Lecture17: Inference for Two Proportions

STAT 515 — Comparing p_1 and p_2

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Outline

1. From one proportion to two
2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$
3. Confidence interval for $p_1 - p_2$
4. Hypothesis tests, and why we pool under H_0
5. Example: youth gambling in Minnesota
6. A one-sided example, and summary

Outline

1. From one proportion to two
2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$
3. Confidence interval for $p_1 - p_2$
4. Hypothesis tests, and why we pool under H_0
5. Example: youth gambling in Minnesota
6. A one-sided example, and summary

Where We Are

- **One sample, one proportion:** is $p = p_0$? We used $\hat{p} = x/n$ and the Central Limit Theorem.
- **Two samples, two means:** is $\mu_1 - \mu_2 = 0$? We used $\overline{X_1} - \overline{X_2}$ and a z or t reference curve.

Today

Put the two together. The response is **categorical** — success or failure, yes or no, improved or not — and we have **two independent groups**. The parameter of interest is the **difference of two proportions**, $p_1 - p_2$.

- **The logic is identical to Lecture 16.** Estimate, standard error, normal curve. Only the formula for the standard error changes.

A Motivating Study: Youth Gambling in Minnesota

Stinchfield, University of Minnesota (Winter 2001)

With the rapid growth of legalized gambling in the United States, there is concern that the involvement of youth in gambling is also increasing. Professor Randy Stinchfield compared rates of gambling among Minnesota public school students. Based on survey data, the table gives the number of ninth-grade boys who gambled **weekly or daily** on any game — cards, sports betting, lotteries — in two survey years.

	1992	**1998**
Ninth-grade boys surveyed	$n_1 = 21{,}484$	$n_2 = 23{,}199$
Gambled weekly or daily	$x_1 = 4{,}684$	$x_2 = 5{,}313$
Sample proportion	$\hat{p}_1 = \frac{4684}{21484} = .218$	$\hat{p}_2 = \frac{5313}{23199} = .229$

Are the two percentages **significantly different**?

Notation: Two Populations, Two Samples

	Population 1	Population 2
True proportion of successes	p_1	p_2
True proportion of failures	$q_1 = 1 - p_1$	$q_2 = 1 - p_2$
Sample size	n_1	n_2
Number of successes	x_1	x_2
Sample proportion	$\hat{p}_1 = x_1/n_1$	$\hat{p}_2 = x_2/n_2$
	$\hat{q}_1 = 1 - \hat{p}_1$	$\hat{q}_2 = 1 - \hat{p}_2$

- The parameter of interest is $p_1 - p_2$; the natural estimate is $\hat{p}_1 - \hat{p}_2$.
- **Decide which population is "1" before you start** and keep it fixed — the sign of every answer depends on it.

Outline

1. From one proportion to two
2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$
3. Confidence interval for $p_1 - p_2$
4. Hypothesis tests, and why we pool under H_0
5. Example: youth gambling in Minnesota
6. A one-sided example, and summary

Recall: One Sample Proportion

The one-sample fact

For a large sample, by the Central Limit Theorem,

$$\hat{p} \dot{\sim} N\left(p, \frac{pq}{n}\right), \qquad q = 1 - p$$

- $\hat{p}$ is **unbiased** for p: $E(\hat{p}) = p$.
- Its variance is pq/n — notice that the variance is **determined by the mean**, which is a peculiarity of proportions and does not happen with means.
- Everything today is built from this one fact, applied twice.

The Two-Sample Fact

Fact

For two **independent** large samples,

$$\hat{p}_1 - \hat{p}_2 \dot{\sim} N\Big(p_1 - p_2,\ \sigma^2_{\hat{p}_1 - \hat{p}_2}\Big)$$

where

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \approx \frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}$$

- $\hat{p}_1 - \hat{p}_2$ is an **unbiased** estimator of $p_1 - p_2$.
- **The variances add, even though we subtract** — as in Lecture 16, the minus sign squares away, $(-1)^2 = +1$.
- p_1 and p_2 are unknown, so we substitute $\hat{p}_1$ and $\hat{p}_2$: legitimate because the samples are large.

