Inference for Proportions and 2×2 Tables

Textbook Sections 8.1, 8.2

Emile Latour, Nicky Wakim, Meike Niederhausen

February 23, 2026

Learning Objectives

By the end of today's lecture, you will be able to:

1. Apply the normal approximation to the sampling distribution of a sample proportion
2. Conduct hypothesis tests for a single proportion
3. Construct and interpret confidence intervals for a single proportion
4. Test for differences between two independent proportions
5. Interpret data displayed in 2×2 tables
6. Choose the appropriate inference method based on study design

Roadmap for Today

Part 1: Moving from Means to Proportions

- Why categorical outcomes matter
- From binomial to proportions
- Sampling distribution of $\hat{p}$
- Success-failure condition

Part 2: Single Proportion Inference

- The melanoma immunotherapy example
- Hypothesis testing for proportions
- Confidence intervals for proportions
- Using `prop.test()` in R

Part 3: Comparing Two Proportions

- Difference in proportions: $p_1 - p_2$
- The aspirin and heart attack example
- Two-sample tests with `prop.test()`
- Interpreting results

Part 4: Understanding 2×2 Tables

- Organizing categorical data
- Risk difference, relative risk, odds ratio
- When to use each measure
- Common mistakes to avoid

CI's and hypothesis testing for different scenarios (1/2)

Day	Section	Population parameter	Symbol	Point estimate	Symbol
9	5.1	Population mean	μ	Sample mean	$\bar{x}$
10	5.2	Population mean of paired differences	μ_d or δ	Sample mean of paired differences	$\bar{x}_d$
10	5.3	Differences in population means	$\mu_1 - \mu_2$	Differences in sample means	$\bar{x}_1 - \bar{x}_2$
13	**8.1**	**Population proportion**	p	**Sample proportion**	$\widehat{p}$
13	**8.2**	**Differences in population proportions**	$p_1 - p_2$	**Differences in sample proportions**	$\widehat{p}_1 - \widehat{p}_2$

Today we add proportions to our inference toolkit!

CI's and hypothesis testing for different scenarios (2/2)

$$\text{point estimate} \pm z^*(or\ t^*) \cdot SE$$

$$\text{test stat} = \frac{\text{point estimate} - \text{null value}}{SE}$$

Part 1: Moving from Means to Proportions

Where are we in the course?

We've been building up our inference toolkit for **numerical outcomes**:

So far:

- One-sample mean: μ
- Paired differences: μ_d
- Difference in means (independent): $\mu_1 - \mu_2$
- Understanding power and sample size

Today we shift to categorical outcomes:

- Single proportion: p
- Difference in proportions: $p_1 - p_2$
- Data organized in 2×2 tables

Why this matters: Many medical outcomes are categorical (disease/no disease, response/no response, alive/dead)!

Why do we care about categorical data?

Categorical Data in Medical Research

Categorical outcomes arise constantly in biomedical research because:

- Disease states are often binary: diabetes vs. no diabetes, cancer vs. no cancer
- Treatment responses: complete response, partial response, no response
- Vital status: alive vs. deceased at follow-up
- Screening results: positive vs. negative test
- Patient characteristics: smoker vs. non-smoker, male vs. female

Examples from recent research:

- Does a new immunotherapy increase the proportion of melanoma patients who respond?
- Do statins reduce the proportion of patients who experience cardiovascular events?
- Is there a difference in COVID-19 infection rates between vaccinated and unvaccinated individuals?

Reminder: The binomial distribution

From Week 4, we learned about binomial random variables:

Binomial Random Variable

X is a binomial random variable if it represents the **number of successes** in n independent trials where:

- Each trial has two possible outcomes: success or failure
- The probability of success is p (same for all trials)
- Trials are independent of each other

Notation: $X \sim \text{Binomial}(n, p)$

Parameters:

- n = number of trials
- p = probability of success
- Mean: $E(X) = np$
- SD: $SD(X) = \sqrt{np(1-p)}$

From counts to proportions

The **sample proportion** is just a rescaled version of the count:

$$\hat{p} = \frac{X}{n} = \frac{\text{number of successes}}{n}$$

Example: In a study of 52 melanoma patients treated with immunotherapy:

- 21 patients responded to treatment (success)
- $\hat{p} = \frac{21}{52} = 0.404$ or 40.4%

Important distinction:

