Lecture 27: Inference for comparing two proportions

Chapter 20

Corinne Riddell (Instructor: Tomer Altman)

November 3, 2025

Learning objectives for today

1. Learn about two methods to make confidence intervals for the difference between two proportions: i) the large sample methods, and ii) the plus four method
2. Conduct a hypothesis test for the difference between two proportions using a z-test
3. Learn how to use `prop.test()`, a function introduced last class, to conduct the hypothesis test in R

Comparing two proportions (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	$\hat{p}_2$

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, that is:

$$var(\hat{p}_1 - \hat{p_2}) = var(\hat{p}_1) + var(\hat{p}_2)$$

  if the data used to construct $\hat{p}_1$ are independent observations from those used to calculate $\hat{p}_2$
- This large-sample interval can have poor **coverage**.
  - If you repeated the method 100 times, fewer than 95 of the 100 created intervals would contain the true value for the difference between the proportions for a 95% CI
- This is the same issue as the large sample method for one proportion, or any situation where sample size is too small for the Central Limit Theorem to take effect

Check your understanding!

Example using the large sample method

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

	DVT	no DVT	Total	$\hat{p}$
Fragamin	42	1476	1518	$\frac{42}{1518} = 2.77\%$
Placebo	73	1400	1473	$\frac{73}{1473} = 4.96\%$

- Check the conditions:
  - We can apply the large study method because the sample sizes are large
  - The number of observed successes and failures are larger than 10 (i.e., 42, 73, 1,476, and 1,400 all larger than 10)
- The estimate of the difference between the two proportions is $4.96\% - 2.77\% = 2.19\%$
- We can use the large sample method to make a confidence interval for this difference

Example using the large sample method

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

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

$0.0219 \pm 1.96 \times 0.0071 = 0.008$ to 0.0358

Thus, the 95% confidence interval for the difference goes from 0.8% to 3.58%

When datasets are small: "Plus 4" method for the comparison of two proportions

- Like done for deriving a CI for a single proportion, researchers have developed small sample adjustments so that one can use the large sample methods with more robust performance (better coverage)
- And like the case with a single proportion, there are more robust methods that do not rely on the CLT
- When the assumptions of the large sample method are not satisfied, we use the "plus four" method

When datasets are small: "Plus 4" method for the comparison of two proportions

- When you have two samples this method says: add four observations total (two to each sample), and one success and one failure to each of the two samples:

$$\tilde{p}_1 = \frac{\text{(no. of successes in population 1)} + 1}{n_1+2} = \frac{\hat{p}_1 n_1 + 1}{n_1+2}$$

$$\tilde{p}_2 = \frac{\text{(no. of successes in population 2)} + 1}{n_2+2} = \frac{\hat{p}_2 n_2 + 1}{n_2+2}$$

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

- 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 (has higher coverage then suggested by the confidence level)

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 large sample method.

Example using the Plus Four Method

$\tilde{p}_1 = \frac{\text{no. of successes in pop1+ 1}}{n_1+2} = \frac{4+1}{100+2} = 0.04901961$