The Conditions for "Large"

Conditions required for valid large-sample inference about $p_1 - p_2$

- The two samples are **randomly selected** from the two target populations, **in an independent manner**.
- The sample sizes n_1 and n_2 are both **large**:

$$n_1\hat{p}_1 \geq 5, \quad n_1\hat{q}_1 \geq 5, \quad n_2\hat{p}_2 \geq 5, \quad n_2\hat{q}_2 \geq 5$$

- In plain words: **each group needs at least about 5 successes *and* at least about 5 failures.**
- Note that $n_1\hat{p}_1 = x_1$ and $n_1\hat{q}_1 = n_1 - x_1$ — so the check is just "count the successes and the failures."
- **A big n is not enough on its own.** With $n = 4000$ and $\hat{p} = .001$ there are only 4 successes, and the normal approximation fails.

Outline

1. From one proportion to two
2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$
3. Confidence interval for $p_1 - p_2$
4. Hypothesis tests, and why we pool under H_0
5. Example: youth gambling in Minnesota
6. A one-sided example, and summary

Confidence Interval for $p_1 - p_2$

$100(1-\alpha)\%$ confidence interval

$$
\begin{aligned}
&(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\, \sigma_{\hat{p}_1 - \hat{p}_2} \\
&\qquad = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \\
&\qquad \approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}
\end{aligned}
$$

- The familiar shape: **estimate $\pm$ (critical value) $\times$ (standard error)**.
- The critical value is always z — **never** t. There is no unknown σ to estimate here, so no extra uncertainty to pay for.

Reading the Signs of the Interval

Interval		Interpretation
$(-, -)$	$\Longrightarrow$	$p_1 - p_2 < 0$, i.e. $p_1 < p_2$
$(+, +)$	$\Longrightarrow$	$p_1 - p_2 > 0$, i.e. $p_1 > p_2$
$(-, +)$	$\Longrightarrow$	0 is plausible; no evidence that $p_1 \neq p_2$

Same rule as for means

An interval that covers 0 is exactly an interval that fails to establish a difference. And as in Lecture 16, comparing two *separate* one-sample intervals and looking at the overlap is **not** a valid substitute — build the interval for the **difference**.

Outline

1. From one proportion to two
2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$
3. Confidence interval for $p_1 - p_2$
4. Hypothesis tests, and why we pool under H_0
5. Example: youth gambling in Minnesota
6. A one-sided example, and summary

Hypothesis Test for $p_1 - p_2$ (General D_0)

Test of $H_0: \; p_1 - p_2 = D_0$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}} \;\dot{\sim}\; N(0,1) \;\text{ under } H_0$$

Alternative	**Rejection region**
$H_a: \; p_1 - p_2 > D_0$	$\{ Z > z_\alpha \}$
$H_a: \; p_1 - p_2 < D_0$	$\{ Z < -z_\alpha \}$
$H_a: \; p_1 - p_2 \neq D_0$	$\{ \lvert Z \rvert > z_{\alpha/2} \}$

The same three shapes as always. But the case we almost always want is $D_0 = 0$, and there we can do **better**.

The Special Case $D_0 = 0$: Why We Pool

When $D_0 = 0$ the null hypothesis says the two proportions share a **common value**:

$$H_0 : \; p_1 - p_2 = 0 \quad \Longleftrightarrow \quad H_0 : \; p_1 = p_2 = p$$

- A test statistic is computed **assuming H_0 is true**. If H_0 is true there is only **one** proportion in the world, not two — so it is wasteful to estimate it twice.
- Both samples are estimating the *same* p. **Pool them into one estimate** and the standard error becomes more precise, which makes the test more powerful.