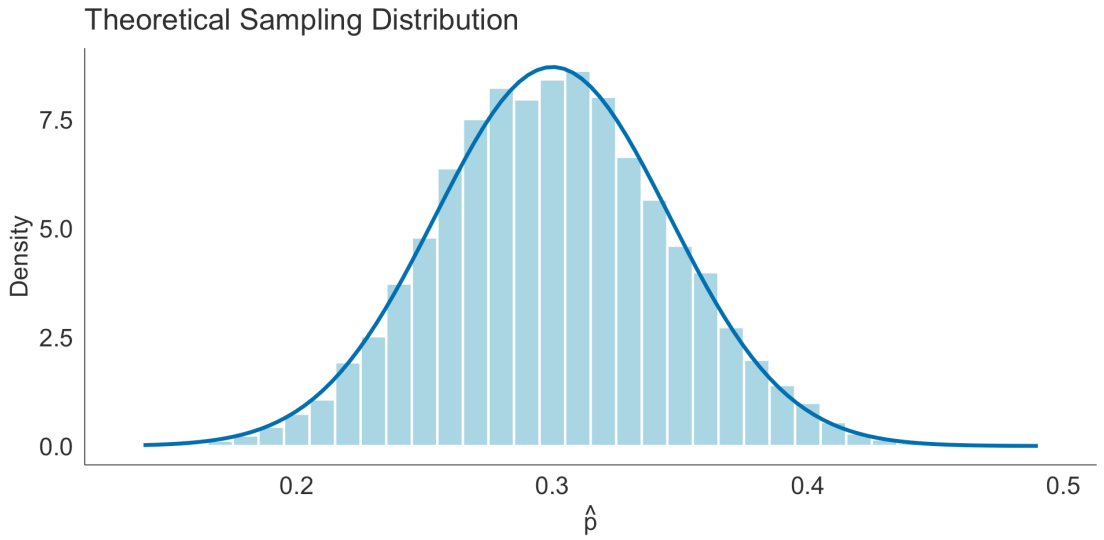
- X = number of successes (a count: 21)
- $\hat{p}$ = proportion of successes (a ratio: 0.404)
- Both give us information about the same thing, just scaled differently

The sampling distribution of $\hat{p}$

Just like with $\bar{x}$, the sample proportion $\hat{p}$ has a sampling distribution!

If we repeated the study many times with different samples:

- Each sample would give us a different $\hat{p}$
- These $\hat{p}$ values would cluster around the true p
- The distribution of $\hat{p}$ would be approximately normal (under certain conditions)



Sampling distribution of $\hat{p}$ (cont.)

- $\hat{p} = \frac{X}{n}$ where X is the number of "successes" and n is the sample size.
- $X \sim Bin(n, p)$, where p is the population proportion.
- For n "big enough", the normal distribution can be used to approximate a binomial distribution:

$$X \sim Bin(n, p) \longrightarrow X \sim N\left(\mu = np, \sigma = \sqrt{np(1-p)}\right)$$

- Since $\hat{p} = \frac{X}{n}$ is a linear transformation of X, we have for large n:

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right)$$

The success-failure condition

When Can We Use the Normal Approximation?

The sampling distribution of $\hat{p}$ is approximately normal when:

1. Independence: Observations are independent of each other

- Usually satisfied with random sampling
- Can be approximately satisfied in well-designed studies

2. Success-failure condition: At least 10 expected successes AND 10 expected failures

- **For confidence intervals:** Check $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
- **For hypothesis tests:** Check $np_0 \geq 10$ and $n(1-p_0) \geq 10$
  - Use the hypothesized p_0 for tests because we're checking conditions under H_0!

Why does the success-failure condition matter?

The binomial distribution is **discrete** (counts: 0, 1, 2, 3, ...)

The normal distribution is **continuous**

When n is small or p is extreme:

- The binomial distribution is skewed
- Normal approximation is poor
- Our p-values and CIs will be inaccurate

When $np \geq 10$ and $n(1-p) \geq 10$:

- The binomial distribution looks approximately symmetric
- Normal approximation works well
- Our inferences are valid

Visual intuition: Think about flipping a coin 5 times ($n = 5, p = 0.5$) vs. 50 times ($n = 50, p = 0.5$). Which will look more normal?

Part 2: Single Proportion Inference

Reminder: The six steps of hypothesis testing

We'll use our familiar framework for proportions:

1. **State hypotheses** (H_0 and H_A)
2. **Set significance level** (usually α = 0.05)
3. **Check assumptions** (independence, success-failure condition)
4. **Calculate test statistic**
   - For proportions: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
5. **Find p-value** (probability of seeing this or more extreme)
6. **Make conclusion** (reject or fail to reject H_0, with context)

Note: We use z not t because we know the population SD exactly under H_0!

Today's example: Advanced melanoma immunotherapy

Research Context

Background:

- Advanced melanoma is an aggressive form of skin cancer
- Until recently, it was almost uniformly fatal
- Researchers noticed rare cases where patients' immune systems successfully attacked the cancer
- This led to development of immunotherapy drugs

Study (Wolchok et al., 2013 in *NEJM*):

- 52 patients with advanced melanoma were treated with **two new therapies concurrently**: nivolumab and ipilimumab
- Both drugs work by releasing "brakes" on the immune system
- **Outcome:** Response to therapy (tumor shrinkage or disappearance)
- **Result:** 21 of 52 patients (40%) responded

Historical context: Previous studies with single-agent therapy showed response rates of 30% or less.

Research question

The Question

Do these results provide evidence that the response probability with **concurrent** therapy is **greater than 30%** (the historical benchmark)?

Why this matters:

- Combination therapy is more toxic and expensive than single-agent
- We need evidence it's better before recommending it widely
- This was an early study to justify larger clinical trials

Let's walk through the six steps!

