Lecture 26: Inference about a population proportion

Chapter 19

Corinne Riddell (Instructor: Tomer Altman)

October 29, 2025

Recap

- So far we've learned the z-test and the t-test that apply when the variable of interest is continuous
- We applied these tests to one-sample (e.g., $H_0 : \mu = 8$) and two-sample settings (e.g., $H_0 : \mu_1 = \mu_2$)
- Today, we will generalize these procedures to binary data, for which we estimate a proportion $\hat{p}$ from a sample and use that as our best guess of the underlying population parameter p
- Notation: $\bar{x}$ is to μ as $\hat{p}$ is to p

Agenda

- Confidence interval for a proportion
- Sample size estimates for a proportion
- Hypothesis tests for a proportion

Recall the sampling distribution for $\hat{p}$

The sampling distribution for $\hat{p}$ is centered on p with a standard error of $\sqrt{\frac{p(1-p)}{n}}$

If we follow the same format for the CI from previous chapters we would get:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is what is known as the **large sample confidence interval for a population proportion**

But. . .

- This confidence interval can perform poorly, meaning that if we repeated the confidence interval 100 times (based on 100 random samples), the coverage could be what is termed
  - overly conservative (e.g., coverage significantly above 95%), or
  - anti-conservative, or permissive, or "poor coverage"
    - Fewer than 95 out of 100 of the 95% confidence intervals, on average, would contain the true value for the proportion p
  - For the Wald approximation, for different combinations of p and n, it can exhibit **both** problems!
- To overcome this, we will modify how we calculate the confidence interval slightly using what is known as the "plus four" method

Introducing: the "Plus Four" Method

- If you add two imaginary successes and two failures to the data set (increasing the sample size by four imaginary trials), the interval can have better performance
- Let $\tilde{p} = \frac{\text{number of successes} + 2}{n+4}$
- Let $SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$
- Then the CI is:

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

- This is called the **"Plus Four" Method**
- Note we use z^* rather than t^*. This is because the standard error of the sampling distribution is completely determined by p and n, we don't need to estimate a second parameter
- In addition, for smaller samples (when one cannot rely on the CLT), one can rely on alternative methods to get inference that do not rely on Normality of the $\hat{p}$
- Use this method when $n \geq 10$

Why does the “Plus Four” Method work?

- It is a simplification of a more complex method known as the Wilson Score Interval
- You don’t need to know why it works, just that it is better to use this “plus four” trick if you’re making a confidence interval for a proportion by hand
- Note: if the number of successes and failures is relatively large, then the “plus four” method will converge with the large sample CI method

Two methods so far...

We have so far introduced the large sample method to calculate the CI for p:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

And the plus four method to calculate the CI for p:

- $\tilde{p} = \frac{\text{number of successes} + 2}{n+4}$
- Let $SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$
- Then the CI is:

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

We are going to talk about two more methods.

What does R use?

- R has two functions to calculate confidence intervals for proportions
- The first function is `prop.test` (analogous to `t.test`) to calculate confidence intervals and hypothesis tests for binomial proportions
- This function uses the **Wilson score** method, specifically, the "Wilson score interval with a continuity correction".
  - Thus, you don't need to "plus 4" the proportion or standard error
  - It will do it for you
  - It does an even better job because of the continuity correction
  - You do not need to know how to calculate the Wilson score method by hand
  - You only need to know how to use R to perform this method.

What does R use?

- There is a fourth often-used method to compute confidence intervals for proportions called the **Clopper Pearson method**
  - Also known as the **"Exact method"**
  - Implemented with the second R function, `binom.test()`
- The exact method is statistically conservative, meaning that it gives better coverage than it suggests
  - A 95% CI computed under this method includes the true proportion in the interval more than 95% of the time
- It can be thought of as an inversion of hypothesis testing
  - It finds the set of all values of the parameter p_0, where we consider the null value in a two-sided hypothesis test, would lead to a p-value $p \geq 0.05$
  - It calculates the p-value using the binomial distribution function
  - As a reminder, this was the cumulative probability function for binomial (Lec 16):

$$P(X \leq x) = \sum^{x} \binom{n}{\cdot} p_0^i (1-p_0)^{n-i}$$

Example applying all the methods

Suppose that 500 elderly individuals suffered hip fractures, of which 100 died within a year of their fracture. Compute the 95% CI for the proportion who died using:

- the large sample method
- the "plus four" method (by hand)
- the Wilson Score method (using `prop.test`)
- the Clopper Pearson Exact method (using `binom.test`)

Example of large sample method to calculate the CI for a proportion (by hand)

```
p.hat <- 100/500 # estimate proportion
se <- sqrt(p.hat*(1-p.hat)/500) # standard error
p.hat - 1.96*se # Lower confidence bound
```

```
## [1] 0.1649385
```

```
p.hat + 1.96*se # Upper confidence bound
```

```
## [1] 0.2350615
```

- Our estimate for the proportion is $\hat{p} = 20\%$
- Using the large sample method, the 95% confidence interval is 16.5% to 23.5%
- Remember, this method has poor coverage, meaning that fewer than 95 of the 100 intervals we would make would contain the true value p on average

Example using the "plus four" method to calculate the CI for a proportion (by hand)

```
p.tilde <- (100 + 2)/(500 + 4)
se <- sqrt(p.tilde * (1 - p.tilde)/(500 + 4)) # standard e
p.tilde - 1.96 * se # Lower confidence bound
```

```
## [1] 0.1673039
```

```
p.tilde + 1.96 * se # Upper confidence bound
```

```
## [1] 0.237458
```

Using the plus 4 method, the confidence interval is 16.7% to 23.7%.