The parallel you already know

This is the same move as the **pooled variance** s_p^2 in the two-sample t test: when the assumption says a quantity is shared, you estimate it once, using all the data.

The Pooled Z Statistic

Test of $H_0: \; p_1 = p_2$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\dfrac{\hat{p}\hat{q}}{n_1} + \dfrac{\hat{p}\hat{q}}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \qquad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- $\hat{p}$ is the **pooled sample proportion**: throw both samples into one pile and count the successes.
- It is a **weighted average** of $\hat{p}_1$ and $\hat{p}_2$, with weights n_1 and n_2 — the larger sample gets more say.
- $\hat{q} = 1 - \hat{p}$.

Pooled for the Test, Unpooled for the Interval

	Standard error used	Why
Test of $H_0 : p_1 = p_2$	$\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	H_0 says there is one p
Confidence interval	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	no H_0 is assumed

- **This is the one place where the test and the interval do not use the same standard error.** A confidence interval makes no assumption about p_1 and p_2, so it has nothing to pool.
- In practice the two standard errors are very close, and the test and the interval essentially always agree. **But write down the right one.**

Outline

1. From one proportion to two
2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$
3. Confidence interval for $p_1 - p_2$
4. Hypothesis tests, and why we pool under H_0
5. Example: youth gambling in Minnesota
6. A one-sided example, and summary

Example 1: Setting Up

1992: $n_1 = 21{,}484,\ x_1 = 4{,}684,\ \hat{p}_1 = .218$

1998: $n_2 = 23{,}199,\ x_2 = 5{,}313,\ \hat{p}_2 = .229$

Define the parameters first

p_1 = the true proportion of Minnesota ninth-grade boys who gambled weekly or daily in **1992**.

p_2 = the same proportion in **1998**.

$$H_0:\ p_1 - p_2 = 0\ \ (p_1 = p_2) \qquad \text{vs.} \qquad H_a:\ p_1 - p_2 \neq 0\ \ (p_1 \neq p_2)$$

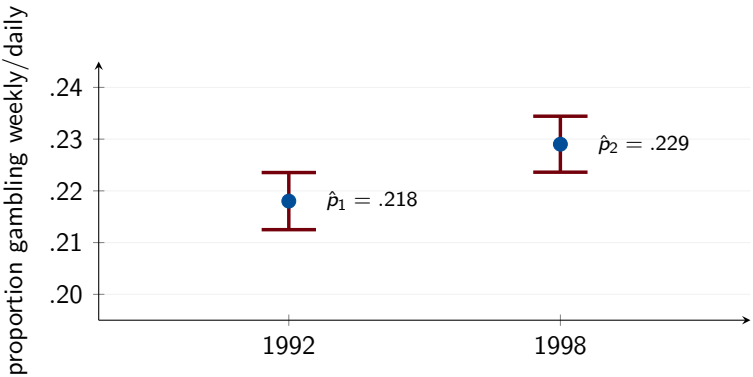
The question asks only whether the rates *differ*, so the test is **two-tailed**. We will use $\alpha = .05$.

Example 1: Checking the Conditions

Quantity	Value	
$n_1\hat{p}_1 = x_1$	4,684	≥ 5 ✓
$n_1\hat{q}_1 = n_1 - x_1$	$21{,}484 - 4{,}684 = 16{,}800$	≥ 5 ✓
$n_2\hat{p}_2 = x_2$	5,313	≥ 5 ✓
$n_2\hat{q}_2 = n_2 - x_2$	$23{,}199 - 5{,}313 = 17{,}886$	≥ 5 ✓

All four counts clear 5 by a wide margin, and the two surveys are independent random samples of ninth-grade boys. **The large-sample normal procedures apply.**

Example 1: What the Data Look Like



Note the stretched vertical scale — the whole picture spans only 4.5 percentage points. The two estimates differ by about **one percentage point**, and the garnet whiskers are their two *separate* 95% one-sample intervals, which only just fail to overlap. **That near-overlap settles nothing** — we need the interval for the difference.