Step 1: State the hypotheses

In symbols:

$$H_0 : p = 0.30$$

$$H_A : p > 0.30$$

In words:

- H_0: The proportion of advanced melanoma patients who respond to concurrent immunotherapy is 30% (same as historical single-agent rate)
- H_A: The proportion who respond to concurrent immunotherapy is **greater than** 30%

Note: This is a one-sided test because we only care if the new treatment is *better*. We wouldn't adopt it if it were worse!

Step 2: Set the significance level

We'll use the conventional $\alpha = 0.05$

Interpretation: We're willing to accept a 5% chance of incorrectly rejecting H_0 (Type I error)

Note: In practice, one-sided tests are sometimes held to $\alpha = 0.025$ to maintain stringency comparable to a two-sided test at $\alpha = 0.05$. We'll use $\alpha = 0.05$ here for simplicity.

Step 3: Check the assumptions

```r
# Study data
n <- 52            # Sample size
x <- 21            # Number of responders
p_hat <- x / n     # Sample proportion
p0 <- 0.30         # Null hypothesis value

# Print values
p_hat
```

```
[1] 0.4038462
```

1. Independence: Patients in a clinical trial are reasonably independent (one patient's response doesn't affect another's)

2. Success-failure condition under H_0:

```r
n * p0             # Expected successes
```

```
[1] 15.6
```

```r
n * (1 - p0)       # Expected failures
```

```
[1] 36.4
```

Both ≥ 10 ✓ — the normal approximation is appropriate!

Step 4: Calculate the test statistic

Formula:

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

```r
# Calculate standard error under H0
SE_p <- sqrt(p0 * (1 - p0) / n)
SE_p
```

```
[1] 0.06354889
```

```r
# Calculate z-statistic
z_stat <- (p_hat - p0) / SE_p
z_stat
```

```
[1] 1.634114
```

Our observed proportion (0.404) is about **1.64 standard errors** above the null value (0.30).

Step 5: Find the p-value

The p-value is the probability of observing $\hat{p} \geq 0.404$ (or more extreme) if H_0 were true.

```r
# One-sided p-value (upper tail)
p_value <- pnorm(abs(z_stat), lower.tail = FALSE)
p_value
```

```
[1] 0.05111742
```

Interpretation:

- If the true response rate were 30.0%, we'd see a sample proportion of 40.4% or higher in about **5.1%** of samples of size 52.

Step 6: Make a conclusion

Statistical conclusion:

At $\alpha = 0.05$, the p-value = 0.051 is right at our threshold. We would barely **fail to reject** H_0.

Contextual conclusion:

The data provide **marginal evidence** that concurrent immunotherapy produces a response rate higher than the historical 30% benchmark. However:

- The result is borderline significant
- This is a small, early-phase study
- Further investigation with larger trials is warranted

Why researchers continued: The 40.4% response rate, even if barely not significant statistically, is **clinically meaningful** for patients with a disease that was previously nearly 100% fatal!

Confidence intervals for proportions

We can also construct a confidence interval for p:

Formula:

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

Important difference from hypothesis testing:

- For **CIs**, use $\hat{p}$ in the SE formula (because we're estimating the SE)
- For **tests**, use p_0 in the SE formula (because we're checking conditions under H_0)

Calculating a 95% CI for the melanoma data

We have a one sided test where we were testing if the proportion who respond to concurrent immunotherapy is *greater than* 0.30, a minimum threshold, not just different than it.

```r
# Calculate SE using p-hat (not p0!)
SE_ci <- sqrt(p_hat * (1 - p_hat) / n)
SE_ci
```

```
[1] 0.06804332
```

```r
# Get z* for 95% CI (one-sided)
z_star <- qnorm(0.95)
z_star
```

```
[1] 1.644854
```

```r
# Calculate margin of error
margin <- z_star * SE_ci
margin
```

```
[1] 0.1119213
```

```r
# Construct one-sided
ci_lower <- p_hat - margin

ci_lower
```

```
[1] 0.2919249
```

Calculating a 95% CI for the melanoma data (two-sided)

If we had wanted a two-sided 95% confidence interval:

```r
# Calculate SE using p-hat (not p0!)
SE_ci <- sqrt(p_hat * (1 - p_hat) / n)
SE_ci
```

```
[1] 0.06804332
```

```r
# Get z* for 95% CI (two-sided)
z_star <- qnorm(0.975)
z_star
```

```
[1] 1.959964
```

```r
# Calculate margin of error
margin <- z_star * SE_ci
margin
```

```
[1] 0.1333625
```

```r
# Construct two-sided 95% CI
ci_lower <- p_hat - margin
ci_upper <- p_hat + margin

c(ci_lower, ci_upper)
```

```
[1] 0.2704837 0.5372086
```

The `prop.test()` function

R has a built in function for performing hypothesis test of proportions:

```
prop.test(x, n,
          p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95,
          correct = TRUE)
```

Key arguments:

- `x` = number of successes
- `n` = number of trials
- `p` = null value (p_0)
- `alternative` = "two.sided", "less", or "greater"
- `conf.level` = confidence level (default 0.95)
- `correct` = whether to use continuity correction or not