Example using the Wilson Score method to calculate the CI for a proportion (using R)

```r
prop.test(x = 100, n = 500, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  100 out of 500, null probability 0.5
## X-squared = 178.8, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1663581 0.2383462
## sample estimates:
##   p
## 0.2
```

- The 95% confidence interval using the Wilson Score method is 16.6% to 23.8%
- Note that the `prop.test` function is also conducting a two-sided hypothesis test (where $H_0: p_0 = 0.5$, unless

Example using the Clopper Pearson "Exact" method to calculate the CI for a proportion (using R)

```r
binom.test(x = 100, n = 500, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  100 and 500
## number of successes = 100, number of trials = 500, p-valu
## alternative hypothesis: true probability of success is no
## 95 percent confidence interval:
##  0.1658001 0.2377918
## sample estimates:
## probability of success
##                    0.2
```

- The 95% confidence interval using the exact binomial test is 16.6% to 23.8%
- Note that this interval is wider than the one made with the large sample method
  - It has larger coverage (contains the true value more often) which necessitates a wider interval
- Note that the `binom.test` function is also conducting a two-sided hypothesis test (where $H_0 : p_0 = 0.5$, unless otherwise specified)

Summary of the confidence intervals across the methods

Method	95% Confidence Interval	R Function
Large sample	16.5% to 23.5%	by hand
Plus four	16.7% to 23.7%	by hand
Wilson Score*	16.6% to 23.8%	`prop.test`
Clopper Pearson**	16.6% to 23.8%	`binom.test`

*with continuity correction

**also known as the exact method

- Only the large sample method is symmetric around $\hat{p} = 20\%$. This is okay. Symmetric confidence intervals are applicable to Normal sampling distributions, which might or might not describe the distribution of $\hat{p}$.
- Non-symmetric CIs make sense because p is bounded between 0 and 1. For example, if p is very small, say 0.012, you would not want a CI that has a lower bound which is negative, this would not make sense.

Another example of the “plus four” method (by hand)

[We are including another example for you to read so you can practice working out the calculations by hand.]

A study examined a random sample of 75 SARS patients, of which 64 developed recurrent fever.

Therefore $\hat{p} = 64/75 = 85.33\%$

Using the plus 4 method: $\tilde{p} = \frac{64+2}{75+4} = 83.54\%$

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{75+4}} = \sqrt{\frac{.8354 \times (1-0.8354)}{79}} = 0.04172$$

Thus the plus four 95% CI is:
$\tilde{p} \pm 1.96 \times SE = 0.8354 \pm 0.04172 = 79.37\%$ to 87.71%

How big should the sample be to estimate a proportion?

Suppose that you want to estimate a sample size for a proportion within a given margin of error. That is, you want to put a maximum bound on the width of the corresponding confidence interval.

Let m denote the desired margin of error. Then $m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

We can solve this equation for n, but we also need to plug in a value for p. To do that we make a guess for p denoted by p^*.

p^* is your best estimate for the underlying proportion. You might gather this estimate from a completed pilot study or based on previous studies published by someone else. If you have no best guess, you can use $p^* = 0.5$. This will produce the most conservative estimate of n. However if the true p is less than 0.3 or greater than 0.7, the sample size estimated may be much larger than you need.

How big should the sample be to estimate a proportion?

Rearranging the formula on the last slide for n, we get:

$m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\sqrt{n}m = z^* \sqrt{p(1-p)}$

$\sqrt{n} = \frac{z^*}{m} \sqrt{p(1-p)}$

$n = (\frac{z^*}{m})^2 p^*(1-p^*)$

This last formula is the one we will use to estimate the required sample size.

Example of estimating sample size

Suppose after the general election, you were interested in estimating the number of STEM undergraduate students who voted. So you want to do a study to estimate this proportion. How many students should you include in your sample?

First you need to decide what margin of error you desire. Suppose it is 4 percentage points or $m = 0.04$ for a 95% CI.

If you had no idea what proportion of STEM students voted then you let $p^* = 0.5$ and solve for n:

$$n = (\frac{z^*}{m})^2 p^*(1 - p^*) = (\frac{1.96}{0.04})^2 \times 0.5 \times 0.5 = 600.25 = 601$$

This implies you would need to sample 601 students to get an estimate with a 95% confidence interval that is +/- 4 percentage points.

However, suppose you found a previous study that estimated the number of STEM students who voted to be 25%. Then what sample size would you need to detect this proportion?

Example of estimating sample size

What if you want the width of the 95% confidence interval to be 6 percentage points. What would m be in this case?

Example of estimating sample size

What if you want the width of the 95% confidence interval to be 6 percentage points. What would m be in this case?