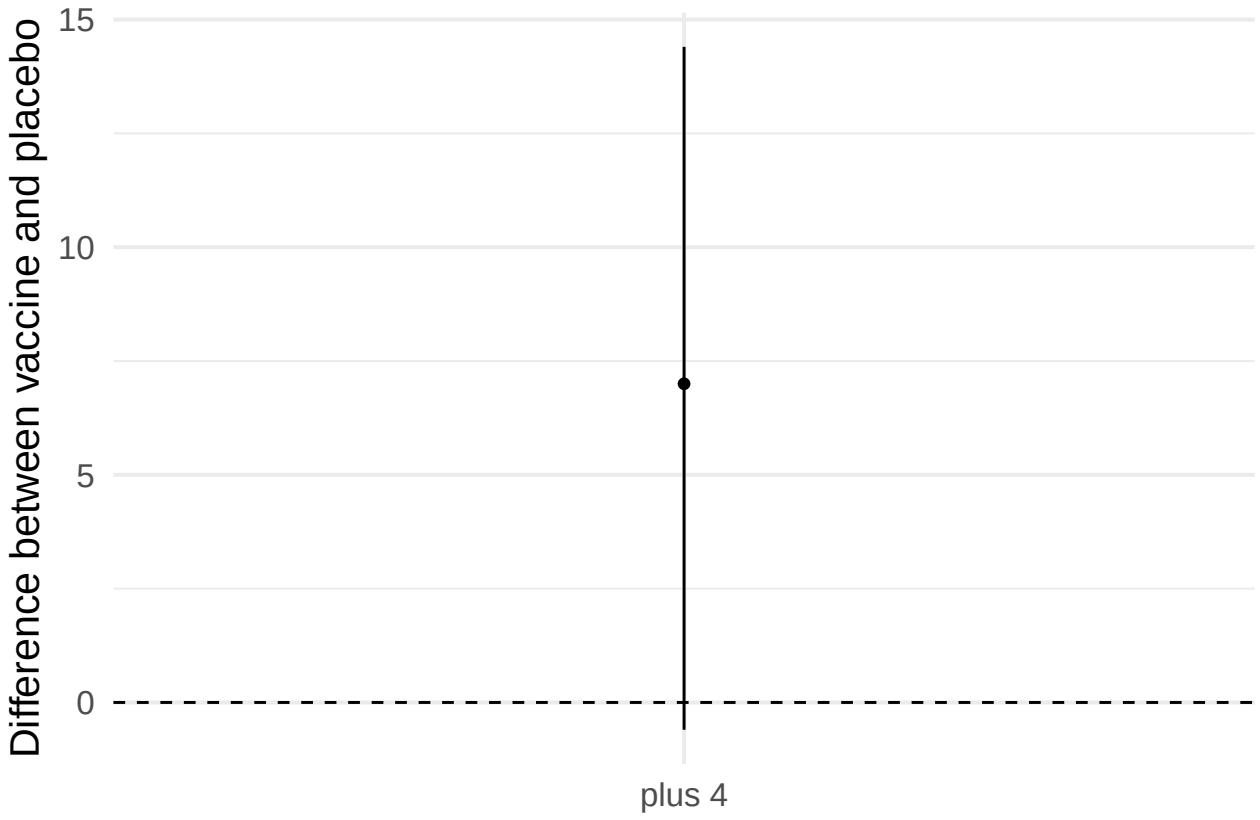
$\tilde{p}_2 = \frac{\text{no. of successes in pop2 + 1}}{n_2+2} = \frac{11+1}{100+2} = 0.1176471$

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

Filling in $\tilde{p}_1 = 0.04901961$, $\tilde{p}_2 = 0.1176471$ and n_1=n_2=100 gives:

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

The 95% CI of the difference ranges 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 the CI 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 when you have two samples and binary data

- H_0:
  - $p_1 = p_2$
  - $p_1 - p_2 = 0$
- H_a :
  - Two-sided:
    - $p_1 \neq p_2$
    - $p_1 - p_2 \neq 0$
  - One-sided upper tail:
    - $p_1 > p_2$
    - $p_1 - p_2 > 0$
  - One-sided lower tail:

* $p_1 < p_2$
* $p_1 - p_2 < 0$

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 two equal underlying proportions is called the pooled sample proportion:

$$\hat{p} = \frac{\text{number of successes in both samples}}{\text{number of individuals in both samples}}$$

- Also, our best guess of the SE for $\hat{p}$ is calculated by taking the formula for the SE for the difference between the two proportions and substituting in $\hat{p}$ for p_1 and p_2:

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

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

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 z-test when the counts of successes and failures are 5 or larger in both samples

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

Recall the RCT data on the occurrence of deep vein thrombosis 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 proportion of DVT is the same between Fragamin and placebo groups.

Suppose you're interested in knowing whether these two groups had different proportions 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: $\sqrt{0.0384(1-0.0384)\left(\frac{1}{1518} + \frac{1}{1473}\right)} = 0.007032308$
3. Compute the test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE}$$

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

4. Calculate the p-value

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

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

The p-value is equal to 0.19%. Under then null hypothesis of no difference between the proportions, there is a 0.19% chance of observing the difference we saw (or more extreme) which provides evidence in favor of the alternative hypothesis that these proportions are different.