Melanoma data with `prop.test()`

Instead of calculating by hand, we can use R's built-in function:

```r
# One-sample proportion test
test_result <- prop.test(
  x = 21,                    # Number of successes
  n = 52,                    # Sample size
  p = 0.30,                  # Null hypothesis value
  alternative = "greater",   # One-sided test
  conf.level = 0.95,
  correct = FALSE            # Default is TRUE, set to FALSE for now
)

test_result
```

```
    1-sample proportions test without continuity correction

data:  21 out of 52, null probability 0.3
X-squared = 2.6703, df = 1, p-value = 0.05112
alternative hypothesis: true p is greater than 0.3
95 percent confidence interval:
 0.2993794 1.0000000
sample estimates:
        p
0.4038462
```

Different confidence intervals

> **Note**
>
> `prop.test()` reports a Wilson score confidence interval, not the Wald interval from our hand calculation — hence the slight difference in the lower bound. Wilson is preferred in practice.
>
> Affects two-sided and one-sided intervals. Two-sided shown here.

From the by hand calculations

```
c(ci_lower, ci_upper)
```

```
[1] 0.2704837 0.5372086
```

From `prop.test()`

```
test_result_2sided <- prop.test(x = 21, n = 52,
                                p = 0.30,
                                alternative = "two.sided",
                                conf.level = 0.95,
                                correct = FALSE)

test_result_2sided %>%
  tidy() %>%
  dplyr::select(conf.low, conf.high)
```

```
# A tibble: 1 × 2
  conf.low conf.high
     <dbl>     <dbl>
1    0.282     0.539
```

Continuity correction: review

When we approximate the binomial distribution with a normal distribution, we apply a **continuity correction (CC)** to account for the fact that the binomial is discrete while the normal is continuous.

We adjust the number of successes by ±0.5 before calculating the probability:

Goal	Binomial	Normal approximation
$P(X \leq k)$	exact	use $P(X \leq k + 0.5)$
$P(X \geq k)$	exact	use $P(X \geq k - 0.5)$

`prop.test()` applies this same correction to the z-test for proportions (`correct = TRUE` is the default).

CC matters more with small samples

Without CC (`correct = FALSE`)

```r
prop.test(x = 21, n = 52, p = 0.30,
          alternative = "greater",
          correct = FALSE)
```

```
    1-sample proportions test without continuity
correction

data:  21 out of 52, null probability 0.3
X-squared = 2.6703, df = 1, p-value = 0.05112
alternative hypothesis: true p is greater than 0.3
95 percent confidence interval:
 0.2993794 1.0000000
sample estimates:
        p
0.4038462
```

With CC (`correct = TRUE`)

```r
prop.test(x = 21, n = 52, p = 0.30,
          alternative = "greater",
          correct = TRUE)
```

```
    1-sample proportions test with continuity
correction

data:  21 out of 52, null probability 0.3
X-squared = 2.1987, df = 1, p-value = 0.06906
alternative hypothesis: true p is greater than 0.3
95 percent confidence interval:
 0.2906582 1.0000000
sample estimates:
        p
0.4038462
```

CC matters less with large samples

Without CC (`correct = FALSE`)

```r
prop.test(x = 210, n = 520, p = 0.30,
          alternative = "greater",
          correct = FALSE)
```

```
    1-sample proportions test without continuity
correction

data:  210 out of 520, null probability 0.3
X-squared = 26.703, df = 1, p-value = 1.186e-07
alternative hypothesis: true p is greater than 0.3
95 percent confidence interval:
 0.3690394 1.0000000
sample estimates:
        p
0.4038462
```

With CC (`correct = TRUE`)

```r
prop.test(x = 210, n = 520, p = 0.30,
          alternative = "greater",
          correct = TRUE)
```

```
    1-sample proportions test with continuity
correction

data:  210 out of 520, null probability 0.3
X-squared = 26.211, df = 1, p-value = 1.53e-07
alternative hypothesis: true p is greater than 0.3
95 percent confidence interval:
 0.3680964 1.0000000
sample estimates:
        p
0.4038462
```

Takeaway:

p-values and CIs converge as n grows — the correction becomes negligible. For small samples, CC makes the test more conservative (larger p-value), which is the safer choice.

What should I use in practice?

Recommendation: Use `prop.test()` defaults

For inference on proportions, use `prop.test()` with its default settings:

- **Wilson score interval** — better than Wald, especially for small samples or extreme proportions
- **Continuity correction** (`correct = TRUE`) — more conservative, accounts for the discrete-to-continuous approximation

```r
# This is what you should use by default
prop.test(x = 21, n = 52,
          p = 0.30,
          alternative = "greater",
          conf.level = 0.95)   # correct = TRUE is the default!
```

Why did we show `correct = FALSE` earlier?

To match our hand calculations and illustrate the method. In practice, always let R apply the continuity correction unless you have a specific reason not to.

