use this test when the counts of successes and failures are ≥ 5 in both samples

Example of hypothesis testing when you have two samples and binary data

Recall the RCT data on the occurrence of DVT between Fragamin vs. placebo groups:

	DVT	no DVT	Total	$\hat{p}$
Fragamin	42	1476	1518	0.0277
Placebo	73	1400	1473	0.0496

$H_0 : p_1 = p_2$, or that the rate of DVT is the same between Fragamin and placebo groups.

Suppose you're interested in knowing whether these two groups had different rates of DVT. Then, $H_a : p_1 \neq p_2$

Example of hypothesis testing when you have two samples and binary data

1. Compute $\hat{p} = \frac{42+73}{1518+1473} = \frac{115}{2991} = 0.03844868$
2. Compute the SE: $\sqrt{0.0384(1-0.0384)(\frac{1}{1518}+\frac{1}{1473})} = 0.007032308$
3. Compute the test statistic:

$$z = \frac{0.04955872 - 0.02766798}{0.007032308} = 3.11$$

4. Calculate the p-value

```
pnorm(q = 3.112881, lower.tail = F)*2
```

```
## [1] 0.001852707
```

Two sample hypothesis testing in R

Example of hypothesis testing when you have two samples and binary data

```
prop.test(x = c(42, 73), # x is a vector of number of successes
          n = c(1518, 1473)) # n is a vector of sample sizes
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(42, 73) out of c(1518, 1473)
## X-squared = 9.107, df = 1, p-value = 0.002546
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.036376917 -0.007404562
## sample estimates:
##     prop 1     prop 2
## 0.02766798 0.04955872
```

Example of hypothesis testing when you have two samples and binary data

- R gives a slightly different p-value because it has a continuity correction.
- This is okay. If you want to use R to check your hand calculation, you need to add the argument `correct = F` to the calculation.

Next week

The next lectures we will introduce the Chi-squared distribution and two tests - Goodness of fit - Chi-squared test of association

This will be the end of new material for the semester. Bonus material on bootstrapping and permutations will be posted as notes, but has been dropped from the required material for the exam. It may be included as extra credit.

Parting Humor