Example 1: The 95% Confidence Interval

$$
\begin{aligned}
&(\hat{p}_1 - \hat{p}_2) \pm z_{.025}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \\
&\qquad = (.218 - .229) \pm 1.96\sqrt{\frac{.218(1-.218)}{21{,}484} + \frac{.229(1-.229)}{23{,}199}} \\
&\qquad = -.011 \pm 1.96(.003943) \;=\; -.011 \pm .00773 \\
&\qquad = (\,-.0187,\; -.00327\,)
\end{aligned}
$$

- We are 95% confident that the 1992 rate was between 0.33 and 1.87 **percentage points lower** than the 1998 rate.
- The interval is of the form $(-,-)$: it **excludes** 0, so $p_1 < p_2$.

Example 1: The Pooled Proportion and Test Statistic

Pooled sample proportion

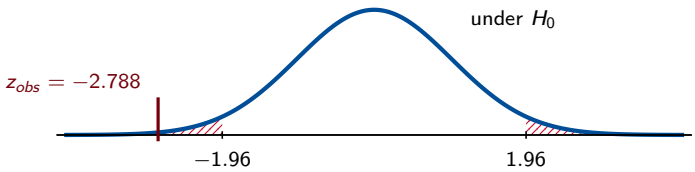
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{4684 + 5313}{21{,}484 + 23{,}199} = \frac{9{,}997}{44{,}683} = \mathbf{.2237}$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.218 - .229}{\sqrt{.2237(1-.2237)\left(\frac{1}{21{,}484} + \frac{1}{23{,}199}\right)}}$$

$$= \frac{-.011}{.003946} = \mathbf{-2.788}$$

Example 1: Decision

Rejection region: $\{\,|Z| > z_{.025}\,\} = \{\,|Z| > 1.96\,\}$



Since $|-2.788| > 1.96$, the statistic **falls in the rejection region**: **reject** H_0 at the .05 level.

$$p\text{-value} = 2\,P_{H_0}(Z \leq -2.788) = 2(0.00266) = \mathbf{0.0053} \qquad 0.0053 < 0.05$$

Example 1: Conclusion

Conclusion

At the .05 significance level we have sufficient evidence that the true percentages of Minnesota ninth-grade boys gambling weekly or daily in 1992 and 1998 are **significantly different**. The 95% interval $(-.0187, -.00327)$ puts the 1998 rate **higher**.

- The test and the interval agree, as they should: the interval excludes 0 and the two-sided test rejects.
- Note how little the two standard errors differ — .003943 unpooled versus .003946 pooled. With samples this large the choice is immaterial numerically, **but the reasoning behind each is still what you must be able to explain**.

Statistically Significant $\neq$ Practically Important

- The difference we just declared significant is **1.1 percentage points** — from 21.8% to 22.9%.
- It is significant because $n_1 + n_2 \approx 44{,}700$. The standard error is .0039, so even a tiny difference is many standard errors from zero.

The lesson

A large enough sample will make almost any difference statistically significant. A small p-value says "this difference is probably real"; it does **not** say "this difference is large or important." That second question is answered by the **confidence interval**, which here says the gap is somewhere between 0.3 and 1.9 percentage points.

Always report the interval, not just the p-value.

Example 1 in R

```
prop.test(x = c(4684, 5313), n = c(21484, 23199),
          correct = FALSE)
```

```
    2-sample test for equality of proportions

data:  c(4684, 5313) out of c(21484, 23199)
X-squared = 7.7653, df = 1, p-value = 0.005326
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.018723916 -0.003267639
sample estimates:
   prop 1    prop 2
0.2180227 0.2290185
```

R reports `X-squared` rather than Z, because $\chi^2 = Z^2$: here $(-2.788)^2 = 7.77$. The p-value and the interval match our hand computation. `correct = FALSE` turns off the continuity correction so the output matches the formula we used.