Back to the Melanoma data with `prop.test()`

Instead of calculating by hand, we can use R's built-in function:

```r
# One-sample proportion test
test_result <- prop.test(
  x = 21,                    # Number of successes
  n = 52,                    # Sample size
  p = 0.30,                  # Null hypothesis value
  alternative = "greater",   # One-sided test
  conf.level = 0.95,
  correct = TRUE             # Use continuity correction
)


library(broom)
tidy(test_result)
```

```
# A tibble: 1 × 8
  estimate statistic p.value parameter conf.low conf.high method     alternative
     <dbl>     <dbl>   <dbl>     <int>    <dbl>     <dbl> <chr>      <chr>
1    0.404      2.20  0.0691         1    0.291         1 1-sample … greater
```

Conclusions from the Melanoma data

Key pieces:

- `estimate`: $\hat{p} = 0.404$ (our sample proportion)
- `statistic`: $\chi^2 = 2.199$ (different test statistic than z, but equivalent)
- `p.value`: $p = 0.069$ (slightly different due to continuity correction)
- `conf.low`, `conf.high`: (0.291, 1.000), one-sided 95% CI for p

Note: The continuity correction makes the test more conservative (higher p-value) than when we calculated $p = 0.051$ before.

Interpretation

The actual percentage of advanced melanoma patients who respond to concurrent immunotherapy is not significantly greater than 30% (one-sided *p*-value = 0.069, one-sample proportions test with continuity correction). Based on the study, it is estimated that 40% of those with advanced melanoma respond to concurrent immunotherapy (95% one-sided CI: at least 29%).

`prop_test` function in `Rstatix`

The `rstatix` package offers a tidy-friendly wrapper that integrates well with pipelines.

```
library(rstatix)
prop_test(
  x = 21,                          # Number of successes
  n = 52,                          # Sample size
  p = 0.30,                        # Null hypothesis value
  alternative = "greater",         # One-sided test
  conf.level = 0.95,
  correct = TRUE,                  # Use continuity correction
  detailed = TRUE
)
```

```
# A tibble: 1 × 11
      n    n1 estimate statistic      p    df conf.low conf.high method
* <dbl> <dbl>    <dbl>     <dbl>  <dbl> <int>    <dbl>     <dbl> <chr>
1    52    21    0.404      2.20 0.0691     1    0.291         1 Prop test
# ℹ 2 more variables: alternative <chr>, p.signif <chr>
```

Exact binomial test: motivation

Our z-test for proportions relies on the **normal approximation** to the binomial.

But what if we want an **exact** result — no approximation needed?

The exact binomial test

Instead of approximating with a normal distribution, `binom.test()` computes the p-value **directly from the binomial distribution**:

$$p\text{-value} = P(X \geq 21 \mid n = 52, p_0 = 0.30)$$

- No success-failure condition required
- Works for any sample size, any p
- Always valid — but less commonly taught because it's harder to do by hand

When to use it: Small samples, extreme proportions, or when you want exact inference without relying on the normal approximation.

Exact binomial test with `binom.test()`

```r
binom_res <- binom.test(
  x = 21,                    # Number of successes
  n = 52,                    # Sample size
  p = 0.30,                  # Null hypothesis value
  alternative = "greater",  # One-sided test
  conf.level = 0.95
)

binom_res
```

```
    Exact binomial test

data:  21 and 52
number of successes = 21, number of trials = 52, p-value = 0.07167
alternative hypothesis: true probability of success is greater than 0.3
95 percent confidence interval:
 0.2889045 1.0000000
sample estimates:
probability of success
             0.4038462
```

Comparing `binom.test()` to `prop.test()`

Exact binomial (`binom.test`)

```
tidy(binom_res) %>%
  dplyr::select(estimate, p.value,
                conf.low, conf.high)
```

```
# A tibble: 1 × 4
  estimate p.value conf.low conf.high
     <dbl>   <dbl>    <dbl>     <dbl>
1    0.404  0.0717    0.289         1
```

- p-value computed directly from binomial
- Uses **Clopper-Pearson** CI (exact)
- Always valid

Normal approximation (`prop.test`)

```
tidy(test_result) %>%
  dplyr::select(estimate, p.value,
                conf.low, conf.high)
```

```
# A tibble: 1 × 4
  estimate p.value conf.low conf.high
     <dbl>   <dbl>    <dbl>     <dbl>
1    0.404  0.0691    0.291         1
```

- p-value based on normal approximation + CC
- Uses **Wilson score** CI
- Requires success-failure condition

Takeaway

Results are similar here because $n = 52$ is large enough for the normal approximation to work well. For small samples or extreme p, they can diverge more — and `binom.test()` is the safer choice.

Confidence intervals: three methods compared

Note: Switching to a two-sided 95% CI here to better illustrate how the three methods differ at both bounds.

Method	Lower	Upper	Width
Wald (by hand)	0.2705	0.5372	0.2667
Wilson (prop.test)	0.2731	0.5487	0.2756
Clopper-Pearson (binom.test)	0.2701	0.5490	0.2789

- **Wald**: simplest formula, least accurate (especially at boundaries)
- **Wilson**: good balance of accuracy and simplicity — `prop.test()` default
- **Clopper-Pearson**: exact, slightly conservative (wider interval)

In practice: Wilson or Clopper-Pearson are both defensible. Wald is taught for conceptual understanding, not for real use.