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

Here is how to conduct the hypothesis test using R's `prop.test()`:

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

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(42, 73) out of c(1518, 1473)
## X-squared = 9.69, df = 1, p-value = 0.001853
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.035708093 -0.008073386
## sample estimates:
##     prop 1     prop 2
## 0.02766798 0.04955872
```

Interpreting this output:

- Focus on the `p-value` output
- R is actually doing a Chi-squared test (which we will learn soon!)
  - provides identical findings to performing a z-test for the equality of two proportions
- In the output `X-squared = 9.69` is the test statistic for the Chi-squared test (where Chi is the Greek letter χ)
- The `X-squared` test statistic is the z-test statistic squared
  - $\sqrt{9.69} = 3.11$, which is what we found on previous slide

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

Here is how to conduct the hypothesis test using R's `prop.test()`:

```r
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
          correct = T)
```

```
##
##  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
```

- Typically, we would set `correct = T` to have R perform a continuity correction, which makes its test a bit better than the one we did by hand or without the correction
- `correct=T` is the default option and the one we will use
- Only set `correct=F` if you want to check work you did by hand

Comparing treatments for UTIs

- In a study of urinary tract infections, patients were randomly assigned to one of two treatment regimes:
  - Treatment 1: trimethoprim / sulfamethoxazole or,
  - Treatment 2: fosfomycin / trometamol
- 92 of the 100 patients assigned to treatment 1 showed bacteriological cure while 61 of 100 assigned to treatment 2 showed bacteriological cure

Comparing treatments for UTIs

i) What is the estimate of the difference in proportions?
ii) Perform a test to provide evidence whether or not this difference in proportions reflects a true difference between treatments

Comparing treatments for UTIs

i) What is the estimate of the difference in proportions?

.92 - .61 = .31. There is a 31 percentage point different between the two proportions.

- In online media "percentage point" is often abbreviated as "point". So a "2 point difference" means a difference of 2 percentage points. Note how this is not the same as saying "a difference of 2 percent"!

Comparing treatments for UTIs

ii) Perform a test to provide evidence whether or not this difference in proportions reflects a true difference between treatments

Large sample z-test of the form $\frac{p_1-p_2}{SE}$, where $SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ and $\hat{p} = \frac{92+61}{200} = 0.765$

$SE = \sqrt{0.765(1-0.765)\left(\frac{1}{100}+\frac{1}{100}\right)} = 0.05996249$

$z = 0.31/0.05996249 = 5.169899$

Without using R, we know the corresponding p-value will be very small.

```
pnorm(5.169899, lower.tail = F)*2
```

```
## [1] 2.342205e-07
```

Interpretation of the p-value: Assuming the null hypothesis of no difference between the proportions, there is less than a 0.0001% chance of seeing a difference of the magnitude that we saw (or larger). This provides evidence in favor of the alternative hypothesis of a difference between the proportions.

CI's for difference of two proportions with very small sample sizes

Consider the following table

	DVT	no DVT	Total	$\hat{p}$
X=0	0	1	1	0.00
X=1	2	2	4	0.50

- This is an extreme example where the it would be dubious to rely on large sample CIs
- However, there are exact confidence intervals available for the difference of two proportions in small sample size
- Like we discussed for single proportion, they are based on an exact hypothesis test and the interval is defined as the set of null values (of the difference) for which the null hypothesis (at level α) is not rejected based on the fixed data
- Not important at this point to understand why exact CIs for difference in proportions work, but good to know they are available if you have very small sample size

CI's for difference of two proportions with very small sample sizes

First, we use the prop.test function to construct a large sample CI based on the table data (note the warning):

```
prop.test(x=c(2,0),n=c(4,1))
```

```
## Warning in prop.test(x = c(2, 0), n = c(4, 1)): Chi-squared approximation may
## be incorrect
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(2, 0) out of c(4, 1)
## X-squared = 1.284e-32, df = 1, p-value = 1
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.489991  1.000000
## sample estimates:
## prop 1 prop 2
##    0.5    0.0
```

- Now we do with the exact method (which requires a different package and uses different syntax)

```
library(ExactCIdiff)
BinomCI(4,1,2,0,CItype="Lower")
```

```
## $conf.level
## [1] 0.95
##
## $CItype
## [1] "Lower"
##
## $estimate
## [1] 0.5
##
## $ExactCI
## [1] -0.57739  1.00000
```

- In this case, one would expect the exact CIs to give better coverage

Check your understanding!