The width of the 95% CI is equal to twice the margin of error. So if you want the width to be 0.06, then this is equivalent to saying you want a margin of error of 0.03.

Hypothesis tests of a proportion

When you only have one sample what is the null hypothesis? You're interested in knowing whether there is evidence against the null hypothesis that the population proportion p is equal to some specified value p_0. That is:

$$H_0 : p = p_0$$

For example, you may want to test whether there is evidence against the null hypothesis that $p = 0.25$.

Hypothesis tests of a proportion

Recall the sampling distribution for the proportion:

- Normally distributed
- Centered at p_0 under the null hypothesis
- Has a standard error of $\sqrt{\frac{p_0(1-p_0)}{n}}$

Hypothesis tests of a proportion

The test statistic for the null hypothesis is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

This is a z-test (not a t-test) so we compare to the standard Normal distribution and ask what is the probability of observing a z value of this magnitude (or more extreme).

Hypothesis tests of a proportion

One sided alternatives:

- $H_a : p > p_0$
- $H_a : p < p_0$

Two-sided alternative:

- $H_a : p \neq p_0$

When to use this test? Use this test when the expected number of successes and failures is ≥ 10. That is, when $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Example of a hypothesis test for a proportion

Consider an SRS of 200 patients undergoing treatment to alleviate side effects from a rigorous drug regimen at a particular hospital, where 33 patients experienced reduced or no side effects.

$\hat{p} = 33/200 = 0.165 = 16.5\%$

Suppose that historically, the rate of patients with little or no side effects is 10%. Does the new treatment increase the rate? That is:

$H_0 : p = 0.10$

$H_a : p > 0.10$

Example of a hypothesis test for a proportion

Step 1: Calculate $\hat{p} = 16.5\%$ from previous slide.

Step 2: Calculate the standard error of the sampling distribution for p under the null hypothesis:
$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.1(1-0.1)}{200}} = 0.0212132$

Step 3: Calculate the z-test for the proportion:

$$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.165-0.10}{0.0212132} = 3.06413$$

Step 4: Calculate the probability of seeing a z-score of this magnitude *or larger*:

```
pnorm(q = 3.06413, lower.tail = F)
```

```
## [1] 0.00109152
```

Step 5: Evaluate the evidence against the null hypothesis. Because the p-value is so small (0.1%), there is little chance of seeing a proportion equal to 16.5% or larger if the true proportion is actually

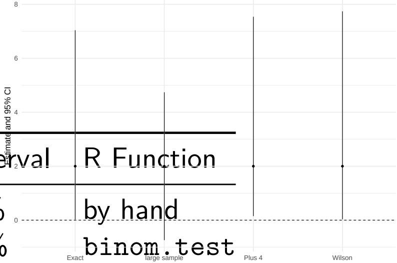
Another Example

Suppose that there were 100 elderly individuals with falls observed, and 2 died. Here are the 95% CIs applying the four different methods:

Method	95% Confidence Interval	R Function
Large sample	-0.74% to 4.74%	by hand
Exact	0.024% to 7.04%	`binom.test`
Wilson Score*	0.034% to 7.74%	`prop.test`
Plus four	0.15% to 7.54%	by hand

*with continuity correction

We can graphically compare the CIs from the previous slide:



Elderly falls example

Findings:

- Notice how different the intervals are, especially large sample vs. others
- Notice that the large sample lower bound is nonsensical (i.e., we can't have negative proportions!)
- The large sample CI differs from the others because the Normal approximation assumptions are not satisfied

Code for elderly falls example

```
p.hat <- 2/100 # estimate pro
se <- sqrt(p.hat*(1-p.hat)/100) # standard error
c(p.hat - 1.96*se, p.hat + 1.96*se) # CI
```

```
## [1] -0.00744  0.04744
```

```
binom.test(x = 2, n = 100, p = 0.5, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  2 and 100
## number of successes = 2, number of trials = 100, p-value
## alternative hypothesis: true probability of success is no
## 95 percent confidence interval:
##  0.002431337 0.070383932
## sample estimates:
## probability of success
##                   0.02
```

```
prop.test(x = 2, n = 100, p =
```

```
##
##  1-sample proportions test
##
## data:  2 out of 100, null p
## X-squared = 90.25, df = 1,
## alternative hypothesis: tru
## 95 percent confidence inter
##  0.003471713 0.077363988
## sample estimates:
##    p
## 0.02
```

```
p.tilde <- (2 + 2)/(100 + 4)
se <- sqrt(p.tilde*(1-p.tilde)
c(p.tilde - 1.96*se, p.tilde +
```

```
## [1] 0.00150119 0.07542189
```

Check your understanding!

Recap

- The binomial distribution has all of the same statistical "procedures":
  - Confidence intervals
  - Sample size estimation
  - Hypothesis testing
- However, there are problems with the coverage of the CIs:
  - Subtle interactions between discrete binomial distribution and Normal-based approximations
  - Use "Plus 4" method when calculating "by hand" and sample is "small"
  - Use `prop.test` when using R functions
  - The coverage problem goes away as the same size gets larger, and more of the Normality assumptions are met