Part 3: Comparing Two Proportions

From one proportion to two

We've learned to make inferences about a **single** proportion.

Now: What if we want to **compare** proportions from two independent groups?

Examples:

- Does aspirin reduce the proportion of people who have heart attacks compared to placebo?
- Is the proportion of smokers higher among lung cancer patients than controls?
- Does a new vaccine produce antibody responses in a higher proportion than the old vaccine?

This is analogous to:

- One-sample mean (μ) → Two independent sample means ($\mu_1 - \mu_2$)

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

Just like with the difference in means, we need to know the sampling distribution of the difference in proportions!

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

When conditions are met, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Mean: $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

Standard error: $SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Note the similarity to the SE for difference in means!

Conditions for comparing two proportions

Success-Failure Condition for Two Proportions

For confidence intervals:

- At least 10 successes and 10 failures in **each** group
- Check:
  - $n_1\hat{p}_1 \geq 10$,
  - $n_1(1-\hat{p}_1) \geq 10$,
  - $n_2\hat{p}_2 \geq 10$,
  - $n_2(1-\hat{p}_2) \geq 10$

For hypothesis tests (testing $H_0 : p_1 = p_2$):

- Use the **pooled proportion** to check conditions
- Pooled: $\hat{p}_{pool} = \frac{x_1+x_2}{n_1+n_2}$ (combines both groups under H_0)
- Check:
  - $n_1\hat{p}_{pool} \geq 10$,
  - $n_1(1-\hat{p}_{pool}) \geq 10$,
  - $n_2\hat{p}_{pool} \geq 10$,
  - $n_2(1-\hat{p}_{pool}) \geq 10$

Independence

- Observations within each group are independent
- The two groups are independent of each other

Example: Aspirin and heart attacks

The Physicians' Health Study

Study design (1989):

- 22,071 male physicians enrolled
- **Random assignment** to one of two groups:
  - Aspirin: 325 mg every other day (n = 11,037)
  - Placebo: Identical-looking pill (n = 11,034)
- Follow-up for heart attacks over ~5 years

Results:

- Aspirin group: 104 heart attacks
- Placebo group: 189 heart attacks

Research question: Does aspirin reduce the risk of heart attack in healthy men?

Setting up the hypothesis test

```r
# Study data
n_aspirin <- 11037
n_placebo <- 11034
x_aspirin <- 104   # Heart attacks in aspirin group
x_placebo <- 189   # Heart attacks in placebo group

# Calculate sample proportions
p_hat_aspirin <- x_aspirin / n_aspirin
p_hat_placebo <- x_placebo / n_placebo

p_hat_aspirin
```

```
[1] 0.00942285
```

```r
p_hat_placebo
```

```
[1] 0.01712887
```

Observed difference:

```r
diff_hat <- p_hat_aspirin - p_hat_placebo
diff_hat
```

```
[1] -0.007706024
```

The aspirin group had about **0.77%** fewer heart attacks (about 8 per 1000 people).

Hypotheses for comparing two proportions

Let:

- $p_{aspirin}$ = proportion who have heart attacks in the aspirin group
- $p_{placebo}$ = proportion who have heart attacks in the placebo group

Hypotheses:

$$H_0 : p_{aspirin} = p_{placebo}$$
$$H_A : p_{aspirin} < p_{placebo}$$

Or equivalently

$$H_0 : p_{aspirin} - p_{placebo} = 0$$
$$H_A : p_{aspirin} - p_{placebo} < 0$$

In words:

- H_0: Aspirin and placebo have the same heart attack rate
- H_A: Aspirin has a **lower** heart attack rate than placebo

This is a **one-sided** test (we're only interested in whether aspirin *reduces* risk).

Check assumptions

1. Independence:

- Random assignment ✓
- Observations within groups are independent ✓

2. Success-failure condition:

For a hypothesis test, we check using the **pooled proportion** under H_0:

```
# Pooled proportion
p_pool <- (x_aspirin + x_placebo) / (n_aspirin + n_placebo)
p_pool
```

```
[1] 0.01327534
```

```
# Check success-failure for each group
```

```
# Expected "successes" (heart attacks) in aspirin
n_aspirin * p_pool
```

```
[1] 146.5199
```

```
# Expected "successes" in placebo
n_placebo * p_pool
```

```
[1] 146.4801
```

```
# Expected "failures" in aspirin
n_aspirin * (1 - p_pool)
```

```
[1] 10890.48
```

```
# Expected "failures" in placebo
n_placebo * (1 - p_pool)
```

```
[1] 10887.52
```

All ≥ 10 ✓ — normal approximation is valid!