Outline

1. From one proportion to two
2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$
3. Confidence interval for $p_1 - p_2$
4. Hypothesis tests, and why we pool under H_0
5. Example: youth gambling in Minnesota
6. A one-sided example, and summary

Example 2: A One-Sided Test

Example 2

A clinical trial compares a new formulation of a drug with the standard formulation. Of 200 patients randomized to the **new** formulation, 36 reported a particular side effect; of 250 patients randomized to the **standard** formulation, 80 did. Test at the .05 level whether the new formulation has a **lower** side-effect rate.

$$n_1 = 200,\ x_1 = 36,\ \hat{p}_1 = .18 \qquad n_2 = 250,\ x_2 = 80,\ \hat{p}_2 = .32$$

p_1 = the true side-effect rate on the new formulation; p_2 = the true rate on the standard formulation.

Counts: 36, 164, 80, 170 — all ≥ 5 ✓

Example 2: Test Statistic

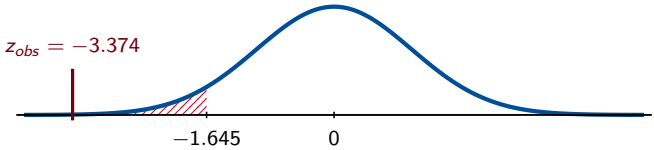
$$H_0: \ p_1 - p_2 = 0 \ \ (p_1 = p_2) \qquad \text{vs.} \qquad H_a: \ p_1 - p_2 < 0 \ \ (p_1 < p_2)$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{36 + 80}{200 + 250} = \frac{116}{450} = \mathbf{.2578}$$

$$Z = \frac{.18 - .32}{\sqrt{.2578(1 - .2578)\left(\frac{1}{200} + \frac{1}{250}\right)}} = \frac{-.14}{.04150} = \mathbf{-3.374}$$

"Lower" puts the whole of α in the **left** tail: the rejection region is $\{ Z < -z_{.05} \} = \{ Z < -1.645 \}$.

Example 2: Decision and Interval



Since $-3.374 < -1.645$ we **reject** H_0; p-value $= P(Z < -3.374) =$ **0.00037**.

$$95\%\ \text{CI}: \quad -.14 \pm 1.96\sqrt{\frac{.18(.82)}{200} + \frac{.32(.68)}{250}} = -.14 \pm .0786 = (\,-.219,\ -.061\,)$$

Conclusion

There is sufficient evidence at the .05 level that the new formulation has a lower side-effect rate; it is lower by somewhere between about 6 and 22 percentage points.

Common Mistakes to Avoid

- **Using the unpooled SE in the test of $H_0 : p_1 = p_2$,** or the pooled SE in the confidence interval. One H_0, one pool.
- **Using t instead of z.** Proportions always use z.
- **Checking only that n is large.** The condition is on the **counts**: at least about 5 successes *and* 5 failures in each group.
- **Pooling percentages instead of counts:** $\hat{p} \neq (\hat{p}_1 + \hat{p}_2)/2$ unless $n_1 = n_2$. Pool the **numerators and denominators**.
- **Losing track of which group is "1."** The sign of $\hat{p}_1 - \hat{p}_2$, of the interval, and of H_a must all agree.
- **Reading a small p-value as a large effect.** Report the interval too.
- **Comparing two separate one-sample intervals** and judging by the overlap.

Summary

- The parameter is $p_1 - p_2$; the estimate is $\hat{p}_1 - \hat{p}_2$; for independent samples the variances **add**.
- **Conditions:** two independent random samples, and at least about 5 successes and 5 failures in each group.
- **Confidence interval:** $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2}$ — the **unpooled** standard error.
- **Test of $H_0 : p_1 = p_2$:** the **pooled** standard error with $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$, referred to $N(0, 1)$.
- An interval that covers 0 says the same thing as a two-sided test that fails to reject.
- **Significance is not importance.** With tens of thousands of observations a one-point difference is significant; the interval is what tells you how big it is.