Using R: `prop.test()` for two proportions

```r
# Two-sample proportion test
aspirin_test <- prop.test(
  x = c(x_aspirin, x_placebo),    # Vector of successes: c(104, 189)
  n = c(n_aspirin, n_placebo),    # Vector of sample sizes: c(11037, 11034)
  alternative = "less",           # Aspirin < Placebo
  conf.level = 0.95,
  correct = TRUE
)

aspirin_test
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(x_aspirin, x_placebo) out of c(n_aspirin, n_placebo)
X-squared = 24.429, df = 1, p-value = 3.855e-07
alternative hypothesis: less
95 percent confidence interval:
 -1.000000000 -0.005082393
sample estimates:
    prop 1     prop 2
0.00942285 0.01712887
```

Interpreting the results

```
1 tidy(aspirin_test)
```

```
# A tibble: 1 × 9
  estimate1 estimate2 statistic     p.value parameter conf.low conf.high method
      <dbl>     <dbl>     <dbl>       <dbl>     <dbl>    <dbl>     <dbl> <chr>
1   0.00942    0.0171      24.4 0.000000385         1       -1  -0.00508 2-samp…
# ℹ 1 more variable: alternative <chr>
```

Conclusion

The percentage of male physicians that experienced a heart attack in the 5-year follow-up was 0.94% in the aspirin group compared to 1.71% in the placebo group. There is strong evidence that aspirin reduces the risk of heart attack ($p < 0.001$, two-sample test of proportions). The true difference in heart attack rates is estimated to be 0.77 percentage points lower in the aspirin group (95% one-sided CI: at least 0.51 percentage points lower).

Clinical vs. statistical significance

Statistical significance: $p < 0.001$ — the difference is real, not due to chance.

Clinical significance: Is an absolute risk reduction of 0.77 percentage points (from 1.71% to 0.94%) clinically meaningful?

Consider:

- **Number Needed to Treat (NNT):** How many people need to take aspirin to prevent one heart attack?

```
# NNT = 1 / absolute risk reduction
NNT <- 1 / abs(diff_hat)
NNT
```

```
[1] 129.7686
```

About **130 people** (men) need to take aspirin for 5 years to prevent **one** heart attack.

Trade-off: Aspirin has side effects (bleeding, ulcers). Is preventing one heart attack per 130 people (men) worth the risk?

For healthy individuals: Guidelines now suggest aspirin primarily for those at higher cardiovascular risk!

Part 4: Understanding 2×2 Tables

Organizing categorical data in tables

When we have two categorical variables, we often display the data in a **contingency table** (also called a **cross-tabulation** or **2×2 table** when each variable has two categories).

General structure:

	Group 1	Group 2	Total
Outcome Y	a	b	$a+b$
Outcome N	c	d	$c+d$
Total	$a+c$	$b+d$	n

Example (Aspirin study):

	Aspirin	Placebo	Total
Heart attack	104	189	293
No heart attack	10,933	10,845	21,778
Total	11,037	11,034	22,071

Creating 2×2 tables in R

```r
# Create the data as a table
aspirin_table <- matrix(
  c(104, 189,      # Heart attacks (aspirin, placebo)
    10933, 10845), # No heart attacks
  nrow = 2,
  byrow = TRUE,
  dimnames = list(
    Outcome = c("Heart attack", "No heart attack"),
    Group = c("Aspirin", "Placebo")
  )
)

aspirin_table
```

```
                Group
Outcome          Aspirin Placebo
  Heart attack       104     189
  No heart attack  10933   10845
```

We can also use `janitor::tabyl()` with raw data, or `table()` for quick summaries.

Three ways to measure association

From a 2×2 table, we can calculate three important measures:

1. Risk Difference (RD) — also called Absolute Risk Reduction (ARR)

$$RD = p_1 - p_2$$

"How much does the risk change in absolute terms?"

2. Relative Risk (RR) — also called Risk Ratio

$$RR = \frac{p_1}{p_2}$$

"How many **times** higher/lower is the risk?"

3. Odds Ratio (OR)

$$OR = \frac{\text{odds}_1}{\text{odds}_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

"How much do the **odds** of the outcome change?"

Calculating measures for the aspirin study

```r
# Proportions
p_aspirin <- 104 / 11037
p_placebo <- 189 / 11034

# Risk Difference
RD <- p_aspirin - p_placebo
RD
```

```
[1] -0.007706024
```

```r
# Relative Risk
RR <- p_aspirin / p_placebo
RR
```

```
[1] 0.550115
```

```r
# Odds Ratio
odds_aspirin <- p_aspirin / (1 - p_aspirin)
odds_placebo <- p_placebo / (1 - p_placebo)
OR <- odds_aspirin / odds_placebo
OR
```

```
[1] 0.5458355
```

Interpreting the measures

Risk Difference = -0.77 percentage points

- Aspirin **reduces** absolute risk by -0.77 percentage points
- Out of 1000 people, about 8 fewer heart attacks with aspirin
- **Not percent or %**

Relative Risk = 0.55 times the risk

- People taking aspirin have **0.55 times** the risk (or 55% of the risk)
- Equivalently: aspirin reduces risk by **45%** (1 - 0.55 = 0.45)
- This sounds more impressive than 0.77 percentage points!

Odds Ratio = 0.55

- The odds of a heart attack among those taking aspirin are 0.55 times the odds among those taking a placebo
- When risks are small (like 1-2%), OR ≈ RR

When to use each measure

Choosing the Right Measure

Risk Difference (Absolute Risk Reduction):

- Most intuitive for patients and clinicians
- Useful for calculating Number Needed to Treat ($NNT = 1/RD$)
- **Use when:** Communicating clinical impact

Relative Risk:

- Shows proportional change in risk
- Often used in epidemiology
- **Use when:** Comparing effect sizes across studies with different baseline risks
- **Be careful:** Can make small absolute differences seem large

Odds Ratio:

- Required for certain study designs (case-control studies)
- Approximates RR when outcome is rare (<10%)
- **Use when:** Doing logistic regression or case-control studies

Common mistakes with 2×2 tables

1. Confusing relative and absolute risk

"Aspirin reduces heart attack risk by 45%!" (Relative)

vs.

"Aspirin reduces heart attack risk from 1.7% to 0.9%" (Absolute)

Both are true, but they feel very different!

2. Not considering baseline risk

- A 50% relative reduction means more when baseline risk is 40% (→20%) than when it's 2% (→1%)
- Always report both relative and absolute effects

3. Misinterpreting odds ratios as relative risks

- When outcomes are common (>10%), OR and RR diverge
- OR always exaggerates effect size when RR > 1 or RR < 1

Example: Misinterpreting odds ratios

Suppose a treatment increases recovery from 40% to 60%:

```r
# Calculate both measures
p_treated <- 0.60
p_control <- 0.40

RR <- p_treated / p_control
RR
```

```
[1] 1.5
```

```r
odds_treated <- p_treated / (1 - p_treated)
odds_control <- p_control / (1 - p_control)
OR <- odds_treated / odds_control
OR
```

```
[1] 2.25
```

- **RR = 1.5:** Treatment increases recovery rate by 50%
- **OR = 2.25:** Odds of recovery are 2.25 times higher

If you report OR = 2.25 as "125% increase in recovery," you're overstating the effect! When outcomes are common, stick with RR.

Practice: Interpreting a 2×2 table

A study investigates whether screening mammography reduces breast cancer mortality:

	Screened	Not screened	Total
Died from breast cancer	500	505	1,005
Did not die	63,500	63,495	126,995
Total	64,000	64,000	128,000

Calculate:

1. The proportion who died in each group
2. Risk difference
3. Relative risk
4. Interpret: Does screening reduce mortality?

Solution on next slide...

Practice solution

```r
# Proportions
p_screened <- 500 / 64000
p_not_screened <- 505 / 64000

p_screened
```

```
[1] 0.0078125
```

```r
p_not_screened
```

```
[1] 0.007890625
```

```r
# Risk Difference
RD_screen <- p_screened - p_not_screened
RD_screen
```

```
[1] -7.8125e-05
```

```r
# Relative Risk
RR_screen <- p_screened / p_not_screened
RR_screen
```

```
[1] 0.990099
```

Interpretation: Screening reduces breast cancer mortality from 0.789% to 0.781%, a very small absolute difference (RD = -0.00008). The relative risk is estimated to be 0.99, meaning screened women have 99% the mortality of unscreened (a 1 percentage point relative reduction). Whether this benefit justifies screening depends on costs, false positives, and patient preferences!

Summary and Key Takeaways

What we learned today

Core concepts:

1. **Single proportion inference**
   - Sampling distribution of $\hat{p}$ is approximately normal when success-failure condition is met
   - Use `prop.test()` for both hypothesis tests and confidence intervals
   - SE for tests uses p_0; SE for CIs uses $\hat{p}$
2. **Two proportion inference**
   - Compare independent proportions with $\hat{p}_1 - \hat{p}_2$
   - Check success-failure condition for **both** groups
   - Pooled proportion used for hypothesis tests under H_0
3. **2×2 tables and measures of association**
   - Risk difference: absolute change
   - Relative risk: proportional change
   - Odds ratio: for case-control studies and regression
   - Each measure tells a different story!

Looking ahead

Next class:

- Chi-squared tests for larger contingency tables (3×3, 4×2, etc.)
- Fisher's exact test for small samples
- Testing independence vs. testing association

Key formulas to know about (not memorize)

Single proportion:

- SE (test): $\sqrt{\frac{p_0(1-p_0)}{n}}$
- SE (CI): $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Test statistic: $z = \frac{\hat{p}-p_0}{SE}$

Two proportions:

- SE: $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- For tests, use pooled $\hat{p}$ in SE

2×2 table measures:

- Risk Difference: $p_1 - p_2$
- Relative Risk: p_1/p_2
- Odds Ratio: $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$

Final thoughts: Why proportions matter in medicine

Many critical health outcomes are binary:

- Survival vs. death
- Disease vs. no disease
- Response vs. no response
- Adverse event vs. no adverse event

Public health impact:

- Even small absolute differences can matter at population scale
- A vaccine that reduces infection risk by 1% (absolute) could prevent millions of cases
- Understanding both relative and absolute effects is essential for informed decision-making

Your job as a researcher:

- Report both absolute and relative effects
- Consider baseline risk and clinical context
- Don’t overstate findings by cherry-picking metrics!